DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DATA MINING ON TEXT DATA AND RELATED APPLICATIONS

by
Bora ÖZGÜL

October, 2011
İZMİR

DATA MINING ON TEXT DATA AND RELATED APPLICATIONS

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science
in Statistics Program

by
Bora ÖZGÜL

October, 2011
İZMİR

M. Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled "**DATA MINING ON TEXT DATA AND RELATED APPLICATIONS**" completed by **BORA ÖZGÜL** under supervision of **PROF. DR. EFENDİ NASİBOĞLU** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Efendi NASİBOĞLU

Supervisor

Y. Doç. Dr. Adil ALPKOÇAK

(Jury Member)

Yrd. Doç. Dr. Emel Kuruoğlu

(Jury Member)

Prof. Dr. Mustafa SABUNCU

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

First of all, I would like to thank my respectful advisor Prof. Dr. Efendi NASİBOĞLU who has guided and supported me in all phases of this study and brightened me in many studies with a great patience.

I also want to thank Asst. Prof. Dr. Adil ALPKOÇAK, who had great contributions and encouraged me in this study by his lecture, with all my gratitude.

I would like to thank my dear mother Muteber ÖZGÜL and my dear brother Cengiz ÖZGÜL who always support me in good and bad times with monetarily and mentally, whom always love me; and my dear father Metin ÖZGÜL who bequeath the biggest inheritance to me with all my gratitude.

I would like to express my sincere gratitude to my dear friends Sezai KAPLAN, Cavit ÇELİK and Numan ZENGİN who had great contributions to my life and never give up on my friendship.

I am grateful to my dear friends Fatma BEKTAŞ, Anıl KORKMAZ and Kerime MATUR for helping me to prepare a test corpus in this study.

DATA MINING ON TEXT DATA AND RELATED APPLICATIONS

ABSTRACT

There is extremely large amount of textual information stored and fast growingly continued to be stored into many storage tools such as database and data warehouse. Thus, reaching needed information is getting slow and hard. Because of this situation, a robust analyzing tool is needed to users. Text mining, which is a branch of data mining, is developed and is still fast developing tool to handle this problem.

Text mining is multidisciplinary that those are "Natural Language Processing, Information retrieval, Statistics and Data Mining". In this study, those areas are defined in detail and what parts of those areas are used in text mining.

There are many applications that text mining tool is used. In this thesis those are mentioned slightly but automatic text summarization. One of the most used applications in text mining area is automatic text summarization. Needed information has to be reached fast but after information reached, user must read whole document for interested information. Automatic text summarization task handles this problem and generates a summary of documents to users for time consuming.

In this study automatic text summarization task is explained in details and a couple of algorithms are mentioned. Finally, a software coded by using one of those algorithms and then ten Turkish news articles are summarized analyzed by the software.

Keywords: Text mining, automatic text summarization

METİN VERİLERİ ÜZERİNDE VERİ MADENCİLİĞİ VE UYGULAMALARI

ÖZ

Çok büyük miktarda depolanmış metinsel veri vardır ve hızlı bir şekilde büyüyerek veritabanı, veri ambarı gibi depolama araçlarına depolanmaya devam edilmektedir. Bu nedenle, ihtiyaç duyulan bilgiye ulaşmak yavaş ve zor bir hal almaktadır. Bu durumdan dolayı, kullanıcılar güçlü bir analiz aracına ihtiyaç duymuştur. Veri madenciliğinin bir dalı olan metin madenciliği, bu problemi ele almak için geliştirilmiş ve hızla geliştirilmekte olan bir araçtır.

Metin madenciliği "Doğal Dil İşleme, Bilgi Çıkarımı, İstatistik ve Veri Madenciliği" olan alanların birleşiminden oluşmuştur. Bu çalışmada, sözü geçen alanlar detaylı bir şekilde açıklanmış ve bu alanların hangi kısımlarının metin madenciliğinde kullanıldığından bahsedilmiştir.

Metin madenciliği alanının kullanıldığı birçok uygulama mevcuttur. Bu tezde, uygulamalar yüzeysel olarak bahsedilmiş fakat otomatik metin özetleme detaya inilmiştir. Metin madenciliği alanında en çok kullanılan uygulamalardan birisi otomatik metin özetlemedir. İhtiyaç duyulan bilgi hızlı bir şekilde ulaşılmalıdır fakat bilgiye ulaşıldıktan sonra, kullanıcı ilgilenilen bilgi için tüm dokümanı okumalıdır. Otomatik metin özetleme görevi bu problemi ele alır ve kullanıcılara zaman kısıtı için özet oluşturur.

Bu çalışmada otomatik metin özetleme görevi detaylı bir şekilde anlatılmış ve birkaç algoritmadan bahsedilmiştir. Son olarak bu algoritmalardan biri kullanılarak bir program kodlanmış ve on Türkçe haber metninin özeti çıkarılarak analiz edilmiştir.

Anahtar Sözcükler: Metin madenciliği, otomatik metin özetleme

CONTENTS

Page

CHAPTER ONE
INTRODUCTION

In the last three decades information has been produced extremely fast and information age ease storing the information electronically. “A recent study indicated that 80% of a company’s information is contained in text documents” (Tan, 1999). Extremely fast growing information storage became as information trash. Reaching to desired information from this information trash is crucial. Thus, an analyze tool is need to reach desired information for transforming information to knowledge. Text mining is the most important tool of the need on this area and it’s also a new area that is just developing.

Labor-intensive manual text mining approaches first surfaced in the mid-1980s, but technological advances have enabled the field to advance during the past decade.

Text mining, in other words text data mining, is roughly text analytics. “Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools” (Feldman & Sanger, 2007, pg. 1). Text mining is, roughly speaking, process of discovering new information that is previously unknown and usually discovering from large amount of unstructured text respitories.

Text mining and data mining has many common attributes. Text mining and data mining both have roughly same steps as preprocessing of data, discovering pattern algorithms and presentation. In contrast text mining discovers patterns from unstructured or semi-structured document collections instead of structured data base sources. For text mining, on preprocessing, document is expressed by natural language. This preprocessing, which is not used for other data mining systems, is responsible for making structured data from document collection.

Text mining has roughly two steps. *Text refine* phase is for transforming free text to moderate form, *knowledge distillation* is for discovering patterns and extracting knowledge from moderate form. Moderate form can be semi-structured or structured. Moderate form may be document based which each unity is a document.

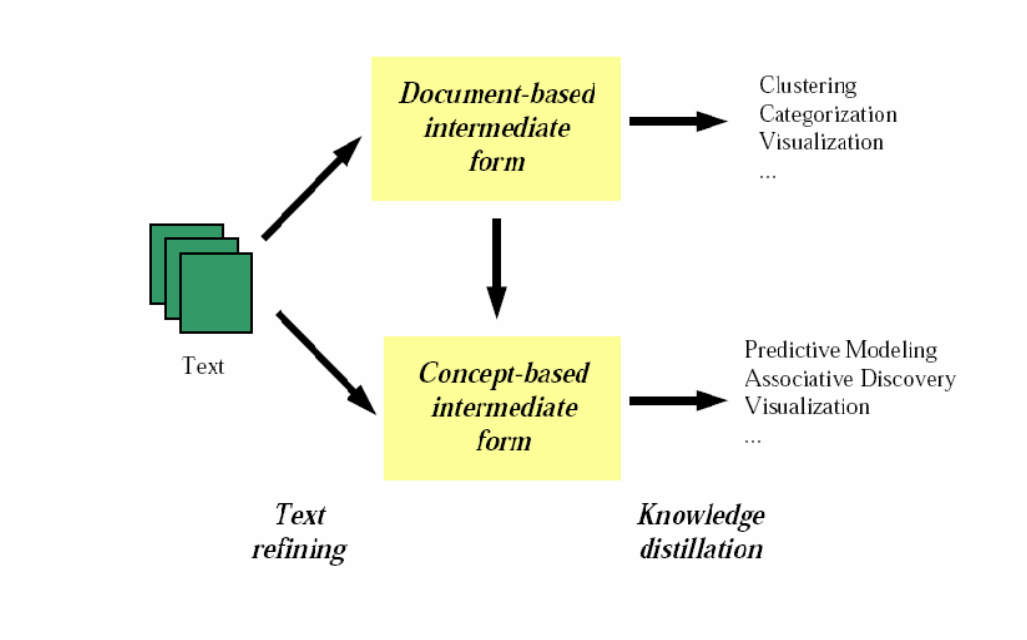


Figure 1.1 The structure and phases of text mining.

Text mining is not simple as shown in figure 1.1. Text mining is an interdisciplinary field that those are natural language processing, information retrieval, statistics and data mining, as shown on figure 1.2.

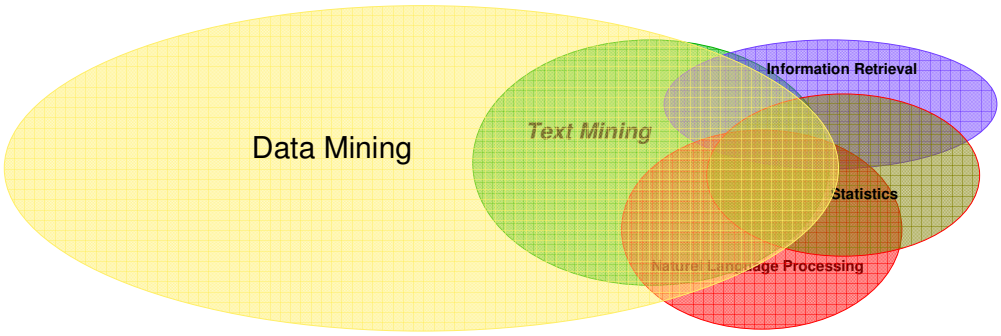


Figure 1.2 Interdisciplinary of text mining.

More specifically text mining steps are compiling documents, text organization and preprocessing, attribute selection for analyzing from organized text, application of data mining algorithms to selected attributes and presentation to the user. Steps are showed precisely in figure 1.3.

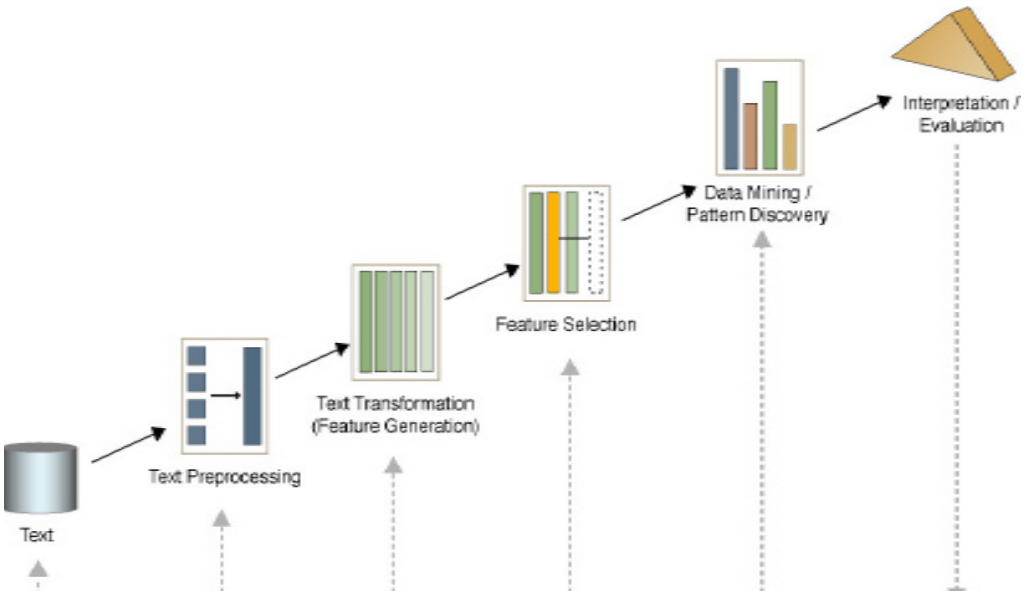


Figure 1.3 Steps of text mining.

Text mining is applied to many fields and is proved that it's a robust tool. Some of the applied fields are security, biomedical, software and applications, online media applications, market applications, academic applications.

The purpose of this study is to build a software and an evaluation set to summarize documents, and to evaluate summarization of documents automatically, which is an adaptation of text mining.

This thesis contains seven chapters. In chapter 1, a short description of text mining and its related applications are mentioned. In chapter 2, Data mining is described by all its methods. In chapter 3, Natural Language Processing (NLP), which is a vital tool to organize documents, is described briefly. In chapter 4, an introduction to Information Retrieval (IR) is described and indexing technology

mentioned to transform unstructured documents into structured form by using IR methods. In chapter 5, text categorization is described too classify or to cluster documents or its components due to its contents. In chapter 6, automatic text summarization is described and three methods to summarize documents described briefly. In chapter 6, application of text summarization is built by using Cover-Coefficient based text summarization method and applied on evaluation set. Finally in chapter 7, a summarization of this thesis is described and result of application is interpreted.

CHAPTER TWO
DATA MINING

Major reason of interestingness to data mining is because of existing large amount of data to be discovered information and knowledge from them. Since the end of 60's, development of computer hardware let people have robust and efficient computers, data collection tools and storage resources. This technology influenced to be developed database and information industry. Information retrieval let information respitories be developed by large amount of databases for data analysis and process management.

Data can be stored in many different types of databases. This leads to form a kind of database architecture. This is called data warehouse. Data warehouse is a respitory that originated from multiple different types of structured databases to manage decision making.

Data warehouse technology includes data cleaning, data integration and Online Analytical Processing (OLAP). OLAP analysis techniques of functionalities such as summarization, consolidation and aggregation to view data from different angles. In this kind of environment many types of databases and data warehouses are occurred and that leads to having large amounts of data. Robust analysis tools are needed because of the large amount of data. "The abundance of data, coupled with need for powerful data analysis tools, has been described as a data rich, information poor situation" (Han & Kamber, 2006, pg. 4).

Large amount of data that are stored in large and countless databases are far exceeded for human ability to be understood. Han & Kamber (2006) described this situation as "data collected in large databases become 'data tombs'- data archives that seldom visited" (pg. 4). The gap of between data and information leads to development of data mining tools that turn data tombs into golden nuggets. From the data, patterns can be found and use it for scientific researches by analyzing data that uses data mining tools.

2.1 Introduction

In simple words, data mining is extracting or mining knowledge from large amount of data. The term *mining* comes from its real meaning. Mining tones of earth and processing it by some kind of chemicals make people to get precious materials such as copper, silver or gold. Same as mining, data mining is extracting knowledge from large amount of data by processed and analyzed.

It also called knowledge mining, knowledge extraction, data pattern analysis, data archeology, but mostly used knowledge discovery on databases (KDD).

Data mining can be explained by iterative steps as:

1) Data Cleaning: getting rid of noisy and inconsistent data from data source.
2) Data Integration: integration of many different data sources.
3) Data Selection: retrieving data from relative database to be analyzed.
4) Data Transformation: by summarization or aggregation process, data are transformed into appropriate form to be mined.
5) Data Mining Algorithms: vital process of data mining that use specific methods to extract data patterns.
6) Pattern Evolution: Process of validating patterns that represent knowledge by interested measures.
7) Knowledge presentation: Process of presenting extracted patterns and knowledge to the user by visualizing.

Process of data mining is shown on figure 2.1.

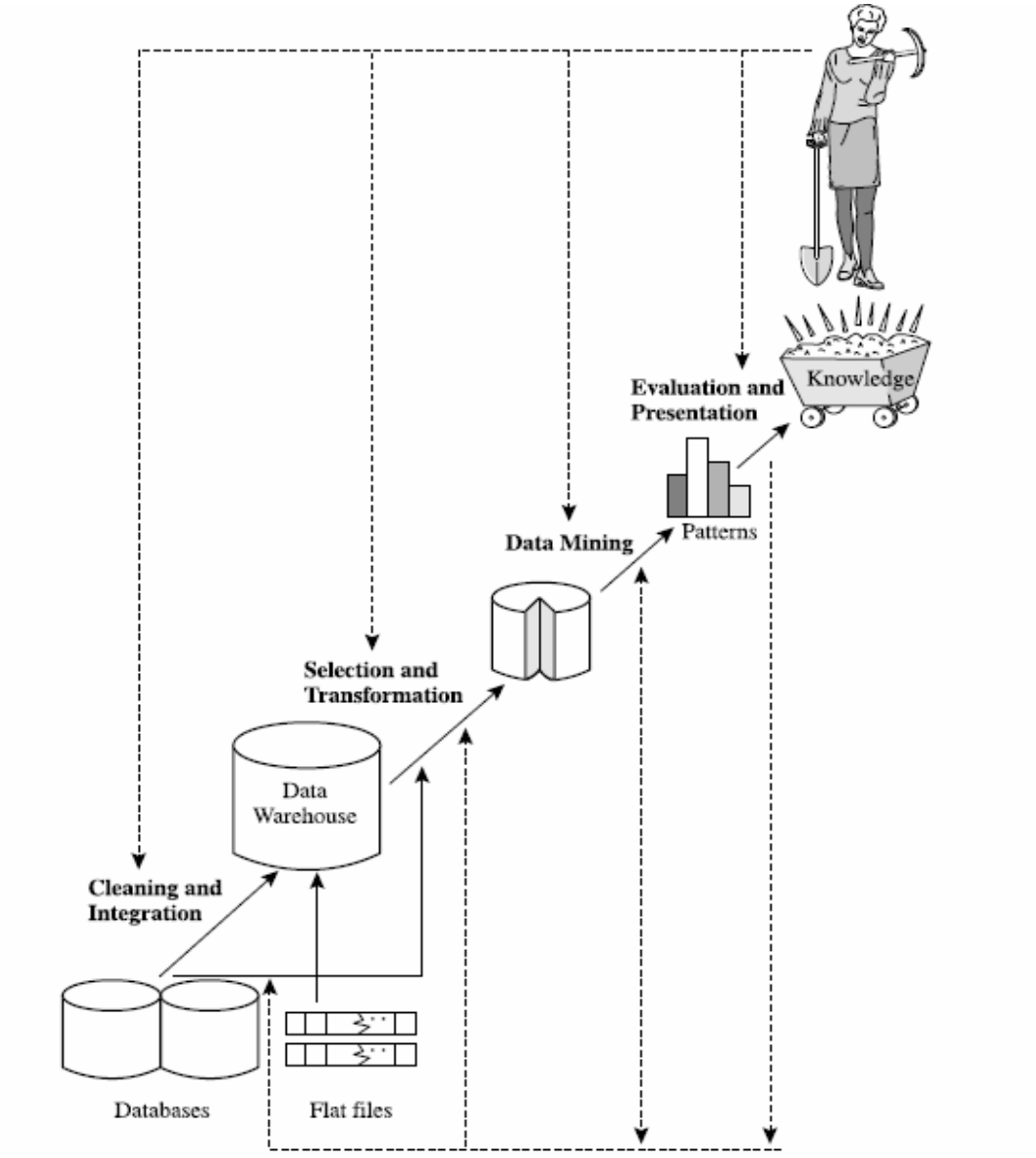


Figure 2.1 Process of data mining

Interested patterns are presented to the user and may be stored in a new knowledge-base. Data mining is just one step in the whole process but it's a vital step that interested knowledge and patterns are extracted. Generally speaking, data mining is discovering knowledge and patterns from large amounts of data source such as databases, data warehouses and data respitories.

Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

2.2 Types of Data in Data Mining

Data mining can be applied on all kinds of data sources such as relational databases, data warehouses, World Wide Web (WWW), advanced databases. Advanced database means object oriented or relational, time series, multimedia and textual database. For each of them, according to their types, different types of mining methods can be applied.

2.1.1 Relational Databases

"A database system, also called a database management system (DBMS), consist of a collection of interrelated data, known as database, and a set of software programs to manage and access the data" (Han & Kamber, 2006, pg. 10). Relational databases are usually data respitories that includes information so, data mining can be applied.

A relational database is a set of tables that given unique names to each other. Each table has a set of colons or attributes and usually includes many rows or records that are tuples. In a relational table each of tuples represents a set of objects that are identified by attribute values and a unique key.

Relational database can be accessed and controlled by a graphical interface or a query language such as SQL. For instance, user can employ a menu by specifying attributes and its constraints by query language or graphical interface. A query can be transformed into a set of operations for an efficient process such as selection, join and projection. A query can retrieve subsets of the table. Some examples are listed below.

- Show all sold units last week.
- Show all sales and wage in last month according to branches.

When data mining applied on relational databases, one can go further by searching trends or patterns. For example, an electronic store can predict customers credit risks by analyzing customers profile data. Data mining system can also extract increase or decrease of sales by deviation analyze. Thus, importance of packing or commercials would be seen by extracting patterns.

2.2.2 Data Warehouses

A data warehouse is a multi-dimensional database that each dimension represents one ore more attributes and its cells include cumulative values such as total sales. Its physical structure can be relational data respitory or multi-dimensional data cube. Thus, data can be viewed from different angles and summary of the data can be seen in a fast way.

Data warehouse is composed of many data sources that unified in a specific scheme and usually exist in just one site. Process of composing a data warehouse has some steps. They are listed as below and shown on figure 2.2.

- Data cleaning
- Data transforming
- Data integration
- Updating data in specific periods

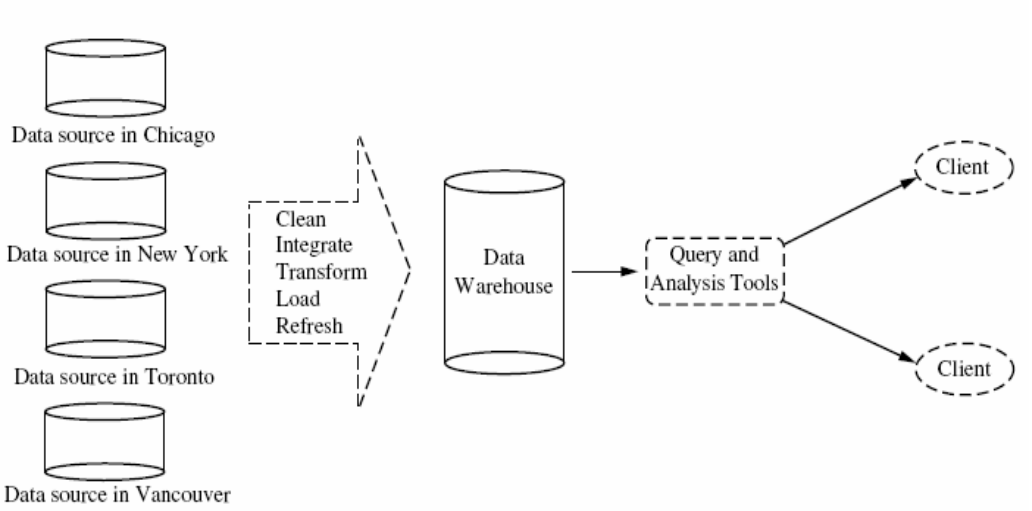


Figure 2.2 Steps and the structure of a data warehouse

Data warehouses are focused on major subjects such as customer and products thus, making decision is much easier. From the past, such as 5-10 years, data are stored as summaries of them instead of details so user can reach the data easily.

2.2.3 Advanced Database Systems And Applications

Advanced database systems deal with spatial data such as maps, engineering design data such as building design, multimedia data such as text, video and sound data. Application of those needs measurable methods that deal with efficient data and complex object structures. For those needs, advanced database systems are developed.

"While information respitories or databases require complex facilities to efficiently store, retrieve and update large amounts of complex data, they also fertile grounds and raise many challenging research and implementation issues for data mining" (Han & Kamber, 2006, pg. 16).

2.2.3.1 Object Oriented Databases

Object oriented databases are related to programming and each of entity seems like an object. This is related to below issues.

- A set of variables that describe the objects.
- A set of messages are exist to communicate between objects and other objects or colons of the database system.
- A set of methods are exist to hold the code to implement a message. “get_photo (employee)” message method would returns photo of a specific employee.

2.2.3.2 Spatial Databases

Spatial databases include dimensional related information. It can be geographic (maps), medical or satellite visual databases. Spatial data may be represented in raster format, consisting of n-dimensional pixels maps. For instance, two dimensional satellite visuals can be raster format. Maps can be represented as vector format. Roads, lakes, buildings, bridges can be shown as basic geometric shapes such as dot, line, polygons and network formed of these shapes.

Geographical database has many application fields. Foresting and ecological planning, locating electrical and phone cables or water supplies for giving people better or different service. Vehicle navigation and delivery systems are also use spatial databases.

So, what can data mining do on spatial databases? Data mining can discover patterns of houses that near specific places such as parks, or can predict weather of different height of mountain areas, or can discover distance between houses and city center, highways to find out poverty ratio in big cities.

2.2.3.3 Text and Multimedia Databases

Text databases take word descriptions as objects. Those word descriptions may be long sentences or paragraph instead of just words. For instance, error reports, report summaries, warning messages or notes. Text data are usually unstructured ones. Sometimes it is semi-structured such as XML or HTML and structured ones like library databases.

What data mining extract from textual databases is general descriptions of object classes by key word or content associations. This can be done by integrating data mining techniques with information retrieval techniques. Documents split into words, sentences or paragraphs to be indexed. Then summarization of documents or similarity of documents to others can be done by extracting the objects that have more weight than others.

On the other hand, dictionaries are also used. A dictionary is about the field that document is relevant. For instance, if database includes just about law words or sentences, dictionary would also be about law.

Multimedia databases are storage of video, voice or visual data. Those are used for some applications such as voice mail systems, picture recognition, and video scan systems. Multimedia databases have large capacities on disks because of including data such as videos. Thus, specific search engines and storages are needed. Standard data mining techniques need to be integrated with specific storage and search engines to mine multimedia databases.

2.2.3.4 World Wide Web (WWW)

"World Wide Web (WWW) and its associated distributed information services such as America Online, Yahoo!, AltaVista, Prodigy, provide rich, world-wide online information services, where data objects are linked together to facilitate interactive access" (Han & Kamber, 2006, pg. 20). User can pass through one link to

another. If link-to-links are recorded, better user/ customer classification can be done. Thus, advertisements that showed will be more attractive and useful to the user.

Web pages are easy to read and surf but totally unstructured. Systematic information retrieval and data mining techniques need to be used because of computers doesn't understand word lexical.

On the web, keyword search would return irrelevant documents to user. As a result user would get limited information fro relevant keyword search. For instance, trying for a keyword to search would return some irrelevant documents or all documents that user can't read all at once so it takes so much time to get the needed information. This problem can be solved by integrating data mining and information retrieval techniques to classify or cluster on better ways.

2.3 Extracting Patterns

Data mining tasks can be classified into two categories, which are descriptive and predictive. Descriptive mining tasks characterize general properties of data in database. Predictive mining tasks perform inference on the current data in order to make predictions.

In some cases user may have no idea what kind of pattern he/she would discover so, tries to extract more than one pattern. Thus, "it is important to have data mining systems that can mine multiple kinds of patterns to accommodate different kind of user expectations and applications" (Han & Kamber, 2006, pg. 21). On the other hand, data mining systems should allow user to specify hints to perform the user focusing on discovering patterns.

Data mining functionalities and patterns that can be discovered are below.

- Characterization and discrimination
- Association analysis

- Classification
- Clustering analysis
- Outlier analysis

2.3.1 Characterization and Discrimination

Data characterization is the summarization of general characterizations of data or target class properties of data. There are many efficient methods to summarize or characterize the data. For instance, data cube can summarize the data according to dimensional of user set.

Data discrimination is comparison of general properties of target class data objects with one or a comparative class. Target and comparative class can be formed by queries. Explaining with an example, a user can extract sales of a specific software in a period that increased by 10% would be a target class, and extract another software sales at the same period that decreased by 30% would be comparative class.

Example: A data mining system can set rules on customers of an electronic store. It can discover customers who regularly buy electronics and who rarely buys electronics. The resulting description could be a general comparative profile of the customers, such as 80% of the customers who frequently purchase electronics are 20-40 years old and have a university education, where as 60% of the customers who infrequently buy such products at the same age interval don’t have university education. Decreasing number of dimension such as occupation or increasing the number of dimension by adding level of income would help to find more discriminative features between two classes.

2.3.2 Association Analysis

Association analysis is extracting frequently seen attribute-value cases in same data. Association analysis mostly used in market applications. Rules are applied on customer data to extract profile of them. Here is an example of a rule:

"Age(X, "20-29") □ Income(X, "20K-29K") → Buy(X, "CD Player")
[Confidence: 60%]"

This rule explains, 60 % of customers buy cd player whom are in 20-29 age interval and have income of 20K-29K interval. This rule has multi-dimensional attribute, thus, it has multi-dimensional association rule.

2.3.3 Classification

"Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use model to predict the class of objects whose class label is unknown" (Han & Kamber, 2006). "In essence the process of classification simply means the grouping together of like things according to some common quality or characteristics" (Hunter, 2009, pg. 1). Models can be shown in different forms such as decision trees, mathematical formulas, classification rules (if…then) or neural networks.

Classification tries to predict data objects classes but in many applications user tries to predict the missing data values instead of class label. This is usually applicable on predicted values that are numeric and sometimes it is known as prediction.

2.3.4 Cluster Analysis

Unlike classification, clustering is analyzing of data without consulting a labeled class. A definition for clustering could be the process of organizing objects into

groups whose members are similar in someway. Objects are grouped or clustered due to minimization of similarity of inter-groups and maximization of similarity of intra-groups. Thus, cluster formed data have maximum similarity in intra-group and have minimum similarity in inter-groups.

There are many algorithms for clustering but the most used and simplest one is K-means algorithm. The procedure follows a simple way to cluster data into certain k number of clusters.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 , \tag{2.1}$$

$\left\| x_i^{(j)} - c_j \right\|^2$ in eq.(2.1) is the distance measure between data point $x_i^{(j)}$ and cluster centroid c_j, k is the number of clusters and n is the number of data points. Steps of K-mean algorithms are below.

1. Place K points into the data point space. These points represent the centroid of clusters.
2. Assign each object to the group that has closest centroid.
3. After all points assigned to a group, recalculate the place of K centroids.
4. Repeat the 2nd and 3rd steps until centroids no longer move.

After process of clustering ends, data would be seen as in figure 2.3.

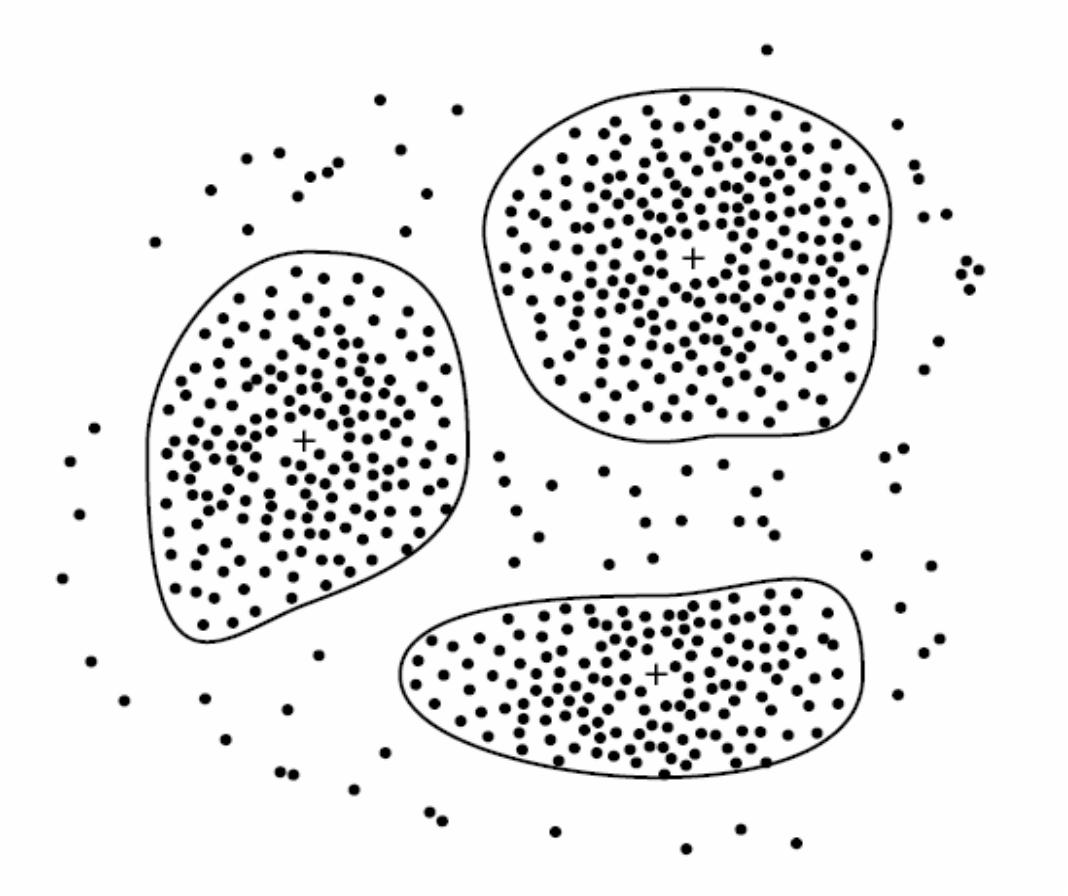


Figure 2.3 An example for a clustering

2.3.5 Outlier Analysis

A database may include objects that comply with the general behavior or model of the data. Those data objects are called outliers. Most of data mining systems ignore or delete because of noisy. On the other hand, in some cases such as fraud, outlier data are more valuable because user can discover behaves that unusual which may mean fraud.

Outliers can be found by applying statistical tests to data according to assuming data fit a distribution or can be found by applying probabilistic models on data. Nevertheless, data that don't are in a cluster are also outliers.

If an expense of a credit card in a specific period is more than ever is a sign of fraud. Outlier analysis may take consideration by adding more dimensions such as area of spending, what kinds of products are bought or how often do products bought.

2.4 Patterns

Data mining systems may discover thousands of patterns but interested patterns are in small percentages. That makes people ask some questions about it.

- What makes pattern important?
- Does a data mining system can discover all patterns exist in data?
- Does a data mining system can discover just the interested patterns?

2.4.1 Importance Of Patterns

This question can be answered by combining some situations which are below.

- Easy to understand by user.
- Valid on new or training data with some degrees of certainty.
- Potentially useful.
- Novel.

An important pattern represents the knowledge. There are some objective measures to identify interestingness of patterns. Usually interestingness of pattern measures is limited by user threshold set. Thus, non-interested, unimportant patterns are eliminated.

Even objective measures help user to identify interesting patterns, it is insufficient to satisfying needs of user unless integrated by subjective measures. For instance, customer data of a supermarket would be important for the manager of supermarket but for an analyst, who analysis the performance of employees, data may not be important.

2.4.2 Discovering All Patterns

This is about completeness of data mining algorithm. It is inefficient for data mining systems to generate all of the possible patterns. Thus, data mining system should be focused on discovering interested patterns and thresholds should be set to eliminate other uninteresting patterns.

2.4.3 Discovering Just Interested Patterns

This question is about optimization of data mining. User desires the data mining system that it only discovers interested patterns because it would be easier to extract interested patterns from all patterns extracted.

CHAPTER THREE
NATURAL LANGUAGE PROCESSING

"The term 'Natural Language Processing' (NLP) is normally used to describe the function of software or hardware components in a computer system which analyze or synthesis spoken or written language" (Jackson & Moulinier, 2002, pg. 2-3). The word "natural" describes the difference between human spoken language and logical, mathematical notations or computer languages such as Java which have a structure of language. Directly speaking, NLP is having computer systems which deal with ambiguous targets that understand as much as a human being.

Oflazer & Bozşahin (2006) describes natural language processing as "ana işlevi doğal bir dili çözümleme, yorumlama ve üretme olan bilgisayar sistemlerinin tasarımını ve gerçekleştirilmesini konu alan bir bilim ve mühendislik alanıdır" which means "analyzing, interpreting and generating a natural language is the major task of a scientific and engineering field that designs and generates computer systems".

It is obvious that machines can be programmed to be comprehended. It is possible that computers can be programmed to solve mathematical or logical puzzles, but it is a truth that analyzing spoken and written language by computer programs is so problematic. Linguistic ambiguous is sometimes because of human being but general word and sentences can be interpreted in many different ways. For instance, "Ali saw the man nearby the telescope in the park" has an ambiguous issue that if telescope belongs to the man or if telescope is property of the park! This sentence is ambiguous.

Information is mostly expressed by language instead of video, sound or picture. Most of the information is in electronic documents or relational databases generated from tables, spreadsheets, articles or books. Thus, language processing is vital on analyzing texts.

Most of texts start with background of linguistics. This can be generalized in order to some steps. Structure of a text starts with syntax, goes on with semantics and end with pragmatics.

3.1 NLP and Linguistics

Some brief definitions of traditional linguistic concepts are necessary, if only to provide an introduction to the literature on NLP.

3.1.1 Syntax and Semantics

In some books there are sentences that are perfectly in syntax but have no semantics. According to those two factors, sentences can be unraveled due to its semantic or syntax. Sentence is first analyzed for syntax, after that analyzed for semantics without consideration of syntax.

Unraveling syntax and semantics is generally applied on languages that are logical form or computer programming languages. In most of unnatural languages, semantic of content is all about the structure of itself. In other words, semantic of the language is understood by its structure without considering linguistics or content of the language. It is different in natural languages. There are some situations like ambiguous of language or conflict of inter-languages.

3.1.2 Pragmatics and Context

Pragmatics is described as the rules of a language. For instance, “you owe me five”; does the sentence mean someone really owes money or does that just an expression? This sentence’s meaning is variable according to situation. When a user searches “natural language processing” in search engine, what actually is user looking for can not be predicted. Does user want to find an expert of it, a course or a reference? A search engine that has artificial intelligent can find what the user wants

to search about by considering past searches. For instance, if previous searches are "what's natural language processing, artificial intelligence book, Dokuz Eylül University" and so search engine can return relevant documents that user wants, with considering previous searches.

"Use and context are inextricably intertwined" (Jackson & Moulinier, 2002, pg. 6). Some contexts affect intensions behind the utterance. For instance, writing Adolf Hitler in the quotes without stating any idea about him or a sentence like "I doubt government will break up Microsoft" will affect the intension behind the utterance that will also affect interpretation of the sentence.

Although there have been some attempts to construct novel theories about using languages. "It has also been argued that patterns of use are so specific to particular domains that a general theory is impossible" (Jackson & Moulinier, 2002, pg. 7). Newspapers, court reports, commercials, CVs have different patterns in their use of language.

3.1.3 Tasks and Super Tasks

NLP's primary application on the web is still document retrieval. It is finding relevant documents according to users' query. Many search engines wasn't use to do NLP so much but in '90s, indexing, identification and presentation got sophisticated. Thus, researchers began to study on NLP more than ever.

The task of automatic routing is to route relevant but not the same document feedback units to user which is called document routing. Document routing is about document classification. On this task, documents are assigned to a specific class and assigning is according to content of the text. Mostly seen general case is one document can be assigned to more than one class and classes may be a part of bigger structures such as topic or hierarchy.

Sometimes the focus may be extracting specific information from document sets or targeted documents which have specific information. For instance, user may want to extract the people who takeover a company from news feedback about corporate takeovers. This is called information extraction. In some forms, document summarization can be called a specific form of information extraction. On document summarization, programs present a surrogate of original text by extracting important sentences.

User can combine the tasks above to generate super tasks. For instance, a computer program can acquire documents by feedback according to their contents, sort by category, extract the important parts and present.

3.2 Linguistic Tools

Analyze of text typically proceeds layered fashion. Documents are splitted into paragraphs, paragraphs into sentences and sentences into words. Then words in a sentence are tagged as part of speech (POS) and parsed to analyzer grammatically. Those parsers are typically based on sentence delimiters, tokenizers, stemmers and POS taggers, but not all applications need all parsers at the same time. All search engines use sentence delimiters but not all use POS tagging.

3.2.1 Sentence Delimiters and Tokenizers

First of all, the components of sentence must be identified to parse sentences from a document.

3.2.1.1 Sentence Delimiters

It is a hard task because of the ambiguity of punctuations that indicates the end of sentence. For instance, is “a full stop or a dot” used for a delimiter of float numbers or indicates an end of a sentence? After a dot, a space and starting with upper case does not mean it is a starting of a sentence. Title of text may be an example of this

case. There are some problems that sentence delimiters are actually not sentence delimiter. “It may appear that using a short list of sentence-final punctuation marks such as ‘.’, ‘?’, ‘!’ is sufficient. However, these punctuation marks are not used exclusively to mark sentence breaks” (Reynar & Ratnaparkhi, 1997).

3.2.1.2 Tokenizers

Sentence delimiters sometimes need help from tokenizers to disambiguate punctuations. Tokenizers are also known as lexical analyzers or word segmenters. Tokenizers make meaningful units by parsing a stream of characters. The units made by tokenizers are called tokens. Tokens can be described by the separation of words by the white spaces.

Simple approaches may be appropriate for some applications but may also cause insufficient situations. For instance, is “data-base” made of one or two tokens? What about “$1000”, does ‘$’ character a token or shall it be taken with both as a token?

Just white space characters shouldn’t be taken as tokenizers because of it depends on languages. For instance, in French “pomme de tere” means potatoes. In some Far East Asian languages there is no space between words. Even in German language has space between words; there are some exceptions like sentence that is made by compound words. “Lebensversicherungsgesellschaft” means “Life Insurance Company”.

3.2.2 Stemmers and Taggers

Just parsing sentences into words is not sufficient to have lexical analysis. Words must be in the root form, too.

3.2.2.1 Stemmers

In linguistic, stemmers are morphological analysis of terms that has same root form. Root can be found as a record in a dictionary. For instance, “go, go-es, go-ing…etc” words are going to be associated with the root form of “go” and those will be assumed that all words are the same. “A system using stemming conflates derived word forms to a common stem…The main reason for the use of stemming is the hope that through the increased number of matches between search terms and documents, the quality of search results is improved” (Braschler & Ripplinger, 2003).

There are two types of morphological analysis. They are called inflectional and derivational. Inflectional morphology is about the syntactic relations between words of the same part of speech such as “inflate, inflates”. More specifically speaking, inflectual morphology applied on different form of the words that if it is singular/ plural or past/future tense to express grammatical properties of the language.

Derivational morphology expresses the creation of the new words from old ones and tries to show the words in one common root form. Derivation usually involves a change in grammatical category of the word and may also involve a modification to its meaning. For instance, “unkind” is created from “kind”, but has the opposite meaning.

Generating a lexicon for morphological analysis is time consuming and expensive. Many applications such as document retrieval do not need morphological analysis to be linguistically correct. The stemmers that used to analyze this case are called heuristic stemmers.

A heuristic stemmer works by removing affixes and suffixes to form the root of the word. The mostly common used heuristic stemmer is called “Porter’s stemmer”. Porter’s stemmer in English removes the suffixes such as “-ing, -ed” and applies derivational rules such as “-ational, -ation” to remove those suffixes so it gets the

root form of the word, but not always. For instance, the word “organiz-ation”‘s root form will be “organiz”.

3.2.2.2 POS Taggers

POS taggers are based on sentence delimiters and tokenizers. It tags each word in the sentence as “name, adverb, adjective…etc”. For an example the sentence below can be considered.

“**Visiting/***adjective* **aunts/***plural noun* **can/***auxiliary verb* **be/***verb* **nuisance/***noun.***”**

“**Visiting/***present cont. tense* **aunts/***plural noun* **can/***auxiliary verb* **be/***verb* **nuisance/** *noun.***”**

In the first sentence “visiting” is adjective which affects the subject aunt; in second one it is a gerund that takes aunt as an object but both of the sentences are written exactly the same.

If all words could assigned to just a POS tag, then POS tagging would be an easier task. However, just like on the example, words can have different tags according to sentence and POS taggers are responsible for the tagging correct ones. In this example case for the POS taggers tag correct ones, it needed to have more content such as association of another sentence. For instance:

“I ought to invite her, but visiting aunts can be nuisance.”
“I ought to visit her, but visiting aunts can be nuisance.”

There are two approaches for POS tagging. Those are rule based and stochastic.

“A rule based taggers try to apply some linguistic knowledge to rule out sequences of tags that are syntactically incorrect” (Jackson & Moulinier, 2002, pg. 13). For instance, “if a name comes after an unknown term, then tag it as adjective”.

Stochastic POS taggers tag words according to frequency of the word tags in the training text data. For this tagger, words must be tagged manually by hand.

3.2.3 Noun Phrase and Name Recognizers

"Name Entity Recognition is an important subject of Natural Language Processing and is used to classify proper nouns into different types such as person, location and organization names in addition to formulae, date and money definitions" (Dalkılıç, Gelişli & Diri, 2010).

Sometimes users need to go further than POS taggers. If an interesting system wanted to be built that recognizes news from business world, a tool needed that recognizes people and company names and also discovers the association between them.

Noun phrase extractors can be symbolic or statistical. Symbolic phrase extractors usually define rules by questions such as "what generates the phrases" and uses relatively simple heuristics. For instance, in English, most of the noun phrases starts with "the, a, this" and mostly used verbs in the sentence are "is, are, have, has".

Name recognizers, recognize the name in the document and it can classify them according to categories such as "company, organization, place, people". Name recognizers ignore the POS to use original form of the word.

CHAPTER FOUR
INFORMATION RETRIEVAL

4.1 Introduction

"Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored in computers)" (anonymous, 2009, pg. 1).

Information Retrieval (IR) is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web.

"Information Retrieval (IR) can be defined as the application of computer technology to the acquisition, organization, storage, retrieval and distribution of information" (Jackson & Moulinier, 2002, pg 26).

A user may write a query such as "information need" and waits for returning of the relevant documents but this query may not be the best one for user's information needed. The reasons user may not get the information he/ she needed are may be incorrect written query, wrong word selection or misusage of the search engine.

Returned documents for the query are generally evaluated upon if they are more or less relevant to the query but this is not correct. User evaluates documents as relevant or not by the information he/she needed instead of the query. Queries such as "British beef imports" may indicate more than one subject to be searched. Does needed information is "the beefs those are imported from other countries" or "the other countries that imports British beef"? This can be known just by asking to the user.

4.2 Indexing Technology

Information retrieval does not begin with query but indexing. Index pages of books, which are generally at the last pages of the book, include words that show which one is at which page; are well known. Indexing of electronic documents for full text search is more complicated than those last pages of books indexing. Some queries that include more than one word may be searched exactly how it's written. This situation can be solved by indexing all words instead of just keywords or titles of documents.

Index of list that includes each word in the document collection is called "inverted list" or "inverted dictionary". Words are stemmed for the root form of them and then added to list. For each token, there are information that are kept which are below:

- Document number, is the number of documents that includes interested tokens. This is used for inverse document frequency (IDF) which is very useful for statistical computation.

- Total frequency number, is the number of a specific token that exist in whole corpora and can be seen how often that specific token exists.

- Frequency is the number of a specific token in a specific document. This number indicates that if the token is relevant to interested subject.

A piece of an inverted list is shown on figure 4.1.

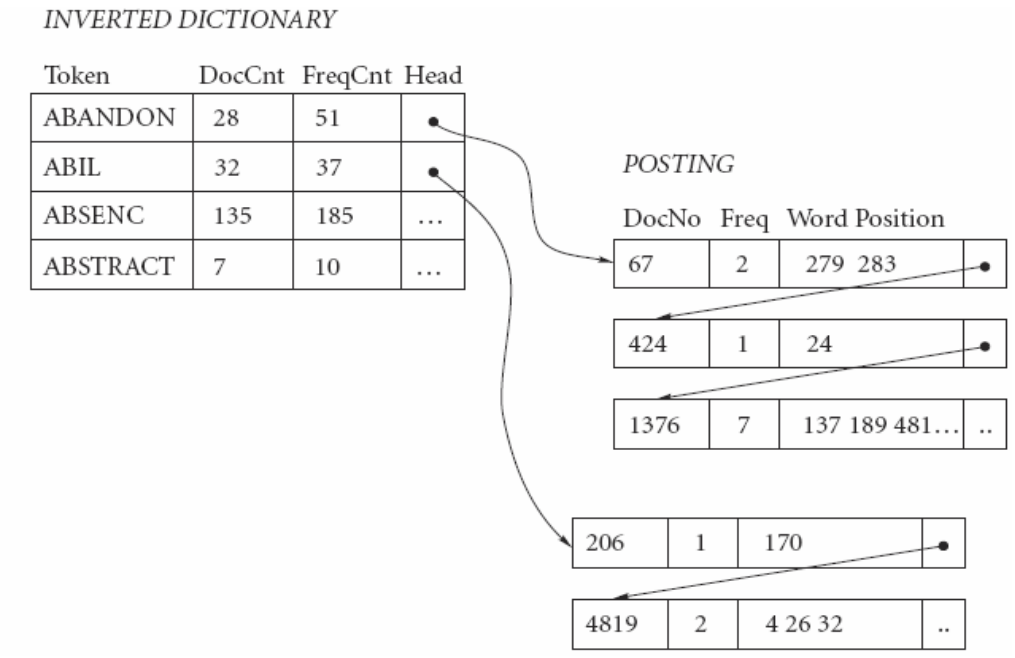


Figure 4.1 A piece of an inverted list

4.3 Query Processing

Query processing first begins with Boolean search but because of its disadvantages it continues with ranked retrieval.

4.3.1 Boolean Search

A Boolean search is searching by the word bonding operators “AND”, “OR” or “NOT” from a database. Operators help to narrow or expand the returned document numbers that includes the words which are searched for. For instance, query “computer AND virus” will return documents that both words exist in the document at the same time.

$$POSTING_{computer} \cap POSTING_{virus}$$

Query “computer OR virus” will return documents that at least one of the words exists in.

$$POSTING_{computer} \cup POSTING_{virus}$$

The operator “NOT” is used when a word does not wanted in from the returned documents. For instance, query “Michael NOT Jordan”.

$$POSTING_{Michael} - POSTING_{Jordan}$$

There should be priority rules in operators because of complication. For instance query “Jordan NOT Michael AND nike” can be interpreted as

$$POSTING_{Jordan} - (POSTING_{Michael} \cap POSTING_{nike})$$

but it must be as below.

$$(POSTING_{Jordan} - POSTING_{Michael}) \cap POSTING_{nike}$$

Most of the boolean systems accept the operators which are not boolean. A query such as “computer /5 virus” will return all documents that 5 words exist between computer and virus. This is useful for name searches. For instance, “President /3 Kennedy” query can be used instead of “President John F. Kennedy” query.

Some of query languages let to use grammatical connectors to search words if it is in the same sentence or paragraph.

When boolean search engines process by those all operators above, some problems come up with it. They can be listed as below.

Large result set contains all documents that satisfies the query, thus, it may be contains so many documents. User can make query by bonding words by operators but it can’t predict the returned document number.

Complex query logic is complexity of the boolean search query that is effective. Simple queries usually return a few or so many documents that can't be examined.

Dichotomous Retrieval is about the unacceptance of relevant degree of the returned documents. Boolean query divides the collection into two subsets: "is relevant to the query" and "is not relevant to the query".

Equal term weight is having equal degree of importance of query words in simple boolean search.

Unordered result set is about the result set is about not ordered due to its relevance degree to query. Documents are ordered some other criterion such as release date of documents. This is useful if the query is about news updates but it is not useful if it is a specific history or information.

4.3.2 Ranked Retrieval

Most of web search engines based on frequency distribution of query terms in the document collection. Roughly speaking, if a term in the query exists in a document more frequently than other documents, it will be assumed that the document is more relevance to the query than others. However, there is a problem occurs in this case. If query includes general words, which is called stop words, such as "and, or, the…etc.", then documents which includes stop words more frequently will have more relevance to the query than others. Usually the stop words don't have meanings.

Boolean interpretation in retrieval task is inadequate and should be developed an alternative model for retrieval task beside Boolean one. Multi-dimensional vector space is generated instead of term set of documents. If each term represents a dimension, and assumed frequency of the term on that dimension is a linear scale, queries can be present as vectors in the result space. For instance, "A dog is an

animal. A dog is a man's best friend. A man is an owner of a dog." Can be shown as vectors as in table 4.1.

Table 4.1 A simple vector presentation of a document

TERM	a	an	animal	best	dog	friend	is	of	man	owner
FREQUENCY	5	2	1	1	3	1	3	1	2	1

This document can be presented as a 10 dimensional vector. (5, 2, 1, 1, 3, 1, 3, 1, 2, 1) vector represents the document.

Similarity of query with documents or similarity of two documents is computed by distance measure. The measure of distance between two vectors such as (5, 2, 1, 1, 3, 1, 3, 1, 2, 1) and (2, 2, 0, 1, 2, 1, 5, 5, 0, 2) can be computed in many ways. The idea of presentation terms in a vector according to their weights caused emerging of many methods in retrieval, indexing and classification tasks. For instance, the similarity of two vectors above can be computed by method of cosine similarity measure shown on eq. (4.1).

$$similarity(V_1, V_2) = \cos(\theta) = \frac{V_1 \cdot V_2}{\|V_1\| \|V_2\|} = \frac{\sum_{i=1}^{n} V_{1,i} \times V_{2,i}}{\sqrt{\sum_{i=1}^{n} (V_{1,i})^2} \times \sqrt{\sum_{i=1}^{n} (V_{2,i})^2}}, \tag{4.1}$$

$$similarity(V_1, V_2) = \frac{44}{\sqrt{55} \times \sqrt{68}} = 0.719$$

The major problem in here is what function will be used for weighting the terms. Three metrics will be taken to solve this problem. One is called *term frequency* which is the number of the query terms that exist in the documents; the other one is called *document frequency* which is the number of documents that includes the query term. The last metric is generated by using document frequency which is showed on eq. (4.2) and called *inverse document frequency (idf).*

$$idf_t = \log\left(\frac{N}{n_t}\right) \tag{4.2}$$

N is the total number of documents in the collection, n_t is number of documents that includes query term *t*. Inverse document frequency measures the sparseness of the term. Weight of term *t* in document vector *d* is computed by eq. (4.3).

$$w_{t,d} = tf_{t,d} \times idf_{t,d}\,, \tag{4.3}$$

The similarity between document vector *d* with the query vector *q* is computed by eq. (4.4).

$$sim(q,d) = \frac{\sum w_{t,d} \cdot w_{t,q}}{\sqrt{\sum \left(w_{t,d}\right)^2} \cdot \sqrt{\sum \left(w_{t,q}\right)^2}}, \tag{4.4}$$

$w_{t,d}$ is the weight of term *t* in document *d*.

$w_{t,q}$ is the weight of term *t* in query *q*

Similarity of the each documents that are retrieved and waiting to be sorted are computed by the similarity formulae above and then rank the documents according to their similarity value from bigger to smaller. The bigger value it has the more relevant it is to the query.

4.3.3 Evaluation of Information Retrieval Systems

For the evaluation of the IR systems, documents are manually read and marked as if it is relevant or not to find out if the returned documents are correct with no missing.

4.3.3.1 Evaluation Studies

There are many test collections are generated. Here are some of those.

- Cranfield collection: Data set that pioneers to evaluate efficiency and accuracy of the IR systems. It is insufficient for the days we are living. It includes 1398 documents and 225 queries. It is generated at the end of 1950's in United Kingdom.
- CLEF (Cross Language Evaluation Forum): This collection was generated for the European languages and intercross language information retrieval systems.
- REUTERS: Reuters-21578 and Reuters-RCV1 is the most used collection for classification of texts. Reuters-21578 includes 21578 news articles. Second version of Reuters, which is Reuters-RCV1, includes 806791 documents.

4.3.3.2 Evaluation Metrics

"Two performance metrics gained currency in the 1960's, when researchers began performing comparative studies of different indexing studies" (Jackson & Moulinier, 2002, pg. 45). Those two metrics are called *precision* and *recall.*

Let it be assumed that there are N documents exist in a collection and n documents are relevant to a specific query. If search of query returns m documents and a of them are relevant then *recall* R and *precision* P is shown on eq. (4.5) and eq. (4.6) according to variables on table 4.2.

$$R = \frac{a}{n}, \tag{4.5}$$

$$P = \frac{a}{m}. \tag{4.6}$$

Table 4.2 The values to compute precision and recall

	Relevant	**Non-Relevant**	**Total**
Retrieved	a (true positive)	b (false positive)	a + b = m
Non-Retrieved	c (true negative)	d (false negative)	c + d = N - m
Total	a + c = n	b + d = N - n	

In other expression *precision* and *recall are:*

$$\Pr ecision = \frac{\#(relevant\_items\_retireved)}{\#(retrieved\_items)},$$

$$\mathrm{Re}\, call = \frac{\#(relevant\_items\_retrieved)}{\#(relevant\_items)}.$$

To express them by words, precision is the number of correctly retrieved documents from all returned documents and recall is the number how many of relevant documents are retrieved out of all relevant ones.

CHAPTER FIVE
TEXT CATEGORIZATION

Internet and electronic mail has become a routine job for people. Sometimes there are messages received that are junk to electronic mails and user may be irritated by them. Those junk-mails were use to categorized manually and to exclude them. Nowadays, it is categorized automatically. Electronic mail programs such as Outlook have rule based categorization which is user can set rules to eliminate junk mails not to receive.

5.1 Classifiers

Text based data are used in classification due to documents' contents. Here some of classifiers are described which are mostly used in text categorization.

5.1.1 Linear Classifiers

Linear classifiers are categorizers that modeled as separators of metric space. It assumes that documents can be sorted in two mutually exclusive classes which are labeled as document is relevant or not. The classifiers correspond to a hyperplane (or a line) that separates negative samples from positive ones. If a document falls one side of the line that means it is relevant; if it falls other side of the line then it is not relevant. If document falls on incorrect side of the line then classification errors occur.

5.1.1.1 Linear Separation in Document Space

"A linear separator can be represented by a vector of weights in the same feature space as documents" (Jackson & Moulinier, 2002, pg. 135). The weights in vector are learned by training data. A general idea in here is to avoid the weight vector from negative sample and make closer to the positive ones.

Documents are represented as feature vectors. Features are typically words of documents in the collection. Sparsely, some methods use expressions or word sequences as features too. Components of a document vector can be 0 or 1 according to if feature exist or not, or can be numeric values such as its frequency in the collection or feature frequency in the document. Mostly term frequency-inverse document frequency (tfidf) weighting is used which is shown on eq. (5.1).

$$tfidf = tf \times idf \tag{5.1}$$

When a new document is classified, user looks how the document close to weight vector. If it is close enough to weight vector then it is classified to the category. This new document's score is obtained by computing dot product of document and vector of weights. For more formerly expressing, if *D* represents the vector of the document which is show in eq. (5.2):

$$\vec{d} = (d_1, d_2, ..., d_n), \tag{5.2}$$

The weight vector as in eq. (5.3)

$$\vec{c} = (w_1, w_2, ..., w_n). \tag{5.3}$$

c represents the class and computing of *D* document's score for class *c* is shown in eq (5.4).

$$f_c(D) = \vec{d} \cdot \vec{c} = \sum_{i=1}^{n} w_i \cdot d_i \tag{5.4}$$

The score computed for membership is numeric value instead of binary such as yes/ no. How assigning document to class *C* is found by setting a threshold θ.

$$f_c(D) \geq \theta$$

If document is close enough determined by the inequality above then it is assigned to relative class.

Weights in category vector are computed by using labeled documents from training data. “Training algorithm for linear classifiers is an adaptation of Rocchio’s formulation of relevance feedback for the vector space model” (Jackson & Moulinier, 2002, pg. 136).

Linear functions are frequently used in information retrieval. Linear functions can be used in probabilistic models, too. An example is shown in eq. (5.5).

$$P(D \mid R_Q = 1) = \sum_{t \in Q} w_{t,d} = \sum_{t \in Q} 1 \cdot w_{t,d} \tag{5.5}$$

$R_Q = 1$ shows that document D is relevant to query Q.

5.1.1.2 Rocchio Algorithm

Rocchio algorithm uses the approach of each document can be assigned only one category. Algorithm is about computing new weight vector *w* by using old one *w’*. The new weight vector’s j^{th} component is computed by eq. (5.6).

$$w_j = \alpha w'_j + \beta \frac{\sum_{D \in c} d_j}{n_c} - \gamma \frac{\sum_{D \notin c} d_c}{n - n_c}, \tag{5.6}$$

n is the number of training examples, *c* is the set of positive examples, n_c is the number of examples in *c*. d_j is the weight of j^{th} feature of document *D*. α, β and γ control weight vector, positive examples and negative examples respectively.

Rocchio algorithm often used in baseline categorization experiments. “One of its drawbacks is that it is not robust when the number of negative instances grows large”

(Jackson & Moulinier, 2002, pg 137). Rocchio algorithm is used when there are a few positive and a few negative examples exist.

In a classification context, there are more documents that do not belong to given class than belong to given class. Many approaches handle this situation by setting arbitrary values to β and γ parameters. For instance, to eliminate negative examples γ is set to "0".

5.1.1.3 Online Learning of Linear Classifiers

Rocchio algorithm is batch learning method which is about the entire set of labeled documents is considered at the same time, thus weight can be computed directly. However, online learning algorithm encounters examples one by one and weights incrementally, computing small changes when a document is presented in. Online learning method is better on dynamic categorizations such as filtering and routing, thus, most of the linear classifiers trained by online learning method.

Generally speaking, online algorithms uses just one example at each time for computing weight vector and updates the weight vector at each step. After i^{th} example included to process, last form of weight vector will be look like eq. (5.7).

$$\overrightarrow{w_i} = \left(w_{i,1}, w_{i,2}, ...w_{i,n}\right) \tag{5.7}$$

In each step new vector $\overrightarrow{w_{i+1}}$ is computed by using old weight vector $\overrightarrow{w_i}$, example $\overrightarrow{x_i}$ and label $\overrightarrow{y_i}$. For all methods, update rule focused on ignoring bad features and prompting good ones.

After linear classifier trained, new document can be classified by using last weight vector $\overrightarrow{w_{n+1}}$. If all weight vectors kept from the beginning, average of the weight vectors can be computed and used for an alternative method instead which is shown in eq. (5.8).

$$\vec{w} = \frac{1}{n+1}\sum_{i=1}^{n+1}\vec{w_i} \tag{5.8}$$

5.1.2 Nearest Neighbor Algorithms

Nearest neighbor algorithm rely on rote learning. At training time, a nearest neighbor classifier remembers each training document and its related features. While classifying a new document D, classifier first gets *k* number of documents from training set which are close to document D. Then one or more categories are picked relevant to *k* documents to assign document D.

Before describing *k-NN* (k nearest neighbor) algorithm, distance metric has to be described which measures how two documents are close to each other. Euclid distance can be used on vector space. The metrics used in search engines to measure how close are the returned documents to the query can be used to measure distance of two documents.

The Euclidean distance $L_E = dist(d^{(A)}, d^{(B)})$ between $d^{(A)}$ and $d^{(B)}$ is which is shown in eq. (5.9).

$$L_E = \sqrt{\sum_{i=1}^{n}\left(d_i^{(A)} - d_i^{(B)}\right)^2} \tag{5.9}$$

(Sojka, 2010, pg. 227).

Let v_1 and v_2 vectors represent two documents. Cosine similarity between two documents is showed in eq. (5.10):

$$sim(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1||v_2|}, \tag{5.10}$$

$$|v_1| = \sqrt{v_1 \cdot v_1}$$

(Han & Kamber, 2006, pg. 619).

Then how documents will be assigned to categories is described. A simple approach is, each document can be assigned to category that the most of closest *k* training documents assigned to.

For more sophisticated classifying documents to single or multiple categories, *weighted k-NN* distance measure can be used. Thus, the more distance between document D with its neighbor, the less probability to be assigned to neighbor's class C_j. This is computed by the eq. (5.11).

$$Score(C_j, D) = \sum_{D_i \in Tr_k(D)} sim(D, D_i) \cdot a_{i,j} \tag{5.11}$$

Score (C_j, D): C_j class score for document D.

Tr_k(D): the set of document D's k nearest neighbors.

sim (D,D_i): similarity between documents.

If document D_i is assigned to class C_j then $a_{i,j} = 1$, else $a_{i,j} = 0$.

"Applying this to binary classification, the best scoring class might be differ from the majority class" (Jackson & Moulinier, 2002, pg. 149).

"k" is usually selected empirically. Generally selection of *k* depends on two cases.

- The case about closeness of the classes in the feature space: if the classes are close to each other then *k* should be a small number.
- The case about what kind of training documents are in a same class: if documents are so heterogeneous, a bigger *k* would be more appropriate.

It is fast to train k-NN classifiers because just one thing to which is representing each document as feature vectors. On the other hand, classification is a slow process because each document has some computations with another one.

5.2 Evaluation of Text Categorization Systems

The methodology for evaluating text classifiers depends on the task that the program is trying to perform. For instance, *routing* and *filtering* may have different evaluation metric. In some routing tasks, if each document has to be sent somewhere, prior metric may be *recall.* In a filtering task, if purpose of the filter is preventing user from seeing certain kinds of document, *precision* metric can be used.

Performance of classification systems usually evaluated according to efficiency. Efficiency metrics for binary classifiers are shown in table 5.1

Table 5.1 Contingency table reflected the assignments performed by a binary classifier

Category C_i	**Expert Assigns Yes**	**Expert Assigns No**	**Total**
Classifier Assigns Yes	TP_i	FP_i	m_i
Classifier Assigns No	FN_i	TN_i	$N - m_i$
Total	n_i	$N - n_i$	N

Precision, which is showed in eq. (5.12), and *recall,* which is showed in eq. (5.13), values are computed by metrics above. Precision is the proportion of documents for which the classifier correctly assigned category C_i. Recall is the proportion of target document correctly classified.

$$P_i = \frac{TP_i}{m_i}, \tag{5.12}$$

$$R_i = \frac{TP_i}{n_i}. \tag{5.13}$$

Recall value can be 100% by assigning yes to all documents, thus classifier wouldn't miss a document for *yes* class but when recall increases, precision decreases and vice versa. The purpose in here is to keep both proportions at high

values. As a result for the evaluation of a classifier, using both metrics will be more adequate.

On this field there are two major measures are used mostly. They are 11 point average precision and F_β measure.

11 point average precision metric is an IR metric and relies on ranking. Its value is the average of precision point taken at the fixed recall values. Recall fixed points are 0 to 1 with the interval of 0.1 values.

This measure is usually used in routing tasks. 11 points average precision is limited due to ranking documents according to categories or vice versa.

F_β Measure is showed in eq. (5.14).

$$F_\beta = \frac{\left(\beta^2 + 1\right) \cdot P_i \cdot R_i}{\beta^2 \cdot P_i + R_i}, \qquad 0 \le \beta \le \infty \tag{5.14}$$

β can be interpreted as relative importance given to precision and recall. Typical value for β is 1, but other values can be used to be biased.

Beside those three metrics, sometimes accuracy metric which is used at machine learning is also used. This metric is proportion of total number of correctly classified to total number of all documents in the collection, which is showed in eq. (5.15).

$$Acc_i = \frac{TP_i + TN_i}{N} \tag{5.15}$$

CHAPTER SIX
AUTOMATIC TEXT SUMMARIZATION

"The process of text summarization can be seen as a data reduction to compress the content of document" (Yu & Ren, 2009). Similarly, automatic text summarization can be described as system that takes original documents as input and makes an output as a short version of document which includes just important parts of original one. "Important" can be interpreted as many ways. The most known is satisfying user's need by a finding relevant documents according to the query.

"Text summarization provides users with summaries of document contents, allowing them to quickly understand the main ideas of documents" (Sornil & Gree-ut, 2006).

"The summaries serve as quick guide to interesting information, providing a short form for each document in the document set; reading summary makes decisions about reading the whole document or not, it also serves as time saver" (Binwahlan, Salim & Suanmali, 2009).

Summarization task is divided to two main categories; abstraction and extraction. Extraction summarization is extracting and joining the most interesting parts of document. Abstract summarization is building summary by different words from original document words. In other words, it is fusion version of extraction summarization.

In both cases, summarization is compressed or data reduction of documents. Extraction approach can be also eliminating the non-interesting parts of document but abstract approach uses more sophisticated methods such as altering specific words with general ones without ignoring the details.

There are two important tasks in text summarization. One is how to extract interesting sentences; the other one is how to order extracted sentences. Edmunson and Luhn pioneered on this field. A general idea is focused on taking structural feature to extract sentences from source document. “Edmunson and Luhn have proposed four sentence features to decide importance of sentences, including the high frequency words (keywords), pragmatic words (cue words), title and heading words, and structural indicators (sentence location)” (Yu & Ren, 2009).

Roughly, structure of text summarization can be described in three steps which are analysis, transformation and synthesis. First step analysis builds an internal representation and that can be formulae or numerical values. Second step internal representation has transformations, which are computations. On the last step natural language is used to represent summary.

Proportion of summary document to the original one is another problem to be solved. Usually 5-30% is used and user may set it according to experiences and length of the document by his/herself. If proportion was always 30% then summary would include so much noisy data. On the other hand if it was always 5% then system would miss important data.

6.1 Summarization by Sentence Selection

A general way to handle tough research task is to degrade it into simpler form. This is done by selecting sentences instead of paragraphs in summarization task. However, paragraphs can be selected to have long summarizes instead short ones. Most of the news articles hide its summary in its first paragraph.

6.1.1 Algorithms for Summarization by Sentence Selection

Automatic text summarization systems usually work by sentence selection methods.

6.1.1.1 A Hybrid Approach to Automatic Text Summarization

K-mixture of connective strength based (KCS) approach is used to enhance the quality of document summary results. K mixture probabilistic model is used to determine the term weights. Then term relationships are computed to find out connective strength of nouns for sentence semantics. Finally sentences with significant connective strengths are extracted to form the summary.

Approach takes noun-noun and noun-verb relations into consideration since the postulated that noun-verb relations are predicate-argument relations within sentence level and noun-noun relations are associated on the discourse level. Word (noun and verb) importance is calculated by inverse document frequency (IDF) metric. Then, the association norms of noun-verb and noun-noun pairs are calculated based on the importance of the words and the distance between each other. The connective strength of a noun is derived by association norm. Finally the average of connection strength of those noun's sentences are computed. Steps of process are below.

1. Documents are parsed into terms; nouns and verbs are taken into root form.
2. The probability of term appearance is computed by Katz's K-mixture probability model. Term *t* in *k* occurrences is showed in eq. (6.1).

$$P_t(k) = (1-\alpha)\cdot\delta_{k,0} + \left(\alpha/B_1\right)\cdot\left(1-\frac{1}{B_1}\right)^{k-1}\cdot(1-\delta_{k,0}) \qquad (6.1)$$

(Chang & Hsiao, 2008)

$\delta_{k,0}$ is set to 1 if k is 0; else $\delta_{k,0}$ is set to 0.

α is the probability of having interested term at least one occurrence

B_1 is the expected number of occurrences among the documents with at least one term occurrence.

Let x be the proportion of documents that includes term t and y be the average number of terms that includes term t. x is used to predict α and y is used to predict (B_1-1). Thus, eq. (6.1) becomes eq. (6.2)

$$P_t(k) = (1-x)\cdot \delta_{k,0} + \left(\frac{x}{y+1}\right)\cdot\left(\frac{y}{y+1}\right)^{k-1}(1-\delta_{k,0}) \tag{6.2}$$

If x and y are showed in TF-IDF form in eq. (6.5) and eq. (6.6). Respectively TF and IDF is also shown on eq. (6.3) and eq. (6.4).

$$TF = \frac{cf}{N}, \tag{6.3}$$

$$IDF = \log_2\left(N/df\right) \tag{6.4}$$

$$y = \frac{cf - df}{df} = TF \times 2^{IDF} - 1 \tag{6.5}$$

$$x = \frac{df}{N} = \frac{TF}{y+1} \tag{6.6}$$

cf is the total number of term *t* occurrences among all documents,
df is the number of documents that includes term *t*,
N is the total number of documents.

After Term occurrence probabilities calculated, term relationship exploration steps are done as well as noun-noun and noun-verb pairs. The term weights are determined by K-mixture probability model.

In addition, distance between terms is also considered because relative terms are usually closer in the same sentences to each other. Distances are computed by the

cardinal numbers. This is done by assigning serial numbers to terms in the sentence. After all, association norms are computed by the eq. (6.7) and eq. (6.8).

$$SNN(N_i) = \sum_j \frac{P_{N_i} \times P_{N_j}}{D(N_i, N_j)} \tag{6.7}$$

$$SNV(N_i) = \sum_j \frac{P_{N_i} \times P_{V_j}}{D(N_i, V_j)} \tag{6.8}$$

Then those norms are sum up as shown in eq. (6.9) to derive connective strengths which are the semantic relationship significance.

$$CS(N_i) = SNN(N_i) + SNV(N_i) \tag{6.9}$$

(Chang & Hsiao, 2008)

Finally importance of sentence is determined by sum of all CS (N_i) which belong to same sentence words.

6.1.1.2 Term Co-occurrence Approach

If two terms are usually exist in the same unit such as same sentence or same paragraph, then they are semantically related. The more they are co-occurred, the more they are related semantically. Relative Co-occurrence degree computed by eq. (6.10).

$$R(w_x \mid w_y) = \frac{f(w_x, w_y)}{f(w_y)} \tag{6.10}$$

(Geng, Zhao, Chen & Cai, 2006)

Where $f(w_x,w_y)$ is the number of word *x* and *y* are in the same unit and $f(w_y)$ is the frequency of the word *y* in document.

$R(w_x \mid w_y)$ and $R(w_y \mid w_x)$ are not the same, thus co-occurrence degree is computed by eq. (6.11).

$$C(w_x, w_y) = \frac{R(w_x \mid w_y) + R(w_y \mid w_x)}{2} \tag{6.11}$$

Here are the short steps of co-occurrence approach to automatic text summarization as shown in the figure 6.1

1. Document is parsed, co-occurrence are computed and subjects are derived.
2. Weights of terms are computed and important terms that represent subjects are extracted.
3. Important terms are used to extract subject sentences and summary is generated by those extracted sentences.

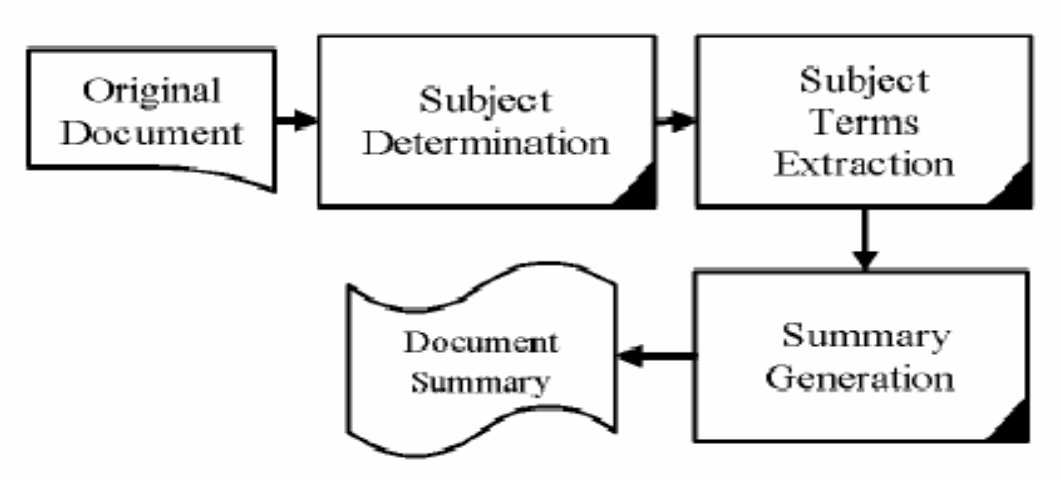


Figure 6.1 Steps of co-occurrence process

After removing meaningless words and stemming the rest, terms are listed to represent document D. Weight of each term is calculated based on the term frequency. First n terms are selected according to their related weights which have higher value of weight. Those n terms are node set of D_{hf} and generate co-occurrence graph G which is shown on figure 6.2.

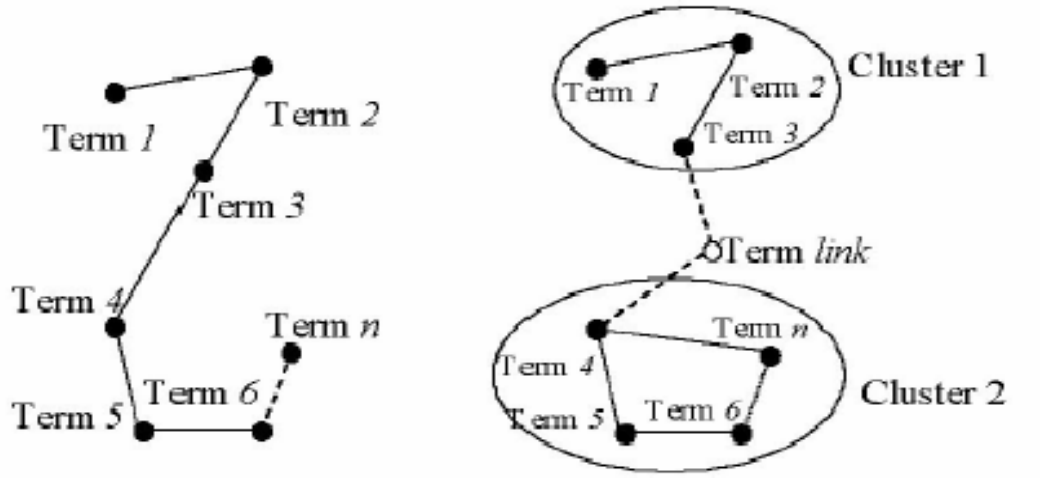


Figure 6.2 Connected graph G and disconnected by clustering graph G'

All co-occurrence degrees between terms in D_{hf} are calculated by eq. (6.11) and then the pairs are linked in graph G. If graph G is singly connected, then document D expresses only one subject; if not, graph G is divided into two or more connected slices. Between different clusters, there could exist some linkage terms called *term link* as bridges to connect different clusters. Those term links are extracted by eq. (6.12) and eq. (6.13).

$$Linkage(w, g) = \sum_{w_g \in g, w_g \neq w} C(w, w_g) \tag{6.12}$$

$$GLinkage(g) = \sum_{w_g \in g} \sum_{w \in D, w \neq w_g} C(w_g, w) \tag{6.13}$$

w_g is a term in cluster g, w is a term in document.

Linkage (w,g) reflects linkage degree between w and g.

GLinkage (g) reflects the significance of cluster in document D.

Significance of each term is also computed by the eq. (6.14).

$$Link(w) = 1 - \prod_{g \subset G} \left(1 - \frac{Linkage(w, g)}{GLinkage(g)} \right) \tag{6.14}$$

The *m* terms with largest Link(w) are chosen to form the term link set D_{lk}. The *m* terms are added to graph G if they are not existed and the new co-occurrence graph G' is generated. $C(w_i,w_j)$, between w_i in D_{lk} and w_j in D_{hf}, is calculated. The *m-1* term pairs are added to graph G' as dotted edge; as shown in figure 6.2.

The next step is to calculate the information gain. *Gain(w)* of each term w in graph G' is calculated by eq. (6.15).

$$Gain(w) = \sum_{(w,w')\in G'} C(w,w') \tag{6.15}$$

E(G') is the edge set (both dotted and solid) in graph G' which is shown on figure 6.2. The top k terms with the largest Gain(w) are chosen as subject terms, Gain(w) is the term weight.

After all steps done, summary is generated by sentences according to their significance. Thus, each sentence in document D needed to be assigned with a weight $w(s_i)$ to measure its significance, $w(s_i)$ is decided by the factors below.

- The bigger the sum of all term weights in a sentence, the greater the possibility of a higher significance attached to the sentence. Here, just the subject terms are taken into consideration. To eliminate the influence of sentence size, the sentence significance should be calculated by dividing the sum of all term weights by the number of terms in a sentence.
- Sentences in preface and conclusion usually have higher significance. If the sentence has prompting terms such as "to conclude" or "in conclusion", then the sentence could be conclusion of the original document. Thus it has higher significance.
- If sentence begins with detailing words such as "for example", then it has a lower significance.

$$w(s_i) = LC(s_i)\times CC(s_i)\times EC(s_i)\times \frac{\sum w_{ki}}{|s_i|} \tag{6.16}$$

$|s_i|$ is the number of terms in i^{th} sentence.

$\sum w_{ki}$ is the sum of all significance of all subject terms in i^{th} sentence.

$\frac{\sum w_{ki}}{|s_i|}$ is the average weight of s_i.

$LC(s_i)$ is the parameter of location weight which is

LC(preface sentence) = LC(conclusion sentence) = 1.3

LC(paragraph begining) = LC(period ending) = 1.1

$CC(s_i)$ is the parameter of term weight which is set to 1.5

$EC(s_i)$ is the proportional factor of detailing term weight and is set to 0.5

By the eq. (6.16), significant sentences are extracted and summary is generated.

6.1.1.3 Cover Coefficient Based Approach

In Cover Coefficient (CC) based method, first of all document is parsed into sentences and sentences are parsed into terms. After stemming procedure done, CC matrix is generated by those sentences and terms. CC matrix, which is denoted by C, rows are sentences and columns are the terms of the document. Each element in C matrix which is denoted by $c_{i,j}$, can be read as how much s_j covers s_i.

$$c_{i,j} = \sum_{l}^{n} \alpha_{i,l} \cdot \beta_{l,j}, \qquad 1 \le i, j \le m \tag{6.17}$$

(Gonenc & Fazli, 2009)

$c_{i,j}$ is the probability of how j^{th} sentence is covering i^{th} sentence.

$\alpha_{i,l}$ is the probability that selecting l^{th} term from i^{th} sentence.

$\beta_{l,j}$ is the probability that l^{th} term occurs in j^{th} sentence.

n is the term frequency in document.

m is the number of sentences in document.

After all members of C matrix computed sum of all $c_{i,j}$ with constant i will be equal to 1. So values of the diagonal of *cover coefficient matrix* are the dissimilarity to the others. To compare i^{th} sentence to others, $(\psi_i=1-c_{ii})$ formulae is used. ψ_i shows how much other sentences cover sentence i. A summary proportion is set. Then, the number of summary sentences is considered by the proportion. After all highest values of ψ_i are selected for the summary.

Example of Cover Coefficient Based Method: Document D has three sentences that those are S_1, S_2 and S_3. After tokenization terms of sentences are below and sparse matrix of terms are shown on table 6.1.

S_1: "document frequency constant"
S_2: "term frequency document"
S_3: "repeated term"

Table 6.1 Sparse matrix of sentences

	Constant	**document**	**frequency**	**repeated**	**term**
S1	1	1	1	0	0
S2	0	1	1	0	1
S3	0	0	0	1	1

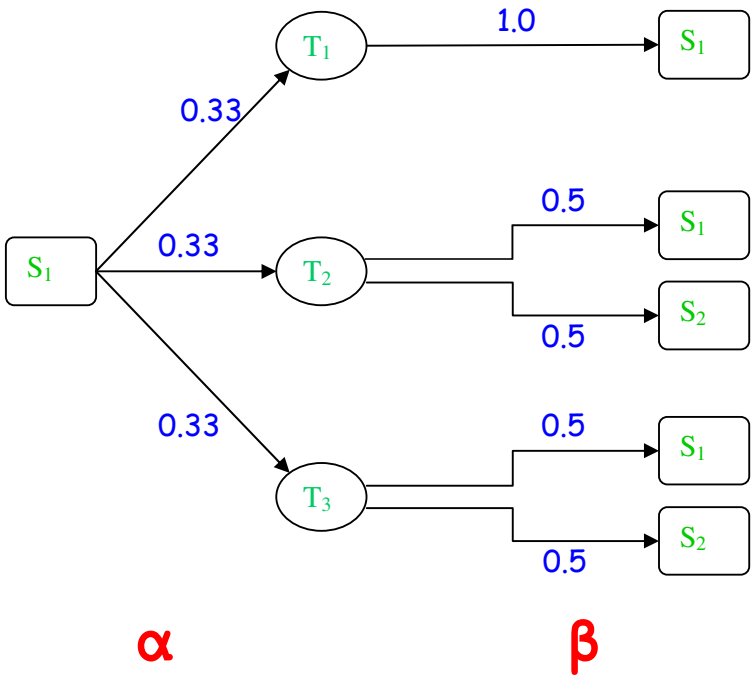


Figure 6.3 Diagram to calculate c_{11} , c_{12} and c_{13}

Calculation of first row of C matrix is shown by a diagram on figure 6.3. By the eq. (6.17), cover coefficient matrix is evaluated as below.

$$C = \begin{bmatrix} 0.66 & 0.33 & 0 \\ 0.33 & 0.5 & 0.17 \\ 0 & 0.25 & 0.75 \end{bmatrix}$$

If the summary proportion set to 0.3 then sentence 2 will be the summary of the document.

6.2 Evaluation of Automatic Text Summarization Programs

It is really hard to evaluate a summary that generated by a computer program. A summary must include the important part of the original document and it has to be read easily. Sentence selected summaries will catch the important part of original document but it is not easy to read. On the other hand paragraph selected summaries can be read easily but it will have noisy data.

Usually two methods are used to evaluate summarizer programs. One is to compare summary that generated by computer program with summary that generated manually by an editor or expert. This is called intrinsic evaluation and more often this evaluation is used. The other evaluation method, which is called extrinsic method that evaluates how summary is helpful to user in information processing task.

When news articles' summary are generated proportion of 10%, the agreement of editors that all agrees the same sentences generate the summary can be as high as %95 but when the proportion is 20 %, agreement ratio can be decreased.

Some experiments showed that terms in two summaries of same document, that generated by two editors manually, fits each other's term about 40-50%. "If humans are unable to agree on which paragraphs best represent an article, it is unreasonable to expect an automatic procedure to identify the best extract, what ever that might be" (Mitra, Singhal & Buckley, 1997). Even though, comparing with a hand-made summary and computer-made summary overlaps more than two hand-made summaries.

Extrinsic method threads summarizer system as last process of the IR engine. Summary is assumed like it is a query and user decide if document is related to query (summary). The performance on this task is evaluated by wasted time, accuracy of decisions and sometimes by reliability in decision makings. On this case there is an assumption which is user can consider the summary if it is relevant to the original document quickly. Evaluation of summarization technology can be relative until the question "how the best summary can be generated?" has an exact answer.

6.3 Application of Automatic Text Summarization

In application, there is a software coded on "C#" by using *Cover Coefficient Based* algorithm to generate Turkish document summaries. However, there is no evaluation corpus for Turkish languages. Therefore, an evaluation study was done on

ten Turkish news articles. Those articles are selected from seven different categories. Three articles from world category, two articles from technology, one from finance, one from sport, one from health, one from politics and one from journal category were selected. Articles are picked from online Turkish news websites such as and www.hurriyet.com.tr. News articles are on appendixes 1-10.

Four people are picked for manually generated summaries. Those four people are students who studies in different branch of sciences. One studies sociology, one studies econometrics, one studies chemistry and one studies statistics science. In other words, there are four people who have different vintages. Summaries are generated by selecting sentences without any changes on sentences. Summaries are on appendixes 11-14.

News articles have many categories and read by people whom have different vintages, thus, manually generated summaries are different from others according to people. This situation can be an explained by the words “even if people’s summaries cover each others’ about 40-50%, how could people expect machines to generate a general summary for the documents” which was mentioned previously.

For an accurate summary to be generated by people, interesting news were selected because people could get bored and want to finish selecting sentences fastly without considering if the sentences are summary sentences that represent the article. On the other hand, some articles are hard to understand because of their contents. However, to evaluate software, those articles had to be considered for summarization.

In journal news article, some people focused on situation or people whom are mentioned in the article, but the other people had read it from different angles. Thus, different sentences are selected to generate summary. However, there were common sentences that all the people selected.

In Technology category, the longest article which is about internet attacks, some people focused on the hackers and the defenders. However, other people focused on the reason and the methods of attacks. On the other news article in technology category is about one of the biggest electronic company's hacked website. News article tells about the victimization of people whom have an account on the hacked website and sadness of company managers' because of the situation. It also tells what must website members do and they will have free game add-ons as gifts. People, who generate summaries, selected sentences according to their vintages and common sentences are considered as summary sentences of the article.

In the world category, one of them's summaries have common sentences mostly according to others which is about the schools of Gülen's. On the other hand, software is also selected the same common summary sentences and the evaluation results were the best for the others'. The shortest news article is also in world category which is about the mixing the rice genes in Japan. Because of its shortness, it seems people whom are generating summaries could not decide which sentence could be the summary sentence. However, there were common sentences within manually summaries and also within automatically generated summary.

In politics, financial and sports category, people whom are generating summaries selected different sentences. Thus, summaries had low number of sentences that are common. Because of this reason, evaluation of this article does not have satisfactory results. This situation can be explained the information that has to be known previously especially for the sports and financial news articles.

As expected, first sentence or paragraph of the articles was selected for the summaries, because usually those first sentences or paragraphs hide a piece of article summary inside of them. The reason of first sentence or paragraph hide a piece of summary is to tell readers about the article in a fast way if readers want to read the whole article or not.

In the software, first of all, articles were splitted into sentences due to punctuations. Articles were also splitted into words as terms. After splitting articles, terms are trimmed to its root form by a stemmer for Turkish words. Root formed terms are stored by their locations and frequencies. By that information, the summarization algorithm is applied on the articles. Summarizations are generated and evaluation metrics are computed by comparing manually generated summaries. For evaluation, *precision, recall* and *F-measure* formulas are used. The code of software is shown on Appendix 16 and the general summaries are shown on Appendix 15.

Summaries were generated by software with and without a stemmer for Turkish languages called *Zemberek.* Degree abbreviations are ignored to reduce the ambiguity of punctuations that indicate end of sentences. Summary proportion was set to 30%. The evaluation was done by comparing manually generated general summaries which includes just the common sentences with automatically generated summaries by the software. Also, People's summaries were compared to automatically generated summary one by one without considering other ones'. The results are shown on Table 6.2. Descriptive statistics such as mean, max and min values of each precision, recall and F-measure are shown in table 6.3 and the graphs of each are shown on figure 6.4, figure 6.5 and figure 6.6 respectively.

Table 6.2 Evaluation results for news articles which are shown in appendix 1-10

Documents	Summaries	With Stemmer			Without Stemmer		
		Precision	Recall	F Meas.	Precision	Recall	F Meas.
World 1	General	0.75	0.60	0.67	0.75	0.60	0.67
	Person 1	0.50	0.50	0.50	0.50	0.50	0.50
	Person 2	0.50	0.67	0.57	0.75	1.00	0.86
	Person 3	0.50	0.33	0.40	0.50	0.33	0.40
	Person 4	0.75	0.43	0.55	0.50	0.29	0.36
World 2	General	0.67	0.43	0.52	0.56	0.36	0.43
	Person 1	0.56	0.45	0.50	0.45	0.36	0.40
	Person 2	0.22	0.33	0.27	0.33	0.50	0.40
	Person 3	0.22	0.18	0.20	0.22	0.18	0.20
	Person 4	0.33	0.27	0.30	0.22	0.18	0.20

Table 6.2 Continue

World 3	General	0.67	0.50	0.57	0.67	0.50	0.57
	Person 1	0.67	0.50	0.57	0.33	0.25	0.28
	Person 2	0.67	0.67	0.67	0.67	0.67	0.67
	Person 3	0.67	0.67	0.67	0.67	0.67	0.67
	Person 4	0.67	0.40	0.50	0.67	0.40	0.50
Finance	General	0.67	0.40	0.50	0.00	0.00	0.00
	Person 1	0.33	0.33	0.33	0.00	0.00	0.00
	Person 2	0.33	0.25	0.29	0.00	0.00	0.00
	Person 3	0.33	0.25	0.29	0.33	0.25	0.29
	Person 4	0.33	0.20	0.25	0.33	0.20	0.25
Journal	General	0.80	0.50	0.62	0.40	0.25	0.31
	Person 1	0.40	0.29	0.33	0.20	0.14	0.17
	Person 2	0.20	0.33	0.25	0.20	0.33	0.25
	Person 3	0.40	0.29	0.33	0.20	0.14	0.17
	Person 4	0.40	0.50	0.44	0.20	0.25	0.22
Sports	General	0.29	0.42	0.34	0.32	0.58	0.41
	Person 1	0.06	0.17	0.09	0.09	0.33	0.14
	Person 2	0.00	0.00	0.00	0.05	0.17	0.07
	Person 3	0.29	0.42	0.34	0.32	0.58	0.41
	Person 4	0.47	0.50	0.48	0.36	0.50	0.42
Health	General	0.44	0.44	0.44	0.57	0.57	0.57
	Person 1	0.22	0.29	0.25	0.22	0.29	0.25
	Person 2	0.11	0.21	0.15	0.11	0.25	0.15
	Person 3	0.67	0.46	0.55	0.67	0.46	0.55
	Person 4	0.33	0.43	0.38	0.33	0.43	0.38
Politics	General	0.40	0.25	0.31	0.60	0.38	0.46
	Person 1	0.20	0.14	0.17	0.40	0.29	0.33
	Person 2	0.40	0.50	0.44	0.40	0.50	0.44
	Person 3	0.20	0.17	0.18	0.40	0.33	0.36
	Person 4	0.40	0.22	0.29	0.6	0.33	0.43
Technology 1	General	0.63	0.33	0.43	0.63	0.33	0.43
	Person 1	0.25	0.20	0.22	0.25	0.20	0.22
	Person 2	0.63	0.50	0.56	0.75	0.60	0.67
	Person 3	0.38	0.25	0.32	0.25	0.17	0.20
	Person 4	0.50	0.27	0.35	0.50	0.27	0.35
Technology 2	General	0.75	0.43	0.55	0.25	0.14	0.18
	Person 1	0.25	0.25	0.25	0.00	0.00	0.00
	Person 2	0.25	0.25	0.25	0.25	0.25	0.25
	Person 3	0.75	0.43	0.55	0.50	0.29	0.36
	Person 4	0.75	0.43	0.55	0.25	0.14	0.18

Table 6.3 Maximum, minimum and mean values of evaluation metrics of general summaries

	Precision With Stemmer	Recall With Stemmer	F-measure with Stemmer	Precision Without Stemmer	Recall Without Stemmer	F-measure Without Stemmer
Mean	0.607	0.43	0.495	0.475	0.371	0.403
Minimum	0.290	0.250	0.310	0.000	0.000	0.000
Maximum	0.800	0.600	0.670	0.750	0.600	0.670

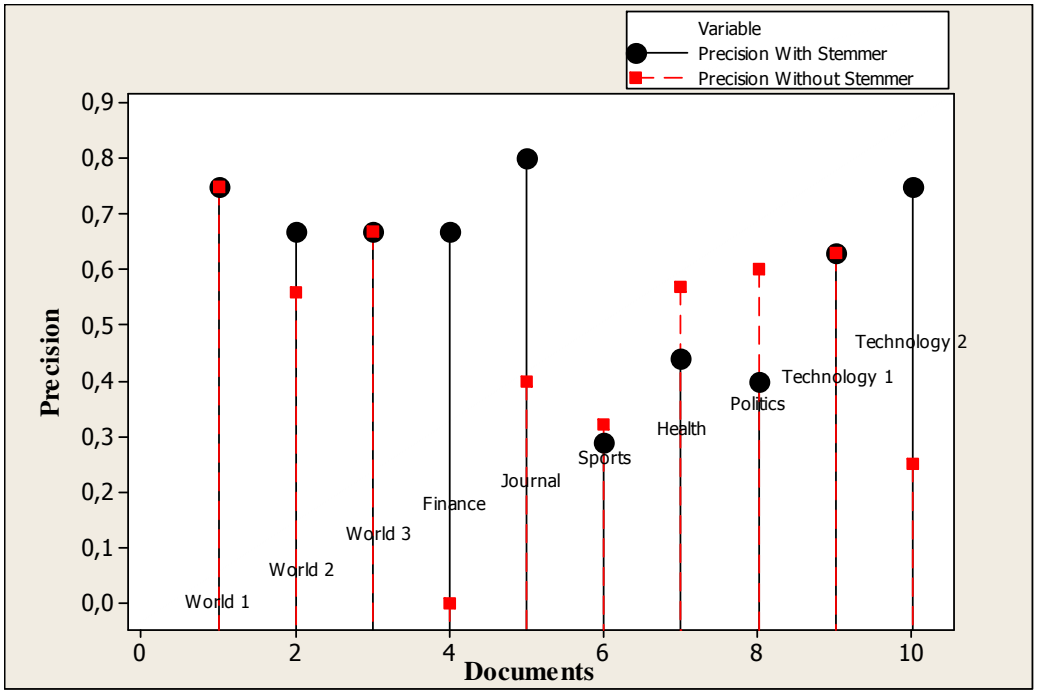


Figure 6.4 Graph of precision values of document summaries with and without stemmer

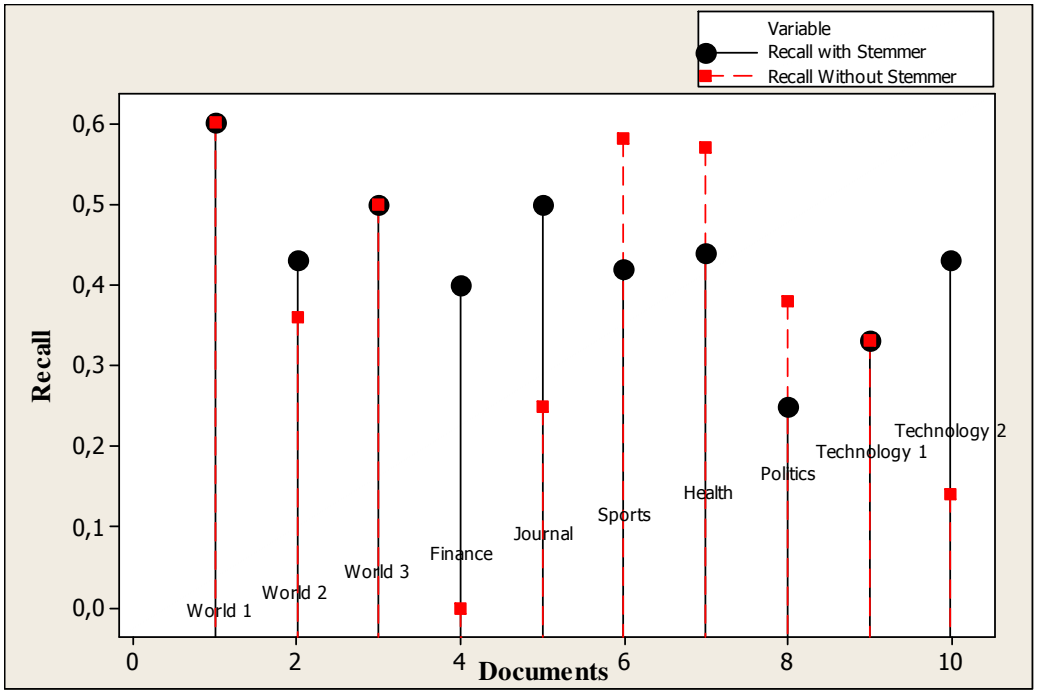


Figure 6.5 Graph of recall values of document summaries with and without stemmer

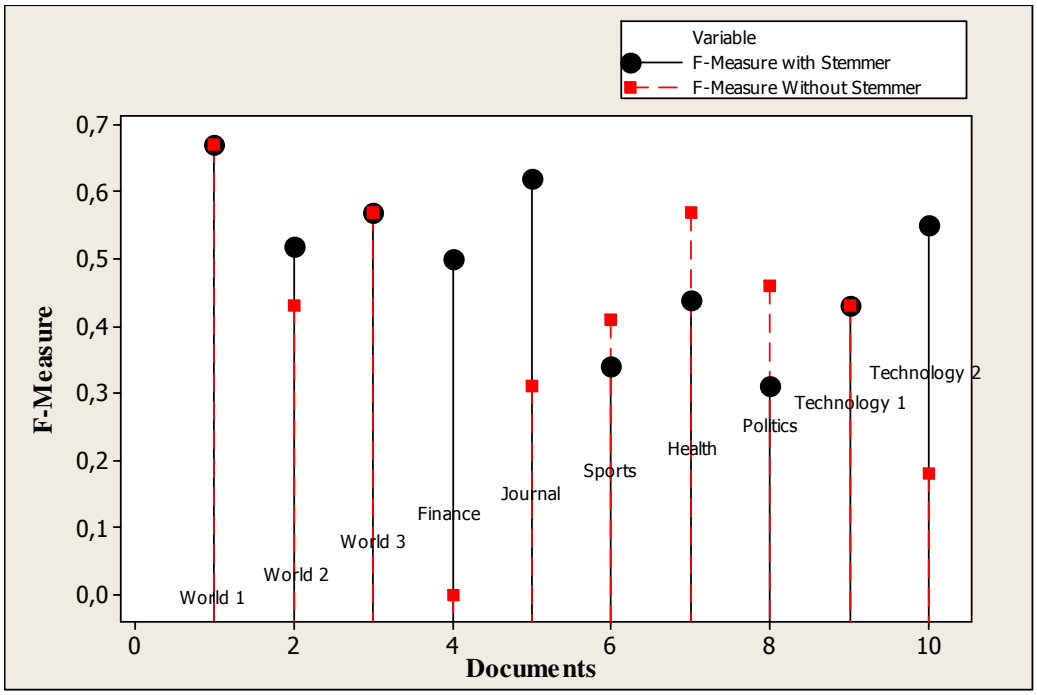


Figure 6.6 Graph of F-measure values of document summaries with and without stemmer

CHAPTER SEVEN
CONCLUSION

In this thesis, the disciplines of text mining, and also text mining, explained in details. The applications of text mining, and how the applications used in text mining text mining, are mentioned. A study about automatic text summarization, which is one of the applications of text mining, is done.

The scope of study is to summarize Turkish text documents and evaluate them to see how the method that used in automatic text summarization is good. There is no evaluation corpus for Turkish texts on automatic text summarization, thus, a small corpus is prepared for the evaluation of the study.

Documents are taken from a couple of online Turkish news websites which are in seven categories. There are ten news articles and four people chosen for the test set. Those documents are summarized by four people by just sentence selecting method. General summaries are generated by selecting common sentences which are selected by at least two people of four.

For the software, *cover coefficient based* method is used and coded on c sharp programming language. There is also a stemmer, which is for Turkish languages called *Zemberek*, integrated into software. Software generated summaries with and without stemmer to see difference of stemmer and compared to general summaries.

Generally, evaluation values of summaries with stemmer are higher than summaries that are generated without stemmer. However, on some news articles, without stemmer generated summaries have better results due to with stemmer generated ones. This reason of situation is about the root form of the words that can not found by the stemmer. For instance, stemmer finds the root form of the word “aylar”, which means “months” in Turkish, as “ayla”, which a female name in Turkey, instead of “ay”, which means “month” in Turkish. Thus, incorrect word is considered in analyze. On some news articles the evaluation values stayed still.

However, it does not mean that software generated summaries with the same sentences, thus, same sentences might not be compared to general summaries.

Articles that values did not change are “World 1, World 3 and Technology 1”. Articles’ summaries that generated without stemmer which have higher values are “Politics, Sports and Health”. The rest of the articles have higher evaluation values with the stemmer. Generally, integrating a stemmer into software makes summaries better.

The best F-measure value with the stemmer is 0.67 which belongs to “World 1” news article. “World 1” s F-measure value without the stemmer used is also the highest one and the same value with 0.67. The lowest value of F-measure with the stemmer is 0.31 which belongs to Politics news article’s summary. Without stemmer, the lowest value of F-measure belongs to Finance article which is 0. However, with the stemmer, Finance’s F-measure value is 0.5. Articles that have lowest values with the stemmer have better values without stemmer generated summaries which are mentioned above.

The article “Journal” s summary generated by stemmer has the highest value of precision which is 0.8. The lowest value of precision with stemmer is 0.29 which belongs to summary of Sports. In without stemmer generated summaries, the lowest value of precision is 0, which belongs to Finance’s summary again. However, Finance’s summary’s precision value with the stemmer is 0.67 which is a good result for automatic text summarization evaluation. In without stemmer generated summaries the best precision value belongs to summary of “World 1”, which is 0.75.

On recall values, the highest value with stemmer used is 0.6 which belongs to “World 1” news article. “World 1” article is also has the highest recall value without the stemmer used which is also 0.6. Politics has the lowest recall value in summaries that generated with the stemmer which is 0.25. Without the stemmer, Finance has the lowest value of recall which is 0 again. However, its’ recall value is 0.4 with the stemmer.

In addition, all people's summaries are also evaluated one by one. Of course, because of not being general summaries of documents, the best and the worst evaluation values belong to people that evaluated just on their own summaries. The worst evaluation values belong to "Technology 2" news article's summary which is generated by "Person 1" without the stemmer, and all the values are 0. Those values are 0.25 with the stemmer generated summary. The best values belong to "World 1" article summary which is generated by "Person 2" without stemmer generated summary.

In conclusion, the evaluation values of summaries that are generated with stemmer have better results. If evaluation corpus was generated by more than four people, results would be better. However, it would cost much more. It is hard to find people that voluntarily do this job.

REFERENCES

Anonymous. (2009). *An Introduction to Information Retrieval* (online ed.). Cambridge, England: Cambridge University Press.

Binwahlan, M. S., Salim, N. & Suanmali, L. (2009). Swarm Based Text Summarization. *Computer and Information Technology-Spring Conference,* 145-150.

Braschler, M. & Ripplinger, B. (2004). How Effective is Stemming, and Decompounding for German Text Retrieval. *Information Retrieval, 7* (3-4), 291-316.

Chang, T. & Hsiao, W. (2008). A Hybrid Approach to Text Summarization. *Computer and Information Technology – IEEE International Conference,* 65-70.

Dalkılıç, F. E., Gelişli, S. & Diri, B. (2010). Named Entity Recognition from Turkish Texts. *Signal Processing and Communications Applications Conference (SIU) – 18th IEEE International Conference,* 918-920.

Feldman, R. & Sanger, J. (2007). *The Text Mining Handbook (1st ed.).* NewYork: Cambridge University Press.

Geng, H., Zhao, P., Chen, E. & Cai, Q. (2006). A Novel Automatic Text Summarization Study Based on Term Co-occurrence. *Cognitive Informatics – 5th IEEE International Conference,* 601-606.

Gonenc, E. & Fazli, C. (2009). Cover Coefficient-Based Multi-document Summarization. *Proceedings of 31st European Conference on IR Research on Advances in Information Retrieval,* 670-674.

Han, J. & Kamber, M. (2006). *Data Mining Concepts and Technics* (2nd ed.), U.S.A.: Elsevier Inc.

Hunter, J. E. (2009). *Classification Make Simple* (3rd ed.), U.S.A.: Ashgate Publishing Company.

Jackson, P. & Moullinier, I. (2002). *Natural Language Processing for Online Applications Text Retrieval, Extraction, Categorization* (1st ed.), Philadelphia: JohnBenjamins Publishing Company.

Oflazer, K. & Bozşahin, H. C. (2006). *Türkçe Doğal Dil İşleme.* Retrieved 20 April, 2011 from http://turkoloji.cu.edu.tr/DILBILIM/dilbilim_ana.php.

Mitra, M., Singhal, A., & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing & Management*, *33* (2), 193-207.

Reynar, J. C. & Ratnaparkhi, A. (1997). A Maximum Entropy Approach to Identfying Sentence Boundaries. *Association for Computational Linguistics - Proceedings of 5th Conference on Applied Natural Language Processing,* 16-19.

Sornil, O. & Gree-ut, K. (2006). An Automatic Text Summarization Approach using Content-Based and Graph-Based Characteristics. *Cybernetics and Intelligent Systems– IEEE Conference*, 1-6.

Tan, A. (1999). *Text Mining: The state of the art and the challenges.* Retrieved Mar. 2, 2011 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.38.7672

Yu, L. & Ren, F. (2009). A Study on Cross-Language Text Summarization Using Supervised Methods. *Natural Language Processing and Knowledge Engineering International Conference - IEEE,* 1-7.

Appendix 1. Turkish News Article from World Category (1)

The New York Times (NYT),'Türkiye bağlantılı okullar Teksas' ta büyüyor' başlıklı haberinde Gülen Hareketi' nin ABD' deki eğitim ağlarını ve öğretmenler için alınan özel vizeleri gündeme getirdi.

ABD' nin en saygın gazetelerinden The New York Times (NYT), birinci sayfasından yayımladığı 'Türkiye bağlantılı okullar Teksas' ta büyüyor' başlıklı haberinde Gülen Hareketi' nin ABD' deki eğitim ağlarını ve öğretmenler için alınan özel vizeleri gündeme getirdi.

NYT, yayımladığı bir şemaya da, cemaatin ABD' de kurulu dernek ve vakıflarını gösterip, hepsini şemanın tepesinde bulunan Fethullah Gülen fotoğrafına bağladı.

Gazetede birinci sayfada verilen haber, 20 sayfada tam, 21. sayfada da yarım sayfa olarak yer alırken, haber NYT' nin internet sitesinde de yaklaşık 7 sayfa yer buldu.

Haberde Gülen Hareketi' nin Charter adı verilen ve kamu kaynaklarıyla işletilen özel okullardan, inşaat faaliyetlerine, özel vize ile getirilen öğretmenlerden, okullardaki öğrencilerin başarısına kadar dek uzanan konular işlendi.

NYT, TMD adlı bir inşaat firmasının kuruluşunun üzerinden bir ay geçmeden Gülen' e yakınlığıyla bilinen Harmony (Uyum) Okulları'nın 8.2 milyonluk inşaat ihalesini kazandığını ve bu firmanın Türkiye ile bağlantılı olduğunu öne sürdü.

TMD' nin 2009 yılından bu yana 50 milyon dolarlık inşaat ihalesi aldığını yazan gazete, Harmony Okulları'nın 16 bin öğrenci ve 33 şube ile Teksas' taki en büyük kamu kaynakları kullanan özel okullar zinciri olduğunu ve yıllık 100 milyon dolar yardım aldığını yazdı.

Stephanie Saul imzalı haberde Fethullah Gülen' in karizmatik bir Türk Vaizi olduğu ve İslam' ın ılımlı yüzünü tüm dünyaya yayarak dinsel, sosyal ve milliyetçi bir hareket kurmak için kendini adadığı dile getirildi.

Gülen hareketiyle doğrudan bağlantılı olduğu belirtilen ve ABD' nin 25 eyaletinde 120 okul bulunduğu yazılan haberde, okulların Amerikan öğrencilerinin genellikle başarısız olduğu bilim ve matematik konularında ağırlıklı olarak eğitim verdiği vurgulandı.

Haberde 'Charter' adı verilen ve Harmony adıyla bilinen bu okulların öteki kamu okullarına göre öğrenci başına 1-2 bin dolar arası daha az maliyet oluşturdukları için devlet tarafından tercih edildiği belirtildi.

Harmony Okulları'nda 2011 yılında 1.500 öğretmenin istihdam edildiği ve bunların 292'sinin 'yüksek nitelikli eleman' olarak nitelenen 'H-1B' vizesi sahibi olduğu yazıldı.

ABD Federal Çalışma Bakanlığı'nın 'H-1B' vizesi sahibi bu Türk öğretmenlerin bir bölümünün yeterince deneyimli olmadığı ve okul yöneticilerinin çevrelerindeki Amerikalı öğretmenleri çalıştırmak istemediklerine yönelik iddiaları incelediği de anımsatılan haberde bazı işçi sendikalarının bu durum nedeniyle okullara tepkili olduğu belirtildi.

Appendix 2. Turkish News Article from World Category (2)

Uluslararası Havayolu Taşıyıcıları Birliği' nin (IATA) bu yılki genel kuruluna, havalimanlarında gelecekte güvenlik kontrolünü hızlandıracak yeni uygulamalar damga vurdu. Yeni uygulamada, yolcular güvenlik risklerine göre üçe ayrılacak. Böylece ayakkabı ve laptop çıkarma son bulacak.

Singapur' da yapılan Uluslararası Havayolu Taşıyıcıları Birliği' nin (IATA) bu yılki genel kurulunda havalimanlarında gelecekte güvenlik kontrolünü hızlandıracak yeni uygulamalar tanıtıldı. 'Geleceğin kontrol noktası' olarak adlandırılan uygulamada, yolcular güvenlik risklerine göre üçe ayrılacak ve 6.1 metrelik koridordan yürüyerek geçecek. Yolcu yürürken, üzerindeki her şey, kabine alacağı çanta, yüksek hassaslıktaki özel cihazlar tarafından taranacak. Sıvı ve elektronik cihazlar çantadan çıkartılmadan kontrol edilebilecek. Bu da güvenlik işlemlerini hızlandıracak. Ayakkabı ve çantadan laptop (dizüstü bilgisayar) çıkarma son bulacak.

IATA' nın planına göre bu teknoloji önümüzdeki yıldan itibaren havalimanlarında kullanılmaya başlanacak. 5 yıl içinde de hızla yaygınlaşacak. Bu sayede uzun güvenlik kuyrukları tarihe karışacak. Yolcular fazla beklemeden hızla güvenlik işlemlerini yaptıracak ve uçağa binebilecek. Yolcular geldiği ülkeye göre, daha önceki bilgileri doğrultusunda bilinen yolcu oluşlarına göre normal veya gelişmiş arama yapılan bölümlerden geçecek. Benzer bir sistemin ABD' de güvenlik kontrollerini yapan TSA tarafından da geliştirildiğine dikkat çeken Amerikan Ulaşım Güvenlik Dairesi Başkanı John Pistole, "Bu sistem, havacılığın ortak kullanımına açılacak. Güvenlik kontrolleri hızlanacak. Hem yolcu konforu artacak, hem de havayollarının geç kalan yolculardan kaynaklanan rötarlarında ciddi azalma olacak" dedi.

IATA ayrıca pasaport kuyruklarının da azaltılması için çipli pasaport uygulamasına hızla geçilmesini tavsiye ediyor. Yolcunun bilgilerinin yanı sıra vize bilgilerinin içinde yer aldığı çipli pasaport da kontrolleri hızlandıracak. Ayrıca yolcunun göz bebeğinin taranmasıyla kimlik bilgileri kontrol edilecek. Bu sistemin

daha fazla havalimanında kullanılması ve sisteme çipli pasaporttaki tüm bilgilerin yüklenmesi planlanıyor. Ancak bilgi paylaşımı konusunda uluslararası anlaşmaların yeniden ele alınması ve ülkelerin diğerlerine vatandaş bilgilerine ulaşım konusunda engelleri kaldırması gerekiyor.

2050'de yolcu sayısı 16 milyara çıkacak. Sektör 2050'de yılda 16 milyar yolcu taşıyacak. Hava kargo 400 milyon ton olacak. Uçuş emniyetine büyük önem veren IATA, bu konuda yüzde 42'lik gelişim yakaladı. Her 1.6 milyon uçuş saatinde 1 olan kaza oranı, 4 milyon saate yükseltildi. e-bilet gibi yeni teknolojiler ve vergilerle ilgili önlemlerle verimlilik yükseldi. Sektör 5 yılda 59 milyar dolar tasarruf etti. Karbon emisyonu konusunda çalışmalar hızlandırıldı. 5 yılda yüzde 1.5 yakıt tasarrufu sağlandı.

Appendix 3. Turkish News Articles from World Category (3)

Çiftçiler, farklı pirinç türlerini seçerek genlerini karıştırdı. Böylece ideal pirinç türünü ortaya çıkarmayı başararak ürünlerinden daha fazla getiri elde etmeyi başardılar.

Japon bilim insanlarının pirincin alt türleri üzerinde yaptığı incelemede, tüm DNA bilgilerini içeren genomlar analiz edildi. Analizin sonucunda, Çinli çiftçilerin SD1 olarak bilinen geni çeşitlendirilerek pirinç bitkisinin gövde uzunluğunu kısalttıklarını ortaya çıktı.

Araştırma ekibinin başındaki isim Dr. Masanori Yamasaki, SD1' in son 50 yılda pirinç yetiştirilmesinde en önemli role sahip gen olduğuna dikkat çekti. SD1 üzerinde yapılan değişiklikler, pirincin daha kısa sürede olgunlaşmasını, irileşmesini sağladığı gibi üretim miktarını da artırdı.

Yamasaki, antik çağlarda yaşamış olan çiftçilerin yapay seleksiyon yöntemiyle SD1 üzerinde değişiklik yaptıklarını ve daha kısa gövdeli pirinç bitkisi üretmeyi başardıklarını belirtti.

Yamasaki, "Antik insanlar yerleşik hayata geçtikleri dönemde yapay seleksiyon yöntemi kullanmaya başladı… Zamanla meyve ve tahıl üretiminde verimlilik ve miktar arttı. Elde ettiğimiz bulgular antik insanların SD1 üzerinde değişiklik yapmayı başardığı ve tarımdaki devrimi sanılandan çok daha önce gerçekleştirdikleri yönünde" dedi.

Appendix 4. Turkish News Article from Financial Category

İngiliz ekonomi gazetesi Financial Times (FT), yüzde 2,42' lik yüksek mayıs enflasyonunun ardından "Türkiye' nin alışılmamış para politikası konusunda yeniden düşünme zamanı mı?" sorusunu ortaya attı.

FT, Merkez Bankası'nın politikasının ekonominin ısınmasını durduramadığı yönündeki korkuyu körükleyeceği yorumunu yaptı.

FT, "Türkiye' nin Alışılmamış Para Politikası: Yeniden Düşünme Zamanı mı?" başlıklı haberinde, "Türk politika yapıcıları, seçimlere bir hafta kalan enflasyon verileriyle hoş olmayan bir sürpriz yaşadıktan sonra seçimler sona erdiğinde faiz oranlarını yükseltme ve maliye politikasını sıkıştırma yönünde artan bir baskı altında kalacak" diye yazdı.

Tüketici enflasyonun mayısta beklenenden bir kat fazla geldiğine işaret edilen haberde şöyle denildi: "Veriler, Merkez Bankası'nın sermaye girişini caydırmaya yönelik düşük faiz ve iç talebi dizginlemeyi amaçlayan yüksek zorunlu karşılıkların karışımından oluşan alışılmamış politikası ile ekonominin ısınmasını durduramadığı korkusunu körükleyecek."

Haberde enflasyonun mayısta yaptığı "sürpriz" in büyük ölçüde gıda fiyatlarından kaynaklandığı belirtilerek, "Gıda fiyatları her zaman değişken. Bu yüzden analistler, cuma verilerinin Merkez Bankası'nın enflasyon görünümünü veya politika tutumunu değiştirmesini beklemiyorlar" denildi.

FT buna karşın, "Ancak zamanlama, siyasi olarak ters. Çünkü gıda fiyatlarının, kamuoyunun enflasyon algısı üzerinde büyük bir etkisi var" yorumunu da yaptı. Cari işlemler açığına dikkat çeken gazete, iç talep ve kredi patlamasına yansıyan hızlı büyümeyi sert bir inişin izleyebileceğinden korkan yatırımcıların tedirgin olduğunu yazdı. FT' ye konuşan Barclays Capital ekonomisti Christian Keller de "Piyasalar, yüksek turizm gelirlerinin Türkiye' nin cari işlemlerini olumlu etkileyeceği

umuduyla, küresel büyüme ve enflasyon gelişmelerini izleyerek ve Merkez Bankası' nın kredi politikasına zaman tanıyarak, yazın beklemeyi kabul edebilir" dedi.

Appendix 5. Turkish News Article from Journal Category

Sahte içkiden zehirlenen Rus rehber Victoria Nikoloeva, yaşam savaşını sürdürüyor.

Bodrum' da sahte içkiden zehirlenen ve Akdeniz Üniversitesi Hastanesi' ne getirilerek yoğun bakım ünitesinde tedaviye alınan 23 yaşındaki Rus rehber Victoria Nikoloeva, yaşam savaşını sürdürüyor.

Akdeniz Üniversitesi Hastanesi' nin yoğun bakımda yaşam destek ünitesine bağlı olarak ölüm kalım savaşı veren Victoria Nikoloeva' nın sağlık durumunda herhangi bir ilerleme olmadığı belirtildi. AÜ Hastanesi Başhekimi Doç. Dr. Abdullah Erdoğan, Rus rehberin beynine kan gelmeye devam ettiğini belirterek şunları söyledi: "Bizim için o bir yoğun bakım hastası. Victoria' nın sağlık durumu ile ilgili olarak uzman heyetimiz ile birlikte her gün düzenli toplantı yapıyoruz. Hastamız bugüne kadar hiçbir tedaviye cevap vermedi. Ancak bitkisel hayata girdi diyebileceğimiz bir durum da henüz oluşmadı. Yani beyninde kan dolaşımı olduğunu görüyoruz. Victoria'nın tedaviye bir an önce cevap vermesi en büyük umudumuz."

Öte yandan Antalya Valiliği' nden bugün yaptığı yazılı açıklamada, alkollü içki denetimlerinin sürdüğü bildirildi. 1 Haziran'da 19 ekiple turizm tesisleri, gezi ve tur tekneleriyle içki satış noktalarında yapılan denetimler rapor haline getirildiği belirtilen açıklamada, Alanya' da 142, Manavgat' ta 103, Kepez' de 111 olmak üzere diğer ilçelerle birlikte Antalya' da toplam 733 turizm tesisi, gezi ve tur teknesiyle içki satış noktasının denetlendiği kaydedildi. Toplanan bin 951 numune yediemine teslim edilirken, 11 numunede olumsuzluğa rastlandığı kaydedildi. Alanya' da, yediemine alınan 16 şişe viskiye savcılık kararıyla el konulduğu bildirilen açıklamada, alkollü içkilerle ilgili eğitimlere başlandığı ve hazırlanan 5 bin bilgilendirme broşürünün dağıtılmaya başlandığı kaydedildi.

Sahte alkolden zehirlenen Rus rehberlerden 28 yaşındaki Maria Shalyapina ve 22 yaşındaki Zalyaeva Auilia Antalya'daki hastanede, 24 yaşındaki Alexandr Zhbckov

Denizli' deki hastanede, 22 yaşındaki Marina Şevelyova ise döndükten sonra ülkesinde yaşamını yitirmişti.

Appendix 6. Turkish News Article from Sports Category

Somalili korsanların elinde tutsak kalan bir denizci, birkaç gün önce Türkiye' ye gelip gazeteleri açsa, "Beşiktaş şampiyon olmuş, Trabzonspor küme düşmüş" sanır! Everest tırmanışından dönen dağcı da öyle, komadan çıkan hasta da. Hatta "şehirden hiç ayrılmayan turp gibi bazı Trabzonsporlular" bile o kanaatte: "Bitirdiler Trabzonspor'u, bitti canım takım"! Ya da ben yanlış anladım.

"Bu işin genel seçimi de var" aşaması bile geride kaldı! Siyaset kesmedi. Trabzonspor Başkanı Sadri Şener, "İnsanlar ahireti de hesaba katsın, bu işin öbür tarafı da var" şeklindeki "ilahi" uyarısını yaptıktan sonra resmen açıkladı: "Fenerbahçe şampiyon değil ki"! Buyurun bakalım! Kim şampiyon o zaman?

Beşiktaş olmalı ki, Başkan Yıldırım Demirören çıkıyor "Teknik direktörümüzden memnunuz, on yıllık mukaveleyi bile tartıştık" diyor. Güleç, keyifli, kendinden emin. En büyük hayalinin, Avrupa'da şampiyonluk olduğunu ifade ediyor. Gidecekleri alınacakları açıklıyor. Geleceğe bakıyor. Bunun adına, yöneticiler ile taraftar arasındaki "pozitif etkileşim" deniyor.

"Tepedeki insan" kendine acıyıp, çaresizlik sergilerse, alttaki kalabalıkların "aldırma gönül aldırma" şarkısı söyleyecek halleri yok tabi. "Tepedeki insan" hesap sorulmasın diye düşman yaratırsa, aklıselime davet edecek olanlar taraftarlar değil elbet.

Dertlenmek eyvallah... Üzüntü tamam. Ama sürekli yenilenen Süper Lig macerasının bir dönemine takılıp orada kalmak, uzatmak, hedefi yeni şampiyonluklardan kin ve intikam rotasına taşımak, sadece Trabzonspor için değil futbolun içindeki her unsur için korkutucudur. Vahimdir.

Trabzonspor yönetiminin yaptığına ne denir peki? Ben söylemeyeyim de "taraflı" falan demesinler. İşte "ezeli ve ebedi" Trabzonspor taraftarı, efsane başkan Mehmet Ali Yılmaz' ın sözleri: "Eğer siz Trabzon' u dövülmüş, elinden kupası alınmış gibi

gösterirseniz sokaktaki insanın tepkisi büyür, sizi sorgular. Çok kötü bir sezon geçirilmedi. Travmayı atlatıp yeni rota çizmek lazım. Yoksa kaos kimseye yaramaz". Kaçan şampiyonluğun acısını unutturmak için kaostan medet umuluyor. "Trabzonspor dövüldü" imajı enjekte ediliyor.

Bu da mümkün elbet! Lakin "yarını" unutmak, hatta yarını "karartmak" riski mevcuttur. Bilen bilir; "kalabalıkların sadece gaz pedalı vardır", "freni" yoktur. Süper Lig' in dörtte üçünü önde götürüp, son anda Fenerbahçe tarafından geçilen ve kendilerinden başka her futbolsever tarafından son derece başarılı görülen Trabzonspor bırakın hedef şaşırtmayı, laf karıştırmayı sadece yarına odaklanan ve "yıldızlarımızı iyi yönetemedik" şeklinde açık açık özeleştiri yapmaktan kaçınmayan Beşiktaş' tan örnek alsın. Yarına baksın.

Appendix 7. Turkish News Articles from Health Category

İmkanı olanlar soluğu spor salonlarında alıyor, açık havada spor yapmayı tercih edenler de açık alanlardaki spor aletleriyle çalışıyor. Ancak bilinçsiz spor yapmak dizlere zarar vermekten başka bir işe yaramıyor. International Hospital Ortopedi ve Travmatoloji Uzmanı Doç. Dr. Sezgin Sarban, spor yapmanın bilinçli yapılırsa faydalı olduğunu, uzay yürüyüşü (ayakta yüksek eğimle pedal çevirmek), step (basamak çıkmak), bisiklet ve spinning (bisikletle yapılan bir egzersiz) sporlarının dizlerde aşırı zorlanma, yüklenmeyle birlikte diz eklemlerine zarar verdiğini söylüyor.

Orta ve ileri yaş grubunda diz kapağına çok yük binmemesini istediklerini, spor salonlarında sık yapıldığı üzere yüksek eğim verilerek yapılan sporların dizdeki "Patellofemoral" eklemine aşırı yük bindirdiğini belirten Doç. Dr. Sezgin Sarban, "Diz kapağının altında bulunan eklemde kıkırdak zorlanmaları oluyor. Dizi bükünce patellofemoral eklemi zorlanıyor, ön diz ağrısı ortaya çıkıyor. Hatta bu zorlanma ve yüklenme nedeniyle kişi antrenmanı bırakmak zorunda kalıyor" diyor.

Aynı şekilde pilates ve yogada da dizi zorlayarak yapılan hareketlerin olduğunu, eğitmenlerin bu konuda gereken uyarıları yapmalarının önemli olduğunu ifade eden Doç. Dr. Sezgin Sarban, şu bilgileri veriyor: "Vücudun kendi ağırlığını kullanarak yapılan germe egzersizlerinde bile eklemler zorlanınca bırakmak, ısrarcı ve inatçı bir tutum içinde olmamak gerekiyor. Aynı şekilde koşu bandında spor yaparken, belli bir tempoda, eğimi çok artırmadan yürüyüş yapılmalı ve saat 5-6 km sınırlarında kalınmalıdır. En doğru egzersiz biçimi, vücudu gerektiği şekilde ısıttıktan sonra, her bir kasın doğru şekilde çalıştırılmasıdır. Her bir kasımız için kondisyoner eşliğinde kaldırabileceğimiz, alabileceğimiz bir yük miktarı var. Belli tekrarlarla bu ağırlıkları kaldırmak faydalı olacaktır."

Spor yapmadan önce vücuttaki olası rahatsızlıkları öğrenmek önem taşıyor. Tepeden tırnağa bir değerlendirmeden geçmek de omurgayı koruyacak şekilde, bilinçli egzersiz yapılmasını sağlıyor. Omurgada bel ve boyunda kireçlenmeler ve

fıtıklar olabiliyor. Özellikle vücutlarının belli bölgelerinde ağrıları bulunan, kireçlemeleri olanların ortopedi ve fizik tedavi bölümlerinden destek alması gerekiyor. Eğer kişinin belirgin fıtığı varsa beyin cerrahisine yönlendirilmesi önem taşıyor. Doç. Dr. Sezgin Sarban, sorunun adı konulduktan sonra da fizik tedavi uzmanları, beyin cerrahisi ve ortopedi uzmanlarının ortak değerlendirmesiyle bu kişilerin de spor akademisi mezunu eğitmenler gözetiminde spor yapabileceğini söylüyor.

Yüzme, eklemlere aşırı yük vermeden kasların güçlenmesini sağlar. Bu sporun yatay şekilde yapılmasından dolayı yerçekimi kuvveti eklemlere dik gelir ve eklem kıkırdaklarına yüklenme oluşturmaz. Bunun yanında yüzme sporu, vücuttaki tüm kasları senkronize şekilde aynı anda çalıştıran tek spordur. Doç. Dr. Sarban, menisküs yırtığında kurbağalama stili, omuz sıkışması hastalığı olanların da serbest stil yüzmeden zarar görebileceğini belirtti.

Hiç spor yapmamış, kasları çalışmamış, hareketsiz kalan kişilerde yaralanmaların olabileceğine değinen Doç. Dr. Sezgin Sarban, şunlara dikkat edilmesi gerektiğini söylüyor: “Kas zorlanmaları, kas iltihapları, ödemler, dejeneratif yırtıklar dediğimiz ve yıpranmaya yatkın menisküslerde oluşan yırtıklar, uzun süre hiç spor yapmayan kişiler birden spora başlayınca görülebiliyor. Özellikle parklarda, açık alanlarda bulunan spor aletlerinin bilinçsizce kullanılması yeni sakatlanmalara neden olabiliyor. Özellikle de sağa sola dönmeyi sağlayan, yuvarlak hareketli diskler üzerinde yapılan egzersizler dizleri çok zorluyor. 50-60 yaşında olup da, hiç aktif spor yapmayanlarda menisküste bulunan yırtıklar daha da artabiliyor. Bu kişilerin vücutlarını çok fazla zorlamadan spor yapmalarında yarar var. Spora yaklaşık 15-20 dakika tempolu yürüyüş yaptıktan sonra başlamak gerekiyor. Spordan önce vücudun yağ - kas oranlarının değerlendirilmesi, sporun da doğru planlanmasını sağlayacaktır. Eğer kişi altı ay süreyle spor yapacaksa; öncesinde, arada ve sonrasında mutlaka bu değerlerin alınması büyük önem taşıyor. Diyetisyen desteği de alınarak spor yapılması sağlıklı olmaya yardımcıdır.”

Appendix 8. Turkish News Article from Politics Category

CHP Genel Başkan Yardımcısı Gülsün Bilgehan, gençlere, CHP ve Türkiye' deki demokratik sistem hakkında bilgi verdi.

Tahrir Meydanında özgürlük hareketini başlatan gençleri ağırlamaktan mutluluk duyduklarını vurgulayan Bilgehan, Mısır' ın, Türkiye için dost ve kardeş bir ülke olduğunu söyledi.

Cumhuriyeti kuran CHP' nin, Türkiye ve dünya siyasetinde önemli bir yeri bulunduğunu dile getiren Bilgehan, CHP' nin, tek parti ve tek adam rejimini değiştiren devrimci bir parti olduğunu ifade etti.

Türkiye' nin ortak değerlere sahip olduğu ülkeler için önemli bir örnek oluşturduğunu anlatan Bilgehan, seçimlerde bir sınavdan geçeceklerini, Tahrir Meydanından gelen gençlerin ziyaretinin partiye uğurlu gelmesini diledi.

Bilgehan, daha sonra Mısırlı gençlerin sorularını yanıtladı. Mısırlı bir gencin, "CHP, Filistin halkına karşı neden İsrail' i destekliyor" sorusuna karşılık Bilgehan, "CHP' nin İsrail' i desteklediğini herhalde AKP ziyaretinde duydunuz. Çünkü biz CHP olarak Ortadoğu'da her zaman dengeli bir politika izlemek gerektiğine inandık. Filistin halkına uygulanan zulmü insanlık suçu olarak görüyoruz ama Ortadoğu'ya barışın gelmesi için bütün tarafların, İsrail dahil, uzlaşması gerektiğine inanıyoruz. Mısır' da da böyle düşünen pek çok aydın, siyasetçi olduğunu biliyoruz" yanıtını verdi.

Seçim sürecinde liderlerin oldukça sert görüntü verdiklerini ifade eden Bilgehan, seçimlerin ardından Mecliste yapıcı ve olumlu bir şeklide çalışacaklarını söyledi. Bilgehan, Türkiye' nin bugünkü ekonomik durumuyla ilgili çekinceleri olduğunu, eşit ve adil bir gelir dağılımı bulunmadığını ifade etti.

Kadın ve gençlik kollarına çok önem verdiklerini, kadınların destekleyeceği partinin seçimi kazanacağını vurgulayan Bilgehan, bu nedenle kadın kollarının çok çalıştığını dile getirdi.

Bilgehan, seçim sonucu ne olursa olsun Türkiye' nin Ortadoğu için model olmaya devam etmesine gayret edeceklerini belirtti.

Mısır halkının tek adam rejiminin sıkıntılarını yaşadığını söyleyen Bilgehan, "Biz de tek adam rejimine geçilmesinden korktuğumuz için seçim çalışmalarında sizin özgürlük hareketinizi örnek gösteriyoruz" dedi.

Mısırlı gençler de ülkede yaşananlar ve bundan sonrasına ilişkin görüşlerini dile getirdi.

Appendix 9. Turkish News Article from Technology Category (1)

İnternette “filtre” iddiaları karşısında temel hak ve özgürlüklerin ihlal edileceğini savunan “Anonymous” (Anonim) adlı bir grup, Türkiye’ deki çeşitli kamu kuruluşları ile bazı medya sitelerine yönelik siber saldırı düzenleyeceği tehdidinde bulundu.

İlk büyük eylemlerini aylar önce Wikileaks' e yönelik ambargo uygulayan Paypal, Visa ve Mastercard gibi online ödeme ve kredi kartı firmalarına karşı gerçekleştiren grup üyeleri, dün kendilerine ait internet sitesinde Türkiye'yi hedef alan bir mesaj yayınladı.

Türkiye'de internet kullanıcılarının “filtre” uygulaması ile sansüre uğrayacağını savunulan mesajda “operationturkey” (Türkiye Operasyonu) adıyla “siber savaş” ilan edilirken, öncelikle “sansür” uygulayan kurumlara karşı harekete geçileceği açıklandı.

Sosyal paylaşım siteleri üzerinden örgütlenen grup üyeleri, “IRC” isimli anlık mesajlaşma kanallarında siber saldırının hedefi ve zamanlaması konusunda bilgiler aktararak, uygulanacak olan yöntem ve stratejiler hakkında çeşitli bilgiler verdi.

Türkiye'den de bazı internet kullanıcılarının destek verdiği grup üyeleri, bu mesajdan bir süre sonra kimi kamu kurumlarının internet sistemlerine yönelik küçük çaplı siber saldırılar düzenledi. Söz konusu saldırılar karşısında bazı internet sitelerine erişim bir süre engellenirken, “hack” leme girişimi başarısızlıkla sonuçlandı.

İlk organize saldırı perşembe saat 18.00'de olacak. Türkiye’ deki kimi kamu kurumlarını hedef alan ilk organize saldırının 9 Haziran Perşembe günü saat 18.00'de (TSİ) gerçekleştirileceğini duyuran Anonymous, eyleme katılmak isteyenlerin bilgisayarlarına bazı yazılımlar indirerek, bu programları aktif hale getirmesini istedi.

Tüm bu gelişmeleri yakından izleyen siber güvenlik uzmanları, söz konusu tehdidin önemine dikkat çekerek, birçok sitenin “DDoS” olarak adlandırılan saldırı ile karşı karşıya kalabileceğini bildirdi. 2007'de nam salan “DDoS” saldırı yöntemi ile hedef alınan bir internet sitesinin aynı anda yoğun bir ziyaretçi akınına uğratılarak, web sitesi veya DNS (Alan Adı Sistemi) sunucularının kullanılmaz hale getirildiğini belirten uzmanlar, “Bu saldırı yöntemi; 50 kişilik bir otobüse 1000 kişinin binmesi gibi bir şey” yorumunda bulundu.

Bu tip saldırılarda amacın, “bilgi çalmak” yerine sistemin erişilemez hale getirilmesi olduğunu kaydeden siber güvenlik uzmanları, burada verilmek istenen mesajın “Bizim internetimize karışırsanız biz de sizin internetinizi kapatırız” anlamı taşıdığını öne sürdü.

Anonymous grubunun saldırılarda özel bir yöntem kullandığını belirten uzmanlar, “Klasik DDoS saldırılarında 'zombi' haline getirilmiş bilgisayarlar kullanılırken, bu grup tamamen 'gönüllü zombi' bilgisayarlarla saldırıyor. Gönüllü zombi olabilmeniz için de size bir program yüklettiriyor ve bu programı IRC üzerinden şuraya saldır buraya saldır şeklinde yönlendiriyorlar” bilgisini verdi.

Siber saldırı sırasında kendi IP adresleri yerine VPN (Sanal Paylaşımlı Ağ) üzerinden IP adreslerini değiştiren grup üyelerinin bu sayede izlerini bazı kişilerin yönlendirmesi ile kendilerine hedefler seçtiklerine dikkat çekti.

DDoS saldırılarından bir sonraki adımın çeşitli devlet kurumlarına ait telefon ve haberleşme sistemlerini çalışamaz hale getirmek olduğunu kaydeden siber güvenlik uzmanları, daha sonra hedefteki kurumlarda çalışan kişilerin bilgisayarlarına “phishing” saldırıları gerçekleştirileceğinin altını çizdi.

“Phishing” yöntemi ile kullanıcılara e-mail göndererek, bilgisayarlarından erişim sağladıkları banka, kredi kartı bilgi ve şifrelerinin yanı sıra kurumlarına ait kimi özel bilgi ve belgelerin sızdırılmaya çalışılacağını anlatan uzmanlar, böylesi bir durumun

yeni bir “Wikileaks” olayına kapı açacağına vurgu yaptı.

Eğer gerekli önlemler alınmazsa ciddi bir siber savaşın başlamış olacağını bildiren uzmanlar, kimi hassas hedeflerin saldırıya uğraması durumunda Türkiye’ deki internet ağının bir süre için çökebileceği uyarısında bulundu.

Bu arada Anonymous grubu üyeleri Perşembe günkü saldırıda çeşitli kamu kurum ve kuruluşlarına ait internet sitelerinin yanı sıra siyasi partiler ve kimi medya sitelerini de kendilerine hedef seçtiklerini açıkladı.

Anonymous' un bugüne kadar gerçekleştirdiği hacker saldırılarını yakından takip eden Siber Güvenlik Uzmanı Huzeyfe Önal, yaptığı açıklamada, hedefteki kurumların kendi bünyelerinde alacağı çeşitli önlemlerin yanı sıra internet servis sağlayıcılarının da ciddi tedbirler alması gerektiğini bildirdi.

Bu tür organize saldırılarda, servis sağlayıcılarının saldırının hedef aldığı web sayfasın yönelik tüm girişimlere anında müdahale edebileceğini anlatan Önal, bunun için saldırı yönteminin çok iyi analiz edilmesi ve engelleme sistemlerinin devreye sokulması gerektiğini anlattı.

Proxy ve benzeri yöntemler ise IP adreslerini değiştirerek saldırıya katılacakların ciddi bir etkiye sahip olmayacağı görüşünü belirten Önal, kendi IP'leri ile eyleme destek verenlerin güvenlik güçleri tarafından kısa sürede tespit edileceğinin altını çizdi.

Anonymous' a destek vermek isteyen çok sayıda kişinin, bu grubun internet sitesinden indirdikleri özel programlarla bilgisayarlarını saldırıda kullanılacak birer 'zombi' haline getirdiğini belirten Önal, sözlerine şöyle devam etti: “Grup üyelerinin Türkiye’ ye yönelik bu saldırı planını 'protesto' olarak gören ve gönüllü olarak katılmayı planlayan çok sayıda Türk vatandaşı ile karşılaştık. Burada dikkat edilmesi gereken en önemli husus protesto ile 'saldırının farklı şeyler olduğu. Protesto amaçlı da olsa böylesi bir saldırıya katılanlara, Türk Ceza Kanunu’ na

(TCK) göre suç işlediklerini hatırlatmak isterim. İnternet sitelerine bu şekilde zarar verenlerin, 5 yıla kadar hapis cezası ile yargılanacaklarını unutmamalı.”

Appendix 10. Turkish News Article from Technology Category (2)

Korsan kriziyle başı ağrıyan Japon elektronik üreticisi Sony, PlayStation şebekesinin bu hafta kısmen açılacağını açıkladı.

Şirket, online oyun kullanıcılarını kredi kartı ve diğer kişisel bilgilerinin çalınmış olabileceği konusunda uyarmış, ardından da PlayStation şebekesini kapatmıştı.

Milliyet'in haberine göre Sony'nin ikinci adamı Kazuo Hirai "Datalarını tehlikeye attığımız, endişelendirdiğimiz ve rahatsızlık verdiğimiz tüm kullanıcılarımızdan özür dileriz" diyerek geleneksel Japon selamı ile eğilerek kullanıcılardan özür diledi.

Hirai hackerların vermiş olduğu rahatsızlıktan dolayı derin üzüntü duyduğunu belirtip yeni bir veri bankası oluşturmaya başladıklarını ve yakın zamanda bu bağımsız birime geçeceklerini söyledi ve ekledi "PlayStation 3'ün yazılımını yükselteceğiz. Bu süreçte, tüm kullanıcılarımız PlayStation şebeke hesaplarının şifrelerini yenileyecek. Bunun onlar için büyük bir sıkıntı olduğunu biliyorum. "Sony rakiplere kaptırmaktan korktuğu kullanıcıları için iyi niyet ve özür göstergesi olarak ücretsiz hediyeler vereceğini açıkladı. Bu hediyeler şöyle: Ücretsiz içerik (İçeriğin detayları henüz belli değil ama ücretsiz bir oyun ve yanında tema, avatar paketi olması bekleniyor). 30 günlük ücretsiz Playstation Plus üyeliği (Mevcut Playstation Plus üyelerine de ekstra 30 günlük ücretsiz kullanım hakkı tanınacak). 30 günlük sınırsız Qriocity kullanımı. Sony ayrıca hackerların kullanıcıların kart bilgilerini çaldığını tam olarak kabul etmese de, yine de bundan dolayı finansal bir zarar gören olursa bu zararı tazmin edeceğini de belirtti.

Kişisel bilgilerin daha iyi korunmasını sağlamak için sistemlerini yeniden kurma konusunda adımlar attıklarını belirten şirket, servislerin en kısa sürede tamamen hizmete gireceğini kaydetti. 59 ülkede 77 milyon kişi PlayStation şebekesini kullanıyor.

Appendix 11. Summarizations of Turkish News Articles by Person 1

Summary for World Category Article (1)

The New York Times (NYT),'Türkiye bağlantılı okullar Teksas' ta büyüyor' başlıklı haberinde Gülen Hareketi' nin ABD' deki eğitim ağlarını ve öğretmenler için alınan özel vizeleri gündeme getirdi. NYT, TMD adlı bir inşaat firmasının kuruluşunun üzerinden bir ay geçmeden Gülen' e yakınlığıyla bilinen Harmony (Uyum) Okulları'nın 8.2 milyonluk inşaat ihalesini kazandığını ve bu firmanın Türkiye ile bağlantılı olduğunu öne sürdü. Gülen hareketiyle doğrudan bağlantılı olduğu belirtilen ve ABD' nin 25 eyaletinde 120 okul bulunduğu yazılan haberde, okulların Amerikan öğrencilerinin genellikle başarısız olduğu bilim ve matematik konularında ağırlıklı olarak eğitim verdiği vurgulandı. Harmony Okulları'nda 2011 yılında 1.500 öğretmenin istihdam edildiği ve bunların 292'sinin 'yüksek nitelikli eleman' olarak nitelenen 'H-1B' vizesi sahibi olduğu yazıldı.

Summary for World Category Article (2)

Uluslararası Havayolu Taşıyıcıları Birliği' nin (IATA) bu yılki genel kuruluna, havalimanlarında gelecekte güvenlik kontrolünü hızlandıracak yeni uygulamalar damga vurdu. Yeni uygulamada, yolcular güvenlik risklerine göre üçe ayrılacak. 'Geleceğin kontrol noktası' olarak adlandırılan uygulamada, yolcular güvenlik risklerine göre üçe ayrılacak ve 6.1 metrelik koridordan yürüyerek geçecek. Yolcu yürürken, üzerindeki her şey, kabine alacağı çanta, yüksek hassaslıktaki özel cihazlar tarafından taranacak. Sıvı ve elektronik cihazlar çantadan çıkartılmadan kontrol edilebilecek. Bu da güvenlik işlemlerini hızlandıracak. Ayakkabı ve çantadan laptop (dizüstü bilgisayar) çıkarma son bulacak. IATA ayrıca pasaport kuyruklarının da azaltılması için çipli pasaport uygulamasına hızla geçilmesini tavsiye ediyor. Yolcunun bilgilerinin yanı sıra vize bilgilerinin içinde yer aldığı çipli pasaport da kontrolleri hızlandıracak. Ayrıca yolcunun göz bebeğinin taranmasıyla kimlik bilgileri kontrol edilecek. Uçuş emniyetine büyük önem veren IATA, bu konuda yüzde 42' lik gelişim yakaladı.

Summary for World Category Article (3)

Çiftçiler, farklı pirinç türlerini seçerek genlerini karıştırdı. Böylece ideal pirinç türünü ortaya çıkarmayı başararak ürünlerinden daha fazla getiri elde etmeyi başardılar. Araştırma ekibinin başındaki isim Dr. Masanori Yamasaki, SD1' in son 50 yılda pirinç yetiştirilmesinde en önemli role sahip gen olduğuna dikkat çekti. SD1 üzerinde yapılan değişiklikler, pirincin daha kısa sürede olgunlaşmasını, irileşmesini sağladığı gibi üretim miktarını da artırdı.

Summary for Financial Category Article

İngiliz ekonomi gazetesi Financial Times (FT), yüzde 2,42' lik yüksek mayıs enflasyonunun ardından "Türkiye' nin alışılmamış para politikası konusunda yeniden düşünme zamanı mı?" sorusunu ortaya attı. FT, "Türkiye' nin Alışılmamış Para Politikası: Yeniden Düşünme Zamanı mı?" başlıklı haberinde, "Türk politika yapıcıları, seçimlere bir hafta kalan enflasyon verileriyle hoş olmayan bir sürpriz yaşadıktan sonra seçimler sona erdiğinde faiz oranlarını yükseltme ve maliye politikasını sıkıştırma yönünde artan bir baskı altında kalacak" diye yazdı. FT' ye konuşan Barclays Capital ekonomisti Christian Keller de "Piyasalar, yüksek turizm gelirlerinin Türkiye' nin cari işlemlerini olumlu etkileyeceği umuduyla, küresel büyüme ve enflasyon gelişmelerini izleyerek ve Merkez Bankası' nın kredi politikasına zaman tanıyarak, yazın beklemeyi kabul edebilir" dedi.

Summary for Journal Category Article

Sahte içkiden zehirlenen Rus rehber Victoria Nikoloeva, yaşam savaşını sürdürüyor. AÜ Hastanesi Başhekimi Doç. Dr. Abdullah Erdoğan, Rus rehberin beynine kan gelmeye devam ettiğini belirterek şunları söyledi: bugüne kadar hiçbir tedaviye cevap vermedi. Ancak bitkisel hayata girdi diyebileceğimiz bir durum da henüz oluşmadı. Yani beyninde kan dolaşımı olduğunu görüyoruz. Öte yandan Antalya Valiliği'nden bugün yaptığı yazılı açıklamada, alkollü içki denetimlerinin sürdüğü bildirildi. Toplanan bin 951 numune yediemine teslim edilirken, 11

numunede olumsuzluğa rastlandığı kaydedildi. Alanya' da, yediemine alınan 16 şişe viskiye savcılık kararıyla el konulduğu bildirilen açıklamada, alkollü içkilerle ilgili eğitimlere başlandığı ve hazırlanan 5 bin bilgilendirme broşürünün dağıtılmaya başlandığı kaydedildi.

Summary for Sports Category Article

Somalili korsanların elinde tutsak kalan bir denizci, birkaç gün önce Türkiye' ye gelip gazeteleri açsa, "Beşiktaş şampiyon olmuş, Trabzonspor küme düşmüş" sanır! Hatta "şehirden hiç ayrılmayan turp gibi bazı Trabzonsporlular" bile o kanaatte: "Bitirdiler Trabzonspor'u, bitti canım takım"! "Tepedeki insan" kendine acıyıp, çaresizlik sergilerse, alttaki kalabalıkların "aldırma gönül aldırma" şarkısı söyleyecek halleri yok tabi. "Tepedeki insan" hesap sorulmasın diye düşman yaratırsa, aklıselime davet edecek olanlar taraftarlar değil elbet. Ama sürekli yenilenen Süper Lig macerasının bir dönemine takılıp orada kalmak, uzatmak, hedefi yeni şampiyonluklardan kin ve intikam rotasına taşımak, sadece Trabzonspor için değil futbolun içindeki her unsur için korkutucudur. Süper Lig'in dörtte üçünü önde götürüp, son anda Fenerbahçe tarafından geçilen ve kendilerinden başka her futbolsever tarafından son derece başarılı görülen Trabzonspor bırakın hedef şaşırtmayı, laf karıştırmayı sadece yarına odaklanan ve "yıldızlarımızı iyi yönetemedik" şeklinde açık açık özeleştiri yapmaktan kaçınmayan Beşiktaş' tan örnek alsın. Yarına baksın.

Summary for Health Category Article

İmkanı olanlar soluğu spor salonlarında alıyor, açık havada spor yapmayı tercih edenler de açık alanlardaki spor aletleriyle çalışıyor. Ancak bilinçsiz spor yapmak dizlere zarar vermekten başka bir işe yaramıyor. International Hospital Ortopedi ve Travmatoloji Uzmanı Doç. Dr. Sezgin Sarban, şu bilgileri veriyor: En doğru egzersiz biçimi, vücudu gerektiği şekilde ısıttıktan sonra, her bir kasın doğru şekilde çalıştırılmasıdır. Her bir kasımız için kondisyoner eşliğinde kaldırabileceğimiz, alabileceğimiz bir yük miktarı var. Belli tekrarlarla bu ağırlıkları kaldırmak faydalı

olacaktır.” Spor yapmadan önce vücuttaki olası rahatsızlıkları öğrenmek önem taşıyor. Tepeden tırnağa bir değerlendirmeden geçmek de omurgayı koruyacak şekilde, bilinçli egzersiz yapılmasını sağlıyor.

Summary for Politics Category Article

CHP Genel Başkan Yardımcısı Gülsün Bilgehan, gençlere, CHP ve Türkiye’ deki demokratik sistem hakkında bilgi verdi. Bilgehan, daha sonra Mısırlı gençlerin sorularını yanıtladı. Mısırlı bir gencin, "CHP, Filistin halkına karşı neden İsrail'i destekliyor" sorusuna karşılık Bilgehan, "CHP’ nin İsrail'i desteklediğini herhalde AKP ziyaretinde duydunuz. Çünkü biz CHP olarak Ortadoğu'da her zaman dengeli bir politika izlemek gerektiğine inandık. Filistin halkına uygulanan zulmü insanlık suçu olarak görüyoruz ama Ortadoğu'ya barışın gelmesi için bütün tarafların, İsrail dahil, uzlaşması gerektiğine inanıyoruz. Mısır'da da böyle düşünen pek çok aydın, siyasetçi olduğunu biliyoruz" yanıtını verdi. Bilgehan, seçim sonucu ne olursa olsun Türkiye’ nin Ortadoğu için model olmaya devam etmesine gayret edeceklerini belirtti.

Summary for Technology Category Article (1)

İnternette “filtre” iddiaları karşısında temel hak ve özgürlüklerin ihlal edileceğini savunan “Anonymous” (Anonim) adlı bir grup, Türkiye’ deki çeşitli kamu kuruluşları ile bazı medya sitelerine yönelik siber saldırı düzenleyeceği tehdidinde bulundu. Türkiye'de internet kullanıcılarının “filtre” uygulaması ile sansüre uğrayacağını savunulan mesajda “operationturkey” (Türkiye Operasyonu) adıyla “siber savaş” ilan edilirken, öncelikle “sansür” uygulayan kurumlara karşı harekete geçileceği açıklandı. Türkiye’ deki kimi kamu kurumlarını hedef alan ilk organize saldırının 9 Haziran Perşembe günü saat 18.00'de (TSİ) gerçekleştirileceğini duyuran Anonymous, eyleme katılmak isteyenlerin bilgisayarlarına bazı yazılımlar indirerek, bu programları aktif hale getirmesini istedi. Bu tip saldırılarda amacın, “bilgi çalmak” yerine sistemin erişilemez hale getirilmesi olduğunu kaydeden siber güvenlik uzmanları, burada verilmek istenen mesajın “Bizim internetimize

karışırsanız biz de sizin internetinizi kapatırız" anlamı taşıdığını öne sürdü. Eğer gerekli önlemler alınmazsa ciddi bir siber savaşın başlamış olacağını bildiren uzmanlar, kimi hassas hedeflerin saldırıya uğraması durumunda Türkiye' deki internet ağının bir süre için çökebileceği uyarısında bulundu. Anonymous' un bugüne kadar gerçekleştirdiği hacker saldırılarını yakından takip eden Siber Güvenlik Uzmanı Huzeyfe Önal, yaptığı açıklamada, hedefteki kurumların kendi bünyelerinde alacağı çeşitli önlemlerin yanı sıra internet servis sağlayıcılarının da ciddi tedbirler alması gerektiğini bildirdi. Önal, sözlerine şöyle devam etti: "Grup üyelerinin Türkiye' ye yönelik bu saldırı planını 'protesto' olarak gören ve gönüllü olarak katılmayı planlayan çok sayıda Türk vatandaşı ile karşılaştık. Burada dikkat edilmesi gereken en önemli husus protesto ile 'saldırının farklı şeyler olduğu. Protesto amaçlı da olsa böylesi bir saldırıya katılanlara, Türk Ceza Kanunu' na (TCK) göre suç işlediklerini hatırlatmak isterim. İnternet sitelerine bu şekilde zarar verenlerin, 5 yıla kadar hapis cezası ile yargılanacaklarını unutmamalı."

Summary for Technology Category Article (2)

Korsan kriziyle başı ağrıyan Japon elektronik üreticisi Sony, PlayStation şebekesinin bu hafta kısmen açılacağını açıkladı. Milliyet'in haberine göre Sony'nin ikinci adamı Kazuo Hirai "Datalarını tehlikeye attığımız, endişelendirdiğimiz ve rahatsızlık verdiğimiz tüm kullanıcılarımızdan özür dileriz" diyerek geleneksel Japon selamı ile eğilerek kullanıcılardan özür diledi. Kişisel bilgilerin daha iyi korunmasını sağlamak için sistemlerini yeniden kurma konusunda adımlar attıklarını belirten şirket, servislerin en kısa sürede tamamen hizmete gireceğini kaydetti. 59 ülkede 77 milyon kişi PlayStation şebekesini kullanıyor.

Appendix 12. Summarizations of Turkish News Articles by Person 2

Summary for World Category Article (1)

The New York Times (NYT),'Türkiye bağlantılı okullar Teksas' ta büyüyor' başlıklı haberinde Gülen Hareketi' nin ABD' deki eğitim ağlarını ve öğretmenler için alınan özel vizeleri gündeme getirdi. ABD' nin en saygın gazetelerinden The New York Times (NYT), birinci sayfasından yayımladığı 'Türkiye bağlantılı okullar Teksas' ta büyüyor' başlıklı haberinde Gülen Hareketi' nin ABD' deki eğitim ağlarını ve öğretmenler için alınan özel vizeleri gündeme getirdi. Haberde 'Charter' adı verilen ve Harmony adıyla bilinen bu okulların öteki kamu okullarına göre öğrenci başına 1-2 bin dolar arası daha az maliyet oluşturdukları için devlet tarafından tercih edildiği belirtildi.

Summary for World Category Article (2)

Uluslararası Havayolu Taşıyıcıları Birliği' nin (IATA) bu yılki genel kuruluna, havalimanlarında gelecekte güvenlik kontrolünü hızlandıracak yeni uygulamalar damga vurdu. Yeni uygulamada, yolcular güvenlik risklerine göre üçe ayrılacak. Böylece ayakkabı ve laptop çıkarma son bulacak. IATA' nın planına göre bu teknoloji önümüzdeki yıldan itibaren havalimanlarında kullanılmaya başlanacak. Yolcular fazla beklemeden hızla güvenlik işlemlerini yaptıracak ve uçağa binebilecek. Yolcunun bilgilerinin yanı sıra vize bilgilerinin içinde yer aldığı çipli pasaport da kontrolleri hızlandıracak.

Summary for World Category Article (3)

Japon bilim insanlarının pirincin alt türleri üzerinde yaptığı incelemede, tüm DNA bilgilerini içeren genomlar analiz edildi. SD1 üzerinde yapılan değişiklikler, pirincin daha kısa sürede olgunlaşmasını, irileşmesini sağladığı gibi üretim miktarını da artırdı. Yamasaki, antik çağlarda yaşamış olan çiftçilerin yapay seleksiyon yöntemiyle SD1 üzerinde değişiklik yaptıklarını ve daha kısa gövdeli pirinç bitkisi üretmeyi başardıklarını belirtti.

Summary for Financial Category Article

İngiliz ekonomi gazetesi Financial Times (FT), yüzde 2,42' lik yüksek mayıs enflasyonunun ardından "Türkiye' nin alışılmamış para politikası konusunda yeniden düşünme zamanı mı?" sorusunu ortaya attı. Haberde enflasyonun mayısta yaptığı "sürpriz" in büyük ölçüde gıda fiyatlarından kaynaklandığı belirtilerek, "Gıda fiyatları her zaman değişken. Bu yüzden analistler, cuma verilerinin Merkez Bankası'nın enflasyon görünümünü veya politika tutumunu değiştirmesini beklemiyorlar" denildi. FT' ye konuşan Barclays Capital ekonomisti Christian Keller de "Piyasalar, yüksek turizm gelirlerinin Türkiye' nin cari işlemlerini olumlu etkileyeceği umuduyla, küresel büyüme ve enflasyon gelişmelerini izleyerek ve Merkez Bankası'nın kredi politikasına zaman tanıyarak, yazın beklemeyi kabul edebilir" dedi.

Summary for Journal Category Article

Bodrum' da sahte içkiden zehirlenen ve Akdeniz Üniversitesi Hastanesi'ne getirilerek yoğun bakım ünitesinde tedaviye alınan 23 yaşındaki Rus rehber Victoria Nikoloeva, yaşam savaşını sürdürüyor. 1 Haziran'da 19 ekiple turizm tesisleri, gezi ve tur tekneleriyle içki satış noktalarında yapılan denetimler rapor haline getirildiği belirtilen açıklamada, Alanya' da 142, Manavgat' ta 103, Kepez' de 111 olmak üzere diğer ilçelerle birlikte Antalya' da toplam 733 turizm tesisi, gezi ve tur teknesiyle içki satış noktasının denetlendiği kaydedildi. Sahte alkolden zehirlenen Rus rehberlerden 28 yaşındaki Maria Shalyapina ve 22 yaşındaki Zalyaeva Auilia Antalya' daki hastanede, 24 yaşındaki Alexandr Zhbckov Denizli' deki hastanede, 22 yaşındaki Marina Şevelyova ise döndükten sonra ülkesinde yaşamını yitirmişti.

Summary for Sports Category Article

Somalili korsanların elinde tutsak kalan bir denizci, birkaç gün önce Türkiye' ye gelip gazeteleri açsa, "Beşiktaş şampiyon olmuş, Trabzonspor küme düşmüş" sanır! Beşiktaş olmalı ki, Başkan Yıldırım Demirören çıkıyor "Teknik direktörümüzden memnunuz, on yıllık mukaveleyi bile tartıştık" diyor. "Tepedeki insan" kendine acıyıp, çaresizlik sergilerse, alttaki kalabalıkların "aldırma gönül aldırma" şarkısı

söyleyecek halleri yok tabi. Ama sürekli yenilenen Süper Lig macerasının bir dönemine takılıp orada kalmak, uzatmak, hedefi yeni şampiyonluklardan kin ve intikam rotasına taşımak, sadece Trabzonspor için değil futbolun içindeki her unsur için korkutucudur. "Eğer siz Trabzon' u dövülmüş, elinden kupası alınmış gibi gösterirseniz sokaktaki insanın tepkisi büyür, sizi sorgular. Süper Lig'in dörtte üçünü önde götürüp, son anda Fenerbahçe tarafından geçilen ve kendilerinden başka her futbolsever tarafından son derece başarılı görülen Trabzonspor bırakın hedef şaşırtmayı, laf karıştırmayı sadece yarına odaklanan ve "yıldızlarımızı iyi yönetemedik" şeklinde açık açık özeleştiri yapmaktan kaçınmayan Beşiktaş' tan örnek alsın. Yarına baksın.

Summary for Health Category Article

International Hospital Ortopedi ve Travmatoloji Uzmanı Doç. Dr. Sezgin Sarban, spor yapmanın bilinçli yapılırsa faydalı olduğunu, uzay yürüyüşü (ayakta yüksek eğimle pedal çevirmek), step (basamak çıkmak), bisiklet ve spinning (bisikletle yapılan bir egzersiz) sporlarının dizlerde aşırı zorlanma, yüklenmeyle birlikte diz eklemlerine zarar verdiğini söylüyor. Doç. Dr. Sezgin Sarban, şu bilgileri veriyor: "Vücudun kendi ağırlığını kullanarak yapılan germe egzersizlerinde bile eklemler zorlanınca bırakmak, ısrarcı ve inatçı bir tutum içinde olmamak gerekiyor. Bu sporun yatay şekilde yapılmasından dolayı yerçekimi kuvveti eklemlere dik gelir ve eklem kıkırdaklarına yüklenme oluşturmaz. Kas zorlanmaları, kas iltihapları, ödemler, dejeneratif yırtıklar dediğimiz ve yıpranmaya yatkın menisküslerde oluşan yırtıklar, uzun süre hiç spor yapmayan kişiler birden spora başlayınca görülebiliyor.

Summary for Politics Category Article

Cumhuriyeti kuran CHP' nin, Türkiye ve dünya siyasetinde önemli bir yeri bulunduğunu dile getiren Bilgehan, CHP' nin, tek parti ve tek adam rejimini değiştiren devrimci bir parti olduğunu ifade etti. Bilgehan, daha sonra Mısırlı gençlerin sorularını yanıtladı. Filistin halkına uygulanan zulmü insanlık suçu olarak görüyoruz ama Ortadoğu'ya barışın gelmesi için bütün tarafların, İsrail dahil, uzlaşması gerektiğine inanıyoruz. Mısır halkının tek adam rejiminin sıkıntılarını yaşadığını söyleyen Bilgehan, "Biz de tek adam rejimine geçilmesinden

korktuğumuz için seçim çalışmalarında sizin özgürlük hareketinizi örnek gösteriyoruz" dedi.

Summary for Technology Category Article (1)

İnternette "filtre" iddiaları karşısında temel hak ve özgürlüklerin ihlal edileceğini savunan "Anonymous" (Anonim) adlı bir grup, Türkiye' deki çeşitli kamu kuruluşları ile bazı medya sitelerine yönelik siber saldırı düzenleyeceği tehdidinde bulundu. İlk büyük eylemlerini aylar önce Wikileaks' e yönelik ambargo uygulayan Paypal, Visa ve Mastercard gibi online ödeme ve kredi kartı firmalarına karşı gerçekleştiren grup üyeleri, dün kendilerine ait internet sitesinde Türkiye'yi hedef alan bir mesaj yayınladı. İlk organize saldırı perşembe saat 18.00'de olacak. Türkiye' deki kimi kamu kurumlarını hedef alan ilk organize saldırının 9 Haziran Perşembe günü saat 18.00' de (TSİ) gerçekleştirileceğini duyuran Anonymous, eyleme katılmak isteyenlerin bilgisayarlarına bazı yazılımlar indirerek, bu programları aktif hale getirmesini istedi. Tüm bu gelişmeleri yakından izleyen siber güvenlik uzmanları, söz konusu tehdidin önemine dikkat çekerek, birçok sitenin "DDoS" olarak adlandırılan saldırı ile karşı karşıya kalabileceğini bildirdi. Anonymous grubunun saldırılarda özel bir yöntem kullandığını belirten uzmanlar, "Klasik DDoS saldırılarında 'zombi' haline getirilmiş bilgisayarlar kullanılırken, bu grup tamamen 'gönüllü zombi' bilgisayarlarla saldırıyor. "Phishing" yöntemi ile kullanıcılara e-mail göndererek, bilgisayarlarından erişim sağladıkları banka, kredi kartı bilgi ve şifrelerinin yanı sıra kurumlarına ait kimi özel bilgi ve belgelerin sızdırılmaya çalışılacağını anlatan uzmanlar, böylesi bir durumun yeni bir "Wikileaks" olayına kapı açacağına vurgu yaptı. Bu arada Anonymous grubu üyeleri Perşembe günkü saldırıda çeşitli kamu kurum ve kuruluşlarına ait internet sitelerinin yanı sıra siyasi partiler ve kimi medya sitelerini de kendilerine hedef seçtiklerini açıkladı. Bu tür organize saldırılarda, servis sağlayıcılarının saldırının hedef aldığı web sayfasın yönelik tüm girişimlere anında müdahale edebileceğini anlatan Önal, bunun için saldırı yönteminin çok iyi analiz edilmesi ve engelleme sistemlerinin devreye sokulması gerektiğini anlattı. Grup üyelerinin Türkiye' ye yönelik bu saldırı planını

'protesto' olarak gören ve gönüllü olarak katılmayı planlayan çok sayıda Türk vatandaşı ile karşılaştık.

Summary for Technology Category Article (2)

Şirket, online oyun kullanıcılarını kredi kartı ve diğer kişisel bilgilerinin çalınmış olabileceği konusunda uyarmış, ardından da PlayStation şebekesini kapatmıştı. Milliyet'in haberine göre Sony'nin ikinci adamı Kazuo Hirai "Datalarını tehlikeye attığımız, endişelendirdiğimiz ve rahatsızlık verdiğimiz tüm kullanıcılarımızdan özür dileriz" diyerek geleneksel Japon selamı ile eğilerek kullanıcılardan özür diledi. 30 günlük ücretsiz Playstation Plus üyeliği (Mevcut Playstation Plus üyelerine de ekstra 30 günlük ücretsiz kullanım hakkı tanınacak). Sony ayrıca hackerların kullanıcıların kart bilgilerini çaldığını tam olarak kabul etmese de, yine de bundan dolayı finansal bir zarar gören olursa bu zararı tazmin edeceğini de belirtti.

Appendix 13. Summarizations of Turkish News Articles by Person 3

Summary for World Category Article (1)

The New York Times (NYT), 'Türkiye bağlantılı okullar Teksas' ta büyüyor' başlıklı haberinde Gülen Hareketi' nin ABD' deki eğitim ağlarını ve öğretmenler için alınan özel vizeleri gündeme getirdi. NYT, yayımladığı bir şemaya da, cemaatin ABD' de kurulu dernek ve vakıflarını gösterip, hepsini şemanın tepesinde bulunan Fethullah Gülen fotoğrafına bağladı. Gazetede birinci sayfada verilen haber, 20 sayfada tam, 21. sayfada da yarım sayfa olarak yer alırken, haber NYT' nin internet sitesinde de yaklaşık 7 sayfa yer buldu. NYT, TMD adlı bir inşaat firmasının kuruluşunun üzerinden bir ay geçmeden Gülen' e yakınlığıyla bilinen Harmony (Uyum) Okulları' nın 8.2 milyonluk inşaat ihalesini kazandığını ve bu firmanın Türkiye ile bağlantılı olduğunu öne sürdü. Stephanie Saul imzalı haberde Fethullah Gülen' in karizmatik bir Türk Vaizi olduğu ve İslam' ın ılımlı yüzünü tüm dünyaya yayarak dinsel, sosyal ve milliyetçi bir hareket kurmak için kendini adadığı dile getirildi. Harmony Okulları' nda 2011 yılında 1.500 öğretmenin istihdam edildiği ve bunların 292' sinin 'yüksek nitelikli eleman' olarak nitelenen 'H-1B' vizesi sahibi olduğu yazıldı.

Summary for World Category Article (2)

Singapur' da yapılan Uluslararası Havayolu Taşıyıcıları (IATA) bu yılki genel kurulunda havalimanlarında gelecekte güvenlik kontrolünü hızlandıracak yeni uygulamalar tanıtıldı. 'Geleceğin kontrol noktası' olarak adlandırılan uygulamada, yolcular güvenlik risklerine göre üçe ayrılacak ve 6.1 metrelik koridordan yürüyerek geçecek. Yolcu yürürken, üzerindeki her şey, kabine alacağı çanta, yüksek hassaslıktaki özel cihazlar tarafından taranacak. IATA' nın planına göre bu teknoloji önümüzdeki yıldan itibaren havalimanlarında kullanılmaya başlanacak. 5 yıl içinde de hızla yaygınlaşacak. Bu sayede uzun güvenlik kuyrukları tarihe karışacak. Yolcunun bilgilerinin yanı sıra vize bilgilerinin içinde yer aldığı çipli pasaport da kontrolleri hızlandıracak. Ancak bilgi paylaşımı konusunda uluslararası anlaşmaların yeniden ele alınması ve ülkelerin diğerlerine vatandaş bilgilerine ulaşım konusunda

engelleri kaldırması gerekiyor. 2050' de yolcu sayısı 16 milyara çıkacak. Uçuş emniyetine büyük önem veren IATA, bu konuda yüzde 42' lik gelişim yakaladı. Her 1.6 milyon uçuş saatinde 1 olan kaza oranı, 4 milyon saate yükseltildi.

Summary for World Category Article (3)

Analizin sonucunda, Çinli çiftçilerin SD1 olarak bilinen geni çeşitlendirilerek pirinç bitkisinin gövde uzunluğunu kısalttıklarını ortaya çıktı. SD1 üzerinde yapılan değişiklikler, pirincin daha kısa sürede olgunlaşmasını, irileşmesini sağladığı gibi üretim miktarını da artırdı. Yamasaki, antik çağlarda yaşamış olan çiftçilerin yapay seleksiyon yöntemiyle SD1 üzerinde değişiklik yaptıklarını ve daha kısa gövdeli pirinç bitkisi üretmeyi başardıklarını belirtti.

Summary for Financial Category Article

FT, Merkez Bankası'nın politikasının ekonominin ısınmasını durduramadığı yönündeki korkuyu körükleyeceği yorumunu yaptı. FT, "Türkiye' nin Alışılmamış Para Politikası: Yeniden Düşünme Zamanı mı?" başlıklı haberinde, "Türk politika yapıcıları, seçimlere bir hafta kalan enflasyon verileriyle hoş olmayan bir sürpriz yaşadıktan sonra seçimler sona erdiğinde faiz oranlarını yükseltme ve maliye politikasını sıkıştırma yönünde artan bir baskı altında kalacak" diye yazdı. Haberde enflasyonun mayısta yaptığı "sürpriz" in büyük ölçüde gıda fiyatlarından kaynaklandığı belirtilerek, "Gıda fiyatları her zaman değişken. Bu yüzden analistler, cuma verilerinin Merkez Bankası'nın enflasyon görünümünü veya politika tutumunu değiştirmesini beklemiyorlar" denildi.

Summary for Journal Category Article

Sahte içkiden zehirlenen Rus rehber Victoria Nikoloeva, yaşam savaşını sürdürüyor. Bizim için o bir yoğun bakım hastası. Hastamız bugüne kadar hiçbir tedaviye cevap vermedi. Ancak bitkisel hayata girdi diyebileceğimiz bir durum da henüz oluşmadı. Yani beyninde kan dolaşımı olduğunu görüyoruz. Alanya' da, yediemine alınan 16 şişe viskiye savcılık kararıyla el konulduğu bildirilen

açıklamada, alkollü içkilerle ilgili eğitimlere başlandığı ve hazırlanan 5 bin bilgilendirme broşürünün dağıtılmaya başlandığı kaydedildi. Sahte alkolden zehirlenen Rus rehberlerden 28 yaşındaki Maria Shalyapina ve 22 yaşındaki Zalyaeva Auilia Antalya' daki hastanede, 24 yaşındaki Alexandr Zhbckov Denizli' deki hastanede, 22 yaşındaki Marina Şevelyova ise döndükten sonra ülkesinde yaşamını yitirmişti.

Summary for Sports Category Article

Somalili korsanların elinde tutsak kalan bir denizci, birkaç gün önce Türkiye' ye gelip gazeteleri açsa, "Beşiktaş şampiyon olmuş, Trabzonspor küme düşmüş" sanır! "Bu işin genel seçimi de var" aşaması bile geride kaldı! Siyaset kesmedi. Trabzonspor Başkanı Sadri Şener, "İnsanlar ahireti de hesaba katsın, bu işin öbür tarafı da var" şeklindeki "ilahi" uyarısını yaptıktan sonra resmen açıkladı: "Fenerbahçe şampiyon değil ki"! Buyurun bakalım! Kim şampiyon o zaman? "Eğer siz Trabzon' u dövülmüş, elinden kupası alınmış gibi gösterirseniz sokaktaki insanın tepkisi büyür, sizi sorgular. Çok kötü bir sezon geçirilmedi. Travmayı atlatıp yeni rota çizmek lazım. Yoksa kaos kimseye yaramaz". Bilen bilir; "kalabalıkların sadece gaz pedalı vardır", "freni" yoktur. Süper Lig'in dörtte üçünü önde götürüp, son anda Fenerbahçe tarafından geçilen ve kendilerinden başka her futbolsever tarafından son derece başarılı görülen Trabzonspor bırakın hedef şaşırtmayı, laf karıştırmayı sadece yarına odaklanan ve "yıldızlarımızı iyi yönetemedik" şeklinde açık açık özeleştiri yapmaktan kaçınmayan Beşiktaş' tan örnek alsın. Yarına baksın.

Summary for Health Category Article

International Hospital Ortopedi ve Travmatoloji Uzmanı Doç. Dr. Sezgin Sarban, spor yapmanın bilinçli yapılırsa faydalı olduğunu, uzay yürüyüşü (ayakta yüksek eğimle pedal çevirmek), step (basamak çıkmak), bisiklet ve spinning (bisikletle yapılan bir egzersiz) sporlarının dizlerde aşırı zorlanma, yüklenmeyle birlikte diz eklemlerine zarar verdiğini söylüyor. Aynı şekilde koşu bandında spor yaparken, belli bir tempoda, eğimi çok artırmadan yürüyüş yapılmalı ve saat 5-6 km

sınırlarında kalınmalıdır. En doğru egzersiz biçimi, vücudu gerektiği şekilde ısıttıktan sonra, her bir kasın doğru şekilde çalıştırılmasıdır. Her bir kasımız için kondisyoner eşliğinde kaldırabileceğimiz, alabileceğimiz bir yük miktarı var. Belli tekrarlarla bu ağırlıkları kaldırmak faydalı olacaktır." Özellikle vücutlarının belli bölgelerinde ağrıları bulunan, kireçlemeleri olanların ortopedi ve fizik tedavi bölümlerinden destek alması gerekiyor. Eğer kişinin belirgin fıtığı varsa beyin cerrahisine yönlendirilmesi önem taşıyor. Yüzme, eklemlere aşırı yük vermeden kasların güçlenmesini sağlar. Bunun yanında yüzme sporu, vücuttaki tüm kasları senkronize şekilde aynı anda çalıştıran tek spordur. Doç. Dr. Sarban, menisküs yırtığında kurbağalama stili, omuz sıkışması hastalığı olanların da serbest stil yüzmeden zarar görebileceğini belirtti. Özellikle parklarda, açık alanlarda bulunan spor aletlerinin bilinçsizce kullanılması yeni sakatlanmalara neden olabiliyor. Bu kişilerin vücutlarını çok fazla zorlamadan spor yapmalarında yarar var. Spora yaklaşık 15-20 dakika tempolu yürüyüş yaptıktan sonra başlamak gerekiyor.

Summary for Politics Category Article

CHP Genel Başkan Yardımcısı Gülsün Bilgehan, gençlere, CHP ve Türkiye' deki demokratik sistem hakkında bilgi verdi. Türkiye' nin ortak değerlere sahip olduğu ülkeler için önemli bir örnek oluşturduğunu anlatan Bilgehan, seçimlerde bir sınavdan geçeceklerini, Tahrir Meydanından gelen gençlerin ziyaretinin partiye uğurlu gelmesini diledi. Mısırlı bir gencin, "CHP, Filistin halkına karşı neden İsrail' i destekliyor" sorusuna karşılık Bilgehan, "CHP' nin İsrail' i desteklediğini herhalde AKP ziyaretinde duydunuz. Çünkü biz CHP olarak Ortadoğu'da her zaman dengeli bir politika izlemek gerektiğine inandık. Bilgehan, Türkiye' nin bugünkü ekonomik durumuyla ilgili çekinceleri olduğunu, eşit ve adil bir gelir dağılımı bulunmadığını ifade etti. Kadın ve gençlik kollarına çok önem verdiklerini, kadınların destekleyeceği partinin seçimi kazanacağını vurgulayan Bilgehan, bu nedenle kadın kollarının çok çalıştığını dile getirdi.

Summary for Technology Category Article (1)

İlk büyük eylemlerini aylar önce Wikileaks' e yönelik ambargo uygulayan Paypal, Visa ve Mastercard gibi online ödeme ve kredi kartı firmalarına karşı gerçekleştiren grup üyeleri, dün kendilerine ait internet sitesinde Türkiye'yi hedef alan bir mesaj yayınladı. Sosyal paylaşım siteleri üzerinden örgütlenen grup üyeleri, "IRC" isimli anlık mesajlaşma kanallarında siber saldırının hedefi ve zamanlaması konusunda bilgiler aktararak, uygulanacak olan yöntem ve stratejiler hakkında çeşitli bilgiler verdi. Söz konusu saldırılar karşısında bazı internet sitelerine erişim bir süre engellenirken, "hack" leme girişimi başarısızlıkla sonuçlandı. "Bu saldırı yöntemi; 50 kişilik bir otobüse 1000 kişinin binmesi gibi bir şey" yorumunda bulundu. Bu tip saldırılarda amacın, "bilgi çalmak" yerine sistemin erişilemez hale getirilmesi olduğunu kaydeden siber güvenlik uzmanları, burada verilmek istenen mesajın "Bizim internetimize karışırsanız biz de sizin internetinizi kapatırız" anlamı taşıdığını öne sürdü. Anonymous grubunun saldırılarda özel bir yöntem kullandığını belirten uzmanlar, "Klasik DDoS saldırılarında 'zombi' haline getirilmiş bilgisayarlar kullanılırken, bu grup tamamen 'gönüllü zombi' bilgisayarlarla saldırıyor. Gönüllü zombi olabilmeniz için de size bir program yüklettiriyor ve bu programı IRC üzerinden şuraya saldır buraya saldır şeklinde yönlendiriyorlar" bilgisini verdi. DDoS saldırılarından bir sonraki adımın çeşitli devlet kurumlarına ait telefon ve haberleşme sistemlerini çalışamaz hale getirmek olduğunu kaydeden siber güvenlik uzmanları, daha sonra hedefteki kurumlarda çalışan kişilerin bilgisayarlarına "phishing" saldırıları gerçekleştirileceğinin altını çizdi. Eğer gerekli önlemler alınmazsa ciddi bir siber savaşın başlamış olacağını bildiren uzmanlar, kimi hassas hedeflerin saldırıya uğraması durumunda Türkiye' deki internet ağının bir süre için çökebileceği uyarısında bulundu. Proxy ve benzeri yöntemler ise IP adreslerini değiştirerek saldırıya katılacakların ciddi bir etkiye sahip olmayacağı görüşünü belirten Önal, kendi IP'leri ile eyleme destek verenlerin güvenlik güçleri tarafından kısa sürede tespit edileceğinin altını çizdi. Protesto amaçlı da olsa böylesi bir saldırıya katılanlara, Türk Ceza Kanunu' na (TCK) göre suç işlediklerini hatırlatmak isterim. İnternet sitelerine bu şekilde zarar verenlerin, 5 yıla kadar hapis cezası ile yargılanacaklarını unutmamalı."

Summary for Technology Category Article (2)

Şirket, online oyun kullanıcılarını kredi kartı ve diğer kişisel bilgilerinin çalınmış olabileceği konusunda uyarmış, ardından da PlayStation şebekesini kapatmıştı. "PlayStation 3' ün yazılımını yükselteceğiz. Bu süreçte, tüm kullanıcılarımız PlayStation şebeke hesaplarının şifrelerini yenileyecek. Bunun onlar için büyük bir sıkıntı olduğunu biliyorum." Sony rakiplere kaptırmaktan korktuğu kullanıcıları için iyi niyet ve özür göstergesi olarak ücretsiz hediyeler vereceğini açıkladı. Sony ayrıca hackerların kullanıcıların kart bilgilerini çaldığını tam olarak kabul etmese de, yine de bundan dolayı finansal bir zarar gören olursa bu zararı tazmin edeceğini de belirtti. 59 ülkede 77 milyon kişi PlayStation şebekesini kullanıyor.

Appendix 14. Summarizations of Turkish News Articles by Person 4

Summary for World Category Article (1)

ABD' nin en saygın gazetelerinden The New York Times (NYT), birinci sayfasından yayımladığı 'Türkiye bağlantılı okullar Teksas' ta büyüyor' başlıklı haberinde Gülen Hareketi' nin ABD' deki eğitim ağlarını ve öğretmenler için alınan özel vizeleri gündeme getirdi. Haberde Gülen Hareketi' nin Charter adı verilen ve kamu kaynaklarıyla işletilen özel okullardan, inşaat faaliyetlerine, özel vize ile getirilen öğretmenlerden, okullardaki öğrencilerin başarısına kadar dek uzanan konular işlendi. NYT, TMD adlı bir inşaat firmasının kuruluşunun üzerinden bir ay geçmeden Gülen' e yakınlığıyla bilinen Harmony (Uyum) Okulları'nın 8.2 milyonluk inşaat ihalesini kazandığını ve bu firmanın Türkiye ile bağlantılı olduğunu öne sürdü. TMD' nin 2009 yılından bu yana 50 milyon dolarlık inşaat ihalesi aldığını yazan gazete, Harmony Okulları'nın 16 bin öğrenci ve 33 şube ile Teksas' taki en büyük kamu kaynakları kullanan özel okullar zinciri olduğunu ve yıllık 100 milyon dolar yardım aldığını yazdı. Gülen hareketiyle doğrudan bağlantılı olduğu belirtilen ve ABD' nin 25 eyaletinde 120 okul bulunduğu yazılan haberde, okulların Amerikan öğrencilerinin genellikle başarısız olduğu bilim ve matematik konularında ağırlıklı olarak eğitim verdiği vurgulandı. Harmony Okulları'nda 2011 yılında 1.500 öğretmenin istihdam edildiği ve bunların 292'sinin 'yüksek nitelikli eleman' olarak nitelenen 'H-1B' vizesi sahibi olduğu yazıldı. ABD Federal Çalışma Bakanlığı'nın 'H-1B' vizesi sahibi bu Türk öğretmenlerin bir bölümünün yeterince deneyimli olmadığı ve okul yöneticilerinin çevrelerindeki Amerikalı öğretmenleri çalıştırmak istemediklerine yönelik iddiaları incelediği de anımsatılan haberde bazı işçi sendikalarının bu durum nedeniyle okullara tepkili olduğu belirtildi.

Summary for World Category Article (2)

Singapur' da yapılan Uluslararası Havayolu Taşıyıcıları (IATA) bu yılki genel kurulunda havalimanlarında gelecekte güvenlik kontrolünü hızlandıracak yeni uygulamalar tanıtıldı. 'Geleceğin kontrol noktası' olarak adlandırılan uygulamada, yolcular güvenlik risklerine göre üçe ayrılacak ve 6.1 metrelik koridordan yürüyerek geçecek. Yolcu yürürken, üzerindeki her şey, kabine alacağı çanta, yüksek

hassaslıktaki özel cihazlar tarafından taranacak. Sıvı ve elektronik cihazlar çantadan çıkartılmadan kontrol edilebilecek. Bu da güvenlik işlemlerini hızlandıracak. IATA' nın planına göre bu teknoloji önümüzdeki yıldan itibaren havalimanlarında kullanılmaya başlanacak. IATA ayrıca pasaport kuyruklarının da azaltılması için çipli pasaport uygulamasına hızla geçilmesini tavsiye ediyor. Yolcunun bilgilerinin yanı sıra vize bilgilerinin içinde yer aldığı çipli pasaport da kontrolleri hızlandıracak. Ayrıca yolcunun göz bebeğinin taranmasıyla kimlik bilgileri kontrol edilecek. Bu sistemin daha fazla havalimanında kullanılması ve sisteme çipli pasaporttaki tüm bilgilerin yüklenmesi planlanıyor. Ancak bilgi paylaşımı konusunda uluslararası anlaşmaların yeniden ele alınması ve ülkelerin diğerlerine vatandaş bilgilerine ulaşım konusunda engelleri kaldırması gerekiyor.

Summary for World Category Article (3)

Japon bilim insanlarının pirincin alt türleri üzerinde yaptığı incelemede, tüm DNA bilgilerini içeren genomlar analiz edildi. Analizin sonucunda, Çinli çiftçilerin SD1 olarak bilinen geni çeşitlendirilerek pirinç bitkisinin gövde uzunluğunu kısalttıklarını ortaya çıktı. Araştırma ekibinin başındaki isim Dr. Masanori Yamasaki, SD1' in son 50 yılda pirinç yetiştirilmesinde en önemli role sahip gen olduğuna dikkat çekti. SD1 üzerinde yapılan değişiklikler, pirincin daha kısa sürede olgunlaşmasını, irileşmesini sağladığı gibi üretim miktarını da artırdı. Yamasaki, antik çağlarda yaşamış olan çiftçilerin yapay seleksiyon yöntemiyle SD1 üzerinde değişiklik yaptıklarını ve daha kısa gövdeli pirinç bitkisi üretmeyi başardıklarını belirtti.

Summary for Financial Category Article

İngiliz ekonomi gazetesi Financial Times (FT), yüzde 2,42'lik yüksek mayıs enflasyonunun ardından "Türkiye' nin alışılmamış para politikası konusunda yeniden düşünme zamanı mı?" sorusunu ortaya attı. FT, "Türkiye' nin Alışılmamış Para Politikası: Yeniden Düşünme Zamanı mı?" başlıklı haberinde, "Türk politika yapıcıları, seçimlere bir hafta kalan enflasyon verileriyle hoş olmayan bir sürpriz yaşadıktan sonra seçimler sona erdiğinde faiz oranlarını yükseltme ve maliye

politikasını sıkıştırma yönünde artan bir baskı altında kalacak" diye yazdı. Tüketici enflasyonun mayısta beklenenden bir kat fazla geldiğine işaret edilen haberde şöyle denildi: "Veriler, Merkez Bankası'nın sermaye girişini caydırmaya yönelik düşük faiz ve iç talebi dizginlemeyi amaçlayan yüksek zorunlu karşılıkların karışımından oluşan alışılmamış politikası ile ekonominin ısınmasını durduramadığı korkusunu körükleyecek." Cari işlemler açığına dikkat çeken gazete, iç talep ve kredi patlamasına yansıyan hızlı büyümeyi sert bir inişin izleyebileceğinden korkan yatırımcıların tedirgin olduğunu yazdı. FT' ye konuşan Barclays Capital ekonomisti Christian Keller de "Piyasalar, yüksek turizm gelirlerinin Türkiye' nin cari işlemlerini olumlu etkileyeceği umuduyla, küresel büyüme ve enflasyon gelişmelerini izleyerek ve Merkez Bankası'nın kredi politikasına zaman tanıyarak, yazın beklemeyi kabul edebilir" dedi.

Summary for Journal Category Article

Bodrum'da sahte içkiden zehirlenen ve Akdeniz Üniversitesi Hastanesi'ne getirilerek yoğun bakım ünitesinde tedaviye alınan 23 yaşındaki Rus rehber Victoria Nikoloeva, yaşam savaşını sürdürüyor. Öte yandan Antalya Valiliği'nden bugün yaptığı yazılı açıklamada, alkollü içki denetimlerinin sürdüğü bildirildi. Alanya' da, yediemine alınan 16 şişe viskiye savcılık kararıyla el konulduğu bildirilen açıklamada, alkollü içkilerle ilgili eğitimlere başlandığı ve hazırlanan 5 bin bilgilendirme broşürünün dağıtılmaya başlandığı kaydedildi. Sahte alkolden zehirlenen Rus rehberlerden 28 yaşındaki Maria Shalyapina ve 22 yaşındaki Zalyaeva Auilia Antalya'daki hastanede, 24 yaşındaki Alexandr Zhbckov Denizli' deki hastanede, 22 yaşındaki Marina Şevelyova ise döndükten sonra ülkesinde yaşamını yitirmişti.

Summary for Sports Category Article

Trabzonspor Başkanı Sadri Şener, "İnsanlar ahireti de hesaba katsın, bu işin öbür tarafı da var" şeklindeki "ilahi" uyarısını yaptıktan sonra resmen açıkladı: "Fenerbahçe şampiyon değil ki"! Buyurun bakalım! Kim şampiyon o zaman? Beşiktaş olmalı ki, Başkan Yıldırım Demirören çıkıyor "Teknik direktörümüzden

memnunuz, on yıllık mukaveleyi bile tartıştık" diyor. Trabzonspor yönetiminin yaptığına ne denir peki? İşte "ezeli ve ebedi" Trabzonspor taraftarı, efsane başkan Mehmet Ali Yılmaz'ın sözleri: "Eğer siz Trabzon' u dövülmüş, elinden kupası alınmış gibi gösterirseniz sokaktaki insanın tepkisi büyür, sizi sorgular. Çok kötü bir sezon geçirilmedi. Travmayı atlatıp yeni rota çizmek lazım. Yoksa kaos kimseye yaramaz". Evet. Kaçan şampiyonluğun acısını unutturmak için kaostan medet umuluyor. "Trabzonspor dövüldü" imajı enjekte ediliyor. Bu da mümkün elbet! Lakin "yarını" unutmak, hatta yarını "karartmak" riski mevcuttur. Süper Lig'in dörtte üçünü önde götürüp, son anda Fenerbahçe tarafından geçilen ve kendilerinden başka her futbolsever tarafından son derece başarılı görülen Trabzonspor - bırakın hedef şaşırtmayı / laf karıştırmayı sadece yarına odaklanan ve "yıldızlarımızı iyi yönetemedik" şeklinde açık açık özeleştiri yapmaktan kaçınmayan Beşiktaş' tan örnek alsın. Yarına baksın.

Summary for Health Category Article

International Hospital Ortopedi ve Travmatoloji Uzmanı Doç. Dr. Sezgin Sarban, spor yapmanın bilinçli yapılırsa faydalı olduğunu, uzay yürüyüşü (ayakta yüksek eğimle pedal çevirmek), step (basamak çıkmak), bisiklet ve spinning (bisikletle yapılan bir egzersiz) sporlarının dizlerde aşırı zorlanma, yüklenmeyle birlikte diz eklemlerine zarar verdiğini söylüyor. Spor yapmadan önce vücuttaki olası rahatsızlıkları öğrenmek önem taşıyor. Tepeden tırnağa bir değerlendirmeden geçmek de omurgayı koruyacak şekilde, bilinçli egzersiz yapılmasını sağlıyor. Hiç spor yapmamış, kasları çalışmamış, hareketsiz kalan kişilerde yaralanmaların olabileceğine değinen Doç. Dr. Sezgin Sarban, şunlara dikkat edilmesi gerektiğini söylüyor: "Kas zorlanmaları, kas iltihapları, ödemler, dejeneratif yırtıklar dediğimiz ve yıpranmaya yatkın menisküslerde oluşan yırtıklar, uzun süre hiç spor yapmayan kişiler birden spora başlayınca görülebiliyor. 50-60 yaşında olup da, hiç aktif spor yapmayanlarda menisküste bulunan yırtıklar daha da artabiliyor. Bu kişilerin vücutlarını çok fazla zorlamadan spor yapmalarında yarar var. Diyetisyen desteği de alınarak spor yapılması sağlıklı olmaya yardımcıdır."

Summary for Politics Category Article

CHP Genel Başkan Yardımcısı Gülsün Bilgehan, gençlere, CHP ve Türkiye' deki demokratik sistem hakkında bilgi verdi. Tahrir Meydanında özgürlük hareketini başlatan gençleri ağırlamaktan mutluluk duyduklarını vurgulayan Bilgehan, Mısır'ın, Türkiye için dost ve kardeş bir ülke olduğunu söyledi. Cumhuriyeti kuran CHP' nin, Türkiye ve dünya siyasetinde önemli bir yeri bulunduğunu dile getiren Bilgehan, CHP' nin, tek parti ve tek adam rejimini değiştiren devrimci bir parti olduğunu ifade etti. Mısırlı bir gencin, "CHP, Filistin halkına karşı neden İsrail'i destekliyor" sorusuna karşılık Bilgehan, "CHP' nin İsrail'i desteklediğini herhalde AKP ziyaretinde duydunuz. Çünkü biz CHP olarak Ortadoğu'da her zaman dengeli bir politika izlemek gerektiğine inandık. Filistin halkına uygulanan zulmü insanlık suçu olarak görüyoruz ama Ortadoğu'ya barışın gelmesi için bütün tarafların, İsrail dahil, uzlaşması gerektiğine inanıyoruz. Mısır'da da böyle düşünen pek çok aydın, siyasetçi olduğunu biliyoruz" yanıtını verdi. Seçim sürecinde liderlerin oldukça sert görüntü verdiklerini ifade eden Bilgehan, seçimlerin ardından Mecliste yapıcı ve olumlu bir şeklide çalışacaklarını söyledi. Bilgehan, seçim sonucu ne olursa olsun Türkiye' nin Ortadoğu için model olmaya devam etmesine gayret edeceklerini belirtti.

Summary for Technology Category Article (1)

İnternette "filtre" iddiaları karşısında temel hak ve özgürlüklerin ihlal edileceğini savunan "Anonymous" (Anonim) adlı bir grup, Türkiye' deki çeşitli kamu kuruluşları ile bazı medya sitelerine yönelik siber saldırı düzenleyeceği tehdidinde bulundu. Türkiye'de internet kullanıcılarının "filtre" uygulaması ile sansüre uğrayacağını savunulan mesajda "operationturkey" (Türkiye Operasyonu) adıyla "siber savaş" ilan edilirken, öncelikle "sansür" uygulayan kurumlara karşı harekete geçileceği açıklandı. İlk organize saldırı perşembe saat 18.00'de olacak. Türkiye' deki kimi kamu kurumlarını hedef alan ilk organize saldırının 9 Haziran Perşembe günü saat 18.00'de (TSİ) gerçekleştirileceğini duyuran Anonymous, eyleme katılmak isteyenlerin bilgisayarlarına bazı yazılımlar indirerek, bu programları aktif hale getirmesini istedi. Tüm bu gelişmeleri yakından izleyen siber güvenlik uzmanları,

söz konusu tehdidin önemine dikkat çekerek, birçok sitenin “DDoS” olarak adlandırılan saldırı ile karşı karşıya kalabileceğini bildirdi. 2007'de nam salan “DDoS” saldırı yöntemi ile hedef alınan bir internet sitesinin aynı anda yoğun bir ziyaretçi akınına uğratılarak, web sitesi veya DNS (Alan Adı Sistemi) sunucularının kullanılmaz hale getirildiğini belirten uzmanlar, “Bu saldırı yöntemi; 50 kişilik bir otobüse 1000 kişinin binmesi gibi bir şey” yorumunda bulundu. Siber saldırı sırasında kendi IP adresleri yerine VPN (Sanal Paylaşımlı Ağ) üzerinden IP adreslerini değiştiren grup üyelerinin bu sayede izlerini bazı kişilerin yönlendirmesi ile kendilerine hedefler seçtiklerine dikkat çekti. DDoS saldırılarından bir sonraki adımın çeşitli devlet kurumlarına ait telefon ve haberleşme sistemlerini çalışamaz hale getirmek olduğunu kaydeden siber güvenlik uzmanları, daha sonra hedefteki kurumlarda çalışan kişilerin bilgisayarlarına “phishing” saldırıları gerçekleştirileceğinin altını çizdi. “Phishing” yöntemi ile kullanıcılara e-mail göndererek, bilgisayarlarından erişim sağladıkları banka, kredi kartı bilgi ve şifrelerinin yanı sıra kurumlarına ait kimi özel bilgi ve belgelerin sızdırılmaya çalışılacağını anlatan uzmanlar, böylesi bir durumun yeni bir “Wikileaks” olayına kapı açacağına vurgu yaptı. Eğer gerekli önlemler alınmazsa ciddi bir siber savaşın başlamış olacağını bildiren uzmanlar, kimi hassas hedeflerin saldırıya uğraması durumunda Türkiye’ deki internet ağının bir süre için çökebileceği uyarısında bulundu. Anonymous' un bugüne kadar gerçekleştirdiği hacker saldırılarını yakından takip eden Siber Güvenlik Uzmanı Huzeyfe Önal, yaptığı açıklamada, hedefteki kurumların kendi bünyelerinde alacağı çeşitli önlemlerin yanı sıra internet servis sağlayıcılarının da ciddi tedbirler alması gerektiğini bildirdi. Anonymous' a destek vermek isteyen çok sayıda kişinin, bu grubun internet sitesinden indirdikleri özel programlarla bilgisayarlarını saldırıda kullanılacak birer 'zombi' haline getirdiğini belirten Önal, sözlerine şöyle devam etti: “Grup üyelerinin Türkiye’ ye yönelik bu saldırı planını 'protesto' olarak gören ve gönüllü olarak katılmayı planlayan çok sayıda Türk vatandaşı ile karşılaştık. Burada dikkat edilmesi gereken en önemli husus protesto ile 'saldırının farklı şeyler olduğu. Protesto amaçlı da olsa böylesi bir saldırıya katılanlara, Türk Ceza Kanunu’ na (TCK) göre suç işlediklerini hatırlatmak isterim. İnternet sitelerine bu şekilde zarar verenlerin, 5 yıla kadar hapis cezası ile yargılanacaklarını unutmamalı.”

Summary for Technology Category Article (2)

Korsan kriziyle başı ağrıyan Japon elektronik üreticisi Sony, PlayStation şebekesinin bu hafta kısmen açılacağını açıkladı. Şirket, online oyun kullanıcılarını kredi kartı ve diğer kişisel bilgilerinin çalınmış olabileceği konusunda uyarmış, ardından da PlayStation şebekesini kapatmıştı. Milliyet'in haberine göre Sony'nin ikinci adamı Kazuo Hirai "Datalarını tehlikeye attığımız, endişelendirdiğimiz ve rahatsızlık verdiğimiz tüm kullanıcılarımızdan özür dileriz" diyerek geleneksel Japon selamı ile eğilerek kullanıcılardan özür diledi. Sony rakiplere kaptırmaktan korktuğu kullanıcıları için iyi niyet ve özür göstergesi olarak ücretsiz hediyeler vereceğini açıkladı. Sony ayrıca hackerların kullanıcıların kart bilgilerini çaldığını tam olarak kabul etmese de, yine de bundan dolayı finansal bir zarar gören olursa bu zararı tazmin edeceğini de belirtti. Kişisel bilgilerin daha iyi korunmasını sağlamak için sistemlerini yeniden kurma konusunda adımlar attıklarını belirten şirket, servislerin en kısa sürede tamamen hizmete gireceğini kaydetti. 59 ülkede 77 milyon kişi PlayStation şebekesini kullanıyor.

Appendix 15. General Summarizations of Turkish News Articles

Summary for World Category Article (1)

The New York Times (NYT), 'Türkiye bağlantılı okullar Teksas' ta büyüyor' başlıklı haberinde Gülen Hareketi' nin ABD' deki eğitim ağlarını ve öğretmenler için alınan özel vizeleri gündeme getirdi. ABD' nin en saygın gazetelerinden The New York Times (NYT), birinci sayfasından yayımladığı 'Türkiye bağlantılı okullar Teksas' ta büyüyor' başlıklı haberinde Gülen Hareketi' nin ABD' deki eğitim ağlarını ve öğretmenler için alınan özel vizeleri gündeme getirdi. NYT, TMD adlı bir inşaat firmasının kuruluşunun üzerinden bir ay geçmeden Gülen' e yakınlığıyla bilinen Harmony (Uyum) Okulları' nın 8.2 milyonluk inşaat ihalesini kazandığını ve bu firmanın Türkiye ile bağlantılı olduğunu öne sürdü. Gülen hareketiyle doğrudan bağlantılı olduğu belirtilen ve ABD' nin 25 eyaletinde 120 okul bulunduğu yazılan haberde, okulların Amerikan öğrencilerinin genellikle başarısız olduğu bilim ve matematik konularında ağırlıklı olarak eğitim verdiği vurgulandı. Harmony Okulları' nda 2011 yılında 1.500 öğretmenin istihdam edildiği ve bunların 292' sinin 'yüksek nitelikli eleman' olarak nitelenen 'H-1B' vizesi sahibi olduğu yazıldı.

Summary for World Category Article (2)

Uluslararası Havayolu Taşıyıcıları Birliği' nin (IATA) bu yılki genel kuruluna, havalimanlarında gelecekte güvenlik kontrolünü hızlandıracak yeni uygulamalar damga vurdu. Yeni uygulamada, yolcular güvenlik risklerine göre üçe ayrılacak. Singapur' da yapılan Uluslararası Havayolu Taşıyıcıları Birliği' nin (IATA) bu yılki genel kurulunda havalimanlarında gelecekte güvenlik kontrolünü hızlandıracak yeni uygulamalar tanıtıldı. 'Geleceğin kontrol noktası' olarak adlandırılan uygulamada, yolcular güvenlik risklerine göre üçe ayrılacak ve 6.1 metrelik koridordan yürüyerek geçecek. Yolcu yürürken, üzerindeki her şey, kabine alacağı çanta, yüksek hassaslıktaki özel cihazlar tarafından taranacak. Sıvı ve elektronik cihazlar çantadan çıkartılmadan kontrol edilebilecek. Bu da güvenlik işlemlerini hızlandıracak. Ayakkabı ve çantadan laptop (dizüstü bilgisayar) çıkarma son bulacak. IATA' nın planına göre bu teknoloji önümüzdeki yıldan itibaren havalimanlarında kullanılmaya

başlanacak. IATA ayrıca pasaport kuyruklarının da azaltılması için çipli pasaport uygulamasına hızla geçilmesini tavsiye ediyor. Yolcunun bilgilerinin yanı sıra vize bilgilerinin içinde yer aldığı çipli pasaport da kontrolleri hızlandıracak. Ayrıca yolcunun göz bebeğinin taranmasıyla kimlik bilgileri kontrol edilecek. Ancak bilgi paylaşımı konusunda uluslararası anlaşmaların yeniden ele alınması ve ülkelerin diğerlerine vatandaş bilgilerine ulaşım konusunda engelleri kaldırması gerekiyor. Uçuş emniyetine büyük önem veren IATA, bu konuda yüzde 42' lik gelişim yakaladı.

Summary for World Category Article (3)

Japon bilim insanlarının pirincin alt türleri üzerinde yaptığı incelemede, tüm DNA bilgilerini içeren genomlar analiz edildi. Analizin sonucunda, Çinli çiftçilerin SD1 olarak bilinen geni çeşitlendirilerek pirinç bitkisinin gövde uzunluğunu kısalttıklarını ortaya çıktı. SD1 üzerinde yapılan değişiklikler, pirincin daha kısa sürede olgunlaşmasını, irileşmesini sağladığı gibi üretim miktarını da artırdı. Yamasaki, antik çağlarda yaşamış olan çiftçilerin yapay seleksiyon yöntemiyle SD1 üzerinde değişiklik yaptıklarını ve daha kısa gövdeli pirinç bitkisi üretmeyi başardıklarını belirtti.

Summary for Financial Category Article

İngiliz ekonomi gazetesi Financial Times (FT), yüzde 2,42' lik yüksek mayıs enflasyonunun ardından "Türkiye' nin alışılmamış para politikası konusunda yeniden düşünme zamanı mı" sorusunu ortaya attı. FT, "Türkiye' nin Alışılmamış Para Politikası: Yeniden Düşünme Zamanı mı" başlıklı haberinde, "Türk politika yapıcıları, seçimlere bir hafta kalan enflasyon verileriyle hoş olmayan bir sürpriz yaşadıktan sonra seçimler sona erdiğinde faiz oranlarını yükseltme ve maliye politikasını sıkıştırma yönünde artan bir baskı altında kalacak" diye yazdı. Haberde enflasyonun mayısta yaptığı "sürpriz" in büyük ölçüde gıda fiyatlarından kaynaklandığı belirtilerek, "Gıda fiyatları her zaman değişken. Bu yüzden analistler, cuma verilerinin Merkez Bankası'nın enflasyon görünümünü veya politika tutumunu

değiştirmesini beklemiyorlar" denildi. FT' ye konuşan Barclays Capital ekonomisti Christian Keller de "Piyasalar, yüksek turizm gelirlerinin Türkiye' nin cari işlemlerini olumlu etkileyeceği umuduyla, küresel büyüme ve enflasyon gelişmelerini izleyerek ve Merkez Bankası' nın kredi politikasına zaman tanıyarak, yazın beklemeyi kabul edebilir" dedi.

Summary for Journal Category Article

Sahte içkiden zehirlenen Rus rehber Victoria Nikoloeva, yaşam savaşını sürdürüyor. Bodrum' da sahte içkiden zehirlenen ve Akdeniz Üniversitesi Hastanesi' ne getirilerek yoğun bakım ünitesinde tedaviye alınan 23 yaşındaki Rus rehber Victoria Nikoloeva, yaşam savaşını sürdürüyor. Hastamız bugüne kadar hiçbir tedaviye cevap vermedi. Ancak bitkisel hayata girdi diyebileceğimiz bir durum da henüz oluşmadı. Yani beyninde kan dolaşımı olduğunu görüyoruz. Öte yandan Antalya Valiliği' nden bugün yaptığı yazılı açıklamada, alkollü içki denetimlerinin sürdüğü bildirildi. Alanya' da, yediemine alınan 16 şişe viskiye savcılık kararıyla el konulduğu bildirilen açıklamada, alkollü içkilerle ilgili eğitimlere başlandığı ve hazırlanan 5 bin bilgilendirme broşürünün dağıtılmaya başlandığı kaydedildi. Sahte alkolden zehirlenen Rus rehberlerden 28 yaşındaki Maria Shalyapina ve 22 yaşındaki Zalyaeva Auilia Antalya' daki hastanede, 24 yaşındaki Alexandr Zhbckov Denizli' deki hastanede, 22 yaşındaki Marina Şevelyova ise döndükten sonra ülkesinde yaşamını yitirmişti.

Summary for Sports Category Article

Somalili korsanların elinde tutsak kalan bir denizci, birkaç gün önce Türkiye' ye gelip gazeteleri açsa, "Beşiktaş şampiyon olmuş, Trabzonspor küme düşmüş" sanır! Trabzonspor Başkanı Sadri Şener, "İnsanlar ahireti de hesaba katsın, bu işin öbür tarafı da var" şeklindeki "ilahi" uyarısını yaptıktan sonra resmen açıkladı: "Fenerbahçe şampiyon değil ki"! Buyurun bakalım! Kim şampiyon o zaman? Beşiktaş olmalı ki, Başkan Yıldırım Demirören çıkıyor "Teknik direktörümüzden memnunuz, on yıllık mukaveleyi bile tartıştık" diyor. Ama sürekli yenilenen Süper

Lig macerasının bir dönemine takılıp orada kalmak, uzatmak, hedefi yeni şampiyonluklardan kin ve intikam rotasına taşımak, sadece Trabzonspor için değil futbolun içindeki her unsur için korkutucudur. İşte "ezeli ve ebedi" Trabzonspor taraftarı, efsane başkan Mehmet Ali Yılmaz' ın sözleri: "Eğer siz Trabzon' u dövülmüş, elinden kupası alınmış gibi gösterirseniz sokaktaki insanın tepkisi büyür, sizi sorgular. Çok kötü bir sezon geçirilmedi. Travmayı atlatıp yeni rota çizmek lazım. Yoksa kaos kimseye yaramaz". Süper Lig' in dörtte üçünü önde götürüp, son anda Fenerbahçe tarafından geçilen ve kendilerinden başka her futbolsever tarafından son derece başarılı görülen Trabzonspor bırakın hedef şaşırtmayı, laf karıştırmayı sadece yarına odaklanan ve "yıldızlarımızı iyi yönetemedik" şeklinde açık açık özeleştiri yapmaktan kaçınmayan Beşiktaş' tan örnek alsın. Yarına baksın.

Summary for Health Category Article

International Hospital Ortopedi ve Travmatoloji Uzmanı Doç Dr Sezgin Sarban, spor yapmanın bilinçli yapılırsa faydalı olduğunu, uzay yürüyüşü (ayakta yüksek eğimle pedal çevirmek), step (basamak çıkmak), bisiklet ve spinning (bisikletle yapılan bir egzersiz) sporlarının dizlerde aşırı zorlanma, yüklenmeyle birlikte diz eklemlerine zarar verdiğini söylüyor. En doğru egzersiz biçimi, vücudu gerektiği şekilde ısıttıktan sonra, her bir kasın doğru şekilde çalıştırılmasıdır. Her bir kasımız için kondisyoner eşliğinde kaldırabileceğimiz, alabileceğimiz bir yük miktarı var. Belli tekrarlarla bu ağırlıkları kaldırmak faydalı olacaktır".

Spor yapmadan önce vücuttaki olası rahatsızlıkları öğrenmek önem taşıyor. Tepeden tırnağa bir değerlendirmeden geçmek de omurgayı koruyacak şekilde, bilinçli egzersiz yapılmasını sağlıyor. Hiç spor yapmamış, kasları çalışmamış, hareketsiz kalan kişilerde yaralanmaların olabileceğine değinen Doç Dr Sezgin Sarban, şunlara dikkat edilmesi gerektiğini söylüyor: "Kas zorlanmaları, kas iltihapları, ödemler, dejeneratif yırtıklar dediğimiz ve yıpranmaya yatkın menisküslerde oluşan yırtıklar, uzun süre hiç spor yapmayan kişiler birden spora başlayınca görülebiliyor. Bu kişilerin vücutlarını çok fazla zorlamadan spor yapmalarında yarar var. Spora yaklaşık 15-20 dakika tempolu yürüyüş yaptıktan sonra başlamak gerekiyor.

Summary for Politics Category Article

CHP Genel Başkan Yardımcısı Gülsün Bilgehan, gençlere, CHP ve Türkiye' deki demokratik sistem hakkında bilgi verdi. Cumhuriyeti kuran CHP' nin, Türkiye ve dünya siyasetinde önemli bir yeri bulunduğunu dile getiren Bilgehan, CHP' nin, tek parti ve tek adam rejimini değiştiren devrimci bir parti olduğunu ifade etti. Bilgehan, daha sonra Mısırlı gençlerin sorularını yanıtladı. Mısırlı bir gencin, "CHP, Filistin halkına karşı neden İsrail' i destekliyor" sorusuna karşılık Bilgehan, "CHP' nin İsrail' i desteklediğini herhalde AKP ziyaretinde duydunuz. Çünkü biz CHP olarak Ortadoğu' da her zaman dengeli bir politika izlemek gerektiğine inandık. Filistin halkına uygulanan zulmü insanlık suçu olarak görüyoruz ama Ortadoğu' ya barışın gelmesi için bütün tarafların, İsrail dahil, uzlaşması gerektiğine inanıyoruz. Mısır' da da böyle düşünen pek çok aydın, siyasetçi olduğunu biliyoruz" yanıtını verdi. Bilgehan, seçim sonucu ne olursa olsun Türkiye' nin Ortadoğu için model olmaya devam etmesine gayret edeceklerini belirtti.

Summary for Technology Category Article (1)

İnternette “filtre” iddiaları karşısında temel hak ve özgürlüklerin ihlal edileceğini savunan “Anonymous” (Anonim) adlı bir grup, Türkiye' deki çeşitli kamu kuruluşları ile bazı medya sitelerine yönelik siber saldırı düzenleyeceği tehdidinde bulundu. İlk büyük eylemlerini aylar önce Wikileaks' e yönelik ambargo uygulayan Paypal, Visa ve Mastercard gibi online ödeme ve kredi kartı firmalarına karşı gerçekleştiren grup üyeleri, dün kendilerine ait internet sitesinde Türkiye' yi hedef alan bir mesaj yayınladı. Türkiye' de internet kullanıcılarının “filtre” uygulaması ile sansüre uğrayacağını savunulan mesajda “operationturkey” (Türkiye Operasyonu) adıyla “siber savaş” ilan edilirken, öncelikle “sansür” uygulayan kurumlara karşı harekete geçileceği açıklandı. İlk organize saldırı perşembe saat 18.00' de olacak. Türkiye' deki kimi kamu kurumlarını hedef alan ilk organize saldırının 9 Haziran Perşembe günü saat 18.00' de (TSİ) gerçekleştirileceğini duyuran Anonymous, eyleme katılmak isteyenlerin bilgisayarlarına bazı yazılımlar indirerek, bu programları aktif hale getirmesini istedi. Tüm bu gelişmeleri yakından izleyen siber güvenlik

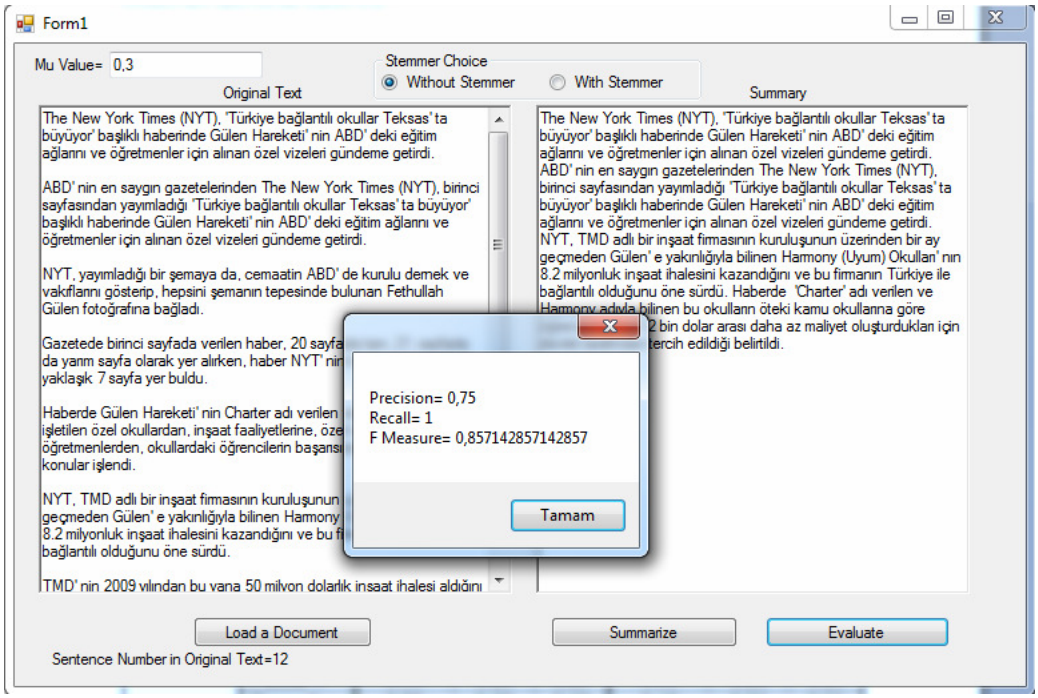
uzmanları, söz konusu tehdidin önemine dikkat çekerek, birçok sitenin “DDoS” olarak adlandırılan saldırı ile karşı karşıya kalabileceğini bildirdi. Bu tip saldırılarda amacın, “bilgi çalmak” yerine sistemin erişilemez hale getirilmesi olduğunu kaydeden siber güvenlik uzmanları, burada verilmek istenen mesajın “Bizim internetimize karışırsanız biz de sizin internetinizi kapatırız” anlamı taşıdığını öne sürdü. Anonymous grubunun saldırılarda özel bir yöntem kullandığını belirten uzmanlar, “Klasik DDoS saldırılarında 'zombi' haline getirilmiş bilgisayarlar kullanılırken, bu grup tamamen 'gönüllü zombi' bilgisayarlarla saldırıyor. DDoS saldırılarından bir sonraki adımın çeşitli devlet kurumlarına ait telefon ve haberleşme sistemlerini çalışamaz hale getirmek olduğunu kaydeden siber güvenlik uzmanları, daha sonra hedefteki kurumlarda çalışan kişilerin bilgisayarlarına “phishing” saldırıları gerçekleştirileceğinin altını çizdi. “Phishing” yöntemi ile kullanıcılara e-mail göndererek, bilgisayarlarından erişim sağladıkları banka, kredi kartı bilgi ve şifrelerinin yanı sıra kurumlarına ait kimi özel bilgi ve belgelerin sızdırılmaya çalışılacağını anlatan uzmanlar, böylesi bir durumun yeni bir “Wikileaks” olayına kapı açacağına vurgu yaptı. Eğer gerekli önlemler alınmazsa ciddi bir siber savaşın başlamış olacağını bildiren uzmanlar, kimi hassas hedeflerin saldırıya uğraması durumunda Türkiye' deki internet ağının bir süre için çökebileceği uyarısında bulundu. Anonymous' un bugüne kadar gerçekleştirdiği hacker saldırılarını yakından takip eden Siber Güvenlik Uzmanı Huzeyfe Önal, yaptığı açıklamada, hedefteki kurumların kendi bünyelerinde alacağı çeşitli önlemlerin yanı sıra internet servis sağlayıcılarının da ciddi tedbirler alması gerektiğini bildirdi. Burada dikkat edilmesi gereken en önemli husus protesto ile saldırının farklı şeyler olduğu. Protesto amaçlı da olsa böylesi bir saldırıya katılanlara, Türk Ceza Kanunu' na (TCK) göre suç işlediklerini hatırlatmak isterim. İnternet sitelerine bu şekilde zarar verenlerin, 5 yıla kadar hapis cezası ile yargılanacaklarını unutmamalı”.

Summary for Technology Category Article (2)

Korsan kriziyle başı ağrıyan Japon elektronik üreticisi Sony, PlayStation şebekesinin bu hafta kısmen açılacağını açıkladı. Şirket, online oyun kullanıcılarını kredi kartı ve diğer kişisel bilgilerinin çalınmış olabileceği konusunda uyarmış,

ardından da PlayStation şebekesini kapatmıştı. Milliyet' in haberine göre Sony' nin ikinci adamı Kazuo Hirai "Datalarını tehlikeye attığımız, endişelendirdiğimiz ve rahatsızlık verdiğimiz tüm kullanıcılarımızdan özür dileriz" diyerek geleneksel Japon selamı ile eğilerek kullanıcılardan özür diledi. "Sony rakiplere kaptırmaktan korktuğu kullanıcıları için iyi niyet ve özür göstergesi olarak ücretsiz hediyeler vereceğini açıkladı. Sony ayrıca hackerların kullanıcıların kart bilgilerini çaldığını tam olarak kabul etmese de, yine de bundan dolayı finansal bir zarar gören olursa bu zararı tazmin edeceğini de belirtti. Kişisel bilgilerin daha iyi korunmasını sağlamak için sistemlerini yeniden kurma konusunda adımlar attıklarını belirten şirket, servislerin en kısa sürede tamamen hizmete gireceğini kaydetti. 59 ülkede 77 milyon kişi PlayStation şebekesini kullanıyor.

Appendix 16.View and Codes of Automatic Text Summarization Software that coded on C#




```
using System;
using System.Collections.Generic;
using System.ComponentModel;
using System.Data;
using System.Drawing;
using System.Linq;
using System.Text;
using System.IO;
using System.Windows.Forms;
using net.zemberek.yapi;
using net.zemberek.tr.yapi;
using net.zemberek.erisim;
using Microsoft.Office.Interop.Word;

namespace WindowsFormsApplication1
{

    public partial class Form1 : Form
    {
        public Form1()
        {
            InitializeComponent();
        }

        private void button1_Click(object sender, EventArgs e)
        {
```

```
            openFileDialog1.Filter    =    "Microsoft    Word    Dosyası
(*.doc)|*.doc|Text Dosyaları (*.txt)|*.txt|Tüm Dosyalar (*.*)|*.*";
            if (openFileDialog1.ShowDialog() == DialogResult.OK)
            {
                string                          ext                          =
Path.GetExtension(openFileDialog1.FileName.ToString());
                if (ext == ".doc")
                {
                    Microsoft.Office.Interop.Word._Application
Worduyg=new Microsoft.Office.Interop.Word.Application();
                    object                    dosya                    =
openFileDialog1.FileName.ToString();
                    object                   bosobje                   =
System.Reflection.Missing.Value;
                    dosya = openFileDialog1.FileName.ToString();
                    _Document    doc    =    Worduyg.Documents.Open(ref
dosya, ref bosobje, ref
bosobje, ref bosobje, ref bosobje, ref bosobje, ref bosobje, ref
bosobje,  ref  bosobje,  ref  bosobje,  ref  bosobje,  ref  bosobje,  ref
bosobje,
                    ref bosobje, ref bosobje, ref bosobje);
                    doc.ActiveWindow.Selection.WholeStory();
                    doc.ActiveWindow.Selection.Copy();
                    IDataObject veri = Clipboard.GetDataObject();
                    richTextBox1.Text                                    =
veri.GetData(DataFormats.Text).ToString();
                    doc.Close(ref    bosobje,    ref    bosobje,    ref
bosobje);
                }
                else if (ext == ".txt")
                {
                    StreamReader          oku          =          new
StreamReader(openFileDialog1.FileName.ToString());
                    richTextBox1.Text = oku.ReadToEnd();
                    oku.Close();
                }
                else
                    MessageBox.Show("Select an Appropriate File!");
            }
        }
        public    static    System.Array    ResizeArray(System.Array
oldArray, int newSize)
        {
            int oldSize = oldArray.Length;
            System.Type                elementType                  =
oldArray.GetType().GetElementType();
            System.Array                   newArray                   =
System.Array.CreateInstance(elementType, newSize);
            int preserveLength = System.Math.Min(oldSize, newSize);
            if (preserveLength > 0)
                System.Array.Copy(oldArray,                newArray,
preserveLength);
            return newArray;
        }
        char[] sc = { '.', '?', '!'};
        char[] st = {' ', '+', '%', '$', '&', '/', '-', '_', '\\',
'\t', '\n', '\r', '.', ',', ':', ';', '\'', '!', '#', '"', '^', '(',
')', '=', '*', '}', ']', '[', '{', '<', '>'};
```

```
        char[] ste = { '+', '%', '$', '&', '/', '-', '_', '\\','.',
',', ':', ';', '\'', '!', '#', '"', '^', '(', ')', '=', '*', '}',
']', '[', '{', '<', '>'};
       private void button2_Click(object sender, EventArgs e)
       {
           if (richTextBox1.Text != "")
           {
               label4.Visible = false;
               DataTable dcumle = new DataTable();
               dcumle.Columns.Add("Index");
               dcumle.Columns.Add("Sentence");
               dcumle.Columns.Add("Terms");
               dcumle.Columns.Add("Term Frequencies");
               dcumle.Columns.Add("Total Frequency");
               DataView vcumle = dcumle.DefaultView;
               string[]                                        cumleler=
SplitToSentence(richTextBox1.Text);

                   int index = 0;
                   foreach (string cumle in cumleler)
                   {
                       if        (cumle        !=        ""        ||
!String.IsNullOrEmpty(cumle))
                       {
                           string cumle1 = cumle;
                           cumle1 = cumle1.Trim().Trim(st);
                           if             (cumle1!=""             &&
Char.IsLetterOrDigit(cumle1[0]))
                           {
                               index++;

                               string[] terimler = cumle1.Split(st);
                               string cumleterimler = "";
                               string terimfrekans = "";
                               int[] tf = new int[terimler.Length];
                               int r = 0;
                               int toplamf = 0;
                               bool[]        tk        =        new
bool[terimler.Length];
                               string[]  stopwords  =  {"bu",  "te",
"ta",  "de",  "da",  "deki",  "daki",  "teki",  "taki",  "ve",  "ile",
"iken",  "ken",  "ki",  "ama",  "fakat",  "lakin",  "ın",  "nin",  "nın",
"in",                          "için",                          "leme",
"lı","li","un","um","u","sun","ya","nda","nde","ndan","nden","ancak"
, " veya", "gibi", "şey", "ise","na" };
                               for (int i = 0; i < terimler.Length;
i++)
                               {
                                   int te=99;
                                   for    (int    j    =    0;    j    <
stopwords.Length; j++)

                                       if       (terimler[i]       ==
stopwords[j])
                                       {
                                           te = j;
                                           tk[i] = true;
                                           break;
```

```
                                    }
                                if (te != 99)
                                    continue;
                                if   (terimler[i]   !=   ""   &&
terimler[i].Length > 2)
                                {
                                    tk[i] = false;
                                    tf[i] = 0;
                                    terimler[i]                 =
terimler[i].Trim().Trim(st).ToLower();
                                    if  (radioButton1.Checked  ==
true)
                                    {
                                        try
                                        {
                                            terimler[i]         =
Kokk(terimler[i]);
                                        }
                                        catch { }
                                    }
                                }
                            }
                            for (int i = 0; i < terimler.Length;
i++)
                            {
                                if
(!String.IsNullOrEmpty(terimler[i]) && tk[i] == false)
                                {
                                    for   (int   j   =   i;   j   <
terimler.Length; j++)
                                    {
                                        if     (terimler[i]     ==
terimler[j])
                                        {
                                            tf[r]++;
                                            tk[j] = true;
                                        }
                                    }
                                    cumleterimler  +=  terimler[i]
+ " ";
                                    terimfrekans += tf[r] + " ";
                                    r++;
                                }
                                else continue;
                            }
                            for (int i = 0; i < r; i++)
                                toplamf += tf[i];
                            dcumle.Rows.Add(index,         cumle1,
cumleterimler, terimfrekans, toplamf);
                        }
                    }
                }

                dataGridView1.Dataource = dcumle;

            DataTable dterm = new DataTable();
            dterm.Columns.Add("Term");
            DataView vterm = dterm.DefaultView;
            richTextBox1.Text = richTextBox1.Text.Trim();
```

```
                char[] sp = { '.', '?', '!', ':', '\n', '\r', '\t', '
', '+', '%', '$', '&', '/', '-', '_', '\\' };
                for (int i = 0; i < dataGridView1.RowCount - 1; i++)
                {
                    string[]                cumpar                =
dcumle.Rows[i]["Terms"].ToString().Trim().Split(' ');
                    foreach (string ter in cumpar)
                    {
                        if (ter!="")
                            dterm.Rows.Add(ter);
                    }
                }

                for (int i = 0; i <= dterm.Rows.Count - 2; i++)
                {
                    for (int j = i + 1; j <= dterm.Rows.Count - 1;
j++)
                    {
                        if    (dterm.Rows[i]["Term"].ToString()    ==
dterm.Rows[j]["Term"].ToString())
                        {
                            dterm.Rows.RemoveAt(j);
                            j--;
                        }
                    }
                }
                vterm.Sort = "Term";
                DataTable dterim = new DataTable();
                DataView vterim = dterim.DefaultView;
                dterim.Columns.Add("Index");
                dterim.Columns.Add("Term");
                dterim.Columns.Add("Found in Sentences");
                dterim.Columns.Add("Term Frequencies");
                dterim.Columns.Add("Total Frequency");
                for (int i = 0; i < dterm.Rows.Count; i++)
                {
                    string bulcum = "";
                    string terfre = "";
                    int topfre = 0;
                    for (int j = 0; j < dcumle.Rows.Count; j++)
                    {
                        string[]               kcumle               =
dcumle.Rows[j]["Terms"].ToString().Trim().Split(sp);

                        for (int k = 0; k < kcumle.Length; k++)
                        {
                            kcumle[k] = kcumle[k].Trim();
                            if             (kcumle[k]             ==
dterm.Rows[i]["Term"].ToString())
                            {
                                string[] freay = dcumle.Rows[j]["Term
Frequencies"].ToString().Trim().Split(' ');
                                try
                                {
                                    bulcum                         +=
dcumle.Rows[j]["Index"].ToString() + " ";
                                    terfre += freay[k] + " ";
                                    topfre                         +=
Convert.ToInt32(freay[k]);
```

```
                                    break;
                                }
                                catch { }
                            }
                        }
                    }
                    dterim.Rows.Add((i        +        1).ToString(),
dterm.Rows[i]["Term"].ToString(),           bulcum,           terfre,
topfre.ToString());

                }
                dataGridView2.Dataource = dterim;
                double[] diag = new double[dataGridView1.Rows.Count -
1];
                double[,] c = new double[dataGridView1.RowCount - 1,
dataGridView1.RowCount - 1];
                for (int i=0;i<dataGridView1.Rows.Count-1;i++)
                for (int j=0;j<dataGridView1.Rows.Count-1;j++)
                    c[i,j]=0.0;
                richTextBox2.Text = "";
               dterim.DefaultView.Sort = "Term";
                double[,]    a=new    double[dataGridView1.Rows.Count-
1,dataGridView2.Rows.Count-1];
                double[,]    b=new    double[dataGridView2.Rows.Count-
1,dataGridView1.Rows.Count-1];
                for (int i=0;i<dataGridView1.Rows.Count-1;i++)
                    for (int j=0;j<dataGridView2.Rows.Count-1;j++)
                    {
                        a[i,j]=0.0;
                        b[j,i]=0.0;
                    }
                for  (int  i  =  0;  i  <  dataGridView1.Rows.Count  -  1;
i++)
                {

                    string[]                 cumpar                 =
dcumle.Rows[i]["Terms"].ToString().Trim().Split(' ');
                    string[]    terfred    =    dcumle.Rows[i]["Term
Frequencies"].ToString().Trim().Split(' ');
                    for (int k=0;k<cumpar.Length;k++)
                    {
                        int
t=Convert.ToInt32(dterim.DefaultView.Find(cumpar[k]).ToString().Trim
());

a[i,t]=Convert.ToDouble(terfred[k].Trim())/Convert.ToDouble(dcumle.R
ows[i]["Total Frequency"].ToString().Trim());
                    }
                }
                dterim.DefaultView.Sort = "Term";
                dataGridView2.Dataource = dterim;
                for  (int  j  =  0;  j  <  dataGridView2.Rows.Count  -  1;
j++)
                {
                    string[]                 tercum                 =
dataGridView2.Rows[j].Cells["Found                                in
Sentences"].Value.ToString().Trim().Split(' ');
```

```
                    string[]                    terfre                    =
dataGridView2.Rows[j].Cells["Term
Frequencies"].Value.ToString().Trim().Split(' ');

                    for (int z = 0; z < terfre.Length; z++)
                    {
                        b[j,  Convert.ToInt32(tercum[z].Trim())-1]  =
Convert.ToDouble(terfre[z].Trim())                                  /
Convert.ToDouble(dataGridView2.Rows[j].Cells["Total
Frequency"].Value.ToString().Trim());

                    }
                }
                for (int i=0;i<dataGridView1.Rows.Count-1;i++)
                    for(int j=0;j<dataGridView1.Rows.Count-1;j++)
                        for(int z=0;z<dataGridView2.Rows.Count-1;z++)
                            if (a[i,z]!=0.0)
                                c[i,j]+=a[i,z]*b[z,j];
                for (int i=0;i<dataGridView1.Rows.Count-1;i++)
                    diag[i]=c[i,i];
                Array.Sort(diag);
                bool[] oc = new bool[dataGridView1.Rows.Count - 1];
                int[] minind = new int[dataGridView1.Rows.Count - 1];
                double mu = Convert.ToDouble(textBox1.Text);
                for (int i = 0; i < dataGridView1.Rows.Count - 1;
i++)
                {
                    for (int j = 0; j < dataGridView1.Rows.Count - 1;
j++)
                    {
                        if (diag[i] == c[j, j])
                        {
                            minind[i] = j;
                            oc[j] = true;
                            break;
                        }
                    }
                }
                richTextBox2.Text = "";
                int                         s                       =
Convert.ToInt32(Math.Ceiling((double)(dataGridView1.Rows.Count  -  1)
* mu));
                MessageBox.Show(s.ToString());
                for (int i = 0; i < s; i++)
                {
                    oc[minind[i]] = false;
                }
                for (int i=0;i<dataGridView1.Rows.Count-1;i++)
                {
                    if (oc[i] == false)
                    {
                        richTextBox2.Text                           +=
dcumle.Rows[i]["Sentence"].ToString() + ". ";
                    }
                }
                label4.Text = "Sentence Number in Original Text=" +
(dataGridView1.Rows.Count - 1).ToString();
                label4.Visible = true;
            }
```

```
            else
                MessageBox.Show("Load a Text into Left Area!");
        }
        static string[] SplitToSentence(string document)
        {
            document = document.Trim();
            char[] sc = { '.', '?', '!' };
            char[] st = { ' ', '+', '%', '$', '&', '/', '-', '_',
'\\', '\t', '\n', '\r', '.', ',', ':', ';', '\'', '!', '#', '"',
'^', '(', ')', '=', '*', '}', ']', '[', '{', '<', '>' };
            char[] ste = { '+', '%', '$', '&', '/', '-', '_', '\\',
'.', ',', ':', ';', '\'', '!', '#', '"', '^', '(', ')', '=', '*',
'}', ']', '[', '{', '<', '>' };
                string[] cumleler = document.Split(sc);
                    for (int i = 1; i < cumleler.Length; i++)
                    {
                        if (cumleler[i] == "")
                            continue;
                        if (!Char.IsWhiteSpace(cumleler[i][0]))
                        {
                            cumleler[i - 1] = cumleler[i - 1] + "." +
cumleler[i];
                            for (int j = i; j < cumleler.Length - 1;
j++)
                            {
                                cumleler[j] = cumleler[j + 1];
                            }
                        }
                        if         (cumleler         [i]!=""         &&
Char.IsNumber(cumleler[i][0]))
                        {
                            cumleler[i - 1] = cumleler[i - 1] + "." +
cumleler[i];
                            for (int j = i; j < cumleler.Length - 1;
j++)
                            {
                                cumleler[j] = cumleler[j + 1];
                            }
                        }
                cumleler[i] = cumleler[i].Trim(st);
                if (cumleler[i] != "" && Char.IsLetter(cumleler[i][0])
&& Char.IsLower(cumleler[i][0]))
                        {
                            cumleler[i - 1] = cumleler[i - 1] + ". "
+ cumleler[i];
                            for (int j = i; j < cumleler.Length - 1;
j++)
                            {
                                cumleler[j] = cumleler[j + 1];
                            }
                        }

                    }
                string sens="";
                for(int i=0;i<cumleler.Length;i++)
                {
                    cumleler[i] = cumleler[i].Trim();
                    if (i != cumleler.Length - 1)
                    {
```

```
                    if (!String.IsNullOrEmpty(cumleler[i]))
                    {
                        sens = sens + cumleler[i] + "|";
                    }
                    else continue;
                }
                else    if    (i    ==    cumleler.Length    -    1    &&
!String.IsNullOrEmpty(cumleler[i]))
                {
                    sens = sens + cumleler[i];
                }
            }
            return (sens.TrimEnd('|').Split('|'));
        }
        static string Kokk(string stri)
        {
            Zemberek sozluk = new Zemberek(new TurkiyeTurkcesi());
            return
sozluk.kelimeCozumle(stri)[0].kok().ToString().Split(' ')[0];
        }
        private void button3_Click(object sender, EventArgs e)
        {
            openFileDialog1.Filter    =    "Microsoft    Word    File
(*.doc)|*.doc|Text File (*.txt)|*.txt|All Files (*.*)|*.*";
            string summary = "";
            if (openFileDialog2.ShowDialog() == DialogResult.OK)
            {
                string                      ext                      =
Path.GetExtension(openFileDialog2.FileName.ToString());
                if (ext == ".doc")
                {
                    Microsoft.Office.Interop.Word._Application
Worduyg2=new            Microsoft.Office.Interop.Word.Application();
object dosya2 = openFileDialog2.FileName.ToString();
                    object                bosobje2                =
System.Reflection.Missing.Value;
                    dosya2 = openFileDialog2.FileName.ToString();
                    _Document   doc2   =   Worduyg2.Documents.Open(ref
dosya2, ref bosobje2, ref
bosobje2,  ref  bosobje2,  ref  bosobje2,  ref  bosobje2,  ref
ref
bosobje2,  ref  bosobje2,  ref  bosobje2,  ref  bosobje2,  ref  bosobje2,
ref bosobje2,
                    ref bosobje2, ref bosobje2, ref bosobje2);
                    doc2.ActiveWindow.Selection.WholeStory();
                    doc2.ActiveWindow.Selection.Copy();
                    IDataObject veri2 = Clipboard.GetDataObject();
                    summary                                          =
veri2.GetData(DataFormats.Text).ToString();
                    doc2.Close(ref   bosobje2,   ref   bosobje2,   ref
bosobje2);
                }
                else if (ext == ".txt")
                {
                    StreamReader          oku          =          new
StreamReader(openFileDialog2.FileName.ToString());
                    summary = oku.ReadToEnd();
                    oku.Close();
                }
```

```
                else
                    MessageBox.Show("Select an appropriate file!");
                string[] docsummsent = SplitToSentence(summary);
                string[]                softsummsent                =
SplitToSentence(richTextBox2.Text);
                int TP=0;
                double P, R, F;
                for (int i = 0; i < docsummsent.Length; i++)
                {
                    for (int j = 0; j < softsummsent.Length; j++)
                    {
                        if (docsummsent[i] == softsummsent[j])
                        {
                            TP++;
                            break;
                        }
                    }
                }
                P          =          (Convert.ToDouble(TP)          /
Convert.ToDouble(softsummsent.Length));
                R          =          (Convert.ToDouble(TP)          /
Convert.ToDouble(docsummsent.Length));
                F = ((2 * P * R) / (P + R));
                MessageBox.Show("Precision=   "   +   P.ToString()   +
"\r\n"  +  "Recall=  "  +  R.ToString()  +  "\r\n"  +  "F  Measure=  "  +
F.ToString());
            }
        }
    }
}
```