ANKARA YILDIRIM BEYAZIT UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

ANALYTICS OF TURKISH OFFICIAL GAZETTE ARCHIVES USING BIG DATA MINING TECHNIQUES

M.Sc. Thesis by

Yasemin CAN

Department of Computer Engineering

December, 2018

ANKARA

ANALYTICS OF TURKISH OFFICIAL GAZETTE ARCHIVES USING BIG DATA MINING TECHNIQUES

A Thesis Submitted to

The Graduate School of Natural and Applied Sciences of

Ankara Yıldırım Beyazıt University

In Partial Fulfillment of the Requirements for the Degree of Master of Science

in Computer Engineering, Department of Computer Engineering

by

Yasemin CAN

December, 2018

ANKARA

M.Sc. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"ANALYTICS OF TURKISH OFFICIAL GAZETTE ARCHIVES USING BIG DATA MINING TECHNIQUES"** completed by **YASEMIN CAN** under supervision of **PROF. DR. FATİH V. ÇELEBİ** 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. Fatih V. ÇELEBİ

Supervisor

Asst. Prof. Dr. Bülent TUĞRUL

Jury Member

Asst. Prof. Dr. Özkan KILIÇ

Jury Member

Prof. Dr. Ergün ERASLAN

Director

Graduate School of Natural and Applied Sciences

I hereby declare that, in this thesis which has been prepared in accordance with the Thesis Writing Manual of Graduate School of Natural and Applied Sciences,

- All data, information and documents are obtained in the framework of academic and ethical rules,

- All information, documents and assessments are presented in accordance with scientific ethics and morals,

- All the materials that have been utilized are fully cited and referenced,

- No change has been made on the utilized materials,

- All the works presented are original,

and in any contrary case of above statements, I accept to renounce all my legal rights.

Date: **Signature:…...……**

Name & Surname:……..

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Prof. Fatih V. ÇELEBİ for his valuable support, expert leading and guidance, high patience, and understanding against me.

Additionally, I would like to thank my family for their continuous support through my education life. Especially, I am grateful to my father for being a motivation source to learn for me.

2018, December **Yasemin CAN**

ANALYTICS OF TURKISH OFFICIAL GAZETTE ARCHIVES USING BIG DATA MINING TECHNIQUES

ABSTRACT

Among being indispensable parts of today's world, rapidly developing information technologies are rising at first and also triggering the production of rapidly growing, unstructured and complicated data. Storing, processing and extracting meaningful information from this huge, complex and unstructured data with dizzying pace cannot be accomplished efficiently and effectively by traditional database management systems and data mining methods. In order to eliminate these deficiencies and to add value to the sector and governments for making meaningful deductions and decisions by processing unstructured data large volume, the big data concepts and technologies has emerged. Today big data is considered to be a serious and valuable field for many governments and private companies. Big data is defined with previously, 3V but recently 5Vof meaning volume, variety, velocity, value and veracity comes from different sources including social networks, sensor records, mobile devices, GPS data etc.

By extracting meaningful information from this unstructured data coming from various sources, governments and companies are developing new policies along with increasing their profits and product and service qualities.

In this thesis study; firstly, the concept of big data and related concepts are explained, then recent approaches and application areas in big data field along with the relationship between knowledge management are discussed. After that, according to the results of literature review we concluded that there is no study on text analysis of Turkish Official Gazette and TRT Haber web news archives. Therefore, these archives selected as a big data source because of being an unstructured text data collection. 714 files consisting of about 61.551.000 words and belonging to July 2015 - July 2017 period from Official Gazette and 542 files from TRT Haber web site consisting of about 3.794.000 words and belonging to the same time interval were selected as corpora and pre-processed, cleaned and analyzed by Apache Spark platform using Python language in terms of generation of term frequency matrices of the two datasets.

According to these, word clouds were generated. After that, comparisons between these two datasets divided by year and by month were done using Tf-IDF and cosine similarity algorithms in order to get some insight about how much similar they are. As a result of these comparisons, it is concluded that the cosine similarity between Official Gazette and TRT Haber web news archives divided by year differs between 0,008 and 0,03. In addition, according to the results of cosine similarity comparison on monthly divided dataset, a correlation between two datasets about how they affect each other is looked for. And it is concluded that the ratio of the effect of agenda of the country covered in TRT News website on Official Gazette is higher than the ratio of the effect of Official Gazette on the TRT News archive.

Finally, a Legislative Referral Engine application, which can be developed by using big data text analysis methods which are considered to be worth working on and which can facilitate the work of lawyers, is discussed as a future work.

Keywords: Big data, text analytics, word frequency, Apache Spark, Turkish Official Gazette, TRT News, correlation, knowledge management, term frequency matrix, cosine similarity

BÜYÜK VERİ TEKNİKLERİ KULLANILARAK RESMİ GAZETE ARŞİVİNİN ANALİZ EDİLMESİ

ÖZ

Günümüz dünyasının vazgeçilmezleri arasında ilk sıralara yükselen ve hızla gelişen teknoloji beraberinde hızla büyüyen ve karmaşıklaşan bir veri üretimini de tetiklemekte. Bu, baş döndürücü bir hızla büyüyen karmaşık yapılı verinin saklanması, işlenebilmesi ve içinde sakladığı anlamlı bilginin çıkarılması geleneksel veri tabanı yönetim sistemleri ve veri madenciliği yöntemleriyle verimli ve etkin bir şekilde gerçekleştirilememektedir. Bu eksikliğin giderilmesinin yanı sıra bu büyük hacimli ve düzensiz yapılı verinin işlenerek anlamlı çıkarımlarla ilgili olduğu sektöre değer katması amacıyla ortaya çıkan büyük veri kavramı ve teknolojileri birçok ülke yönetimi ve özel şirketlerce ciddi ve değer katan bir çalışma alanı olarak kabul edilmektedir. Büyük verinin 3V olarak başlayan ve günümüzde 5V olarak kabul gören hacim, hız, çeşitlilik, değer ve gerçeklik özellikleri kapsamına giren veriler sosyal ağlar, sensor kayıtları, mobil araçların ürettikleri, GPS verileri gibi çok çeşitli kaynaklardan gelmektedir. Belli bir formatta olmayan bu çok çeşitli verilerden anlamlı bilgilerin çıkarılması ile devletler yeni politikalar geliştirmekte, şirketler karlarını ve ürün kalitelerini artırmaktadırlar.

Bu tez çalışmasında; öncelikle büyük veri kavramı ve ilişkili kavramlar açıklanmıştır, ayrıca büyük veri alanında güncel yaklaşımlar ve uygulama alanları konusu ile bilgi yönetimi ile olan ilişkilerine değinilmiştir. Sonrasında ise literatür taraması sonucu daha önce üzerinde bir metin analizi çalışması yapılmadığı sonucuna varılan T.C. Resmi Gazete ve TRT Haber sitesinde yayımlanan haber arşivleri düzensiz metin verisi yapısında olması nedeniyle büyük veri kaynağı olarak ele alınmış, çeşitli ön işleme ve veri temizleme sürecinden geçirildikten sonra Temmuz 2015 – Temmuz 2017 dönemine ait Resmi Gazete arşivinden yaklaşık 61.551.000 kelime içiren 714 dosya, TRT Haber web arşivinden yaklaşık 3.794.000 kelime içeren 542 dosya alınarak kütüphaneler oluşturulmuş, Apache Spark platformu ve Python dili ile kelime sıklığı işlemi ile analiz edilmiş olup, söz konusu döneme ait Resmi Gazete ve TRT Haber arşivinde geçen kelimelerin azalan sırada terim sıklığı matrisi çıkarılmış ve

bunlardan kelime bulutları oluşturulmuştur. Daha sonra bu iki veri setinin benzerlik ilişkisini tespit etmek amacıyla Tf-IDF ve kosinüs benzerliği algoritmaları kullanılarak yıllık ve aylık olarak bölünmüş alt veri setleri üzerinde karşılaştırma çalışmaları yapılmıştır. Yapılan bu çalışma sonucunda yıllık olarak gruplanmış Resmi Gazete ve TRT Haber web arşivinden alınan aynı zaman aralığına ait içeriklerin kosinüs benzerliği 0,008 ile 0,03 aralığında bulunmuştur. Ayrıca, aylık olarak bölünmüş veri setleri üzerinde Resmi Gazete ve TRT Haber arşivlerinin birbirlerini nasıl etkilediklerini tesbit etmek amacıyla bir kosinüs benzerliği karşılaştırması yapılmış ve sonucunda ülke gündeminin yansıdığı TRT Haber arşivlerinin Resmi Gazete içerikleri üzerindeki etkisi, Resmi Gazete içeriğinin TRT Haber arşivi ve dolayısıyla ülke gündemi üzerine etkisinden daha yüksek olduğu sonucuna varılmıştır.

Son olarak ise gelecekte üzerinde çalışmaya değer olarak görülen ve hukuk alanında çalışanların işini kolaylaştırabilecek büyük veri metin analizi yöntemleri kullanılarak geliştirilebilecek bir Mevzuat Tavsiye Motoru uygulaması ele alınmıştır.

Anahtar Kelimeler: Büyük veri, metin analizi, kelime sıklığı, Apache Spark, Resmi Gazete, TRT Haber, korelasyon, bilgi yönetimi, terim sıklığı matrisi, kelime bulutu, kosinüs benzerliği

CONTENTS **Page**

ABBREVIATIONS

API	Application Programming Interface
ACID	Atomicity, Consistency, Isolation, Durability
BD	Big Data
BI	Business Intelligence
CAP	Consistency, Availability, Partition Tolerance Theorem
CDR	Call Detail Record
CPU	Central Processing Unit
CRM	Customer Relationship Management
CRUD	Create Read Update Delete
DAS	Direct Access Storage
DBMS	Database Management Systems
DSS	Decision Support Systems
ERP	Enterprise Resource Planning
GB	Giga Bytes
GPS	Global Positioning System
HDFS	Hadoop Distributed File System
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IDC	International Data Corporation
IDE	Integrated Development Environment
IoTs	Internet of Things
JBOD	Just a Branch of Discs
JDK	Java Development Kit
JSON	JavaScript Object Notation
KBV	Knowledge Based View
KM	Knowledge Management
NAS	Network Attached Storage
NLP	Natural Language Processing
NoSQL	Not Only Structured Query Language
OCR	Optical Character Recognition
OLTP	Online Transaction Processing

OS	Operating System
PDF	Portable Document Format
RAID	Redundant Array of Independent Discs
RDBMS	Relational Database management Systems
RDD	Resilient Distributed Dataset
REST	Representational State Transfer
SAN	Storage Attached Network
SQL	Structured Query Language
SSD	Solid-State Drive
TB	Terra Bytes
TDM	Text and Data Mining
XML	Extensible Markup language
ZB	Zeta Bytes

LIST of FIGURES

LIST of TABLES

CHAPTER 1

INTRODUCTION

1.1 Data and Types

In Computer science terminology the term data is used for defining distinct and raw pieces of fact especially formatted in a special form, [1] and needs to be processed and organized. However, information refers to processed, well organized and structured data that is meaningful and useful.

This special form will be categorized in 3 main groups as structured, semi-structured and unstructured. Structured data refers to any data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets [2] Relational Database management systems (RDBMS) are being used to handle structured data.

As Beal states that “Unstructured data is all those things that can't be so readily classified and fit into a neat box: photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents.

Semi-structured data is a cross between the two. It is a type of structured data, but lacks the strict data model structure. With semi-structured data, tags or other types of markers are used to identify certain elements within the data, but the data doesn't have a rigid structure. For example, word processing software now can include metadata showing the author's name and the date created, with the bulk of the document just being unstructured text. Emails have the sender, recipient, date, time and other fixed fields added to the unstructured data of the email message content and any attachments.” [2].

Historically, data was generated only by worker by hand. Later on users started to contribute data generation by using social media etc. And then machine generated data has become a crucial part of the unstructured data like sensor data, security camera records or log data. It is known that 80% of all data is in unstructured form. This change

in the data sources also changes the requirements about storage and analysis. Since, conventional RDBMS do not meet the needs, multiple CPU usage, parallel processing, distributed systems and clouds are some results of the effort for handling growing amount of data. Although these improvements, excessive growth of data requires new data storage systems and analysis methods.

1.2 Definition of Big Data

The term Big Data was used to describe huge amounts of data sets up to 2000s. Namely, volume of the data was the prominent measurement factor. During 2000s experts stated to define big data with 2 more characteristics as velocity and variety. Briefly, velocity defines the formation speed of the data and variety is for range of data types included. Recent researchers add new features to these 3Vs, like veracity and value to define a data set whether it is big data or not. In 2013 Ohlhorst defines big data which is huge in volume and could not be processed using traditional data analytics and processing methods [3].

Palmer states that data resembles raw petrol, it is valuable but if we do not refine it we cannot take advantage of it [4] Big Data is like huge amount of raw petrol. Refining big data means taking out meaningful information from it. Refining process of big data results in huge amount of valuable treasury. This treasury appeals more attention of the gurus from economy to technology including politics, health sector and so on. Because analyzing and mining big data results in exploration of new job areas, more accurate and quick results for surveys, decrease in criminal incidents and traffic accidents, prevention from epidemic illnesses etc.

About ten years ago leader data analyses companies like Oracle, IBM, Teradata and HP was offering “data warehouses” to the companies those have terabytes of data because nearly all data was stored in relational database management systems. However, according to IDC statistics today data comes from 44 different sources from sensors to planes, super computers to automobiles. This results in wide range of structured and unstructured data types including text, voice, image and digital [5]. This means if we use only structured data to get meaningful inference to lead our new steps, we will miss eighty percent of the treasury.

MasterCard company made an analysis study on 65 B shopping transactions by 1.5 B card users among 210 countries in order to establish business and consumer trends. And came up with the result that says; customers who dashed out into gas station, generally spent $35 – 50$ at restaurants or supermarkets in following one hour. Market owners use this information and send shopping stamps to their customers and make profit by increasing their sales. MasterCard shares this big data analytic results with other companies by money [64].

Additionally, if we make a comparison between traditional data and big data, we will conclude with the following table;

Table 1.1: Comparison between traditional data and big data

Metric	**Traditional Data**	**Big Data**
SCHEMA TYPE	Schema on write	Schema on read, data is simply stored
DATA VOLUME AND GROWTH RATE	Gigabytes) to Terabytes(10*12)	Terabytes to (10*12)zettabytes(10*21)
DATA VELOCITY	Relatively Slow	Very High
STORAGE	Brand Redundant Servers	Cheap HW/White Boxes
BACKUP SYSTEM	Well designed	No need
VENDORS	Traditional vendors	Open source solutions
PROGRAMMING	Traditional programming, SQL	Non-traditional (Map - Reduce), NoSQL
DATA MODEL	Strict schema based	Flat Schema

DATA VELOCITY	Batch or near-real time	Generally real-time
DATA SOURCE	Generally internal (transactional data, web transactions, CRM Data etc.)	Both internal and external (Social media, sensor data, log data, device data, video and images etc.)
DATA VALUE	Business Intelligence, analysis and reporting	Complex, advanced, predictive business analysis and insights
DATA STRUCTURE	Predefined structured	Unstructured, structured or semi structured
DATA RELATIONSHIP	Stable and complex interrelationship	Unknown, almost flat with few relationship
DATA LOCATION	Centralized	Highly distributed
DATA ANALYSIS	After the complete build	On-stream, intermediate analysis
DATA REPORTING/DISTILLATION	Limited with pre-defined interaction paths in structured database	Possible in all directions of the data also in real time according to the demand
COST FACTOR	Specialized high end hardware and software	Inexpensive commodity boxes in cluster mode

CAP THEOREM (CAP=consistency, availability, partition tolerance)	Consistency- Top Priority	Availability – Top Priority

1.3 Characteristics of Big Data

As its name inspires the term Big Data firstly used to define huge amounts of data in volume before 2000s. However, in 2001 Doug Laney is an analyst from META Group (now Gartner Research Company) defines Big Data as 3D data by considering big data's components, variations and the opportunities in a report of the company. 3D term here corresponds the 3V of Big data those are increasing volume, high speed namely velocity and range of data as variety [6-7].

Today, most definitions of Big Data still depend on this 3V [8]. But in 2012 Gartner updated its definition as extreme volume, extreme velocity and overmuch variety [9]. Later on some other companies added a new Vs to this definition as veracity and value in order to describe different aspects of big data. Along with all formal definitions the "big data" is now being used for predictive analysis, user behavior analysis and any other analytics aiming to extract value from data [10].

In accordance with the rapid development of big data technologies the number of specific characteristics of big data is increasing rapidly, following table summarizes the main characteristics;

Table 1.2: Characteristics of Big Data

9 V of Big Data	Description	Landscape
Volume	How big is your data?	GB, TB, ExaB, PetaB, TerraB
Variety	How many types of data composed together?	Image, log, text, sensor data etc.
Velocity	How speedy is your data?	Doubling every 18 months, Near real time response needed
Veracity	How poor is the quality?	Untrusted, trusted, uncleansed, cleansed
Value	What is the return of your data?	Return of data into Money or other benefit
Vigilance	Is data in motion or rest?	Real-time, batch, Hybrid use of SQL and NoSQL
Visualization	Does your data trigger a valuable decision?	Concretization of data
Viscosity	Does it stick with you? Does it call for action?	Complex event processing
Virality	Ability of data to be distributed over networks	Time and number of crosslinks

The most accepted five characteristics of big data will be mentioned in the following section briefly.

1.3.1 Volume

Volume is used to explain the amount of structured and unstructured data with size of gigabytes to zettabytes. Several data sources from social media to sensors generates huge amounts of data every minute. For example, users upload 400 hours of new video to YouTube, Facebook users share 216.302 photos and Google Translate translates 69.500.000 word in every minute by 2016. Additionally, 69,444 hours of video streamed from Netflix, and Amazon makes $258,751.90 in sales, and 103,447,520 spam mails were sent, and LinkedIn gained more than 120 are professionals, and Americans use 2,657,700 GB of internet data in every minute by 2017. These are data related with internet, besides these huge amounts are being generated from logs, sensors, and other devices.

Gartner and IDC state that data is doubling every 18 months, current estimation says there is over 4 Zettabytes of data in the World and if these speed continues the amount data will be 40 Zettabytes by 2020. For today, 4 Zettabytes equals 1 million 4 Terabytes hard drives, if all this data is printed on 8”x10” paper and laid end to end is 210Trillion Miles that is equal to 35.8 Light years. And all that printed data would need 16.4 trillion trees which is equal to 4 times of summation of all trees over the world [11].

This amazing numbers gives a preview about how the generated data is big [12].

Following table summarizes the impact of social media on data growth;

Table 1.3: Data production rates of Social media [19]

Data Source	Data Production
YouTube	▪ Users upload 100 hours of new videos per minute ▪ Each month, more than 1 billion unique users access YouTube

	▪ Over 6 billion hours of video are watched each month, which corresponds to almost an hour for every person on Earth. This figure is 50% higher than that generated in the previous year
Facebook	▪ Every minute, 34,722 Likes are registered ▪ 100 terabytes (TB) of data are uploaded daily ▪ Currently, the site has 1.4 billion users ▪ The site has been translated into 70 languages
Twitter	▪ The site has over 645 million users ▪ The site generates 175 million tweets per day
Foursquare	▪ This site is used by 45 million people worldwide ▪ This site gets over 5 billion check-ins per day ▪ Every minute, 571 new websites are launched
Google+	▪ 1 billion accounts have been created
Google	▪ The site gets over 2 million search queries per minute Every day, 25 petabytes (PB) are processed
Apple	▪ Approximately 47,000 applications are downloaded per minute
Brands	▪ More than 34,000 Likes are registered per minute
Tumblr	▪ Blog owners publish 27,000 new posts per minute
Instagram	▪ Users share 40 million photos per day

Flickr	▪ Users upload 3,125 new photos per minute
LinkedIn	▪ 2.1 million groups have been created
WordPress	▪ Bloggers publish near 350 new blogs per minute

1.3.2 Velocity

As mentioned above large datasets produced by transactions with high refresh rate results in data streams with a very high pace. Parallel with data growth hardware features are been developing rapidly however, this development cannot catch the speed of data. Gartner and IDC state that data is doubling every 18 months [11].

At this point the big data technologies play a key role in order to analyze and get meaningful and valuable information from row data.

1.3.3 Variety

Today, data is being produced by a wide range of sources. In addition to human generated data, the term "Internet of Things" namely IoTs emerges at this point which refers to the network of intelligent devices which include sensors to measure the environment around them, actuators which physically act back into their environment such as opening a door, processors to handle and store the vast data generated, nodes to relay the information and coordinates to help manage sets of these components [13].

IoTs consists of multiple technologies, including ubiquitous wireless communication, real-time analytics, machine learning, commodity sensors, and embedded systems [13] of wireless sensor networks, integrated and discrete control systems, all automation systems(including home and building automation), and more [14].

The following figure depicts a summary of IoTs landscape;

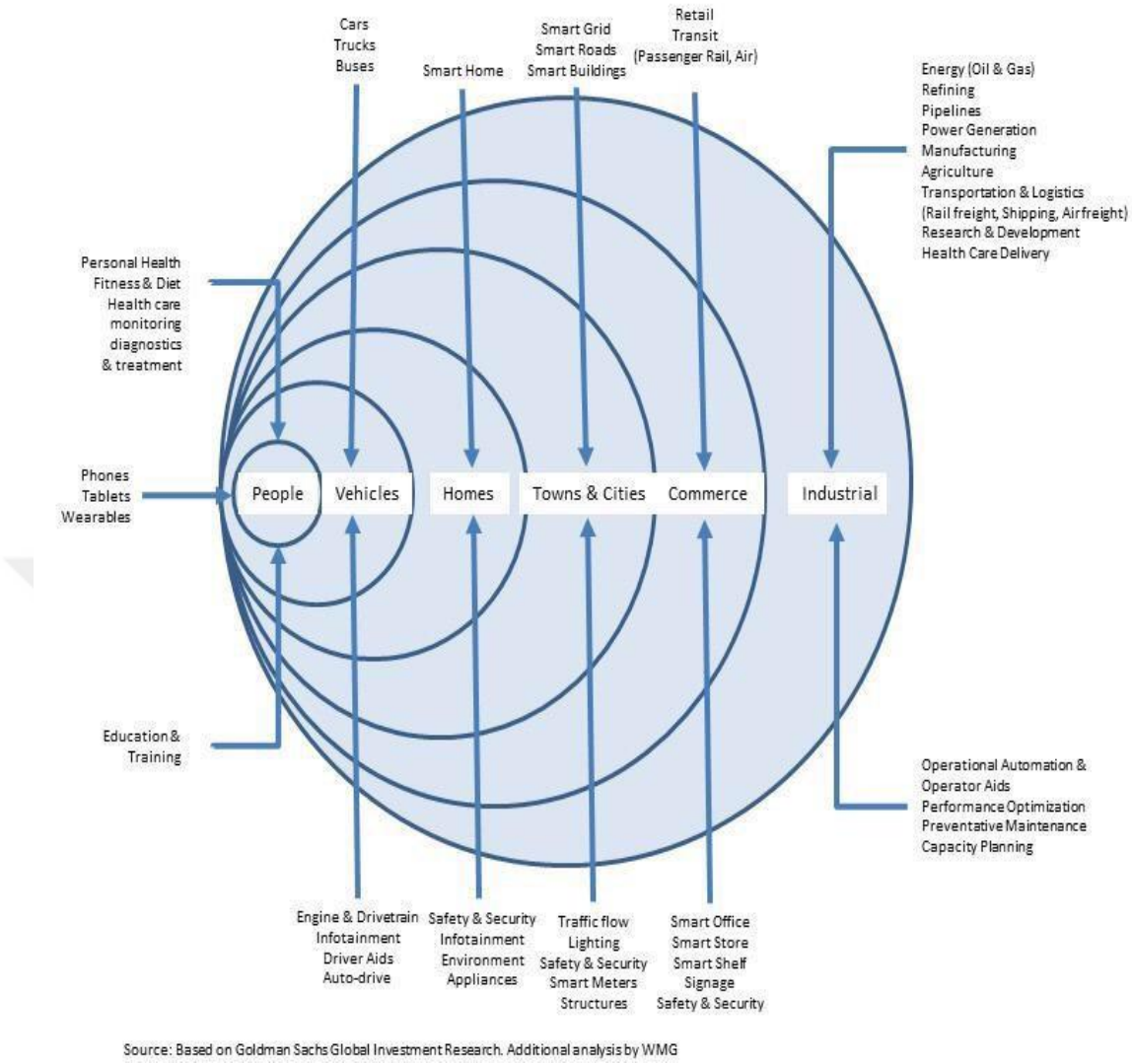


Figure 1.1: IoTs landscape [15]

Both human generated and machine generated data will be in different formats as structured, semi-structured, and unstructured.

Structured data refers to any data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets while Unstructured data is all those things that can't be so readily classified and fit into a neat box: photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents. Additionally, Semi-structured data is a cross between the two. It is a type of structured data, but lacks the strict data model structure. With semi-structured data,

tags or other types of markers are used to identify certain elements within the data, but the data doesn't have a rigid structure. For example, word processing software now can include metadata showing the author's name and the date created, with the bulk of the document just being unstructured text. Emails have the sender, recipient, date, time and other fixed fields added to the unstructured data of the email message content and any attachments. Photos or other graphics can be tagged with keywords such as the creator, date, location and keywords, making it possible to organize and locate graphics. XML and other markup languages are often used to manage semi-structured data [2].

A brief comparison of main characteristics of this three data types is given in the following section;

Table 1.4: Comparison of structured, semi-structured and unstructured data

Features	**Structured Data**	**Semi-Structured Data**	**Unstructured Data**
Technology	Relational database table	XML, RDF, CSV…	Character and binary data
Storage	RDBMS, spreadsheets	XML Repository…	Unmanaged file structures
Metadata	Syntax	Semantics, Syntax	Semantics
Integration Tools	ETL, ELT	Legacy Systems	Batch Processing, manual data entry
Standard	SQL, ADO.NET, ODBC	OpenXMl, JSON,SMTP, SMS,CSV…	JPEG, Social Media inputs, Twitter records …

Databases	MySQL, MsSQL, Oracle	MongoDB, STORED etc..	Hadoop, HDInsight, MongoDB
Context	Text, digit, BLOB..	XML, HTML, JSON..	Text, image, audio, video, documents
Scalability	Difficult	Simple	Very Scalable
Robustness	Very Robust	Not widely spread	Robust
Flexibility	Schema dependent	Flexible tolerant schema	Very flexible, absence of schema
Transaction Management	Matured, various concurrency techniques	Adapted from RDBMS not yet Matured	No transaction management, no concurrency
Version Management	Versioning over tuples, rows, tables …	Versioning over tuples or graphs is possible	Versioned as a whole
Query Performance	Structured query allows complex joins	Queries over anonymous nodes are possible	Only textual queries possible

1.3.4 Value

Data value measures the usefulness of data in making decisions. It has been noted that "the purpose of computing is insight, not numbers". Data science is exploratory and useful in getting to know the data, but "analytic science" encompasses the predictive power of big data [16]. Data especially the unstructured data does not give valuable insights unless it is processed by using several techniques and tools. Big data makes

this process more difficult because of its volume. Fortunately, hardware's processing power and storage capacity are also increasing in very a high speed and parallel computing techniques coupled with Big Data technologies like Hadoop, HDFS, MapReduce, Spark etc. makes big data mining possible to gain value from huge amounts of messy data.

1.3.5 Veracity

A lot of data generated are noisy, e.g., data from sensors. Data are often incorrect. For example, many websites you access may not have the correct information. It is difficult to be absolutely certain about the veracity of big data [17].

In other words, veracity is about the completeness and accuracy of data. The main causes can be listed as; data inconsistency and incompleteness, ambiguities, latency, deception and model approximations. IBM reports that more than 12 terabytes of Tweets created daily and more than 5 million trade evets is generated per second, all these operations result in more than 100 different types of data and only 1 of 3 decision makers trust their information [18].

This shows the importance of veracity of the data and as expected unstructured big data is more likely to have less veracity than structured data. However, big data is more likely to include useful information hidden in it.

1.4 Big Data Taxonomy and Lifecycle of Big Data

As mentioned above, today data generation speed is very high and getting higher day by day. Proportional to this high speed data types constituting big data spread a very wide range. Following part summarizes the main classification of sources of big data;

- **Electronic Device Records and Sensor Data:** Especially this kind of data is being produced on real-time for example industrial and biological sensors
- **Social Network Data**: This kind of data is generated as a result of human interactions on internet. Nearly all social networking data consists of unstructured text in natural languages and needs to be analyzed via special analytic methods like sentiment analysis, trend topics analysis in order to take out valuable information and meaningful insights.

- **Business and Financial Transaction Data**: This kind of data will be structured, semi- structured or unstructured form and data is produced at a very fast pace since every commercial organization keeps its activities' and transactions' logs. These logs may be stored in relational databases or in plain text, PDF, Excel, picture forms etc.
- **Electronic Files**: These type of data consist of unstructured documents, statically or dynamically produced which are stored or published as electronic files, like Internet pages, videos, audios, PDF files, etc.
- **Broadcasting Data**: Refers to video and audio produced on real time, getting statistical data from the contents of this kind of electronic data. Broadcasting data is huge in amount and hard to handle. İt requires different analytic methods.
- **Large Scale Science:** Refers to outputs or inputs of scientific researches. May be in structured, semi structured or unstructured form like digital measurement data or medical images.
- **Network Security:** Refers to logs of network activities in order to keep the system secure generally in structured form [20].

Giving the taxonomy the following section mentions about the big data lifecycle. All data requires a well-designed lifecycle management process. The traditional data lifecycle consists of mainly 4 steps as; collection, storage, analyze and consume phases. However, big data is more complicated so is the most demanding type. As Khan stated [19]; 'This problem was first raised in the initiatives of UK e-Science a decade ago. In this case, data were geographically distributed, managed, and owned by multiple entities [4]. The new approach to data management and handling required in e-Science is reflected in the scientific data life cycle management (SDLM) model. In this model, existing practices are analyzed in different scientific communities. The generic life cycle of scientific data is composed of sequential stages, including experiment planning (research project), data collection and processing, discussion, feedback, and archiving [58–60]. The following section presents a general data life cycle that uses the technology and terminology of Big Data. The proposed data life cycle consists of the following stages: collection, filtering & classification, data

analysis, storing, sharing & publishing, and data retrieval & discovery. The following sections briefly describe each stage as exhibited in Figure 1.2.

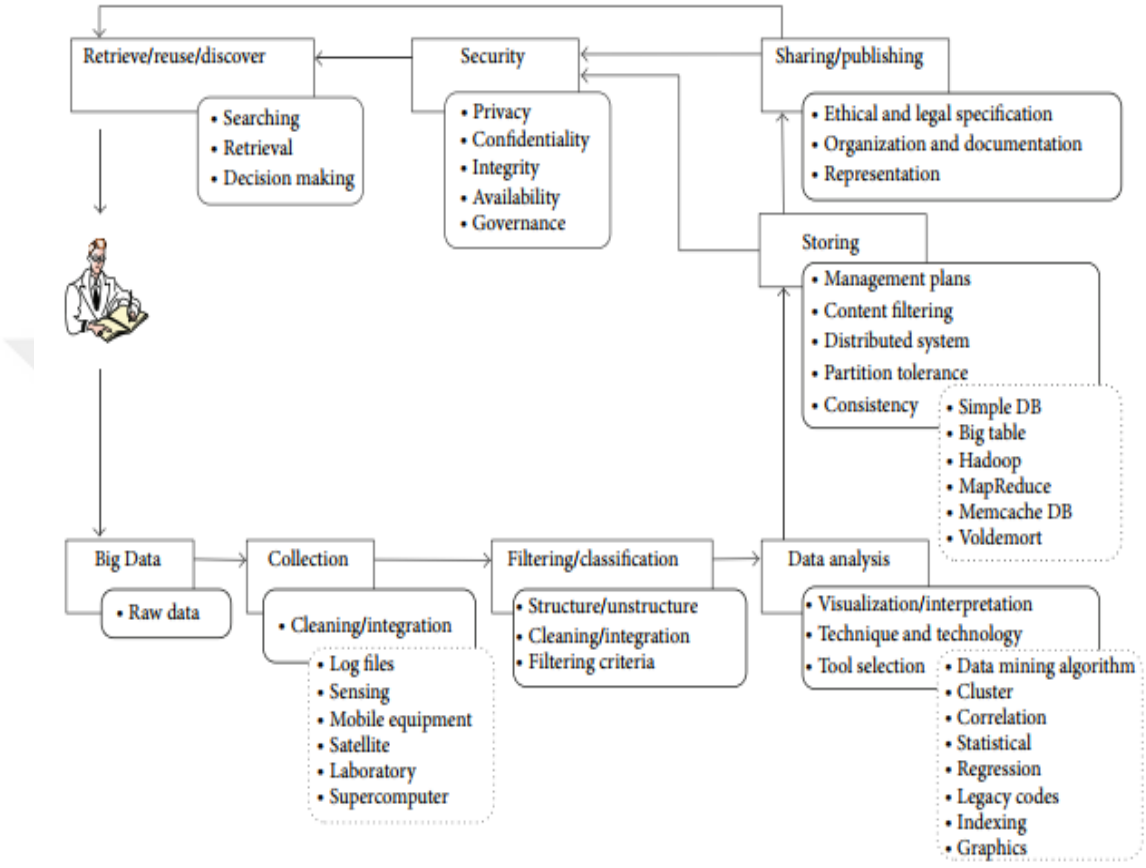


Figure 1.2: Proposed data life cycle using the technologies of Big Data [19]

Here, the raw data implies the bare data generated by the systems listed above in structured, semi-structured or unstructured forms. Collection, filtering and integration phase is the first stage of big data lifecycle. Then filtering and classification phase aims to sort data according to its structure and other filtering criteria. Data analysis phase aims to interpret data using data mining algorithms like: clustering, correlation, statistical, regression, legacy codes, indexing or graphics. After data analysis data is stored in suitable database like: SimpleDB, Big Table, Hadoop, MapReduce, Memcache DB or Voldemort. Then optionally, publishing the findings will be considered parallel with the security stage. Because some risks threaten big data like

data leakage and additionally, confidentially, integrity, availability and governance issues must be handled. After all these eight steps data retrieval and decision making phase come and the lifecycle will be completed according to Khan's method. However, there are some other models proposed some other researchers, an accepted 9 step of big data lifecycle model is presented by the Figure 1.3

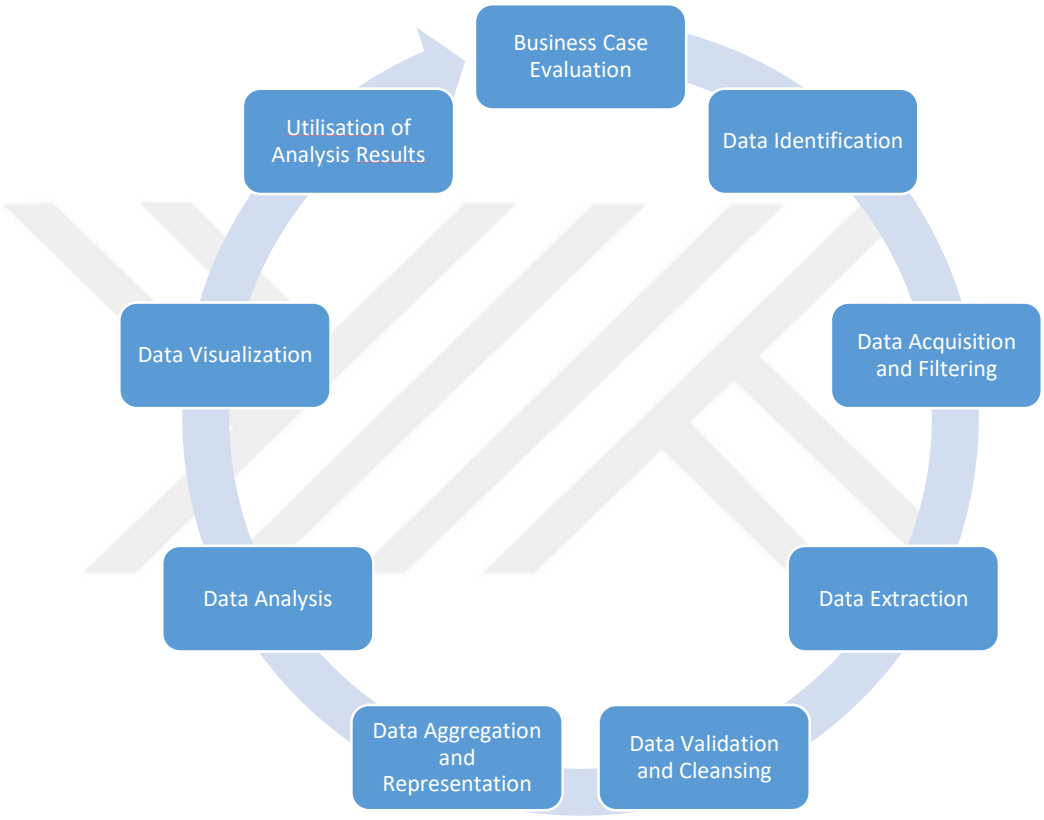


Figure 1.3: Big Data Lifecycle [21]

Although the differences between models the common inspiration says that big data analysis differs from traditional data analysis in many aspects. Therefore, big data analysis requires a more structural approach, so given lifecycle helps personals or organizations dealing with big data for a systematical and efficient analysis.

1.5 Big Data Storage Management and Security

Big data is a different data concept therefore it requires a different storage and management procedure. This procedure has to meet the following key requirements; [22].

- Scalability
- Tiered storage support
- Self-managing
- High availability
- Wide accessibility
- Security
- Self-healing
- Integration with legacy applications
- Enable integration with public, private and hybrid cloud ecosystems [65]
- Work flow automation support [65]
- Both analytical and content applications support [65]

Reckon with these requirements the following physical storage options will be considered to store big data;

- Traditional Storage Systems(RDBMS)
- NAS, SAN
- DAS, Tape
- Object-based Storage
- Distributed Nodes
- Scale Out NAS
- Hyper Scale Storage (Google, Facebook, Apple etc.) [22]

Traditional storage systems are not an efficient way to store big data in terms of satisfaction of the requirements listed above.

A small comparison of mostly-used storage systems seen below;

Table 1.5: Comparison of mostly used data storage systems [24]

Architecture	Description	Best Use Case
Distributed nodes	Most often implemented as low-cost commodity JBOD directly attached to the compute server: can be true direct-attached storage or even server memory	Hadoop, small distributed files
Scale-out network-attached storage(NAS)	NAS capable of scaling throughput and capacity in tandem or separately: usually has its own distributed or clustered file system	Large files processing; more traditional extract, transform, load implementations of big data
All-Solid-State (SSD) arrays	All-SSD arrays can be implemented like JDOB and distributed nodes, or as a traditional fully featured array	High performance processing where time is of the essence
Object-based storage	Stores data in flexible containers, not blocks; uses hash tables and replication instead of RAID; allows peer-to-peer file sharing across distributed nodes	Organizations willing to experiment, but in search of a real competitive advantage

Giving the physical storage options briefly, following part summarizes available database management tools and technologies for big data;

1.5.1 NoSQL (not Only Structured Query Language)

Big data by its nature could not fit traditional RDBMS. Therefore, as a different approach NoSQL has been developed by the big companies like Google, Amazon, Yahoo etc. in order to meet the need of propagating data storage. As Adiba stated the main objectives of NoSQL systems are to increase scalability and extensibility, using classic hardware. They also ensure reliability, fault tolerance, and good performance despite the increasing number of requests. The common characteristic of these systems is that they enable data manipulation without having to define a schema, so data are mainly semi-structured. Thereby, they avoid schema definition and database loading after data cleaning. The characteristics of these systems architecture are: horizontal extensibility for executing distributed simple operations on a multitude of servers; data partitioning and duplication on several servers (data shading); low level interfaces for accessing the systems; concurrent access model more flexible than ACID transactions; distributed indexing and in memory storage; easy data structure evolution. They provide basic CRUD functions adapted for efficiently manipulating data structures according to their underlying data model [25].

NoSQL now leads the way for the popular internet companies such as LinkedIn, Google, Amazon, and Facebook - to overcome the drawbacks of the 40-year-old RDBMS. Following figure summarizes the market places and the economic value of NoSQL.

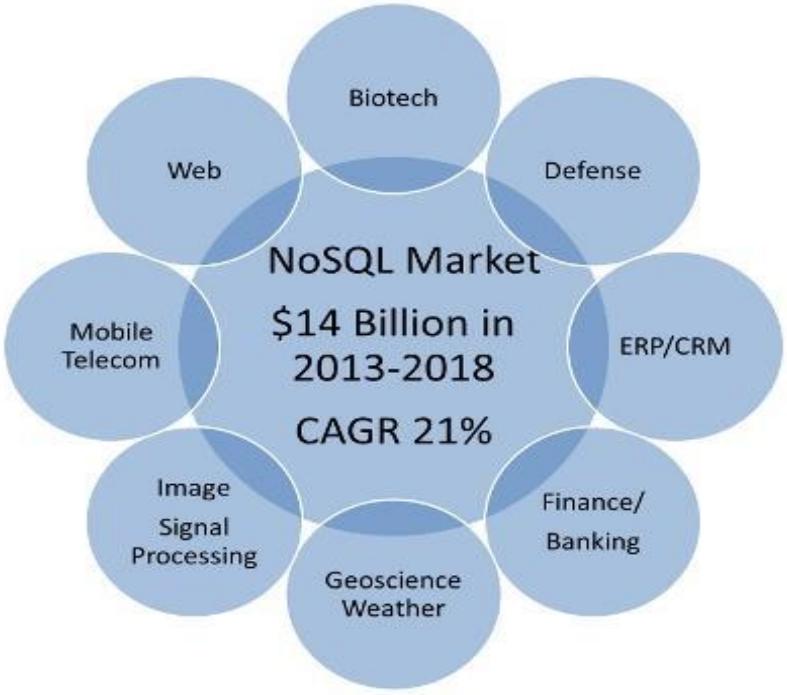


Figure 1.4: Big data applications landscape [26]

1.5.2 NewSQL

NewSQL is next-generation Scalable RDBMS for OLTP that provide scalable performance of NoSQL systems for read-write workloads, as well as maintaining the ACID guarantees of a traditional database system [27].

Learning NewSQL is easier than SQL and it is more elegant, consistent, and well defined.

NewSQL is not a variety of Object Database and has a different approach based on top of the cross database library LDBC, therefore it should not be considered a subset or extension of SQL.

In SQL data is stored in RDBMS. In NOSQL data is stored in graphs and trees and NewSQL gives both advantages. In SQL goes in one machine where as in NoSQL data is distributed and NEWSQL gives advantages of both. Here, a comparative study shows the main difference of traditional SQL, NoSQL and NewSQL databases [28].

Figure 1.5: Relationship of Traditional SQL, NoSQL and NewSQL

1.5.3 Cassandra

Cassandra is an open source data storage system suitable for large amounts of data spread over the World served by Apache Foundation. It is distributed and decentralized and provides highly available service with no single point of failure [29].

Cassandra is scalable, fault-tolerant and consistent. It serves key-value and column-oriented databases created at Facebook and is being used by many biggest companies like Facebook, Twitter, Cisco, Rackspace, eBay, Twitter, Netflix, and more.

Additionally, Cassandra serves the following facilities;

- Elastic scalability
- Always on architecture
- Fast linear-scale performance
- Flexible data storage
- Easy data distribution
- Transaction support (ACID)
- Fast writes [29]

1.5.4 MongoDB

MongoDB is a cross-platform, document oriented database that provides, high performance, high availability, and easy scalability. MongoDB works on concept of

collection and document. A document is a set of key-value pairs. Documents have dynamic schema. Dynamic schema means that documents in the same collection do not need to have the same set of fields or structure, and common fields in a collection's documents may hold different types of data [30].

Following table shows the relationship of RDBMS terminology with MongoDB;

Table 1.6: RDBMS and MongoDB terminology [30]

RDBMS		MongpDB
Database	⟺	Database
Table	⟺	Collection
Tuple/Row	⟺	Document
Column	⟺	Field
Table Join	⟺	Embedded Documents
Primary Key	⟺	Default key_id
Database Server And Client		
Mysqld/Oracle	⟺	mongod
mysql/sqlplus	⟺	mongo

In MongoDB the data can have a hierarchical structure, named JSON. Here is a sample of JSON document:

```
{
   Firstname:  'John',
   Lastname:   'Steven',
   Hobbies:
   {
      {
         Hobbyname:  'Tennis',
         HoursPerDate  :  5
      },
      {
         Hobbyname:  'Swimming',
         HoursPerDate  :  3
      }
  }
```

Figure 1.6: MongoDB JSON code sample [31]

1.5.5 HBase

HBase is an open source, non-relational big data storage system developed in Java programming language and served by Apache Foundation. HBase is widely used when

random and real-time access to Big Data is required and is operates on the top of HDFS [32].

1.5.6 Hadoop

Hadoop is open source big data processing software served by Apache foundation again. Hadoop provides reliability and scalability. Using distributed programming techniques Hadoop enables processing the large volume of data across clusters of machines. Hadoop uses batch processing paradigm and MapReduce framework, in order to process data in small batches [33].

1.5.7 Apache Spark

Spark is open source Apache product which enables fast processing of data with large volume using cluster computing methods. Spark resembles Hadoop in operating on batches but its batch window size is remarkable small than Hadoop. It supports Java, Python and Scala programming languages to develop new modules according to the needs. Additionally, spark supports SparkSQL for SQL processing, MLlib for machine learning, GraphX for graph processing and Spark Streaming for stream analysis [34].

1.5.8 Yahoo S4

Yahoo S4 is a real-time data stream processing tool developed by Yahoo and later joined Apache. Yahoo S4 enables modular programming and design of new applications according to needs. It uses a distributed architecture inspired by MapReduce [35].

1.5.9 Storm

Strom is the result of the idea of developing a stream processing system which can be presented as a single program of Nathan Marz. Spark joined Apache Foundation in 2014. Apache Storm supports developers to build real-time distributed processing systems, which can process the unbounded streams of data very fast. It is also called Hadoop for real-time data.

Apache Storm is highly scalable, easy to use, and offers low latency with guaranteed data processing. It provides a very simple architecture to build applications called Topologies along with enabling developers to develop their logic virtually in any programming language. Also it supports communication over a JSON-based protocol over stdin/stdout [36].

1.5.10 SimpleDB

SimpleDB is a distributed document-oriented database developed by Amazon in Erlang programming language. Although SimpleDB has some similarities it differs from relational databases in many aspects. For instance, like a relational database, SimpleDB is designed for storing tuples of related information. Unlike a relational database, SimpleDB does not provide data ordering services—that is left to the programmer. Instead of tables, SimpleDB offers domains, which are schema-less. A domain is filled with items— similar to a row. Each item must have a unique name (which is not generated by SimpleDB) and may contain up to 256 attributes. Each attribute can have multiple values [37].

A weaker version of consistency of traditional database management systems is provided by SimpleDB called eventual consistency. This is often considered a limitation, because it is harder to reason about, which makes it harder to write correct programs that make use of SimpleDB. This limitation is the result of a fundamental design trade-off. By foregoing consistency, the system is able to achieve two other highly desirable properties:

1. Availability – components of the system may fail, but the service will continue to operate correctly.
2. Partition tolerance – components in the system are connected to one another by a computer network. If components are not able to contact one another using the network (a condition known as a network partition), operation of the system will continue [38].

1.5.11 CouchDB

CouchDB is an open source NoSQL document based storage database developed by Apache software foundation. CouchDB uses JSON, to store data (documents), java script as its query language to transform the documents, http protocol for API to access the documents, query the indices with the web browser. It is a multi-master application released in 2005 and became an apache project in 2008. The main pros of CouchDB can be listed as follow;

- Uses HTTP-based REST API, which helps to communicate with the database easily

- Provides a simple structure of HTTP resources and methods (GET, PUT, DELETE) are easy to understand and use.
- Provides a flexible document-based structure, there is no need to worry about the structure of the data.
- Provides with powerful data mapping, which allows querying, combining, and filtering the information.
- Provides easy-to-use replication, using which you can copy, share, and synchronize the data between databases and machines [39].

1.5.12 Redis

Redis is an open-source in-memory database project implementing a distributed, in-memory key-value storage with optional durability. Redis supports different kinds of abstract data structures, such as strings, lists, maps, sets, sorted sets, hyper logs, bitmaps and spatial indexes. The project is mainly developed by Salvatore Sanfilippo and is currently sponsored by Redis Labs [40].

1.5.13 Elasticsearch

Elasticsearch is a search engine based on Lucene which is an again Apache's high-scalable open source, text search engine library. Elasticsearch provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents. It is developed in Java and is released as open source under the terms of the Apache License. Official clients are available in Java, .NET (C#), PHP, Python, Apache Groovy and many other languages. According to the DB-Engines ranking, Elasticsearch is the most popular enterprise search engine followed by Apache Solr, also based on Lucene [41].

Elasticsearch uses Kibana which is an open source data visualization tool. On huge volumes of data users can generate scatter, bar and line plots along with maps and pilot charts Using Kibana plugin.

CHAPTER 2

IMPORTANCE and BENEFITS of BIG DATA ANALYTICS

Although it requires some more budget and work power, many big companies and governments interest in big data because of its function of extracting meaningful information and other valuable returns from huge mess of raw data. A brief list of remarkable benefits of big data analysis given below in following main categories;

2.1 Business Aspect Benefits

- Getting insights about the customer activities
- Making more meaningful business decisions
- Improving sales and service quality
- Integrated customer behavior modeling
- Support for CDR (call details records) processing
- Support for churn prediction (prediction of the customers who are more likely to cancel their subscription, product or service.)
- Support for operations and failure analysis from device, sensor and GPS inputs
- Support for Voice-to-text mining for understanding customer behavior
- Analyzing internet behavior and buying patterns [42]
- Increasing overall revenue and decreasing costs
- Understanding customer demands and expectations
- Improving product quality and features
- Effective utilization of workers and production lines
- Improving the efficiency of advertisements
- Increasing company reputation and customer satisfaction
- Increase systems security
- Preventing possible losses
- Detecting possible fraud in online transactions
- Having idea about new product and service ideas

2.2 Governmental Benefits

- Enabling public security by analyzing sensor data, networking logs and social network data against to terror and cyber attacks
- Predicting epidemic contagious diseases
- Decreasing crime rates
- Enabling crime early warning systems
- Enabling intelligent traffic management systems [42]
- Detecting incidents relationships, take precautions and making meaningful forecasting in healthcare, education, military, agriculture and other fields
- Building and managing smart cities
- Improving transparency and decision-making while reducing costs [43]
- Understanding citizens' demands and take possible precautions
- Detecting tax frauds and minimize it
- Improving mission outcomes
- Make better decisions more quickly
- Identifying and reduce inefficiencies
- Eliminating waste, fraud and abuse
- Boasting return of investment and cut total cost of ownership [44]

  and so on.

CHAPTER 3

TEXT ANALYTICS and BIG DATA TEXT ANALYTICS

Like mining process of raw materials, data needs to be processed in order to get the core and valuable essence of it. Information is the first upper level of the processing output of raw data. Knowledge is the next level and wisdom is on the top. The following figure show the old pyramid of data.

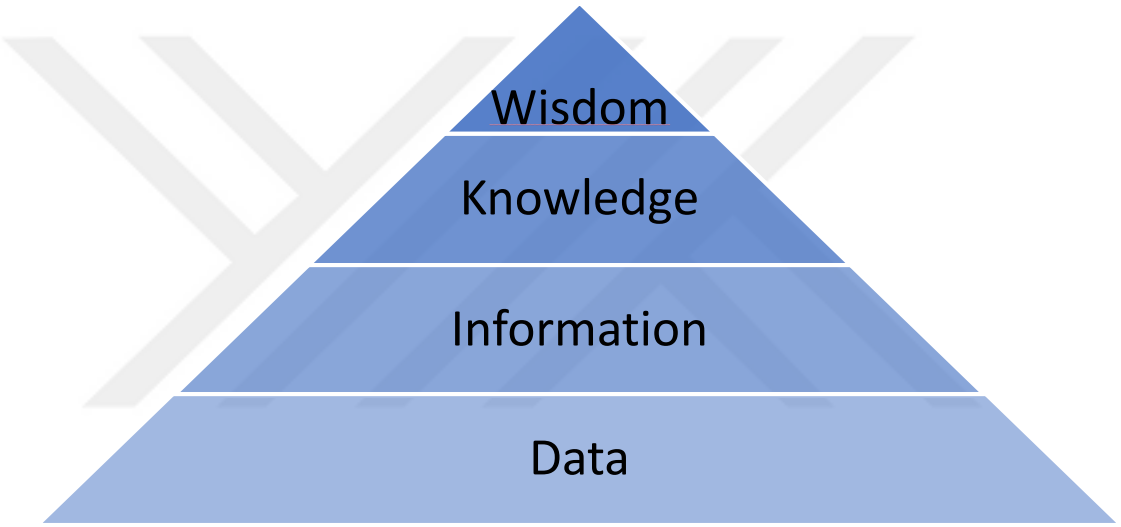


Figure 3.1: Data-Wisdom Pyramid [45]

As stated above, big data hides the wisdom in it. In order to reach that wisdom, we need to analyze it and maintain the knowledge management process. The following section summarizes the knowledge management concept briefly.

3.1 The Knowledge Management Concept

Knowledge management is an important term for all types of organizations. It is used for explaining the deliberate, systematic and synchronized approach to ensure the full utilization of the company's knowledge base, paired with the potential of individual skills, competencies, thoughts, innovations, and ideas to create a more efficient and effective company [46].

In other words, as King states knowledge management is the planning, organizing, motivating, and controlling of people, processes and systems in the organization to ensure that its knowledge-related assets are improved and effectively employed.

The processes of KM can be listed as below;

- Knowledge Acquisition
- Creation
- Refinement
- Storage
- Transfer
- Sharing
- Utilization

Following figure illustrates the Knowledge management process model;

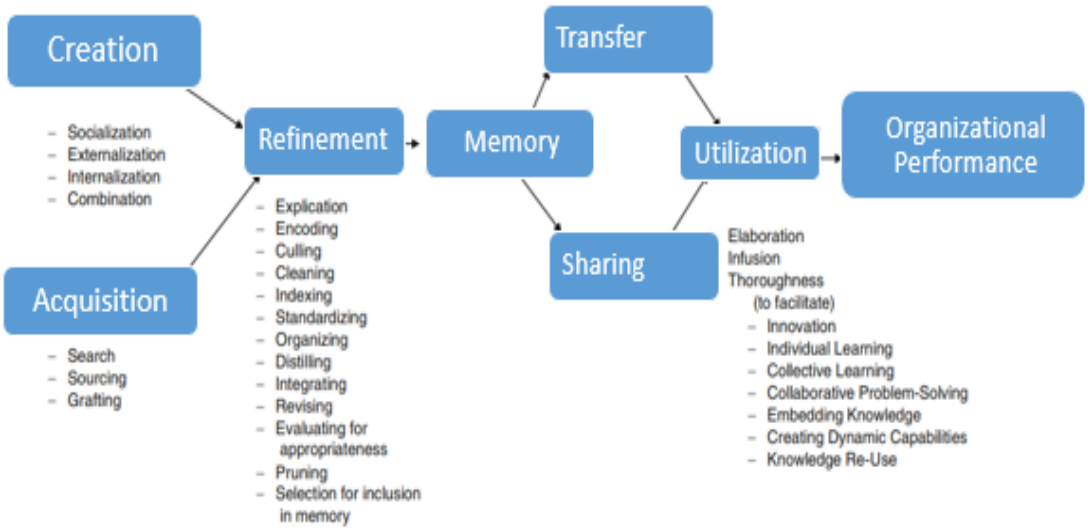


Figure 3.2: Knowledge Management Process [47]

KM function in the organization operates these processes, develops methodologies and systems to support them, and motivates people to participate in them. The main objectives of KM are the leveraging and improvement of the organization's knowledge

assets to effectuate better knowledge practices, improved organizational behaviors, better decisions and improved organizational performance [47]

3.2 Big Data and Knowledge Management

Being in a fast changing technologic world, we need to understand and make meaningful decisions for future from the data in hands. Especially, organizations should create and maintain the business knowledge process healthy. Today, there are two main approaches for knowledge management as traditional KM and big data-based KM methods. As Shorfuzzaman states and the following figure illustrates; "Traditional approach typically focuses on conversion of tacit knowledge into explicit knowledge. The normal flow is to capture people's know-how and good practices and then to codify them and put them into repository. On the contrary, big data-based KM deals with discovering the rise of new knowledge based on the huge volume of data that are amassed today. This data is mostly collected from internal sources in addition to external sources, especially from the clouds that we have access to. The focus of big data-based KM will be to do knowledge predictions, knowledge navigation, and knowledge discovery to support enhanced operations and decision making in organizations. These two approaches are in fact not mutually exclusive and can be applied to help businesses and societies at the same time." [48].

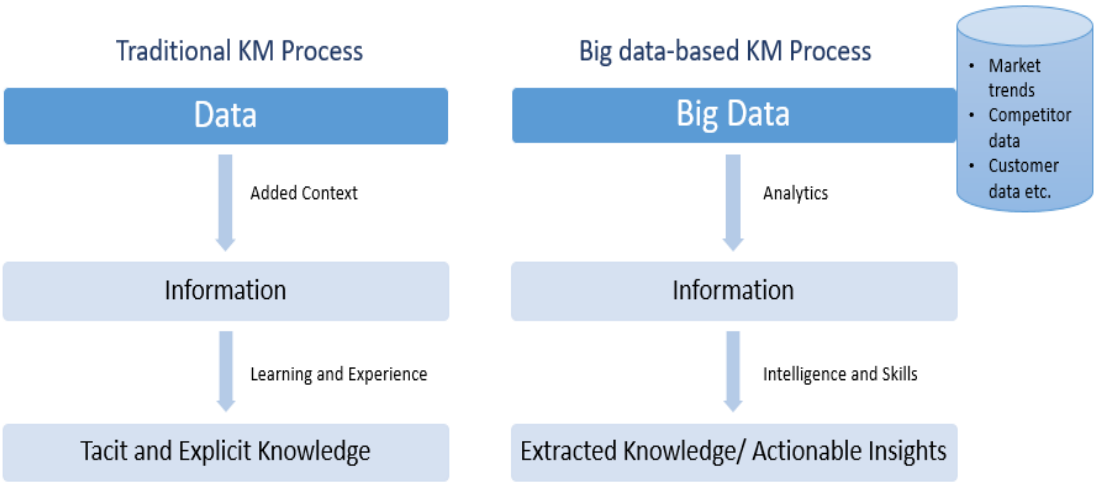


Figure 3.3: Comparison of Traditional KM process vs Big Data based KM process [48]

At this point we can say that Big data helps us to handle and analyze raw structured and unstructured data comes from various channels with large volumes, so from acquisition to utilization phases big data can be considered as an enabler of knowledge management.

Following table summarized the differences between traditional KM and big data-based KM methodologies;

Table 3.1: Comparison of Traditional KM vs Big Data Based KM [48]

Factor	Traditional Knowledge Management	Big Data based Knowledge Management
Knowledge Type	Both tacit and codified knowledge	Data is processed in real time to extract useful knowledge
Necessary Skills	Apparently not necessary	Analytical skills are highly necessary
Orientation	Highly person oriented	Depends on more on machine, less on person
Interaction	Requires frequent face to face interaction with people	Requires minimal face to face interaction with people
Knowledge Creation and Storage	Tacit knowledge repository is mostly people' brain. Tangible storage of huge size.	Cloud storage is mostly used. Knowledge is created through perpetual flow and processing.

A new generation of data management for decision support process is being created by big data today. And businesses are aware of the opportunities that big data held. A crucial component of deriving valuable information from data and also big data is use of analytics [49]. Therefore, the analytics phase of big data is an important component of knowledge extraction and decision making for business and governments. According to main purpose of this thesis study a review of big data analytics concept and a deep look at text analytics concept is going to be handled.

3.3 What is Analytics?

Collection and storing data is the first step and easy job but do not serve applicable insights by itself. So the collected data needed to be processed. In 1970s "Decision Support Systems" (DSS) was established in order to meet the need of value extraction from data. In 1990s "Business Intelligence" (BI) concept emerged and take the place

of DSS. Up to 2010s BI had an important function for organizations. From 2010s the concept of "Analytics" is the main functional actor of information extraction mechanism. While dictionary explanation of analytics is "Information resulting from the systematic analysis of data or statistics", it is defined as "Discovery, interpretation, and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance." [50].

At this point it can be said that "Big data analytics examines large and different types of data to uncover hidden patterns, correlations and other insights." [51].

Giving the necessary definitions the main methods used in big data analytics listed below;

3.3.1 Descriptive Analytics

Helps the extraction of information about past and present in an easily understandable form. The presentation of population data which classifies population across a country by sex, job, income, tax, ethnic root etc. can be given as an example [17].

3.3.2 Predictive Analytics

By using statistical methods, neural networks, and machine learning algorithms predictive analysis serves the ability of foreseeing of expected event that can happen in the near future from available data in hand [17].

3.3.3 Exploratory / Discovery Analytics

This type of analytics enables figuring out the unexpected relationships between different parameters in collections of big data [17].

3.3.4 Prescriptive Analytics

Prescriptive analysis identifies opportunities to optimize the solutions to existing problems from previous data [17].

3.3.5 Golden Path Analysis

Golden path analysis is a new and extraordinary type of predictive analysis. It focuses on the large volume of data associated with the activities or actions of people (behavioral data) in order to identify patterns of evets or activities that foretell customer actions such as not renewing a cell phone contract. The main motivation

point of golden path analysis is predicting customer behavior and by using some offers change the anticipated behavior [49].

3.4 What is Text Analytics?

Hurwits and Association Company defines the term text analytics; “Text analytics is the process of analyzing unstructured text, extracting relevant information, and transforming it into structured information that can be leveraged in various ways.” [52].

The following list summarizes the benefits of text analytics;

• Monitor and analyze brand reputation

• Determine purchase behavior

• Identify product issues

• Summarize surveys, customer reviews

• Improve customer service and customer experience management

• Understand customer feedback

• Improve customer retention

• Predict and reduce churn

• Identify and reduce frauds

• Develop cross-sell, upsell strategies

• Design next best offer strategies [52]

And according to Zaratsian common techniques of text analysis can be summarized as follow;

- “Text Mining (i.e. Text clustering, data-driven topics)
- Categorization (i.e. Tagging unstructured data into categories and sub-categories; hierarchies; taxonomies)
- Entity Extraction (i.e. Extracting patterns such as phrases, addresses, product codes, phone numbers, etc.)

- Sentiment Analysis (i.e. Tagging positive, negative, or neutral with varying levels of sentiment)
- Deep Linguistics (i.e Semantics. Understanding causality, purpose, time, etc.)" [66]

As expected text analysis has a process flow illustrated as below;

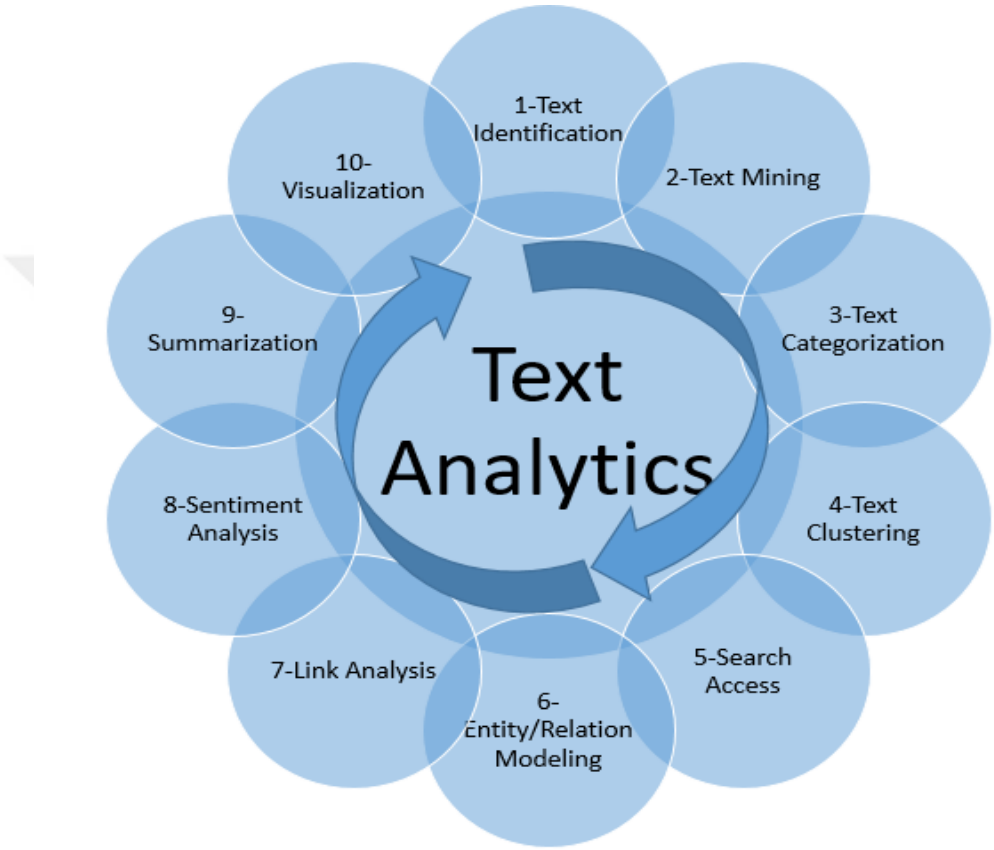


Figure 3.4: Text Analytics Process flow [53]

3.4.1 Data Cleaning and Preprocessing

In general, the analyst needs to clean and preprocess text data before applying analytics process. The main aim of preprocessing is to make the inputs to a given analysis less complex in a way that does not adversely affect the interpretability or substantive conclusions of the subsequent model. In practice, perfecting this tradeoff- simpler data, but not too much information loss- is a non-trivial matter, and scholars have invested considerable energies in exploring the optimal way to proceed [54]. Most common methods used in preprocessing are;

- Decapitalization
- Tokenization
- Pruning words back to their stems i.e. stemming
- Removing very common words (like of, the, an etc.)
- Removing numbers (if numbers are meaningless)
- Removing punctuation characters
- Removing Infrequently Used Terms
- Inclusion of n-grams
- Striping white spaces

Today, the importance of text analytics is accepted and companies along with governments give attention to it. This means more focus and work power are given to this issue. Following figure illustrates a general view of the application areas of text analytics and changes the percentages according to years;

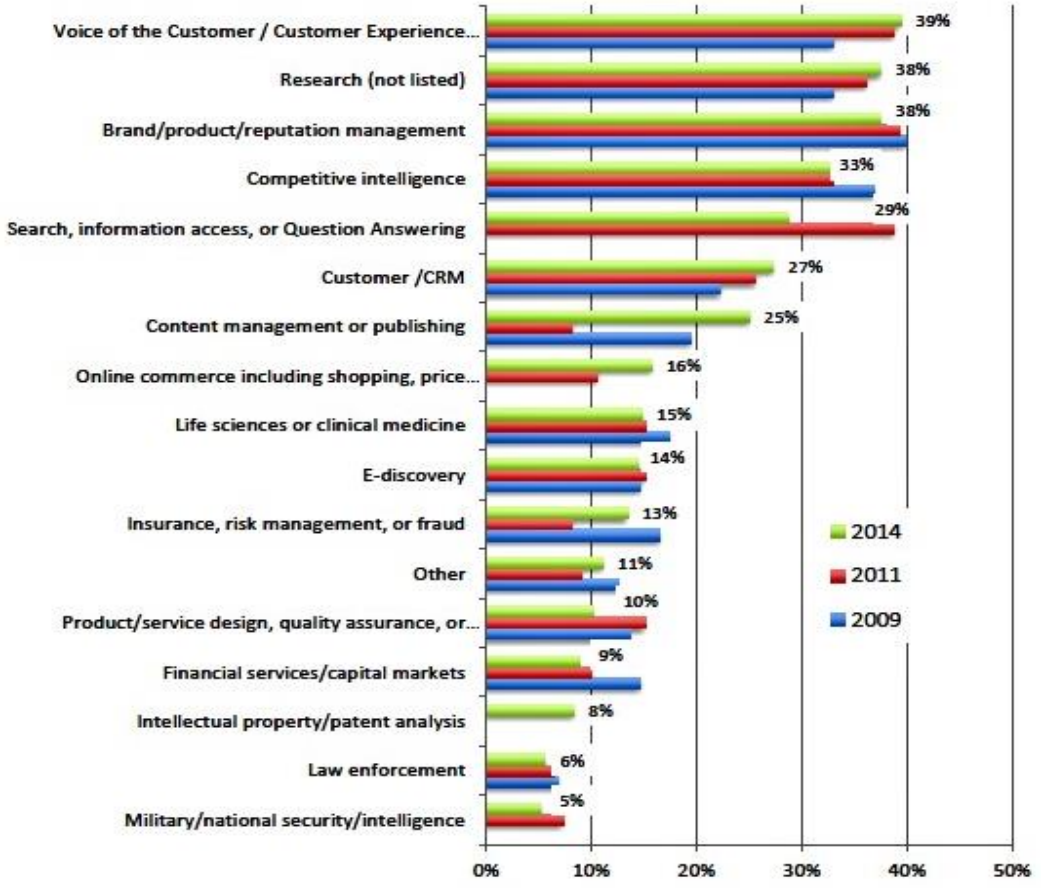


Figure 3.5: Application landscape of text analytics [55]

3.5 Why Big Text Data Analytics and Big Data Text Analytics Process Flow

Giving the benefits and applications of text analytics the importance of big data text analytics takes attention, because the amount of unstructured text data including emails, surveys, digital documents, call center logs, claim records, customer forms, customer letters, public requests and recommendations, public complaints, blogs, social media entries, tweets, forums, articles, reports, all other web entries etc. is one of the main sources of big data. In order to get benefit from all these sources we need to apply text analytics on big data. As is known big text data differs from traditional text data stored in RDBMs, so it needs a modified approach of text analytics. An accepted big data text analytics model is given below;

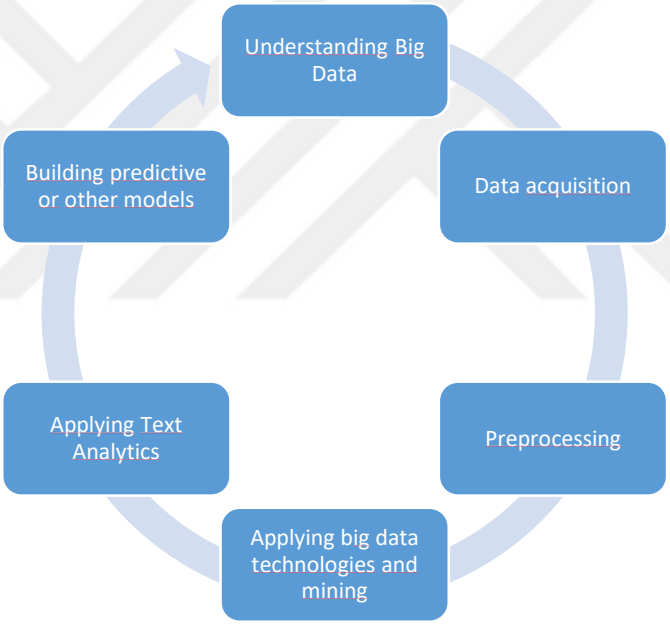


Figure 3.6: Big Data text analytics process flow [53]

3.6 Main Terminology of Text Analytics

In the following section the main and most commonly used terms of text mining are given according to the Gartner's Text and Data Mining glossary [56].

3.6.1 Application Programming Interface (API): The abbreviation API is used to explain the technically developed environment that enables the users to reach huge

amount of information varying from text to objects in a format that readable by machine [56].

3.6.2 Corpus: Generally used as “text corpus” and it means a set of documents, like web pages or journal articles [56].

3.6.3 Crawling: For scraping information from a website the automatic technique is called crawling, by finding and following links [56].

3.6.4 Entity: Used for defining e real world thing such as dog [56].

3.6.5 Extensible Mark-up Language (XML): Maybe the mostly used document markup standard of web is XML. That is designed for enabling simplicity and flexibility for web authorship and design. In addition, XML is not in a fixed-format so it differs from HTML [56].

3.6.5 Hypertext Mark-up Language (HTML): HTML is developed for building web pages, that is in text based format and can be translated by web browsers in to its own structure [56].

3.6.6 Information Extraction: In order to automatically discard some specific words or information from the going to be analyzed unstructured text [56].

3.6.7 Machine Learning: The overall of the statistical methods and mathematical algorithms used for automatically extract patterns from data [56].

3.6.8 Natural Language Processing (NLP) Tools: NLP tools are some software services and systems that helps the analyzer in order to do automatic analysis [56].

3.6.9 Ontology: An ontology consists of a domain of entities and their relationships as an organization. For example, a domain could be “automobile” or “chemistry”, the entity could be a specific car brand e.g. Mercedes and all the forms that it might show up in a particular text [56].

3.6.10 Parsing: (Linguistic) To define the syntactic analysis of some text namely identifying how a sentence traces the grammar rules of a language, the word parsing is used. It breaks down a sentence into its sub components like words [56].

3.6.11 Relationship Extraction: Refers to the process of automatically finding relationships between two or more entities within a text. Usually used for finding semantic relationship [56].

3.6.12 Semantic Relationship: Semantic relationship is used for defining the linguistic relationship between at least two entities for enabling the machine to understand the relationship between those two or more entity [56].

3.6.13 Sentiment Analysis: The extraction of words or phrases which have emotional meanings. For example, the sentence "The meal was tasteless and low-quality" indicates a negative sentiment as the word "tasteless" have emotional meanings [56].

3.6.14 Scraping: Refers to the method of automatic data extraction from human-readable output that is coming from another program, web page or another computer program [56].

3.6.15 Taxonomy: Taxonomy refers to a vocabulary that is organized in hierarchical control, or enriched with synonyms and non-hierarchical relationships e.g. human is a homo-sapiens, is a mammal, etc. [56].

3.6.16 Text and Data Mining (TDM): Text mining is the process of data analysis of the text extracted from natural language products like newspapers, books etc. assuming the text as a data form. Generally merged with data mining and referred as text and data mining (TDM) [56].

3.6.17 Treebank: Refers to a complete works of syntactically parsed documents used to train text and data mining models [56].

3.7 Common Text Analytics Tools

Although the plenty of data analytics tools, at this part of the thesis a list of mostly used commercial and open source text analytics tools is given in tabular form;

Table 3.2: Summary of text analytics tools [57]

Category	Product	Text Analytics Abilities	Target Buyer	Use Case Example
Analytical DBMS	IBM DB2, HPE Vertica, Teradata Aster Analytics, Oracle (via OEM of Lexalytics), SAP Hana, and SAP IQ	Broad text mining	Tech Management	Improving performance of high number of records processing using scalable DBMS engines.
Predictive Analytics	Alpine Data Labs, Alteryx, Angoss Software, Dell, FICO, IBM, KNIME, Microsoft, Oracle, RapidMiner, SAP, SAS, and Teradata Aster Analytics	Broad text mining capabilities	Business	Leveraging results of text mining into predictive models.
Contact Center Platforms	CallMiner and Genesys	Mostly speech -to-text mining and analytics	Business	Improving contact center agent quality

Data integration	Ab Initio, Informatica, and SAP	Broad text mining capabilities	Tech Management	Preparing data for loading into data warehouses, datamarts, or to be used by apps.
Document and eDiscovery Classification	Active Navigation, Exterro, FTI Consulting, HPE Records Manager, IBM StoredIQ eDiscovery, Knowliah, Megaputer, Nuix, and OpenText Discovery	Focused on document classification	Business	Classifying documents by categories (e.g., risk, compliance, etc.).
Enterprise Content Management (ECM)	IBM, Lexmark Enterprise Software, OpenText	Focused on document classification	Business	Classifying documents by categories (e.g., risk, compliance, etc.).
Natural Language Processing (NLP)	AlchemyAPI, Apache OpenNLP, Basis Technology, Content Analyst,	Natural language processing	Tech management	Embedding NLP into other platforms

	Google Cloud Machine Learning and HPE Haven OnDemand			and customer apps.
Search	AddStructure, Attivio, Coveo Solutions, Google, HPE, IBM, Lexmark International, Lucidworks, Mindbreeze, Oracle BigData Discovery, and Squirro	Broad text mining capabilities	Tech management	Improving keyword search results with semantic search and knowledge discovery.
Social Media Listening	Brandwatch, Clarabridge, EPAM, Infegy, NetBase, Synthesio, and Sysomos	Focused on sentiment analysis	Business	Garnering insights from social and mainstream media

CHAPTER 4

In this section of the thesis, the main dataset will be introduced.

OFFICIAL GAZETTE

4.1 About Turkish Official Gazette i.e. Resmi Gazete

Official Gazette is the official publication organ of Turkish Republic that the validity of many official transactions depends on. Roots of Official Gazette depends on "Takvim-i Vekai" which was a journal published weekly during the administration of Sultan II. Mahmud.

However, The Turkish Republic's official journal has been started its life at 10/07/1920 by the decision of Grand National Assembly. This date still exists on the first page of Official Gazette as the year of foundation. Firstly, the name of this official publication was "Ceride-i Resmiyye" until 10/09/1923, after that date the name was updated and used as "Resmi Ceride" up to 17/12/1927. At 17/12/1927 the name was being updated as "Resmi Gazete" from 763rd issue and still this name is being used.

Official Gazette was published weekly up to 15th issue inclusive, 16th,17th and 18th issues were published biweekly. 19th, 20th and 21st issues also published weekly. However, because of the wars the 22nd issue was published after two years from 21st issue. Therefore, the 134th - 338th laws come into force without published in the Official Gazette.

Until 01/12/1928 Official Gazette was published with ottoman letters, from that date it was published with the current Latin letters. From 1929, Official Gazette was being published daily except national festivals, general holidays and Sundays until 09/05/1970. According to the changes made on 09/05/1970 it is being published on Sundays and 18/05/2009 it will be published on national festivals and general holidays if it is necessary for the maintenance of the service. Additionally, if there is a compulsory and urgent situation Official Gazette will be published repeatedly with the "Mükerrer" commentary and the same number in the same day.

جريدهٔ رسميه

بازار ايرتسى
٢/ جمازى‌الاول ١٣٣٩
٧ شباط ٣٣٧
تاريخ تأسيسى
٧ تشرين اول ١٣٣٦

آبونه شرائطى
يوز نسخه‌سى ٢٥٠ غروش
الى ٠ ١٢٥ ٠
بير نسخه‌سى ١٠٠ پاره‌در

نومروسى : ١ تورکيه بيوك ملت مجلسى حكومتك رسمى غزتەسى اولوب شمديلك هفته‌ده بر دفعه نشر اولنور . سنه : ١

هيئت مرخصه‌نك عزيتى

لوندره‌ده ٢١ شباط: انعقادى مقرر قونفرانسده تورکيا بيوك ملت مجلسى حكومتى تحيل اينك اوزره خارجيه وكيلى بكر سامى بك افندينك رياستلرى آلتنده اعضاى تفرر ايدن هيئت مرخصه پازار كونى آقشامى نصف‌الليله خصوصى بر ترەنله آنقره‌دن حركت ايلەمشدر.

توجيهات ملكيه

آطه واليلكنه ايغايل متصرفى حلمى، ايغايل متصرفلغنه قيره‌شهرى متصرفى اكرم، قيره‌شهرى مصرفلغنه اسكى‌شهر متصرفى سعيد، اسكى شهر متصرفلغنه قرق‌كليسا متصرف سابق حلمى، حكارى متصرفلغنه موصل مبعوث اسبق ابراهيم فوزى، بايزيد متصرفى طاهر بكك‌عزيله بايزيد متصرفلغنه اونى متصرفى صبرى، ايكنجى صنف ملكيه مفتشلكنه داخليه وكالتى مأمورين مميزى صبرى، قره‌كوسه قائمقاملغنه كرده قائمقاملغندن آجيقده جلال‌الدين، قولب قائمقاملغنه الشكرد قائمقام وكيلى حسن ضيا، بور

صالح، الوجره قائمقاملغنه صرشهرى قائمقامى احسان، صو شهرى قائمقاملغنه سورمه مكتوبجيسى يوسف‌ضيا، رشاديه قائمقامى منيرالف بكك عزيله رشاديه قائمقاملغنه پروازى قائمقامى عبدالهادى بك واغديار تعيين ايدلمشلردر.

قانونلر

[تورکيا بيوك ملت مجلسنك اندائى تشكلندن اعتباراً موقع مرعيته وضع ايتديكى قانونلر:]

نومرو: ٨٥

تشكيلاتى اساسيه قانونى

مواد اساسيه

ماده ١ ــ حاكميت بلا قيدوشرط ملتكدر. اداره اصولى خلقك مقدراتى بالذات وفاعل اداره ايتمسى اساسنه مستنددر.

ماده ٢ ــ اجرا قدرتى وتشريع صلاحيتى ملتك يكانه وحقيقى ممثلى اولان بيوك ملت مجلسنده تجلى وتمركز ايدر.

ماده ٣ ــ تورکيا دولتى بيوك‌ملت مجلسى طرفندن اداره‌اولنور وحكومتى «تورکيا بيوك ملت مجلسى حكومتى» عنوانى طاشير.

ماده ٤ ــ بيوك ملت مجلسى ولايتلر خلفنجه منتخب اعضادن مركبدر.

ماده ٥ ــ بيوك‌ملت مجلسنك انتخابى ايكى سنه‌ده‌بر كره‌اجرا اولنور. انتخاب اولنان اعضانك اعضالغى مدتى ايكى سنه‌دن عبارت اولوب

Figure 4.1: First Official Gazette (Ceride-i resmiyye) with Ottoman letters

Official Gazette is formed by three main parts as legislation, execution and administration. The content of Official Gazette is explained as “To regulate the provisions concerning the cooperation, authority and duty fields of the Prime Ministry, ministries and public legal entities, provided that they are not concerned with national safety and national security and not a degree of confidentiality; regulations that cover general provisions of public staff or public interest” in the 3011th Official Gazette published at 24/05/1984.

T. C.
Resmî Gazete

Kuruluş Tarihi : (7 Teşrinievvel 1336) - 7 Ekim 1920

Yönetim ve Yazı İşleri İçin Başbakanlık Neşriyat Daire Başkanlığına başvurulur	24 Mayıs 1984 PERŞEMBE	Sayı : 18410

YASAMA BÖLÜMÜ

KANUNLAR

Kütahya İli Gediz İlçesinin Merkezinin Değiştirilmesi Hakkında Kanun

Kanun No. 3006 — Kabul Tarihi : 16/5/1984

MADDE 1. — Kütahya İli Gediz İlçesinin merkezi yeni yerleşim merkezi Yenigediz'e nakledilmiştir.

MADDE 2. — Bu Kanun 5/8/1970 tarihinden geçerli olmak üzere yayımı tarihinde yürürlüğe girer.

MADDE 3. — Bu Kanun hükümlerini Bakanlar Kurulu yürütür.

22/5/1984

Gümrük Mevzuatına Göre Tasfiye Edilecek Eşya Hakkında Döner Sermaye Kanunu

Kanun No. 3007 — Kabul Tarihi : 16/5/1984

BİRİNCİ BÖLÜM
Amaç, Kapsam ve Kuruluş

Amaç ve kapsam

MADDE 1. — Gümrük denetimindeki sundurma, antrepo ve depolarda bulunan ve Gümrük Kanunu ile Kaçakçılığın Men ve Takibine Dair Kanuna göre tasfiyesi öngörülen eşyanın satış ve tasfiyesi amacıyla gerekli yerlerde ve sayıda tasfiye işleri döner sermaye işletmeleri kurmak ve 100 milyon lirası cari yılda genel bütçeden ödenmek üzere Maliye ve Gümrük Bakanlığına bir milyar lira sermaye tahsis edilmiştir.

Yasama Bölümü Sayfa : 1

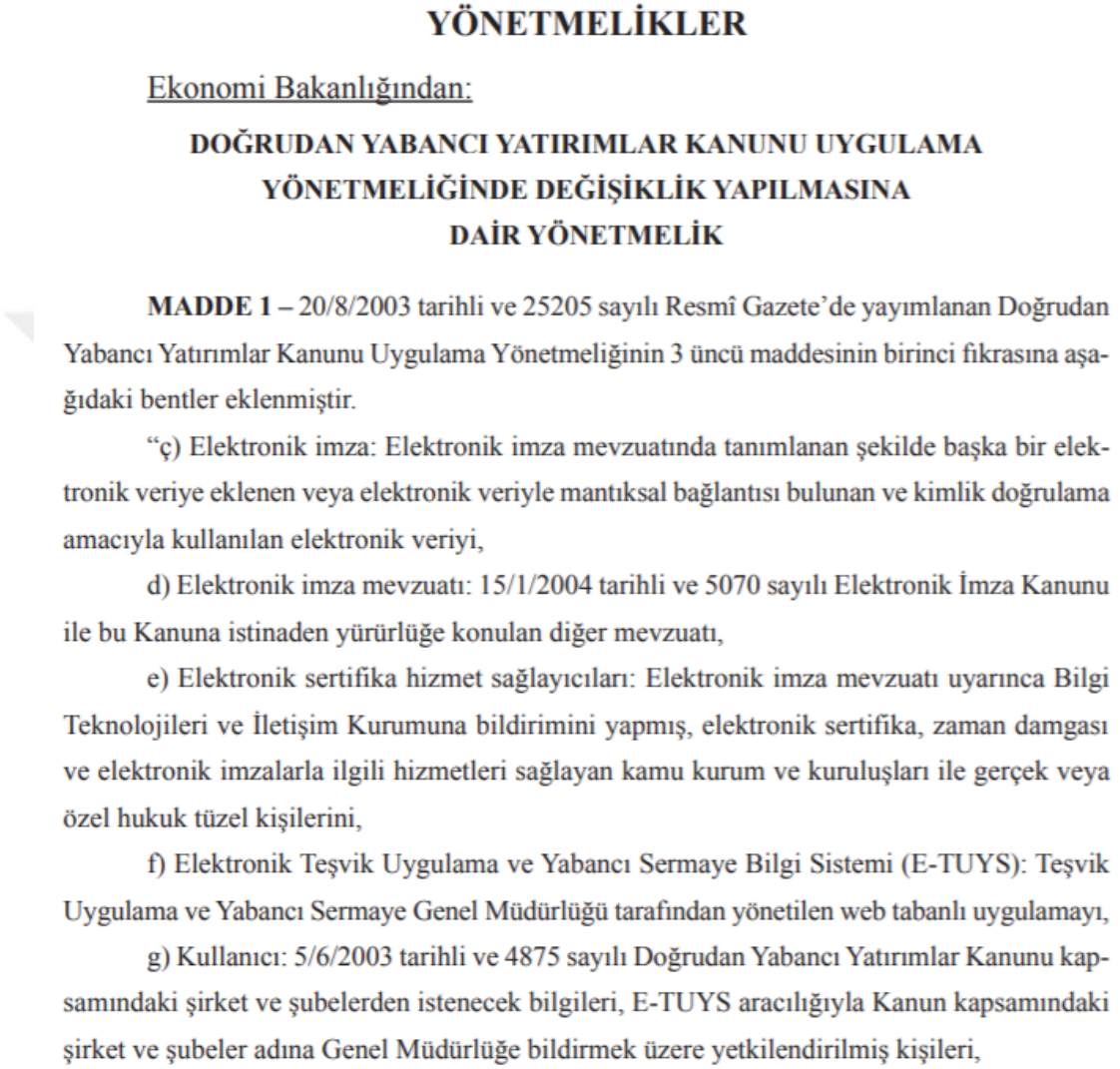
Resmî Gazete Kodu : 840584 — Resmî Gazete Fihristi 64. Sayfadadır.

Figure 4.2: Official Gazette with Turkish letters

At present, the distribution of the Official Gazette is done by posting to the subscribers, and everybody who wants also can buy Official Gazette from the Prime Ministry

Printing House. In addition, it can be subscribed to the Official Gazette from domestic and abroad for at least one-year period.

Furthermore, from 2011, all issues from 07/02/1921 of Official Gazette can be accessed at http://www.resmigazete.gov.tr/default.aspx [58].



YÖNETMELİKLER

Ekonomi Bakanlığından:

DOĞRUDAN YABANCI YATIRIMLAR KANUNU UYGULAMA YÖNETMELİĞİNDE DEĞİŞİKLİK YAPILMASINA DAİR YÖNETMELİK

MADDE 1 – 20/8/2003 tarihli ve 25205 sayılı Resmî Gazete'de yayımlanan Doğrudan Yabancı Yatırımlar Kanunu Uygulama Yönetmeliğinin 3 üncü maddesinin birinci fıkrasına aşağıdaki bentler eklenmiştir.

"ç) Elektronik imza: Elektronik imza mevzuatında tanımlanan şekilde başka bir elektronik veriye eklenen veya elektronik veriyle mantıksal bağlantısı bulunan ve kimlik doğrulama amacıyla kullanılan elektronik veriyi,

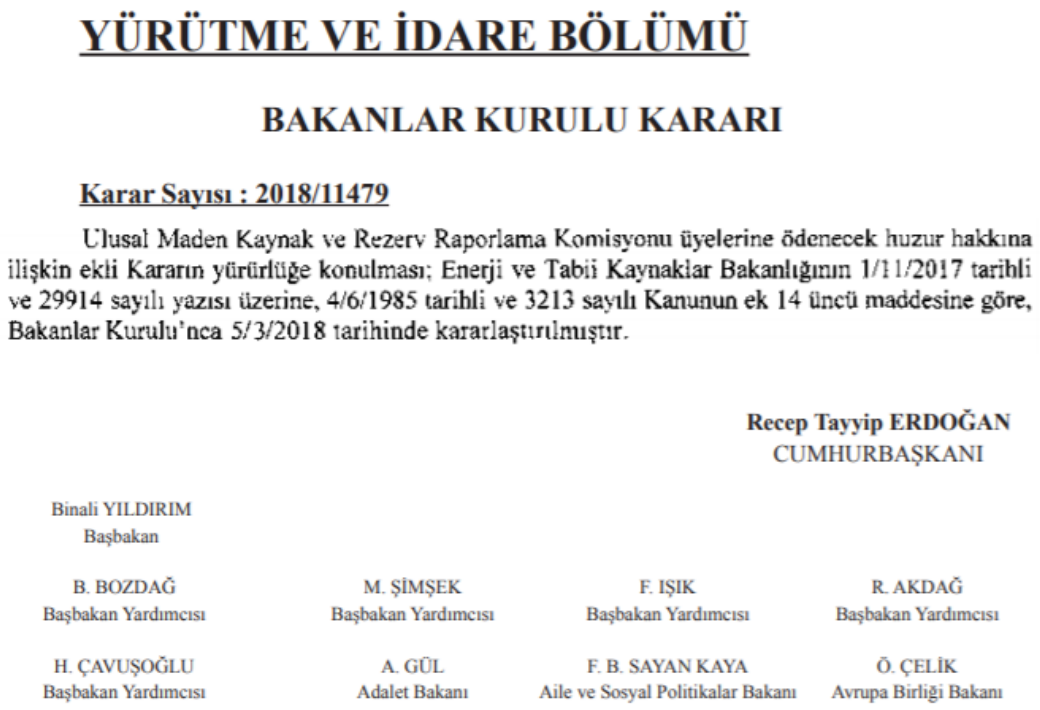
d) Elektronik imza mevzuatı: 15/1/2004 tarihli ve 5070 sayılı Elektronik İmza Kanunu ile bu Kanuna istinaden yürürlüğe konulan diğer mevzuatı,

e) Elektronik sertifika hizmet sağlayıcıları: Elektronik imza mevzuatı uyarınca Bilgi Teknolojileri ve İletişim Kurumuna bildirimini yapmış, elektronik sertifika, zaman damgası ve elektronik imzalarla ilgili hizmetleri sağlayan kamu kurum ve kuruluşları ile gerçek veya özel hukuk tüzel kişilerini,

f) Elektronik Teşvik Uygulama ve Yabancı Sermaye Bilgi Sistemi (E-TUYS): Teşvik Uygulama ve Yabancı Sermaye Genel Müdürlüğü tarafından yönetilen web tabanlı uygulamayı,

g) Kullanıcı: 5/6/2003 tarihli ve 4875 sayılı Doğrudan Yabancı Yatırımlar Kanunu kapsamındaki şirket ve şubelerden istenecek bilgileri, E-TUYS aracılığıyla Kanun kapsamındaki şirket ve şubeler adına Genel Müdürlüğe bildirmek üzere yetkilendirilmiş kişileri,

Figure 4.3: A sample from an Official Gazette belonging to 2018 with less noise



YÜRÜTME VE İDARE BÖLÜMÜ

BAKANLAR KURULU KARARI

Karar Sayısı : 2018/11479

Ulusal Maden Kaynak ve Rezerv Raporlama Komisyonu üyelerine ödenecek huzur hakkına ilişkin ekli Kararın yürürlüğe konulması; Enerji ve Tabii Kaynaklar Bakanlığının 1/11/2017 tarihli ve 29914 sayılı yazısı üzerine, 4/6/1985 tarihli ve 3213 sayılı Kanunun ek 14 üncü maddesine göre, Bakanlar Kurulu'nca 5/3/2018 tarihinde kararlaştırılmıştır.

Recep Tayyip ERDOĞAN
CUMHURBAŞKANI

Binali YILDIRIM
Başbakan

B. BOZDAĞ Başbakan Yardımcısı	M. ŞİMŞEK Başbakan Yardımcısı	F. IŞIK Başbakan Yardımcısı	R. AKDAĞ Başbakan Yardımcısı
H. ÇAVUŞOĞLU Başbakan Yardımcısı	A. GÜL Adalet Bakanı	F. B. SAYAN KAYA Aile ve Sosyal Politikalar Bakanı	Ö. ÇELİK Avrupa Birliği Bakanı

Figure 4.4: Sample from an Official Gazette belonging to 2018 with more noise

4.2 Why Official Gazette as Data Source

The Official Gazette archive was selected as big data source for this thesis in 2013. At the conditions of that year big data text analytics was comparatively a new concept and data sources were very limited. After a literature review it is concluded that there was not any study on Official Gazette in terms of text analytics.

The entire digital archive is nearly 80 GB in size. This amount may seem small for being a big data source, however, size namely volume is only one V of big data, the V for variety encapsulates big data as a source. Since the archive consists of unstructured PDF text files, scanned documents as images and also maps varying in size and number of pages it is suitable to handle with big data tools.

However, since the archive consists of scanned files the data quality is seriously low for handling data as text. This is an unexpected and unwanted situation, although the trials to preprocess a selected part of the archive to prepare for big data analysis, the results have deflection from ideal.

CHAPTER 5

APPLICATION

5.1 Introduction

The main purpose of this application is to generate a term-frequency matrix of a subset of Official Gazette archive using Apache Spark as big data technology in order to extract some value from lexical features of Official Gazette archives and get some insights about the correlation between Official Gazette and TRT News archives in terms of how they affect each other.

5.2 Technical Requirements

In this section a brief information going to be given below about development tools used in this study.

5.2.1 Java JDK

Java is a need for HDFS by default, therefore I installed the current (JDK 1.8) as the first step of setting up the development environment. It can be found at: http://www.oracle.com/technetwork/java/javase/downloads/index.html

5.2.2 Apache Spark

Spark has started its life in 2009 as a subproject of Hadoop Matei Zaharia in UC Berkeley's AMPLab. It was donated by Apache in 2013 and in 2014 became a top-level popular Apache project. It is designed for fast computation and can be defined as a very fast cluster computation technology. Spark developed on Hadoop MapReduce technology, but it extends MapReduce. The key feature makes Spark faster is in-memory cluster computing which enables efficient computation of different types like interactive queries, iterative algorithms and stream processing. Second key advantage is the reduced cost of management of maintaining storage tools [59].

There are four main features makes Spark important, these are;

- **Speed:** In memory, Spark runs 100 times faster an application than Hadoop Cluster and 10 times faster while running on disk. It reduces the number of read

and write operations to disk by storing the intermediate processing data in memory.

- **Advanced Analytics:** Along with support of MapReduce Spark also supports SQL queries and machine learning libraries, graph algorithms and streaming data processing.
- **Multiple language Support:** Scala, Python and Java languages are supported by Spark. So we can develop our applications in different languages according to our needs by taking advantages of each one.
- **Ease of Installation:** Installation of Apache Spark is considerable easy than Hadoop. Although, Hadoop need more and notably complex configurations, Spark serves a simple installation experience [59].

There are three ways that one can use Spark on Hadoop, as depicted in following figure;

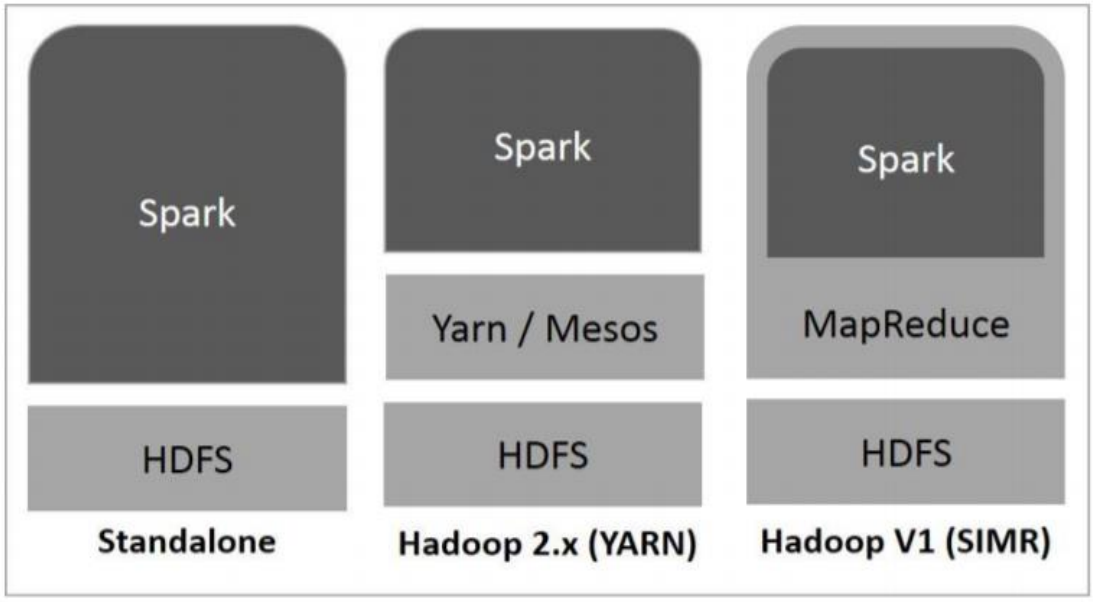


Figure 5.1: Spark installation and working environments [59]

Standalone Apache Spark distribution is selected for this study because of its features listed above like multiple language support, easy installation and fault-tolerance.

5.2.2.1 Resilient Distributed Datasets

In this thesis study Resilient Distributed Datasets i.e. RDD is used that is an important data structure that Spark uses. An RDD is a stable distributed collection of objects. In an RDD every dataset is divided into logical parts, that can be computed on a different cluster or node. Also an RDD can handle any type of Java, Scala or Python objects along with the user defined classes [59].

In this thesis, RDD is used in Python language, related codes/scripts can be found in Appendix A.

5.2.3 Anaconda

Wikipedia describes "Anaconda is a free and open source distribution of the Python and R programming languages for data science and machine learning related applications (large-scale data processing, predictive analytics, scientific computing), that aims to simplify package management and deployment. Package versions are managed by the package management system called conda. Anaconda Distribution is used by over 6 million users, and it includes more than 250 popular data science packages suitable for Windows, Linux, and MacOS." [60].

Anaconda also supports Apache Spark. In order to work with Spark via Anaconda, the pyspark library has to be installed, to do this; following command can used in Windows shell under the location of Anaconda's scripts folder.

-conda install pyspark

Anaconda can be downloaded from https://www.anaconda.com/download/

5.2.4 Python

Python is a higher-level programming language that enables programmers to write codes resembling human language [61].

Python has following advantages that make programming with python more logical;

- Object oriented, procedural, and functional
- Open source general-purpose language
- Easy to generate interface with C/Objective C/Java/Fortran and with C++ (via SWIG)
- Advanced interactive environment [62]

In this study the Python 3.6 distribution that comes with Anaconda was used.

5.2.5 Jupyter-Notebook

Jupyter notebook is a free and light software tool which enables to generate readable Python, Scala, Java or machine learning codes. It also helps to keep code along with images, formulas, comments or plots. It comes with an integrated python version. So we can code and compile easily [63].

In order to install jupyter-notebook conda package manager was used with the following command on shell.

-conda install jupyter-notebook

5.2.6 Renee PDF Aide

Renee PDF Aide is a commercial software developed for converting PDF files to different formats. It includes an OCR engine, therefore used to convert scanned archive files to text document. The free version was used in this study, which can be found at https://www.reneelab.com

5.3 Application Development Activity Diagram

Here, the following diagram depicts the process flow followed in this study;

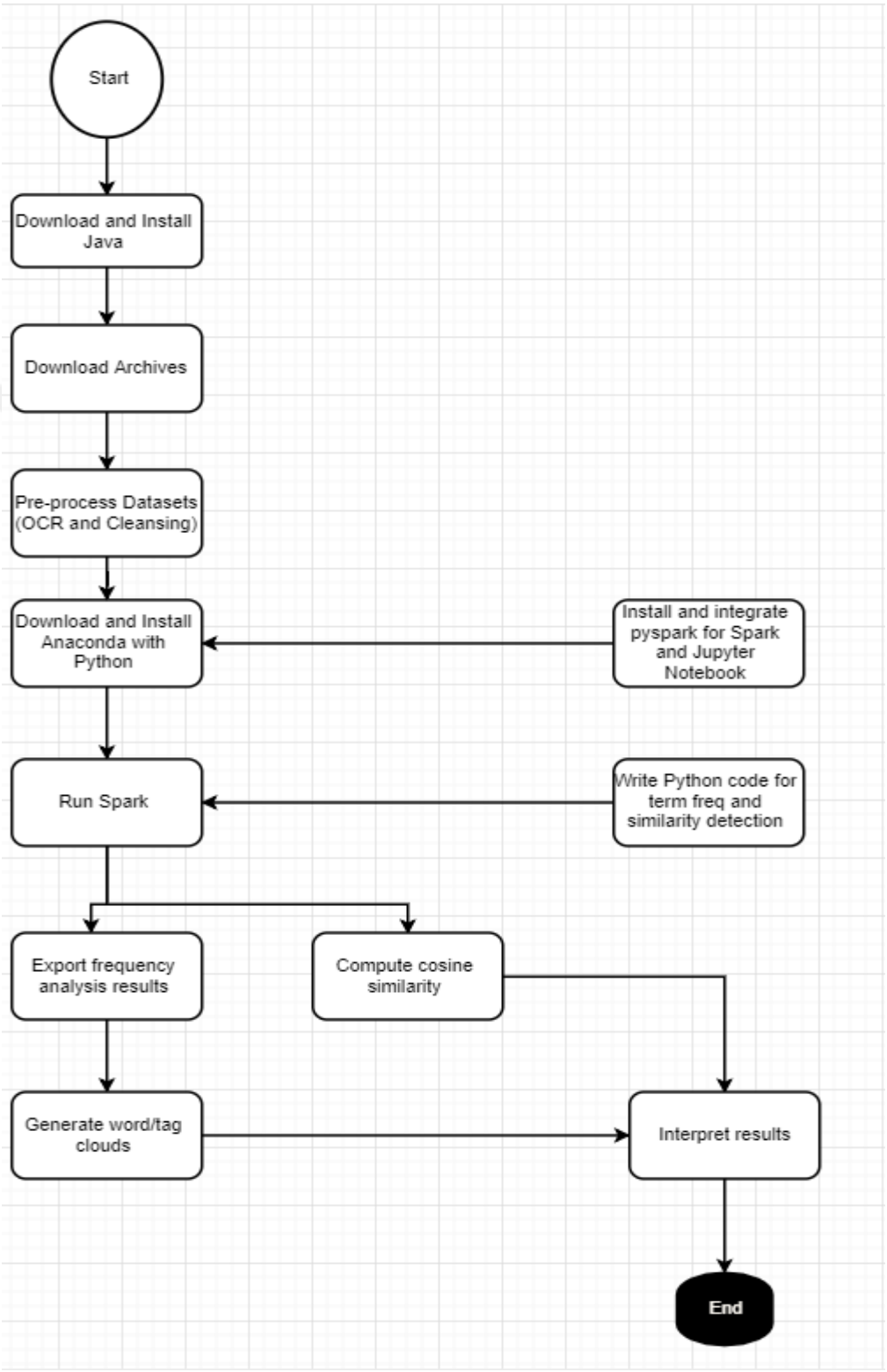


Figure 5.2: Application development Activity Diagram

5.4 Implementation

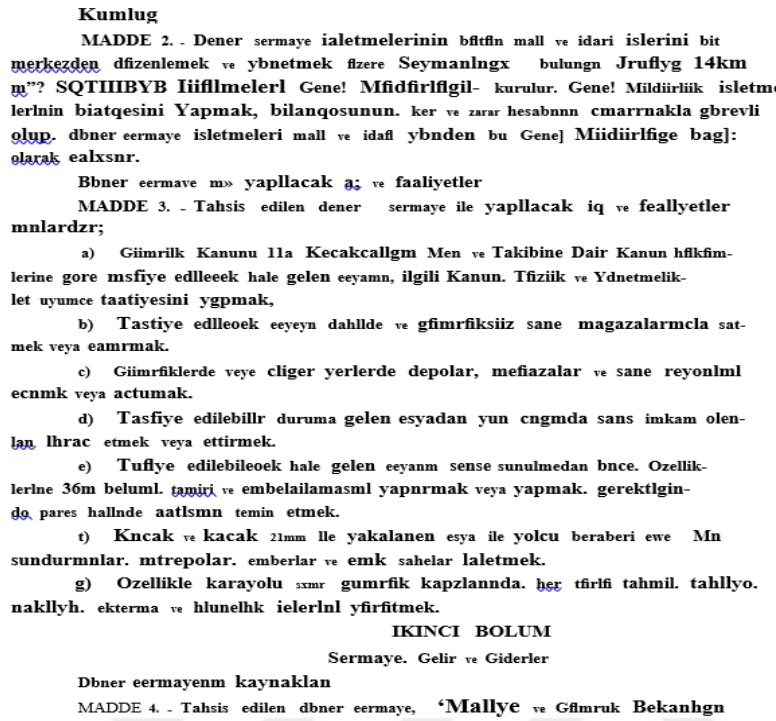
As mentioned above in this thesis study, a subset of Official Gazette and TRT Haber web news archives were selected as data source for gaining experience in big data text analytics, extracting valuable information about both archives and get insight about the similarity between these data sets.

Although the entire archive of Official Gazette was downloaded via a self-written script, because of the very low quality of the scanned documents and the limited ability of freeware OCR converters lead the study to shrink the size of the data set. To get a meaningful insight the archive between July 2015 and July 2017 was assigned sample data including 714 mixed (scanned, image etc.) pdf files each includes more than 1000 lines and nearly 61.551.000 words.

Additionally, in order to retrieve TRT Haber web news archive, again a self written Python code -can be found in Appendix A- was used. This news archive was retrieved from https://www.trthaber.com/haber/gundem/ all links between pages 353 and 895 corresponding to the selected time interval extracted firstly, then all paragraphs attached to these links stored to disk and 542 text files consisting of about 3.794.000 words were generated.

After that, in order to prepare Official Gazette data set for textual analysis, Renee PDF Aide (a freeware software) was used for OCR and conversion to text. The results of conversion were considerable noisy and low quality because of the quality of scanned files. Sample conversion results with Ottoman letter, early gazettes with low quality and recent gazette can be seen below.

Figure 5.3: Sample conversion result of gazette with Ottoman letters



Kumlug

MADDE 2. - Dener sermaye ialetmelerinin bfitfln mall ve idari islerini bit merkezden dfizenlemek ve ybnetmek fizere Seymanlngx bulungn Jruflyg 14km m"? SQTIIIBYB Iiifllmelerl Gene! Mfidfirlflgil- kurulur. Gene! Mildiirliik isletmelerlnin biatqesini Yapmak, bilanqosunun. ker ve zarar hesabnnn cmarrnakla gbrevli olup. dbner eermaye isletmeleri mall ve idafl ybnden bu Gene] Miidiirlfige bag]: olarak ealxsnr.

Bbner eermave m» yapllacak a; ve faaliyetler

MADDE 3. - Tahsis edilen dener sermaye ile yapllacak iq ve feallyetler mnlardzr;

a) Giimrilk Kanunu 11a Kecakcallgm Men ve Takibine Dair Kanun hflkfimlerine gore msfiye edlleeek hale gelen eeyamn, ilgili Kanun. Tfiziik ve Ydnetmeliklet uyumce taatiyesini ygpmak,

b) Tastiye edlleoek eeyeyn dahllde ve gfimrfiksiiz sane magazalarmcla satmek veya eamrmak.

c) Giimrfiklerde veye cliger yerlerde depolar, mefiazalar ve sane reyonlml ecnmk veya actumak.

d) Tasfiye edilebillr duruma gelen esyadan yun cngmda sans imkam olenlan lhrac etmek veya ettirmek.

e) Tuflye edilebileoek hale gelen eeyanm sense sunulmedan bnce. Ozelliklerine 36m beluml. tamiri ve embelailamasml yapnrmak veya yapmak. gerektlgindo pares hallnde aatlsmn temin etmek.

t) Kncak ve kacak 21mm lle yakalanen esya ile yolcu beraberi ewe Mn sundurmnlar. mtrepolar. emberlar ve emk sahelar laletmek.

g) Ozellikle karayolu sxmr gumrfik kapzlannda. her tfirlfi tahmil. tahllyo. nakllyh. ekterma ve hlunelhk ielerlnl yfirfitmek.

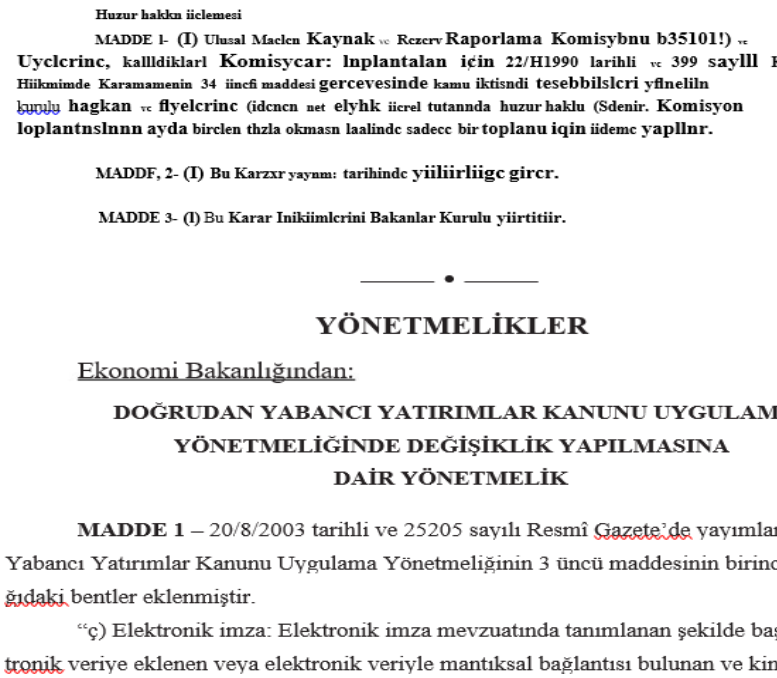
IKINCI BOLUM

Sermaye. Gelir ve Giderler

Dbner eermayenm kaynaklan

MADDE 4. - Tahsis edilen dbner eermaye, 'Mallye ve Gfimruk Bekanhgn

Figure 5.4: Sample conversion result of early gazettes with low quality



Huzur hakkn iiclemesi

MADDE l- (I) Ulusal Maclcn Kaynak vc Rczcrv Raporlama Komisybnu b35101!) ve Uyclcrinc, kallldiklarl Komisycar: lnplantalan için 22/H1990 larihli vc 399 saylll Kanun Hiikmimde Karamamenin 34 iincfi maddesi gercevesinde kamu iktisndi tesebbilslcri yfineliln kurulu hagkan vc flyelcrinc (idcncn net elyhk iicrel tutannda huzur haklu (Sdenir. Komisyon loplantnslnnn ayda birclen thzla okmasn laalindc sadecc bir toplanu iqin iidemc yapllnr.

MADDF, 2- (I) Bu Karzxr yaynm: tarihindc yiiliirliigc gircr.

MADDE 3- (I) Bu Karar Inikiimlcrini Bakanlar Kurulu yiirtitiir.

YÖNETMELİKLER

Ekonomi Bakanlığından:

DOĞRUDAN YABANCI YATIRIMLAR KANUNU UYGULAMA YÖNETMELİĞİNDE DEĞİŞİKLİK YAPILMASINA DAİR YÖNETMELİK

MADDE 1 – 20/8/2003 tarihli ve 25205 sayılı Resmî Gazete'de yayımlanan Doğrudan Yabancı Yatırımlar Kanunu Uygulama Yönetmeliğinin 3 üncü maddesinin birinci fıkrasına aşağıdaki bentler eklenmiştir.

"ç) Elektronik imza: Elektronik imza mevzuatında tanımlanan şekilde başka bir elektronik veriye eklenen veya elektronik veriyle mantıksal bağlantısı bulunan ve kimlik doğrulama amacıyla kullanılan elektronik veriyi,

Figure 5.5: Sample conversion result of recent a gazette

After that, via Windows PowerShell scripts all text files' encoding was updated to UTF-8 with the aim of handling Turkish specific characters and the data sets of both corpora merged into files that have the data last six months of 2015 (from July), whole of 2016 and first seven months of 2017 to examine by year, and corpora also merged into monthly separated files and generated 25 files, belonging each month between July 2015 and July 2017.

After data preparation, development environment was settled. Because of the listed advantages Apache Spark is the main platform.

Anaconda3 with Python3.6 and Jupyter-notebook was used as IDE.

Self-written python code was used for cleaning the data set from punctuation characters and numbers along with pruning the globally defined Turkish stop words. By the way, the stop words of a language mean mostly used words but they have no meaning for the big picture. I used the list that can be found at https://github.com/ahmetax/trstop uploaded by a Turk user, and some other meaningless words were added to the list since they were wrongly converted by PDF converter in order to prune them from the main texts. The list of stop words can be found in Appendix C. And a word frequency application was coded in jupyter-notebook using pyspark library. Used python source codes and Windows shell scripts are attached to Appendix A.

The organization of the resulting excel files can be explained as; each sheet stores all year's data (2015, 2016, and 2017). Every part in a column lists the mostly used word in descending order, so every word/term get a weight. Then, word clouds were generated using the most frequent words of the year2016 for both data sets via an online word cloud site. In addition, in order to have some insights about the similarity of these data sets, Tf-IDF vectorization and then cosine similarity algorithms were applied to datasets. While calculating the cosine similarity, the texts files of each year divided into 4 (quarter by quarter) from two data sets given as parameters, and a sample mostly used words list from each text was used as training string in order, and the similarity value of cross document (for example if training string is from doc1 then the similarity value of doc2 is focused for getting insight about how similar them.)

Moreover, as a deeper analysis, again cosine similarity algorithm run on the monthly separated datasets. This time, all archive from both corpora given as parameters to the cosine similarity compassion method and every month's data was compared to every month from other corpus. In other words, txt document consisting of the archive of June 2016 from Official Gazette was compared to every month's document from TRT News.

5.5 Results

After implementation the results imported to Excel in descending order. Since the lists are too long, the first parts of the lists attached here in order to enhance readability. The whole results can be found in the CD attached to this thesis. In the excel file, each column stores a period of the selected year. In each cell the word and its frequency is stored. For example; the following figure tells us he word "lisans" is used 21698 times in the period between January and July in 2017.

1	2015	2016	2017
2			
3	('sayı', 47552)	('sayılı', 86287)	('sayılı', 50888)
4	('sayılı', 45554)	('sayı', 85199)	('sayı', 49086)
5	('the', 32884)	('the', 81610)	('için', 32078)
6	('ihale', 29936)	('ihale', 59918)	('ihale', 25922)
7	('için', 27650)	('yer', 46253)	('yer', 25388)
8	('yer', 26502)	('için', 43056)	('halinde', 24496)
9	('tdk', 25764)	('halinde', 40970)	('tarafından', 24048)
10	('halinde', 23122)	('genel', 35750)	('kurulu', 22516)
11	('teklif', 21424)	('ticaret', 34837)	('lisans', 21698)
12	('genel', 21136)	('teklif', 34337)	('the', 21294)
13	('alan', 19180)	('alan', 33896)	('göre', 20652)
14	('konusu', 18856)	('konusu', 33318)	('genel', 20236)
15	('uygun', 18334)	('göre', 31950)	('kayıtlı', 19350)

Figure 5.3: Word frequency matrix sample from Official Gazette

2	2015	2016	2017
3	('terör', 3690)	('terör', 9105)	('terör', 4225)
4	('davutoğlu', 3408)	('türkiye', 5713)	('türkiye', 3603)
5	('genel', 3292)	('olduğunu', 4429)	('erdoğan', 3068)
6	('parti', 2981)	('türkiyenin', 4387)	('olduğunu', 3054)
7	('türkiye', 2949)	('devam', 4381)	('türkiyenin', 2955)
8	('türkiyenin', 2882)	('erdoğan', 4318)	('devam', 2918)
9	('başbakan', 2732)	('genel', 4250)	('bakanı', 2907)
10	('erdoğan', 2486)	('bakanı', 4200)	('cumhurbaşkanı', 2868)
11	('olduğunu', 2444)	('darbe', 4088)	('genel', 2584)
12	('devam', 2351)	('cumhurbaşkanı', 3919)	('yıldırım', 2392)
13	('başkanı', 2325)	('ilişkin', 3877)	('büyük', 2330)
14	('dedi', 2225)	('başkanı', 3629)	('başkanı', 2299)
15	('cumhurbaşkanı', 2115)	('türk', 3568)	('dedi', 2268)

Figure 5.4: Word frequency matrix sample from TRT Haber web news

Here, if the resulting tables analyzed deeply, it can be found that the most common words have similar weights and they do not give more meaningful information about the agenda of its time for Official Gazette, but for TRT Haber archive there are some words they affect the agenda appears at the top levels in the frequency list, for instance the word “terör” is an agenda term for Turkey. It was concluded that to get an insight about the agenda of the Official Gazette the words with smaller weights should be focused while for TRT Haber archives the words with higher weights give insights about the agenda.

In Appendix B, lists of mostly used 150 words and their frequencies of both data sets can be found. Furthermore, following word/tag clouds gives us some clue about the similarity of the datasets. As it can be seen the dictionaries of the datasets significantly different from each other. This result is supported by the results of the cosine similarity algorithm between Official Gazette and TRT Haber web news archives, since the comparison results of eight textual data from Official Gazette and eight from TRT Haber web news belonging to eight quarter of the selected time interval, differ between 0,008 and 0,03. By the way, if the training string for cosine similarity is selected from the words with higher frequency the similarity value is found higher than the similarity value of the words with lower weights, as expected.

In addition, according to the resulting matrix of comparison of monthly divided data attached in Appendix C, it can be said that;

The similarity of Official Gazette to the TRT News archive belonging to the previous months is higher than the similarity to the latter months. The weight of the similarity to the previous months is about 65% while the weight of the similarity to latter months is about 35% according to the matrix. This information can be read as the agenda of the country covered in media triggers the legislation and leads new legal results. However, after July 2016 the similarity index to latter months is raising relatively. This means, Official Gazette can also affect the agenda of country and reflected on media.

Furthermore, the similarity index between the same month's data is considerably low, and from this information, it can be concluded that the reciprocal influence takes some time. By the way, highest similarity indexes can be seen up to one year before.

Figure 5.5: Word cloud of Official Gazette of the year 2016 generated via https://www.wordclouds.com

Figure 5.6: Word cloud of TRT Haber web news of the year 2016 generated via https://www.wordclouds.com

CHAPTER 6

RESULTS

6.1 Conclusion

When the topic of this thesis study selected, concept of big data was relatively a new research area, Apache Hadoop was the leading platform in limited big data world, and the reference sources were limited and finding data source to analyze is a difficult part of such studies.

The Official Gazette archive and for comparison TRT Haber web news archive were selected as data source since they are both open to the public, was not studied on before and in unstructured text format, so could not handle in traditional RDBMS. However, there were some problems faced during the preparation phase of the datasets like the quality of the scanned documents, Turkish character problems and the limited free OCR converter's abilities etc. Though these problems, applying text cleansing techniques and pruning unrelated elements like throwing stop words, punctuation characters the quality of the final data source improved considerably.

Although the first selected big data frameworks for this study were Hadoop and Mahout and a sample implementation environment was settled on Linux (Ubuntu) OS with one Hadoop cluster and Mahout, that was too hard to maintain the implementation environment stable, it also required too many configurations. Later on it was tried to settle up an implementation environment on Windows OS, it was again so difficult to settle up and maintain the system, and too many configurations needed. The incompatibility between distributions of frameworks and need of different configurations for every distribution makes the study extremely complex and difficult. After all these undesirable experience and consumed time, implementation environment is replaced with the newly developed technology Apache Spark, because of its advantages like speed, support for various languages and libraries along with the ease of installation. Spark is again works on Hadoop file system HDFS. In order to use it on Window OS, only the installation of winutils executable file and adding it to environment variable is needed. It brings and enables the use of HDFS on Windows.

Moreover, during implementation one of the most useful tools were Anaconda and Jupyter Notebook since it was more easy to setup and supports Apache Spark and too many other tools by only downloading the related package via its package manager named “conda”. There was no need for extra configurations and confusion.

The most time consuming phase of this study was preparation the data source retrieved from Official Gazette for application. Applying OCR to raw source took considerably long time. But this study is a first and has a role of being first on analysis of Official Gazette and TRT Haber archives can be considered an introduction to whom wants to learn about big data concept, its related technologies, advantages, text analysis, Apache Spark, Anaconda, Jupyter-Notebook, Tf-IDF and cosine similarity terms.

At the beginning, the target of this study was analyzing the entire archive from 1923 to today however, working with scanned documents (most of them includes images, ottoman letter, maps etc.) made study harder. Because of this reason, the data set was shrink in size and time interval. And it was seen that the resulting corpus of Official Gazette of every period resembles each other, while corpus of TRT Haber archive differs relatively from year to year. Additionally, the most important output of this thesis, is extraction of the similarity indexes between datasets and according to them the result of the agenda of the country affects the legislation represented by Official Gazette, more than the effect of Official Gazette on agenda of the country represented by TRT News website is reached.

To sum up, this study is the first attempt for analyzing Official Gazette and TRT Haber archives in terms of their textual and similarity features using big data techniques. So, results of the study give new information about the lexical characteristic and similarity indexes of both databases along with the information about the correlation between them, namely between agenda of the country and the legislative organ of it.

6.2 Future Work

As mentioned above, Apache Spark supports machine learning library MlLib that enables to development recommendation applications. As a future work, a Turkish regulations recommendation engine can be developed for whom related with laws using Apache Spark with MlLib using the data source can be reached from

http://www.mevzuat.gov.tr/Kanunlar.aspx, not directly official gazette archives, since mevzuat.gov.tr stores the latest and current prevail versions of the laws while resmigazete.gov.tr stores all forms of the achieve whether prevail or not.

Different form Official Gazette Mevzuat.gov.tr serves laws in high quality documents in PDF and MS Word (.doc) formats. Therefore, working with the final regulations and good quality document will make the job easier. Anaconda with Jupyter-Notebook will be the best development environments again for such a study because of their various advantages. By the way, today there are several commercial tools used by whom works with laws, but the proposal of this thesis may differ from other in terms of its big data technology at background.

APPENDICIES

APPENDIX A

Source Codes

Sample Java Code used for downloading Official Gazette archive;

```
package pdfdownloader;

import java.io.File;
import java.io.IOException;
import java.net.URL;
import org.apache.commons.io.FileUtils;

/**
 * @author YBU
 */
public class PdfDownloaderApache {

    public static void main(String[] args) throws IOException {

        for (int t = 1; t <2; t++) {
            String urlQueue = t + "";

            String uri = "http://www.resmigazete.gov.tr/arsiv/" + urlQueue + ".pdf";
            //String uri = "http://www.jpl.nasa.gov/about_JPL/jpl101.pdf";

            URL url = new URL(uri);

            try {
                // Contacting the URL
                System.out.print("Connecting to " + url.toString() + " ... \n");
                url = new URL(uri);
                String destAdd = "C:\\Users\\YBU\\Desktop\\resmigazeteDownloaded\\"
+urlQueue;
                File destination = new File(destAdd+".pdf");
                FileUtils.copyURLToFile(url, destination);
                System.out.print("Saved as " + urlQueue + " ... \n");
            } catch (Exception e) {
                System.out.println("FAILED.\n[" + e.getMessage() + "]");
            }
        }

    }
}
```

Python Code used for retrieving TRT Haber web news archive;

```
import requests
from bs4 import BeautifulSoup


def getArticle(url):
  # print("--Basla--")
   #url = "https://www.trthaber.com/haber/gundem/meclis-baskani-bugun-belli-olacak-192256.html"
   result = requests.get(url)
   print(url)
   c = result.content
   soup = BeautifulSoup(c)

   article_text = ''
   try:
      article = soup.find("div", {"class":"editorPart blackle"}).findAll('p')
      for element in article:

            article_text += '\n' + ''.join(element.findAll(text = True))
   except:
         print('An error occured.')

   return article_text


urlStart=353 #temmuz 2017 gundem linkleri baslangıc sayfa numarası 353
urlEnd=895 #temmuz 2015 gundem linkleri bitis sayfa numarası 895
print("yazma BASLADI")
#file = open("C:\\Users\\yasemin.can\\Desktop\\trtHaberIcerik.txt","w", encoding="utf-8")
while urlStart < urlEnd:
```

```
    file =
open("C:\\Users\\yasemin.can\\Desktop\\trtHaberArsiv\\trtHaberIcerik"+str(urlStart)
+".txt","w", encoding="utf-8")
    url = "https://www.trthaber.com/haber/gundem/"+str(urlStart)+".sayfa.html"
    r = requests.get(url)
    soup = BeautifulSoup(r.content)
    data = soup.findAll('ul',attrs={'class':'katListe2'})


    for ul in data:
        data2 = soup.findAll('h2')
        for h2 in data2:
            links = h2.findAll('a')
            for a in links:
                guncelLink = "https://www.trthaber.com/" + a.get("href")
                file.write(getArticle(guncelLink))
                #file.write("\n")
    file.close()
    urlStart = urlStart+1
print("TRT ARSIV BITTI")
```

Python Code used for document cleansing and calculating word frequency;

```
import pyspark
import re
import string

def removePunctuation(text):
    translator=text.maketrans("",string.punctuation)
    text=text.translate(translator)
    return re.sub('[^a-zA-ZçÇğĞıİöÖşŞüÜ]',"",text.lower().strip(r'\\'))

if not 'sc' in globals():
    sc = pyspark.SparkContext()
text_file =
sc.textFile("C:\\Users\\yasemin.can\\Desktop\\resmigazete\\tez16102018\\resmigazet
earsiv\\201507a.txt")
stopwords =
open('C:\\Users\\yasemin.can\\Desktop\\resmigazete\\tez16102018\\kaynak
liste\\turkceStopWords.txt').read().split()
markeredtext=[]
for x in stopwords:
    counts=text_file.flatMap(lambda line: line.split(" "))\
    .map(lambda line: removePunctuation(line))\
    .map(lambda word: (word, 1))\
    .reduceByKey(lambda a, b: a+b)
c = counts.sortBy(lambda a: -a[1]) #minus sign is for descending ordering
for t in c.collect():
    temp = t[0].strip('\n\r')
    if temp not in stopwords and len(temp) > 1:
        markeredtext.append(t)

for x in markeredtext:
    print (x)
```

Sample Python codes used for applying TF vectorization and calculating cosine similarity [67]:

```
import os
import glob
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import nltk
import numpy as np

def process(file):
    raw = open(file, encoding="utf-8").read()
    tokens = word_tokenize(raw)
    words = [w.lower() for w in tokens]

    porter = nltk.PorterStemmer()
    stemmed_tokens = [porter.stem(t) for t in words]

    stop_words = set(stopwords.words('turkish'))
    filtered_tokens = [w for w in stemmed_tokens if not w in stop_words]

    count = nltk.defaultdict(int)
    for word in filtered_tokens:
        count[word] +=1
    return count;

def cos_sim(a,b):
    dot_product = np.dot(a,b)
    norm_a = np.linalg.norm(a)
    norm_b = np.linalg.norm(b)
    return dot_product / (norm_a*norm_b)
```

```
def getSimilarity(dict1,dict2):
    all_words_list =[]
    for key in dict1:
        all_words_list.append(key)
    for key in dict2:
        all_words_list.append(key)
    all_words_list_size=len(all_words_list)

    v1 = np.zeros(all_words_list_size, dtype=np.int)
    v2 = np.zeros(all_words_list_size, dtype=np.int)

    i= 0
    for (key) in all_words_list:
      v1[i] = dict1.get(key,0)
       v2[i] = dict2.get(key,0)
       i = i+1
    return cos_sim(v1,v2);

#if __name__ == '__main__':
dict1 = process("F:\\AAResmiGazeteAylık\\rg201512.txt")

trtRoot = "F:\\AAtrtHaberAylık\\"
rgRoot = "F:\\AAResmiGazeteAylık\\1\\"

for root, dirs, files in os.walk(rgRoot):
    for file in files:
        if file.endswith(".txt"):
            #print(os.path.join(root, file))
            print(file)
            dict1 = process(os.path.join(root, file))
```

```
        for root1, dirs1, files1 in os.walk(trtRoot):
            for file1 in files1:
                if file1.endswith(".txt"):

                    dict2 = process(os.path.join(root1, file1))
                    print(getSimilarity(dict1,dict2))

                else:
                    print("arsiv dosyası degil")
                    print("**")
print("---------------bitti")
```

Another approach for calculating cosine similarity;

```
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.feature_extraction.text import TfidfVectorizer
from nltk.corpus import stopwords

# Bring in standard stopwords

stopWords  = set('ve ile diger ancak'.split()) #stopwords.words('turkish')

print ("\nCalculating document similarity scores...")


f = open('F:\\trtHaberArsiv\\trthaberArsivByquarter\\trt20163rdQuarter.txt',
encoding='utf8')
doc2 = str(f.read())

f = open('F:\\resmiGazeteArsivByquarter\\20163rdQuarter.txt', encoding='utf8')
doc3 = str(f.read())

train_string = 'ihale genel ticaret bakan yasaklama sicil lisans karar kamu'

train_set = [train_string, doc2, doc3]

# Set up the vectoriser, passing in the stop words
tfidf_vectorizer = TfidfVectorizer(stop_words=stopWords)

# Apply the vectoriser to the training set
tfidf_matrix_train = tfidf_vectorizer.fit_transform(train_set)

# Print the score
```

```
print ("\nSimilarity Score [*] ",cosine_similarity(tfidf_matrix_train[0:1],
tfidf_matrix_train))
```

Output:

```
Calculating document similarity scores...

Similarity Score [*] [[1. 0.03432468 0.1017993 ]]
```

Sample Scripts for merging files into periods and converting their encoding to UTF without BOM for eliminating Turkish character problem are:

Get-ChildItem C:\Users\yasemin.can\desktop\resmigazete\tez16102018\resmigazetearsiv -include 201509*.txt -rec | ForEach-Object {gc $_; ""} | Add-Content C:\Users\yasemin.can\desktop\resmigazete\tez16102018\resmigazetearsiv\quarter\20153rdQuarter3.txt

Get-ChildItem C:\Users\yasemin.can\Desktop\trtHaberArsiv\temp -include trt*.txt -rec | ForEach-Object {gc $_; ""} |

Add-Content C:\Users\yasemin.can\Desktop\trtHaberArsiv\quarter\trt20153rdQuarter.txt

```
$files = [IO.Directory]::GetFiles("F:\TEZLE İLGİLİ\renee deneme conversion\merge")
foreach($file in $files)
{
$content = Get-Content -path $file
$Utf8NoBomEncoding = New-Object System.Text.UTF8Encoding $False
[System.IO.File]::WriteAllLines($file, $content, $Utf8NoBomEncoding)
}
```

APPENDIX B

Term Frequency Lists of Top 150 Words from Official Gazette

(Whole list can be found in CD attached to this thesis)

Official Gazette 2015	Official Gazette 2016	Official Gazette 2017
('sayı', 47552)	('sayılı', 86287)	('sayılı', 50888)
('sayılı', 45554)	('sayı', 85199)	('sayı', 49086)
('the', 32884)	('the', 81610)	('için', 32078)
('ihale', 29936)	('ihale', 59918)	('ihale', 25922)
('için', 27650)	('yer', 46253)	('yer', 25388)
('yer', 26502)	('için', 43056)	('halinde', 24496)
('tdk', 25764)	('halinde', 40970)	('tarafından', 24048)
('halinde', 23122)	('genel', 35750)	('kurulu', 22516)
('teklif', 21424)	('ticaret', 34837)	('lisans', 21698)
('genel', 21136)	('teklif', 34337)	('the', 21294)
('alan', 19180)	('alan', 33896)	('göre', 20652)
('konusu', 18856)	('konusu', 33318)	('genel', 20236)
('uygun', 18334)	('göre', 31950)	('kayıtlı', 19350)
('tarafından', 17836)	('kayıtlı', 31756)	('teklif', 18720)
('göre', 17652)	('tdk', 31508)	('alan', 18570)
('ticaret', 17578)	('and', 31506)	('olması', 18496)
('edilen', 17112)	('bakanı', 30775)	('ticaret', 18474)
('kurulu', 17022)	('tarafından', 30357)	('hale', 18118)
('kayıtlı', 16902)	('olması', 30013)	('uygun', 17236)
('olması', 16756)	('kurulu', 29940)	('kabul', 17046)
('kabul', 15908)	('uygun', 29777)	('doldurulacaktır', 17018)
('doldurulacaktır', 14926)	('edilen', 29120)	('bakanı', 16998)
('and', 14570)	('doldurulacaktır', 28523)	('edilen', 16196)
('belirtilen', 13928)	('kabul', 25999)	('tez', 15740)
('hale', 13424)	('nci', 24711)	('yönetim', 15692)
('bakanı', 13212)	('ilişkin', 24217)	('yüksek', 15312)
('tarihi', 12686)	('veveya', 23866)	('tarihi', 14700)
('veveya', 12668)	('belirtilen', 23320)	('konusu', 14556)
('kamu', 12620)	('tarihi', 23026)	('kararı', 14392)
('yapılan', 12386)	('yasaklama', 21839)	('nci', 14074)
('kararı', 12228)	('ortak', 21741)	('veveya', 13842)
('ilişkin', 12208)	('kararı', 21683)	('üzere', 13806)
('yasaklama', 11920)	('yapılan', 21255)	('belirtilen', 13208)

('diğer', 11898)	('lisans', 21070)	('öğretim', 12940)
('nci', 11802)	('bulunan', 20722)	('bulunan', 12916)
('üzere', 11744)	('kamu', 20597)	('yapılan', 12908)
('ortak', 11724)	('gün', 20512)	('diğer', 12802)
('yönetim', 11334)	('verilen', 20243)	('ortak', 12786)
('olduğu', 11112)	('sicil', 20219)	('sicil', 12700)
('verilen', 10724)	('üzere', 20127)	('öğrenci', 12604)
('müdürlüğü', 10720)	('olduğu', 19879)	('ilişkin', 12564)
('lisans', 10368)	('müdürlüğü', 19612)	('aynı', 12448)
('kimlik', 10324)	('kültür', 18969)	('kültür', 12288)
('bulunan', 10322)	('koruma', 18877)	('yapı', 11816)
('durumunda', 10256)	('yönetim', 18766)	('verilen', 11656)
('ilikin', 10236)	('durumunda', 18516)	('durumunda', 11544)
('kasım', 10176)	('diğer', 18512)	('enstitü', 11448)
('maddesinde', 10052)	('kimlik', 18040)	('olduğu', 11334)
('iqin', 10024)	('tüzel', 17632)	('içinde', 11316)
('dahil', 9888)	('dahil', 17097)	('müdürlüğü', 11284)
('belgeler', 9798)	('idari', 16911)	('yasaklama', 11072)
('tüzel', 9758)	('ait', 16622)	('süresi', 10950)
('gün', 9614)	('sanayi', 16157)	('gün', 10764)
('koruma', 9504)	('birinci', 15948)	('denetim', 10670)
('sicil', 9496)	('tebliğ', 15917)	('koruma', 10482)
('esas', 9388)	('kurum', 15890)	('ders', 10472)
('dikkate', 9376)	('ihaleye', 15806)	('kanun', 10352)
('ait', 9208)	('belgeler', 15594)	('kayıt', 10308)
('içinde', 9152)	('inci', 15388)	('kamu', 10218)
('öğretim', 9078)	('aynı', 15369)	('fazla', 10106)
('birinci', 9050)	('kanun', 15013)	('ilikin', 10104)
('teknik', 9024)	('ilan', 15011)	('birinci', 9998)
('ceza', 9002)	('içinde', 14993)	('eğitim', 9994)
('aynı', 8870)	('maddesinde', 14989)	('dahil', 9724)
('gerekli', 8822)	('fazla', 14945)	('yeterlik', 9650)
('son', 8728)	('süresi', 14821)	('uygulama', 9442)
('ağustos', 8726)	('tarihinden', 14821)	('iki', 9394)
('iki', 8666)	('dikkate', 14711)	('tarihinden', 9284)
('sınav', 8538)	('gerekli', 14686)	('içerisinde', 9240)
('kültür', 8520)	('teknik', 14620)	('inci', 9214)
('fazla', 8496)	('içerisinde', 14612)	('tüzel', 9178)
('sanayi', 8462)	('hale', 14569)	('ait', 8998)
('aralık', 8454)	('tespit', 14488)	('ilan', 8884)
('tespit', 8452)	('son', 14444)	('hakkında', 8838)

('bilgi', 8180)	('esas', 14423)	('kurum', 8832)
('ihaleye', 8134)	('iqin', 14391)	('yıl', 8828)
('kurum', 8082)	('hakkında', 14375)	('yönetmelik', 8712)
('ilan', 7994)	('iki', 14353)	('teknik', 8690)
('yazılı', 7986)	('ilikin', 14315)	('kimlik', 8602)
('üniversitesi', 7894)	('ceza', 14301)	('yeni', 8584)
('geçici', 7874)	('kanununun', 14157)	('sınav', 8534)
('yüksek', 7814)	('öğretim', 14022)	('bölüm', 8484)
('süresi', 7802)	('odası', 14002)	('belgeler', 8466)
('anayasa', 7754)	('geçici', 13829)	('sit', 8422)
('kayıt', 7704)	('yüksek', 13662)	('tebliğ', 8416)
('içerisinde', 7658)	('gerçek', 13648)	('ikinci', 8266)
('eğitim', 7646)	('adresi', 13619)	('gerekli', 8206)
('yıl', 7624)	('esnaf', 13335)	('idari', 8172)
('yönetmelik', 7606)	('bölge', 13179)	('öğrencinin', 8136)
('günü', 7602)	('bilgi', 13153)	('doktora', 8124)
('tarihinden', 7586)	('kayıt', 13122)	('üniversitesi', 8104)
('ikinci', 7528)	('yıl', 12908)	('dikkate', 8064)
('sit', 7480)	('mevzuat', 12602)	('and', 8044)
('kanun', 7462)	('tltk', 12553)	('gerçek', 8014)
('tebliğ', 7428)	('for', 12532)	('son', 8014)
('inci', 7424)	('sit', 12491)	('mayıs', 7970)
('odası', 7392)	('ankara', 12479)	('yazılı', 7896)
('birlikte', 7384)	('tez', 12346)	('odası', 7846)
('hakkında', 7340)	('hizmet', 12284)	('geçici', 7822)
('gerekir', 7308)	('yazılı', 12280)	('yerine', 7796)
('kanununun', 7274)	('kapsamında', 12257)	('sanayi', 7746)
('cumhuriyet', 7260)	('bölüm', 12228)	('süre', 7718)
('merkez', 7228)	('birlikte', 12219)	('esnaf', 7614)
('bölüm', 7218)	('günü', 12206)	('bölge', 7590)
('adresi', 7210)	('belge', 12202)	('maddesinde', 7544)
('gerçek', 7166)	('yerine', 11944)	('belirlenen', 7504)
('bölge', 7112)	('ikinci', 11906)	('kurulunun', 7406)
('öğrenci', 7038)	('yönetmelik', 11821)	('kii', 7402)
('saat', 6986)	('eğitim', 11808)	('kapsamında', 7378)
('nolu', 6978)	('saat', 11803)	('iii', 7312)
('adet', 6936)	('aşağıdaki', 11679)	('nolu', 7276)
('ders', 6936)	('adet', 11677)	('bilgi', 7226)
('esnaf', 6910)	('kişi', 11626)	('yönetmeliğin', 7104)
('ilk', 6900)	('üniversitesi', 11566)	('adresi', 6956)
('idari', 6884)	('nolu', 11539)	('birlikte', 6882)

('belirlenen', 6874)	('belirlenen', 11402)	('mevzuat', 6878)
('belge', 6830)	('yasaklamalarda', 11400)	('günü', 6866)
('mevzuat', 6720)	('yapı', 11372)	('merkezin', 6858)
('yerine', 6692)	('shall', 11349)	('yasaklamalarda', 6802)
('ankara', 6648)	('öğrenci', 11336)	('esas', 6800)
('uygulama', 6636)	('uygulama', 11304)	('kapsamındakiler', 6756)
('kapsamında', 6596)	('belgesi', 11233)	('belgesi', 6742)
('kii', 6570)	('vergi', 11113)	('aşağıdaki', 6718)
('vergi', 6478)	('kapsamındakiler', 11032)	('şekilde', 6708)
('nedeniyle', 6400)	('ticaretesnaf', 10735)	('sahip', 6686)
('yönetmeliğin', 6364)	('sahip', 10591)	('kanununun', 6676)
('hak', 6362)	('teminat', 10502)	('lisansüstü', 6674)
('derece', 6288)	('posta', 10395)	('ankara', 6658)
('hizmet', 6286)	('sosyal', 10380)	('tespit', 6632)
('numarası', 6266)	('yeni', 10359)	('teslim', 6614)
('aşağıdaki', 6252)	('numarası', 10359)	('ihaleye', 6592)
('şekilde', 6198)	('ilk', 10284)	('bakanlığı', 6576)
('belgesi', 6162)	('ders', 10241)	('hizmet', 6570)
('igin', 6154)	('tüm', 10219)	('bilimsel', 6568)
('tüm', 6108)	('denetim', 10172)	('belge', 6566)
('sahip', 6106)	('karan', 10087)	('yabancı', 6514)
('ihlal', 6056)	('şekilde', 10046)	('edilir', 6410)
('teminat', 6002)	('merkez', 9997)	('özel', 6278)
('tez', 5990)	('uyarınca', 9925)	('proje', 6262)
('yasaklamalarda', 5966)	('ili', 9912)	('merkez', 6250)
('devam', 5940)	('satın', 9898)	('teminat', 6218)
('bakanlığı', 5866)	('sınav', 9883)	('yarıyıl', 6140)
('posta', 5792)	('mah', 9860)	('saat', 6120)
('merkezin', 5746)	('başbakan', 9860)	('uyarınca', 6102)
('mali', 5732)	('kasım', 9808)	('alınarak', 6078)
('karan', 5724)	('kurulunun', 9786)	('dalı', 5888)
('kapsamındakiler', 5698)	('imza', 9785)	('ada', 5880)
('satın', 5660)	('yönetmeliğin', 9765)	('öğrenciler', 5784)

Term Frequency Lists of Top 150 Words from TRT Haber Archive

(Whole list can be found in CD attached to this thesis)

TRT Haber Web News	TRT Haber Web News	TRT Haber Web News
2015	**2016**	**2017**
('terör', 3690)	('terör', 9105)	('terör', 4225)
('davutoğlu', 3408)	('türkiye', 5713)	('türkiye', 3603)
('genel', 3292)	('olduğunu', 4429)	('erdoğan', 3068)
('parti', 2981)	('türkiyenin', 4387)	('olduğunu', 3054)
('türkiye', 2949)	('devam', 4381)	('türkiyenin', 2955)
('türkiyenin', 2882)	('erdoğan', 4318)	('devam', 2918)
('başbakan', 2732)	('genel', 4250)	('bakanı', 2907)
('erdoğan', 2486)	('bakanı', 4200)	('cumhurbaşkanı', 2868)
('olduğunu', 2444)	('darbe', 4088)	('genel', 2584)
('devam', 2351)	('cumhurbaşkanı', 3919)	('yıldırım', 2392)
('başkanı', 2325)	('ilişkin', 3877)	('büyük', 2330)
('dedi', 2225)	('başkanı', 3629)	('başkanı', 2299)
('cumhurbaşkanı', 2115)	('türk', 3568)	('dedi', 2268)
('büyük', 1921)	('başbakan', 3481)	('ilişkin', 2162)
('seçim', 1900)	('dedi', 3433)	('türk', 2141)
('ifade', 1871)	('yıldırım', 3409)	('darbe', 2118)
('konuştu', 1804)	('büyük', 3200)	('olduğu', 1956)
('bütün', 1757)	('karşı', 3139)	('başbakan', 1945)
('bakanı', 1753)	('ifade', 3090)	('yer', 1930)
('ilişkin', 1645)	('şekilde', 3008)	('ifade', 1893)
('şekilde', 1627)	('örgütü', 2989)	('konuştu', 1841)
('karşı', 1586)	('konuştu', 2874)	('tarafından', 1829)
('siyasi', 1541)	('olduğu', 2815)	('söyledi', 1752)
('söyledi', 1412)	('yönelik', 2807)	('şekilde', 1716)
('son', 1409)	('tarafından', 2804)	('yeni', 1709)
('şöyle', 1397)	('ardından', 2771)	('yönelik', 1695)
('milletvekili', 1385)	('yer', 2673)	('şöyle', 1684)
('yeni', 1370)	('yapılan', 2519)	('yapılan', 1678)
('önemli', 1357)	('göre', 2494)	('önemli', 1634)
('değil', 1355)	('parti', 2433)	('değil', 1623)
('içinde', 1317)	('şöyle', 2390)	('parti', 1611)
('zaman', 1294)	('söyledi', 2390)	('karşı', 1600)
('ardından', 1270)	('fetö', 2364)	('bütün', 1575)
('olduğu', 1232)	('yeni', 2341)	('son', 1552)

('göre', 1193)	('üzerine', 2335)	('örgütü', 1521)
('ahmet', 1190)	('kapsamında', 2330)	('ardından', 1492)
('güvenlik', 1184)	('güvenlik', 2326)	('üzerine', 1465)
('kaydetti', 1158)	('son', 2308)	('göre', 1459)
('üzerine', 1150)	('önemli', 2287)	('bakan', 1444)
('devlet', 1139)	('bütün', 2268)	('fetö', 1411)
('tüm', 1132)	('değil', 2238)	('milli', 1380)
('yer', 1127)	('mücadele', 2164)	('recep', 1355)
('iki', 1117)	('kurtulmuş', 2114)	('tüm', 1350)
('tarafından', 1115)	('hakkında', 2076)	('zaman', 1346)
('türk', 1113)	('devlet', 2063)	('güvenlik', 1341)
('bugün', 1097)	('iki', 2002)	('yok', 1337)
('yardımcısı', 1058)	('kabul', 1987)	('tayyip', 1291)
('yapılan', 1054)	('zaman', 1967)	('aynı', 1290)
('örgütü', 1046)	('tüm', 1944)	('kabul', 1287)
('chp', 1018)	('yaptığı', 1910)	('iki', 1265)
('yönelik', 1014)	('üzere', 1879)	('kaydetti', 1263)
('mücadele', 1010)	('arasında', 1879)	('kullandı', 1262)
('ilk', 1000)	('milli', 1853)	('hakkında', 1249)
('ortaya', 994)	('ilk', 1812)	('ilk', 1248)
('yaptığı', 988)	('yok', 1811)	('yaptığı', 1246)
('şunları', 979)	('kaydetti', 1803)	('üzere', 1244)
('hiçbir', 979)	('aynı', 1790)	('şehit', 1226)
('türkiyede', 977)	('recep', 1787)	('şunları', 1212)
('kullandı', 958)	('önce', 1786)	('kapsamında', 1197)
('önce', 941)	('içinde', 1768)	('milletin', 1176)
('şehit', 929)	('bazı', 1759)	('bazı', 1164)
('akdoğan', 927)	('bugün', 1720)	('yıl', 1159)
('kabul', 922)	('kullandı', 1702)	('arasında', 1157)
('üzere', 916)	('tayyip', 1676)	('bugün', 1139)
('bazı', 904)	('yardımcısı', 1674)	('içinde', 1136)
('dile', 900)	('edilen', 1673)	('avrupa', 1130)
('birlikte', 898)	('istanbul', 1671)	('evet', 1128)
('yok', 891)	('cumhuriyet', 1666)	('edilen', 1117)
('tayyip', 888)	('ortaya', 1660)	('önce', 1114)
('recep', 886)	('anayasa', 1655)	('dile', 1106)
('aynı', 882)	('şunları', 1646)	('birlikte', 1105)
('partinin', 880)	('alınan', 1632)	('mücadele', 1096)
('bulundu', 863)	('silahlı', 1603)	('böyle', 1092)
('yılmaz', 856)	('bulundu', 1601)	('çavuşoğlu', 1086)
('hükümet', 854)	('diğer', 1591)	('yüzde', 1083)

('kendi', 851)	('hava', 1591)	('bulundu', 1067)
('sayın', 850)	('siyasi', 1589)	('devlet', 1062)
('sadece', 849)	('avrupa', 1580)	('istanbul', 1060)
('arasında', 839)	('edildi', 1570)	('kurtulmuş', 1057)
('konusunda', 832)	('hiçbir', 1544)	('millet', 1057)
('belirten', 832)	('yıl', 1541)	('kendi', 1025)
('yüzde', 830)	('açıklamada', 1537)	('diğer', 1022)
('belirterek', 808)	('çavuşoğlu', 1535)	('halk', 995)
('uluslararası', 796)	('türkiyede', 1526)	('güçlü', 992)
('milli', 791)	('emniyet', 1520)	('ortaya', 990)
('ülke', 789)	('dile', 1496)	('yardımcısı', 978)
('kurtulmuş', 786)	('şehit', 1485)	('olacak', 970)
('böyle', 781)	('milletin', 1480)	('iyi', 962)
('diğer', 777)	('bulunan', 1479)	('hiçbir', 958)
('iyi', 777)	('birlikte', 1476)	('cumhuriyet', 955)
('yıl', 764)	('bakan', 1474)	('gün', 946)
('gelen', 759)	('sadece', 1460)	('düzenlenen', 946)
('nasıl', 757)	('gözaltına', 1456)	('türkiyede', 944)
('başkan', 743)	('içerisinde', 1429)	('özel', 940)
('kasım', 730)	('böyle', 1425)	('soylu', 934)
('hava', 707)	('özel', 1418)	('belirten', 928)
('olacak', 707)	('kendi', 1417)	('ülke', 926)
('daiş', 705)	('ele', 1412)	('anayasa', 916)
('bakan', 700)	('dışişleri', 1407)	('hava', 916)
('verdi', 680)	('söz', 1400)	('sadece', 910)
('lazım', 666)	('eğitim', 1392)	('siyasi', 903)
('içerisinde', 663)	('kişi', 1388)	('gelen', 888)
('araya', 663)	('ülke', 1363)	('eski', 885)
('saat', 648)	('örgütünün', 1363)	('hayır', 880)
('milletin', 646)	('millet', 1350)	('bulunan', 875)
('açıklamada', 643)	('belirten', 1333)	('söz', 875)
('kurulu', 643)	('düzenlenen', 1329)	('ele', 864)
('terörle', 641)	('ankara', 1328)	('artık', 862)
('türkiyeye', 640)	('bulunduğu', 1325)	('eğitim', 861)
('takip', 633)	('konusunda', 1307)	('açıklamada', 852)
('basın', 630)	('gün', 1302)	('hale', 852)
('başka', 629)	('soruşturma', 1301)	('şey', 847)
('çavuşoğlu', 628)	('verdi', 1296)	('iddianamede', 842)
('kılıçdaroğlu', 624)	('gelen', 1289)	('emniyet', 832)
('söz', 617)	('türkiyeye', 1287)	('cumhurbaşkanlığı', 825)
('pkk', 616)	('destek', 1283)	('edildi', 822)

('ülkenin', 615)	('olacak', 1276)	('mehmet', 817)
('güçlü', 612)	('başka', 1271)	('başka', 813)
('alan', 612)	('suriye', 1258)	('belirterek', 807)
('özel', 612)	('uluslararası', 1256)	('ankara', 805)
('belirtti', 610)	('alan', 1234)	('içerisinde', 805)
('dolayısıyla', 606)	('sosyal', 1228)	('ziyaret', 800)
('avrupa', 603)	('fetullahçı', 1225)	('dışişleri', 798)
('günü', 602)	('ceza', 1222)	('kişi', 786)
('ortak', 600)	('iyi', 1212)	('ülkenin', 786)
('şey', 599)	('kararı', 1211)	('silahlı', 784)
('ankara', 598)	('cumhurbaşkanlığı', 1205)	('araya', 783)
('sosyal', 597)	('tbmm', 1196)	('verdi', 782)
('suriye', 594)	('terörle', 1196)	('saat', 781)
('gün', 594)	('konusu', 1185)	('şimdi', 779)
('tbmm', 590)	('bozdağ', 1149)	('ceza', 774)
('mehmet', 589)	('belirterek', 1148)	('sahip', 767)
('milyon', 588)	('ziyaret', 1148)	('bulunduğu', 766)
('istanbul', 584)	('hale', 1142)	('nasıl', 762)
('bulunan', 582)	('fetönün', 1141)	('ifadelerini', 758)
('şimdi', 580)	('belirtti', 1137)	('işaret', 757)
('düzenlenen', 580)	('nasıl', 1104)	('yüksek', 753)
('inşallah', 577)	('askeri', 1090)	('üzerinde', 745)
('ziyaret', 577)	('anda', 1089)	('uluslararası', 745)
('özellikle', 575)	('şey', 1088)	('jandarma', 741)
('yapılacak', 575)	('binali', 1082)	('gelecek', 740)
('koalisyon', 573)	('eski', 1080)	('özellikle', 738)
('dikkati', 571)	('ifadelerini', 1068)	('birliği', 733)
('işaret', 565)	('basın', 1058)	('alan', 732)
('gelişmeler', 559)	('ait', 1055)	('belirtti', 727)
('anda', 558)	('özellikle', 1048)	('dünya', 720)
('millet', 556)	('milyon', 1048)	('hükümet', 717)
('cumhurbaşkanlığı', 556)	('işaret', 1045)	('milyon', 711)

APPENDIX C

Resulting cosine similarity matrix of monthly divided datasets. Here, rows represents TRT News and coloumns represent Official Gazette month by month.

TRT/ Resmi Gazete	rg201507	rg201508	rg201509	rg201510	rg201511	rg201512
Trt201507	-0.12359066636040425	-0.18189240672222584	0.44103557462813825	-0.35424418162243143	-0.16052257365030975	-0.2831948081254188
Trt201508	-0.21265107356674795	-0.28039165033746477	0.36550624361192763	0.3535831504359094	-0.24386944083595416	0.37238733009607544
Trt201509	0.06844460319627037	0.03715963586001209	-0.3230916483812571	-0.09336175956298066	0.02921405995388358	-0.04129966301081409
Trt201510	0.07156255592230736	0.049633880285535045	-0.20964199501822448	-0.04369144829573688	0.04382612673861995	-0.004985024308330955
Trt201511	-0.044911361464352914	-0.09281982857235814	-0.05729230516106597	0.0870908306766251	-0.08193922093847046	0.1276909292071896
Trt201512	0.02587465277317932	-0.011731303955457056	-0.4043797971917435	-0.14821507332761616	-0.009768691702997824	-0.09145615862555787
Trt201601	-0.01844346526602986	-0.06255620236776605	0.43143704622574436	-0.20908668834586	-0.055939847200271356	-0.14868871401691122
Trt201602	0.0554359009200852	0.021222641904803195	-0.3421205893489875	-0.10726872263431635	0.018299604891552024	-0.05411013028237231
Trt201603	0.013342534087731948	-0.026938411294576633	-0.42783363762836213	-0.1648101125889474	-0.0230399561643098	-0.10703219434302942
Trt201604	-0.24085497860041702	-0.3164141051211315	0.43001791310816373	0.4175192563766885	-0.27960547711364003	-0.43140314156488097
Trt201605	-0.15208725505051446	-0.2142892717625895	0.4371599289841931	-0.3941517339853031	-0.1905529897606086	-0.32063629813716565
Trt201606	-0.15849394902119213	-0.2218205341774141	0.43943842538053474	-0.40141135666988326	-0.19593400763206145	-0.32709519921197006
Trt201607	0.034408478650233436	-0.0020978578612180277	-0.3908440716649536	-0.1379113611644347	-0.0013702046633765929	-0.08103316462277947
Trt201608	0.0051408757239576825	-0.03610034347731081	-0.005280730411829376	0.12341405194351186	-0.03137594975317367	-0.11864788891463683
Trt201609	-0.08367490746075368	-0.13730760923106083	-0.09364751704055013	0.06357767610546611	-0.12043751087693788	0.10703065225869225
Trt201610	-0.010826574256393105	-0.05435081097630992	-0.01935179581499499	0.11439375016839338	-0.04647559503519061	-0.13792648097039553
Trt201611	-0.09891724142758686	-0.15434410940482804	-0.10802138251775546	0.05434580479277448	-0.13536359472165682	0.09885736208193613
Trt201612	-0.12479109570406167	0.1414921874003252	-0.13646927336362025	0.03414728241025433	0.12474177826017448	0.0792452113003995
Trt201701	-0.05132272678646655	-0.09987548481212347	-0.06027500128438082	0.08657955486127827	-0.08749648258104582	0.1286340569744262
Trt201702	0.0860117679867549	0.056053097855405444	0.07861686263964378	-0.0656764743852312	0.051695414065415696	-0.013435598350736962
Trt201703	-0.00019723590143907302	-0.04241281337121413	0.14396221833041117	0.021533918501646272	-0.03404024146448562	0.06831663654644538
Trt201704	0.047865563142528995	0.01282752309815467	0.03916131703087215	-0.11852573097783693	0.01257397871140488	-0.06307853097106716
Trt201705	0.003626145701412258	-0.02844043005377171	-0.005375143213260704	0.09581979253474752	-0.025140893567319147	-0.09526773640186258
Trt201706	0.010983592449786393	-0.02894381162040609	-4,08E+10	0.12694897934820285	-0.026032794272799583	-0.11157620704642793
Trt201707	-0.014538141516118852	-0.05887208231327931	-0.024693792781183943	0.11038784607805778	-0.05139876789573273	-0.14409006145951359

TRT/ Resmi Gazete	rg201601	rg201602	rg201603	rg201604	rg201605	rg201606
Trt201507	0.3717722383643915	-0.39279829620637163	-0.1976604505463672	-0.22461852189361692	-0.2934446280003272	-0.2622518458714766
Trt201508	0.30621228650733723	0.3676481804792237	-0.2867678438931504	-0.29983418514037236	0.3416634818474602	-0.3510945482821828
Trt201509	-0.2400284616285502	-0.11505821401551382	0.01325506812903122	-0.020930282965149625	-0.05505087825940958	-0.03058804765708743
Trt201510	-0.15023393974704874	-0.05850184656207664	0.033717493380272946	0.006893488250537052	-0.01763528045964965	0.0011511521755676677
Trt201511	-0.028777212517741766	0.07849757090943037	-0.10915618471868818	0.12177229603432861	0.10647133712390609	0.12858920817387967
Trt201512	-0.31017659901344186	-0.1722330601120951	-0.03033133696435993	-0.06249954400006717	-0.10652180606907333	-0.07962708359189136
Trt201601	0.36537978793766096	-0.2375696619506784	-0.08015724755214071	-0.1109362769764028	-0.16110945248530734	-0.133764349307173
Trt201602	-0.25822482263700347	-0.12918369867533053	0.001911039041165321	-0.03141607498131526	-0.06883283562323154	-0.043588239413739426
Trt201603	-0.3287184289141154	-0.1903795949740243	-0.04421475980770737	-0.07615545405699095	-0.12128188227053273	-0.09443406779303325
Trt201604	0.364461846838907	0.43205853315262693	-0.3263863560895985	-0.3502831622202599	0.40551921157263465	-0.40292152862794134
Trt201605	0.37036124650988295	-0.4353826747547959	-0.22950606206762403	-0.2564448775623618	-0.32800403716607	-0.29699171939722213
Trt201606	0.3702325318834866	0.44072729114379816	-0.23512069171832634	-0.2622688461239082	-0.33545766826121254	-0.3034789248524151
Trt201607	-0.2991393713787093	-0.16120463601030288	-0.021470685032334503	-0.05465761801957268	-0.09745159504312036	-0.06998385449783305
Trt201608	0.011945436940365251	0.11775193797885054	-0.054277849536609454	-0.08635501627235012	-0.13358334973894292	-0.10566818156835972
Trt201609	-0.06008813737093426	0.05405503307016658	0.14199378356621473	0.1039706996491048	0.08472686782573711	0.1085101665169207
Trt201610	-0.0005064595377998031	0.10833225434422557	-0.07095244680220487	-0.10257138684730618	0.13044820885382657	-0.12447839942647115
Trt201611	-0.07215540058777282	0.044464731704269844	0.13521410098855774	0.09711161971309175	0.07646141348755463	0.10043730122067587
Trt201612	-0.09364096920939576	0.022689952626265667	0.11838311275219136	0.0803130456715069	0.05874847673664981	0.08264797294273533
Trt201701	-0.03341277129033711	0.07887188074096105	-0.11598705691916776	0.12216804470531561	0.10552452238668428	0.12900525767693416
Trt201702	0.07844879541088269	-0.08386813760983232	0.037413230008496494	0.0026241779272254333	-0.03246774675884152	-0.006454717342392218
Trt201703	-0.11242574084020239	0.009336894822523584	-0.057361309594044375	0.07142737041540825	0.0456322839713912	0.07061009437459691
Trt201704	0.04704097953922914	-0.14055054868663544	-0.005778490539993361	-0.03921724428969074	-0.08011017954536559	-0.05321738276456516
Trt201705	0.009442801546610004	0.0918510661925082	-0.042844160391618556	-0.06712798899075424	-0.10348718099946706	-0.08294997846334316
Trt201706	0.016361904759111912	0.12157617736605475	-0.04764953210413701	-0.08067042189994619	-0.1261492463700823	-0.09868203240075371
Trt201707	-0.004314405692939406	0.10408793668373574	-0.07605867462408496	-0.1077127309858559	0.12736849618573823	-0.12981077262081522

TRT/ Resmi Gazete	rg201607	rg201608	rg201609	rg201610	rg201611	rg201612
Trt201507	0.41280754881242304	-0.289486016841746	0.42831564671129235	-0.12437515958543288	-0.04656274376284246	-0.2740643522588498
Trt201508	0.3401302013101859	0.37295557756202824	0.35353432948076896	-0.2125380854839559	-0.1275094108037262	0.34876403106142473
Trt201509	-0.2462334667484665	-0.041114843835788874	-0.2923915102520627	0.06067588147604348	0.10897088878040179	-0.04221681717293703
Trt201510	-0.15484761761862803	-0.0073244048956352574	-0.18843968933644398	0.0669269629198734	0.09840020884766014	-0.008097533816405551
Trt201511	-0.019231921692533658	0.1261671262063608	-0.0442467065187841	-0.048295634268754606	0.01485469878924543	0.11685707938464425
Trt201512	-0.32020418563168834	-0.095200344966552	-0.37430277523964856	0.021180130604771544	0.074874224982751	-0.09171080964707709
Trt201601	-0.39635275387059926	-0.15307017587441524	0.4199854207467391	-0.022624865485252412	0.03685543263860677	-0.14591077840192726
Trt201602	-0.26533533132595977	-0.05738508026337409	-0.31357528384512345	0.04907567199309825	0.09623623839653009	-0.055877370966792926
Trt201603	-0.34034814132892977	-0.1120009989032334	-0.3966058552491456	0.008588042503627947	0.06227029459167626	-0.10696990115392907
Trt201604	0.4027301886272332	0.4383427863390773	0.41829317070004296	-0.23925916128381966	-0.14720229574900254	0.41057229402372114
Trt201605	0.41022375234139713	-0.32631042268895405	0.42559080774028557	-0.15263222176668417	-0.07123295479495782	-0.309389896934544
Trt201606	0.4106639590725008	-0.33420756892613324	0.4258628588151344	-0.15826360271575832	-0.07713241290204505	-0.31625911269814866
Trt201607	-0.3073420647142511	-0.08497333864945947	-0.3612710401447502	0.029863902179602983	0.08241099529140593	-0.08174633336161621
Trt201608	0.026267370340352408	-0.12331820027886953	0.004267984270778484	0.0006699534539675605	0.05730726117861059	-0.11774958706445383
Trt201609	-0.05086908419477724	0.10480259512636779	-0.08027614539733928	-0.08554270057768223	-0.015597409575546029	0.09677918072963507
Trt201610	0.013051799572196906	-0.14384472839378057	-0.010390228688339097	-0.01411701417633503	0.04440985262469221	-0.13672820232738442
Trt201611	-0.06365880502557311	0.09625467255934446	-0.09448791866676697	-0.09999658365639208	-0.027607855886760423	0.0887013786237635
Trt201612	-0.08709962634267399	0.07808742038613427	-0.11966739876515176	-0.12580036309243442	-0.049325260756083456	0.07101737149987605
Trt201701	-0.022160921906302432	0.1269413980155738	-0.0488426320678971	-0.05350259128341386	0.011764094240052298	0.11738710362158408
Trt201702	0.09744368766048009	-0.018284777711974714	0.082157236272359	0.08042200126383128	-0.02405701481482869	-0.018510216116799903
Trt201703	-0.10714688965732247	0.06311956883050107	-0.1423158918008704	-0.002973996146104784	0.05323241864626227	0.058619854616550685
Trt201704	0.06385831781926947	-0.06757688636125267	0.045665599566773574	0.04304403708799547	-0.07084801455089856	-0.06515914210030221
Trt201705	0.020100561008044783	-0.09695119142607861	0.003097532799963406	0.0001076045945142 1009	0.04469048671836503	-0.09246193745822007
Trt201706	0.03082337382183845	-0.11559000465068056	0.009617234457573192	0.006328047445420268	0.06195793423287244	-0.11097669930298326
Trt201707	0.009372683029383428	0.1492246026771586	-0.014666101661294235	-0.018622118760695697	0.04088323011496175	0.13843355996586124

TRT/ Resmi Gazete	rg201701	rg201702	rg201703	rg201704	rg201705	rg201706	rg201707
Trt201507	-0.30151951212924943	0.4079282938562104	-0.23371122404069242	-0.36331294336211756	-0.3678690686025136	-0.3098069305031509	0.432813150920679
Trt201508	0.34150635198010787	0.33637334533860525	-0.31420566453871424	0.30452936074589665	0.32989807284728007	0.34935942945177006	0.3595683073666406
Trt201509	-0.06025778084154725	-0.20026983972536863	-0.021457619902721513	-0.11712300706443764	-0.11159686808681656	-0.06336781470572855	0.4248050003845051
Trt201510	-0.022512275325376852	-0.12281667476385856	0.006231580249560169	-0.0639771054041417	-0.05804298884548984	-0.023594744791762365	0.3288048065261501 3
Trt201511	0.10156357412044387	0.008261743923073456	0.12504122319530617	0.04956577411386315	0.06497128263274202	0.10351538335645208	-0.381923198039931
Trt201512	-0.11310272414106635	-0.2708155448239281	-0.06645654971045413	-0.17072018376191853	-0.16579132161981114	-0.11684234572101423	0.42808338484927944
Trt201601	-0.16895256847643222	-0.3409080217553713	-0.1164796313886989	-0.2270842935486606	-0.22543829525973086	-0.17376665908038172	0.4245718665118523
Trt201602	-0.07603681668200533	-0.2196222512362455	-0.03389344184607227	-0.13173280144889732	-0.12553171962320164	-0.07854393156010443	0.42104999027142487
Trt201603	-0.1291849216623162	-0.2896836255770004	-0.08057896243998247	-0.1863483112378498	-0.18160176953034146	-0.1325569682578095	0.4244815654137425
Trt201604	0.40372397933682863	0.3986742302287244	-0.36478755156235443	0.3585273665913192	0.38905880932748443	0.412846559700363	0.4224629114490703
Trt201605	-0.3371955416385658	0.40640213058866376	-0.26682445099874647	0.36482499023138193	0.39558092063771944	-0.34611076538054797	0.4308570174910544
Trt201606	-0.3448968216514294	0.40590664551598465	-0.2732733379556362	0.3648914250379877	0.3960538474209414	-0.3536083744806646	0.4306009439095932
Trt201607	-0.1037357173783512	-0.2596895329985052	-0.05739168110565073	-0.16142942030520546	-0.15525440241148517	-0.10674121527190665	0.4323998439949324
Trt201608	0.1352109070058856	0.04937582641393577	-0.09044487385873243	0.08283342416841197	0.10058289492661819	0.13767676330593825	-0.3025279985736764
Trt201609	0.08021467179768359	-0.02221474183347657	0.10696763416228774	0.02661751732407277	0.04184738085388376	0.08144304833790429	0.43143009643268293
Trt201610	0.12629646316645915	0.036847179847478595	-0.10772909537146533	0.07366204231119347	0.09145009208835345	0.1286861086329007	-0.3295943023835219
Trt201611	0.07168476637009445	-0.03396296068687584	0.09952074056781629	0.01758277794400725	0.03283534882831262	0.07298348661183847	0.43363979787529283
Trt201612	0.054728434705975236	-0.05503028707718477	0.08292107855922178	-0.0003418980816142264	0.013085455765936595	0.054676090929781984	0.43072642403014877
Trt201701	0.10153285254142327	0.004668912066079238	0.12596808617900612	0.04798493742761374	0.06439547981235268	0.10312746440630201	-0.3966327755376348
Trt201702	-0.03914795728598107	0.11504341450123794	0.0011996793587773125	-0.09473105773908852	-0.08463339831468475	-0.04001524863180325	-0.16780478259709733
Trt201703	0.03943297294361573	-0.07529330413874298	0.0721444514990933	-0.014053253622061758	0.000716384322278109 5	0.04088530223293247	-0.06368063240909196
Trt201704	-0.08656427413091669	0.0843719245597893	-0.0419446419133223	0.1110043147277287	0.13032630254079405	-0.08920068456158654	-0.23093129908856716
Trt201705	0.10492138469587008	0.03855738103837461	-0.07058596278464965	0.06404045603296515	0.07791818787159302	0.10721930052897585	-0.23615793186896458
Trt201706	-0.13301219444692156	0.054245594559212955	-0.08441773161323732	0.08602010762467524	0.10400049895727864	-0.13673018584410784	-0.29187222600580404
Trt201707	0.12344984065131394	0.033526136392107525	-0.11285029354350205	0.07033652608493703	0.08773206300201837	0.12546557937739355	-0.3354241920864616

REFERENCES

[1] Beal, V. , Data, Retrieved December 1, 2017, from https://www.webopedia.com/TERM/D/data.html, 2017

[2] Beal, V. , Data, Retrieved December 1, 2017, from https://www.webopedia.com/TERM/S/structured_data.html, 2017

[3] Ohlhorst, F., Big Data Analytics: Turning Big Data into Big Money, New Jersey, Wiley, 2013

[4] Palmer, M., Data is the New Oil, Retrieved July 30, 2016, from http://ana.blogs.com/maestros/2006/11/data_is_the_new.html, 2006

[5] Barrenechea, M., Big Data: Big Hype, Retrieved November 21, 2017, from https://www.forbes.com/sites/ciocentral/2013/02/04/big-data-big-hype/#a9d641236665, 2013

[6] Douglas, L., 3D Data Management: Controlling Data Volume, Velocity and Variety, Application Delivery Strategies, META Group, 2001

[7] Börteçin, E., Rastlantının Bittiği Yer: Big Data, Bilim ve Teknik, (550), 25, 2013

[8] Karabey, B., Büyük Veri (Big Data) ve Kişisel Verilerin Korunması, Retrieved January 1, 2018, from http://by2012.bilgiyonetimi.net/cagrili-konusma-buyuk-veri-big-data-ve-kisisel-verilerin-korunmasi/, 2012

[9] Agrawal D., Bernstein P., Bertino E., Davidson S., Dayal U., Franklin M., ... Widom J. ,Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association, Retrieved January 5, 2018, from http://cra.org/ccc/resources/ccc-led-whitepapers/, 2012

[10] Big Data, Retrieved November 30, 2017, from https://www.wikizero.com/en/Big_data, 2017

[11] Larson, D., Data Analysis Structured vs. Unstructured Data, Retrieved February 2, 2018 from, http://www.freeitdata.com/wp-content/uploads/Structured-v-unstructured-data.pdf

[12] Data Never Sleeps 5.0, Retrieved January 25, from, https://www.domo.com/learn/data-never-sleeps-5, 2017

[13] Morabito, V., Big data and Analytics - Strategic and Organizational Impacts, Switzerland, Springer International Publishing, 2015

[14] Rouse, M., Wigmore, I.,Internet of Things (IoTs), Retrieved 21 January, 2018, from, https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT, 2018

[15]Nicolas, N., IOT Landscape – Big Data Innovation, Retrieved February 2, 2018, from https://www.nsuchaud.fr/2015/08/iot-landscape-big-data-innovation, 2015

[16] Swapnil K. W., Anil Y., Gupta S., Big Data: Characteristics, Challenges and Data Mining, International journal of computer applications, 2016

[17] Rajaraman, V., Big Data Analytics, Resonance, Indian Academy of Sciences, 2016

[18] Venkata, Subramaniam L., Big Data and Veracity Challenges: Text Mining Workshop, IBM Research India, 2014

[19] Khan, N., Yaqoon, I., Hashem, I., ….. Gani, A., Article, Big Data: Survey, Technologies, Opportunities, and Challenges, The Scientific World Journal 2014, 2014

[20] Vale, S., Classification of Types of Big Data, Retrieved January 21, 2018, from https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data, 2013

[21]Elephant, P., The Big Data Lifecycle, Retrieved January 13, 2018, from https://www.pinkelephantasia.com/big-data-lifecycle, 2017

[22] Vaidya, A., Kosarwal, A., Big Data Storage, Santa Clara, IGATE, 2014

[23] Woods, D., Ten Properties of the Perfect Big Data Storage Architecture, retrieved January 28, 2018, from https://www.forbes.com/sites/danwoods/2012/07/23/ten-properties-of-the-perfect-big-data-storage-architecture/#3447b0b3799e, 2012

[24] Rouse, M., Big Data Storage, Retrieved January 28, 2018, from http://searchstorage.techtarget.com/definition/big-data-storage, 2017

[25] Adiba, M., Castrejon-Castillo, J., Espinosa Oviedo J.A., …., Zechinelli-Martini, J., Big Data Management Challenges, Approaches, Tools and Their Limitations, Chapman and Hall, 2016

[26] NoSQL vs SQL- 4 Reasons Why NoSQL is better for Big Data applications, Retrieved March 8, 2018, from https://www.dezyre.com/article/nosql-vs-sql-4-reasons-why-nosql-is-better-for-big-data-applications/86, 2015

[27] Mohmmed, A.G.M., Osman, S. E. F., Study on SQL vs. NoSQL vs. NewSQL, Journal of Multidisciplinary Engineering Science Studies, Vol. 3 Issue 6, 2017

[28] Tailor, H., Choudhary , S., Jain, V., Rise Of NewSql, International Journal For Research in Emerging Science and Technology, Vol. 2 Issue 1, 2015

[29] Cassandra Query Language, Retrieved March 07, 2018, from https://www.tutorialspoint.com/cassandra/cassandra_tutorial.pdf, 2015

[30] MongoDB Tutorial, Retrieved March 05, 2018, from http://mdslab.unime.it/sites/default/files/mongodb_tutorial.pdf

[31] MongoDB Tutorial, Retrieved March 05, 2018, from http://www.dbschema.com/MongoDB-Tutorial.pdf

[32] Apache Hbase, Retrieved March 06, 2018, from http://hbase.apache.org/, Apache, 2014

[33] What Is Apache Hadoop?, Retrieved March 06, 2018, from http://hadoop.apache.org/, 2018

[34] Apache Spark, Retrieved March 06, 2018, from https://spark.apache.org/, 2018

[35] Apache Incubator - S4 Project Incubation Status, Retrieved March 06, 2018, from http://incubator.apache.org/s4/, 2018

[36] Iqbal, M.H., Soomro, T.R., Big Data Analysis: Apache Storm Perspective, International Journal of Computer Trends and Technology, Vol. 19 Issue 1, 2015

[37] Data with SimpleDB, Retrieved March 07, 2018, from http://awsmedia.s3.amazonaws.com/pdf/simpledb.pdf

[38] Amazon SimpleDB, Retrieved March 07, 2018, from https://w1kipedia.org/p.php?https://en.wikipedia.org/wiki/Amazon_SimpleDB, 2018

[39] Couch DB- NoSQL Document Store, Retrieved March 07, 2018, from https://www.tutorialspoint.com/couchdb/couchdb_tutorial.pdf, 2015

[40] Redis, Retrieved March 07, 2018, from https://w1kipedia.org/p.php?https://en.wikipedia.org/wiki/Redis, 2018

[41] Elastic Search , Retrieved March 07, 2018, from https://w1kipedia.org/p.php?https://en.wikipedia.org/wiki/Elasticsearch, 2018

[42] Enterprise Applications, Analytics and Knowledge Management Trends 2013, Retrieved February 12, 2018, from https://www.slideshare.net/Einats/einat-applications-summit2013final, 2013

[43] Rijmenam V. M., Benefits For The Public Sector When Governments Start Using Big Data, Retrieved February 27, 2018, from https://datafloq.com/read/4-benefits-public-sector-governments-start-big-dat/171, 2012

[44] 8 Benefits of Big Data for State and Local Governments, Retrieved February 28, 2018, from https://statetechmagazine.com/article/2013/05/8-benefits-big-data-state-and-local-governments , 2013

[45] DIKW pyramid, Retrieved February 28, 2018, from https://en.wikipedia.org/wiki/DIKW_pyramid

[46] Knowledge Management, Retrieved March 09, 2018, from https://www.tutorialspoint.com/knowledge_management/knowledge_management_tutorial.pdf, 2015

[47] King W. R., Knowledge Management and Organizational Learning, 3 Annals of Information Systems 4, DOI 10.1007/978-1-4419-0011-1_1, Springer Science and Business Media, 2009

[48] Shorfuzzaman M., Leveraging Cloud Based Big Data Analytics In Knowledge Management For Enhanced Decision Making In Organizations, International Journal of Distributed and Parallel Systems (IJDPS) Vol.8, Issue 1, 2017

[49] Watson H. J., Tutorial: Big Data Analytics: Concepts, Technologies, and Applications, Communications of the Association for Information Systems, Retrieved March 19, 2018, from http://aisel.aisnet.org/cais/vol34/iss1/65, Vol. 34, Article 65, 2014

[50] Analytics, Retrieved March 19, 2018, from https://w1kipedia.org/p.php?https://en.wikipedia.org/wiki/Analytics, 2018

[51] Awanish, Edureka Big data Analytics Tutorial: Big Data Analytics Tutorial | Big Data Analytics for Beginners | Hadoop Tutorial, Retrieved March 19, 2018, from https://www.edureka.co/blog/big-data-tutorial, 2016

[52] Text Analytics Beginner's Guide, Retrieved March 20, 2018, from http://www.angoss.com/wp-content/uploads/2013/04/eBook-Text-Analytics-Beginners-Guide.pdf, Angoss Software Corporation, 2013

[53] What is Text Analytics, Retrieved March 20, 2018, from https://www.predictiveanalyticstoday.com/text-analytics/#wysija, 2018

[54] Denny M. J., Spirling A., Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It, 2017

[55] Thompson B., Study: Voice of Customer, Customer Experience Tops in Primary Text Analytics Applications, Retrieved March 20, 2018, from http://customerthink.com/study-voice-of-customer-customer-experience-tops-in-primary-text-analytics-applications/, 2014

[56] Text and Data Mining Glossary, Retrieved March 20, 2018, from https://www.elsevier.com/__data/assets/pdf_file/0018/102906/TDM-Glossary.pdf, 2018

[57] Evelson B., The Forrester Wave™: Big Data Text Analytics Platforms, Retrieved March 20, 2018, from https://www.forrester.com/report/The+Forrester+Wave+Big+Data+Text+Analytics+Platforms+Q2+2016/-/E-RES122667, 2016

[58] Resmi Gazete Tarihçesi, Retrieved March 26, 2018, from http://www.mevzuat.gov.tr/RegaTarihce.aspx

[59] Spark Core Programming, Retrieved March 26, 2018, from https://www.tutorialspoint.com/apache_spark/apache_spark_tutorial.pdf, 2015

[60] Anaconda, Retrieved March 26, 2018, from https://w1kipedia.org/p.php?https://en.wikipedia.org/wiki/Anaconda_(Python_distribution), 2018

[61] Halterman R. L., Learning to Program with Python, 2011

[62] Introduction to Python, Retrieved March 27, 2018, from http://tdc-www.harvard.edu/Python.pdf, 2009

[63] Ahemad F., Machine Learning with Jupyter using Scala, Spark and Python: The Setup, Retrieved March 26, 2018, from https://medium.com/@faizanahemad/machine-learning-with-jupyter-using-scala-spark-and-python-the-setup-62d05b0c7f56 , 2017

[64] Bean R., Mastercard's Big Data For Good Initiative: Data Philanthropy On The Front Lines, Retrieved November 27, 2018, from https://www.forbes.com/sites/ciocentral/2017/08/07/mastercards-big-data-for-good-initiative-data-philanthropy-on-the-front-lines/#159fdea420dc, 2017

[65] Woods D., Ten Properties of the Perfect Big Data Storage Architecture, Retrieved February 27, 2018, from https://www.forbes.com/sites/danwoods/2012/07/23/ten-properties-of-the-perfect-big-data-storage-architecture/#4c8f0721799e, 2012

[66] Zaratsian D., Spark Text Analytics - Uncovering Data-Driven Topics, Retrieved October 14, 2018, from https://community.hortonworks.com/articles/84781/spark-text-analytics-uncovering-data-driven-topics.html, 2017

[67] Hoadley D., Calculating Cosine Similarity Between Documents, Retrieved October 23, 2018, from http://carrefax.com/new-blog/2017/7/4/cosine-similarity, 2017

CURRICULUM VITAE

Yasemin CAN

E-mail: yaseminguder@gmail.com, yasemin.can@saglik.gov.tr

Birth Place: Eskişehir, TURKEY

EDUCATION

- **Graduate:** Yıldırım Beyazıt University, ANKARA**,** Computer Engineering, Master

  Student (2012 - Present)

  - **Interest:** Big Data, Artificial Neural Networks
  - **Master Thesis:** Analytics of Turkish Official Gazette Archives Using Big Data Mining Techniques

- **Undergraduate:** Bilkent University (Full Scholarship), ANKARA

  Computer Engineering, 2003 - 2008

EMPLOYMENT

- T.C. Ministry of Health, ANKARA, Health Information Systems Department, Assistant Health Specialist (2014 - Present)
- Yıldırım Beyazıt University, ANKARA**,** Computer Engineering Department Research/Teaching Assistant (2012-2014)

LANGUAGES

• **Turkish** – Native language

• **English** – Advanced level reading, writing, listening and high level at speaking. YDS: 92,5 / September 2017