Department of Computer Engineering

HACETTEPE
ÜNİVERSİTESİ

Internship report

Mohammed Ali
Student ID: 21403227

Advisor :Asst. Prof. Dr. Adnan ÖZSOY

From July 03 , 2017 to August 11, 2017
(**30 working Days**)

Acknowledgements

The internship opportunity I had in Hacettepe University, Computer Engineering Department was a great chance for learning and professional development. Therefore, I consider myself very lucky that I got indulged in work regarding the field I loved so much.I express my deepest gratitude to my dear advisor :Asst. Prof. Dr. Adnan ÖZSOY who enlightened me with valuable guidance and continuous advises that facilitate my effort all period of internship .

I perceive as this opportunity as a big milestone in my career development and this internship was just the beginning to start my journey in world of data analysis and web scraping. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work and think in order to attain desired career objectives.I hope to continue work in the future and develop this project.

Last, I would like to thank my parents and family for supporting my stay in Turkey financially and emotionally and many thanks to all my friends in Turkey as well.

Sincerely,

Contents

1 Introduction:

Web scraping (web data extraction) is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler.[1]

Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when you view the page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed and extract to get the targets links in these pages.

For this project, I have made six Turkish news websites, Is a form of copying, in which specific data is gathered and copied from the pages, typically into a central local mongo database for later retrieval and plotting graphs for increasing links by time. Main steps are explained in image that it can be seen in appendix A.

Many external programs and tools are used to accomplish this project and make working easy such as Robomongo[2] and Bitbucket[3]. Robomongo is used to check, delete and edit the content of databases. Bitbucket is used as a web-based hosting service tool which help manage the code for the project as it changes over time.

2 Web Scraping

Extracting data from any website starts by passing URL to jsoup library to parse HTML page to get links then passing these links to a boilerpipe to extract just text data and ignore photos,videos and advertisements after that links should be stored contents in files or databases.

2.1 Creating Objects

For this internship, Most popular Turkish newspaper websites have been scraped and extracted for the main news in that websites and ignored any things else. Eight objects have been created for each website, Milliyet[4], Hurriyet[5], Sabah[6], Odatv[7], BBC[8], CNN[9], sozcu[10] and Al jazeera[11]. All objects have created integratedly.Classes and methods have been illustrated in the image provided as UML diagram in the appendix B.

2.2 HTML Parsing

HTML parsing is basically is taking in HTML code and extracting relevant information like the title of the page, paragraphs in the page, headings in the page, links or texts. For this project, we need just Texts data. Many libraries are available in java to use. Jsoup library was used because it's easier to use and learn[12].

2.3 Using Jsoup To Get Links.

jsoup is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods. Jsoup library can do:

- Scrape and parse HTML from a URL, file or string
- Find and extract data, using DOM traversal or CSS selectors
- Manipulate the HTML elements, attributes, and text
- Output tidy HTML

2.4 How To Choose CSS Selectors

A CSS selectors are the part of a CSS rule set that actually selects the content you want to style. In this project, selectors are used to it choose the target news in the HTML pages, Google chrome has tools which helps to choose the selectors fast without tracing HTML code line by line. Each website has its own css selectors. As long as we get no links , that means they update the code for that website.CSS selectors should be checked and updated periodically; otherwise we can not figure out where links are located inside HTML page.

2.4.1 Css Selectors For Sabah Site

- body > section > div > div:nth-child(1) > div > div > div.col-sm-7.col-sm-12.side.left > div > div a
- body > section > div > div:nth-child(2) > div > div > ul a
- body > section > div > div:nth-child(3) > div > div > div a

2.4.2 Css Selectors For BBC Site

- #comp-top-story-1 > div > div a
- #comp-top-story-2 > div a
- #comp-top-story-3 > div > div a

2.4.3 Css Selectors For CNN Site

- *body > div.main-container > div.container.headline-container > div:nth-child(1) > div a*
- *body > div.main-container > div.container.headline-container > div:nth-child(3) a*
- *body > div.main-container > div.container.headline-container > div.row.flex-order-1 > div.col-md-8.col-sm-12 > div > ol a*

For Al Jazeera the css is selectors

- *#block-boxes-aljazeera-main-promo-box a*

2.4.5 Css Selectors for Sozcu Site

- *#sz_manset a*

2.4.6 Css Selectors For Hurriyet Site

- *body > main > div > div > div:nth-child(1) > div a*
- *body > main > div > div > div:nth-child(4) a*
- *body > main > div > div > div:nth-child(5) > div.col-xs-8.col-sm-8.col-md-8.col-lg-8 a*

2.4.8 Css Selectors For Odatv Site

- *#boxed-wrap > section.container.page-content > div:nth-child(2) a*

2.4.7 Css Selectors For Milliyet Site

- *.flashbar1 . a*
- .flashbar1 .top_p1 a
- .flashbar1 .top_p2 a
- .flashbar1 .tnw a
- .mnst11 a

2.2.3 Using Boilerpipe

Extracting the main content (‘body’) text from a web page is difficult for the general case. It would appear to lend itself to machine learning. Here, Boilerpipe was used to get news articles.Boilerpipe implements a number of extraction algorithms for different circumstances. ArticleExtractor is recommended for most cases, but it is specifically intended for news articles[13][14].

3 Software usage

The project was designed to be continually compiled every three hours for one day, starting from creating objects till plotting graphs. In addition to showing the information of software usage, tools and mavens should be mentioned as following:

3.1 Tools, IDES and Maven

The project has been achieved on Ubuntu operating system by using java programming language,version "1.8.0_131" and Intellij is used as a Java integrated development environment (IDE) for developing this project. External libraries were added as maven dependencies so external libraries have been installed to laptop. Five maven were used in this project such as *jsoup ,syncthemall[15] ,mongodb[16], jfree[17] and codehaus.jackson[18]* respectively. Chrome developer tools are used to choose CSS selectors.

3.2 Mogno Database

3.2.1 Introduction

MongoDB is an open-source database developed by MongoDB, Inc. MongoDB stores data in JSON-like documents that can vary in structure. Related information is stored together for fast query access through the MongoDB query language. MongoDB uses dynamic schemas, meaning that I can create records without first defining the structure, such as the fields or the types of their values. I can change the structure of records (which we call documents) simply by adding new fields or deleting existing ones. Data model is given the ability to store many objects easily[19].

3.2.2 Why MongoDB?

MongoDB is a NoSQL database which it has been found faster in dealing with data than doing so with files that were used before but in vain. It was quite easy to use, to learn and faster to save or retrieve data for those who did not deal with databases before. Mongodb was used to save news links and contents of these links inside the databases.

3.2.3 Creating Two MongoDB

Two databases have been created one is for news and the others for links counter over time. Each database contains eight collections;one for each website.

- NewsDB that contains, eight collections; one for each website. It has been illustrated in the image provided in the appendix C.

- linksCounter Mongodb, inside this database there a collection for each website .It has been illustrated in the image provided in the appendix D.

3.2.4 Using Robomongo

RoboMongo is a visual tool helping you manage Database MongoDB. It is a part of free open source software supporting all of three operating systems: Windows, Linux, Mac OS. Connecting to a MongoDB Database Using Robomongo has been used to show, edit, delete what is inside database collections easily.

3.3 JFreeChart

JFreeChart is a free 100% Java chart library that makes it easy for developers to display professional quality charts in their applications[20]. I have used this library to draw eight line graphs according increasing links with eight period of time. We can see more explanation in the graph section below.

3.4 About Bitbucket

Bitbucket is a web-based hosting service used as distributed version control system for source code and development projects.It used to manage the code for the project as it changes over time. It is allowed past versions of the project to be saved in case new changes break things.Bitbucket is used to make collaboration with advisor more easier.

3.5 Data Structures and Algorithm

3.5.1 Data Structures

Arraylist was used as a data structure because it has more flexibility and features for examples, It is dynamic if we don't know the size of array and it works efficiently.
Two arraylists were used to in this project:

- One to hold links after parsing HTML page
- Second to hold content of links after using boilerpipe

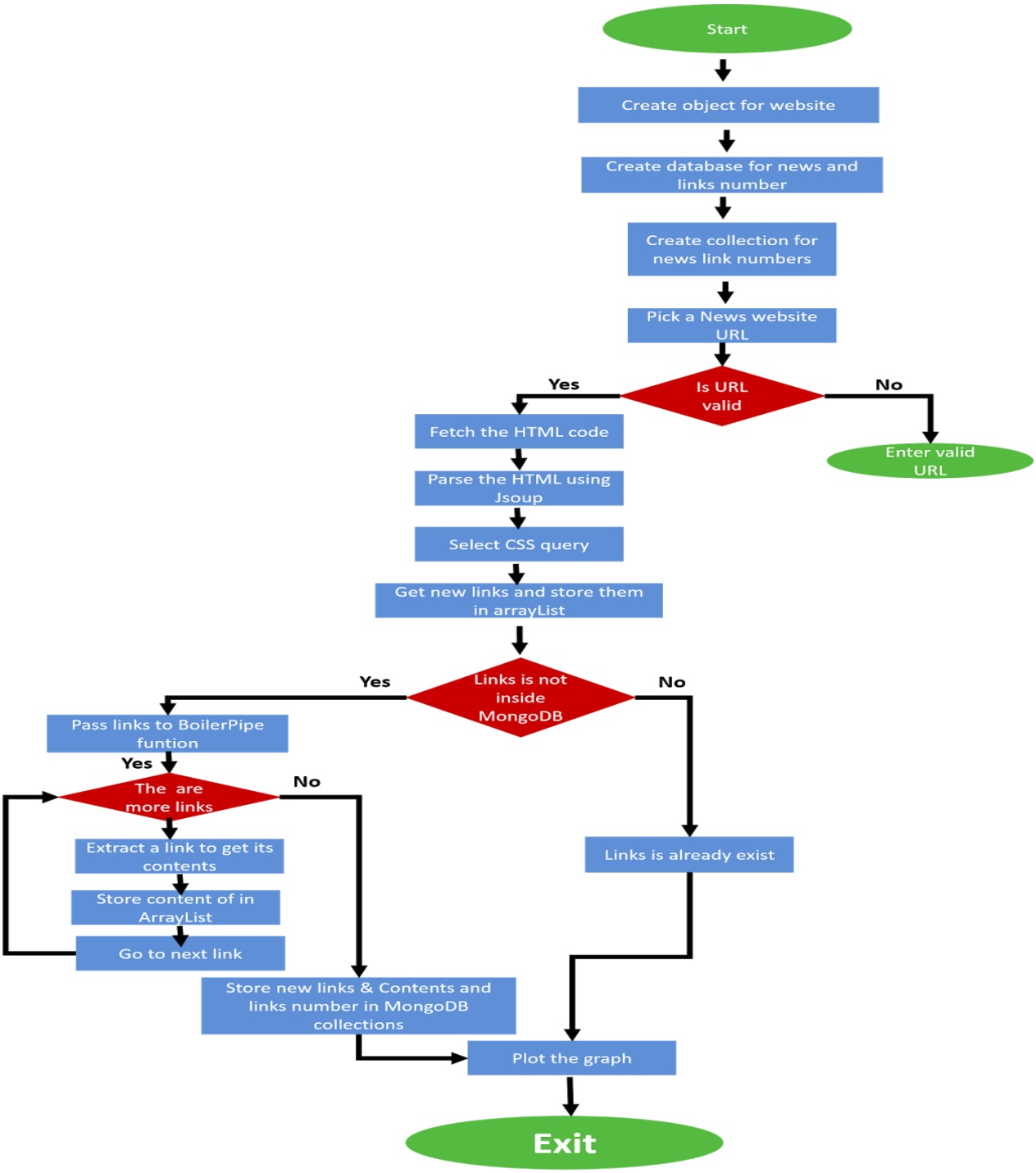
3.5.2 Algorithm

Algorithm is illustrated the operations for web scraping for a Turkish news site .
Pseudocode and flowchart were explained as following:

a) Pseudocode

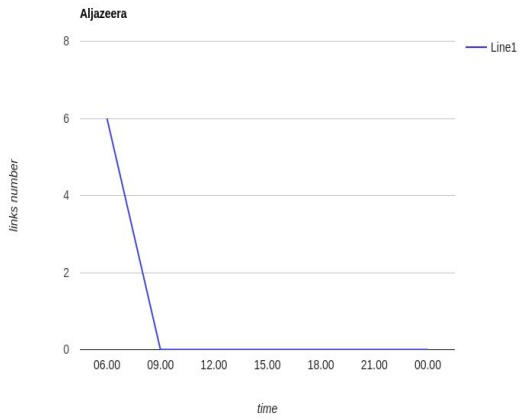
```
Create object for website;
Create databases for news and links number;
Create collection for news and link numbers;
Pick a News website URL;
        if(url exist?){
               Fetch the HTML code;
               Parse the HTML using jsoup ;
               select css query;
               Get new links and store them in arrayList ;
                if(Links is not inside MongoDB){
                        Pass links to BoilerPipe function;
                              if (The  are more links){
                                    Extract link to get  content;
                                    Store content of in ArrayList;
                                    Go to next link;
                                    }
                              }
                  else {
                    Links is already exist;
                    }
                  plot the graph on screen;
               }
         else
              Enter valid url;
        }
```

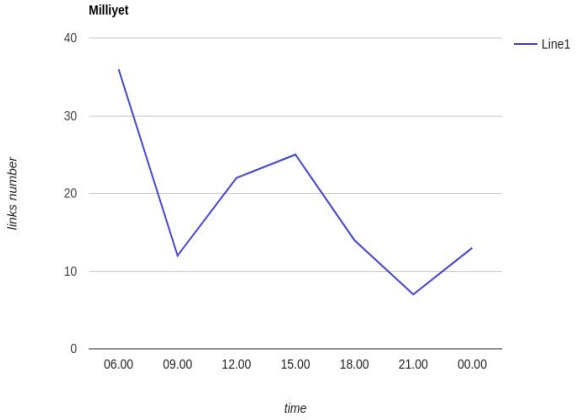
b) Flowchart Diagram

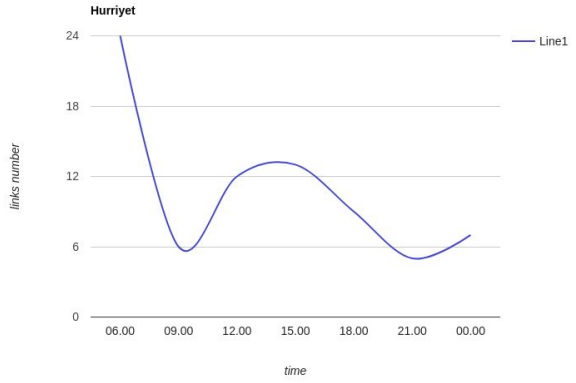


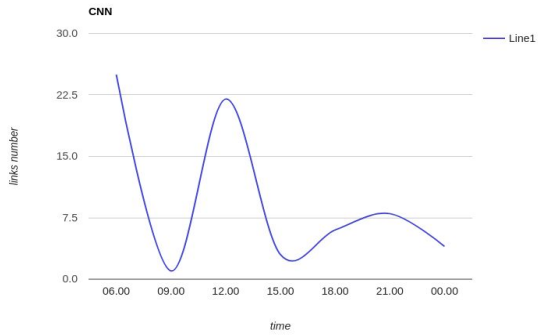
4 Graphs

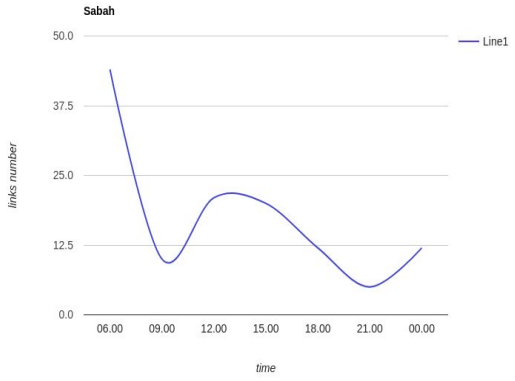
Eight graphs have been plotted for each website to show increasing of links by time during 24 hours. X-axis is shown the time for compiling program and y-axis is shown the number of new links at that time.Graphs are plotting as following:

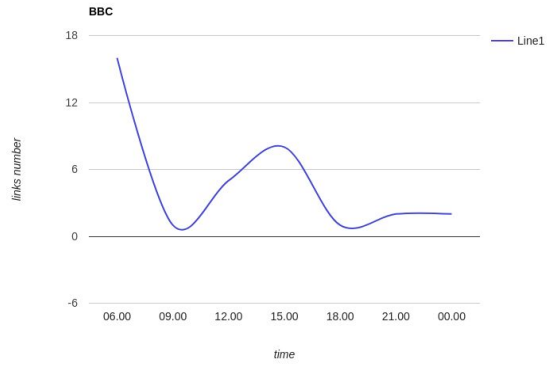


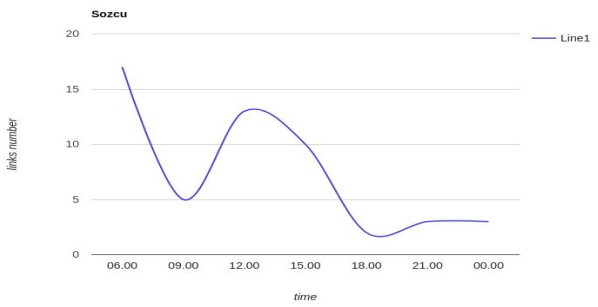


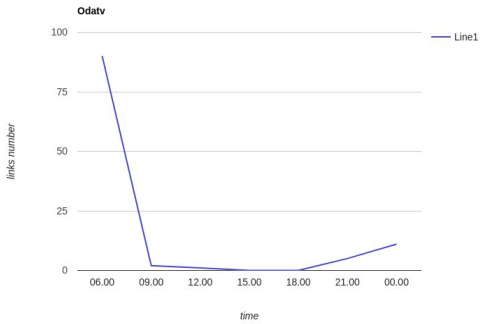












5 Issues and Open Problems

1. How to figure out automatically if the source code of a website has been changed?
   My suggestion:
   I think that I should compile program and check if I did not find any link in the list , that means updating CSS selectors.

2. How scraping for a website can be avoided?
   My suggestion:
   We can not prevent scraping, we can figure out who is scrapping if someone access to website more than twice at the same time then we can block next requests.

3. What can be done with data?
   For future work !

6 Project Improvements

- It would have been better to utilize Ubuntu operating system or any alike operating system to have the script written, rather than being written using Timer in Java.
- Two databases have being used to to store data. I was thinking we could create one for all and inside that one we may have put two of database.
- It would have been better to create a website for scraping Turkish newspapers every day.
- We could make analysis for news data and make studies about turkish news

7 About Institution and Advisor

About Institution: I have done my internship in hacettepe university, department of computer Engineering[21]. The Department of Computer Engineering at Hacettepe University was established as a graduate school in 1974 under the name The Institute of Informatics. Three years after its creation, it underwent a transformation and opened its doors to the first intake of 20 undergraduate students, making Hacettepe University the first Turkish university to offer a Bachelor's degree in computer engineering. The three main divisions of the department are Computer Science, Computer Software, Computer Hardware. Within the last 5 years, the Department has expanded its faculty to include 15 new members which strengthens the department's research ability across multiple diverse areas.

About Avisor:

Adnan Ozsoy is a Ph.D. candidate at the School of Informatics and Computing of Indiana University — Bloomington. He received his B.Sc. in Computer Science from Virginia Polytechnic Institute and State University in 2005, and his M.Sc. in Computer Science degree from University of Texas at Austin in 2007. His research interests include parallel programming, high performance computing with GPUs, and application parallelism problems. [22]

8 Conclusion

During a period of six weeks working on his project, It was really great chance for me to meet a new work field.I become more familiar with linux as a new operating system, Git and also databases commands lines. I have learned more about web scraping, how it works and some libraries for scrapping in java. Moreover, I have learned how to use boilerpipe library and how it works. Mongo database is used to store data without any repeated data. Mongo made work faster and easier. Finally I learned about drawing graphs in java using JFreeChart to draw line graphs.

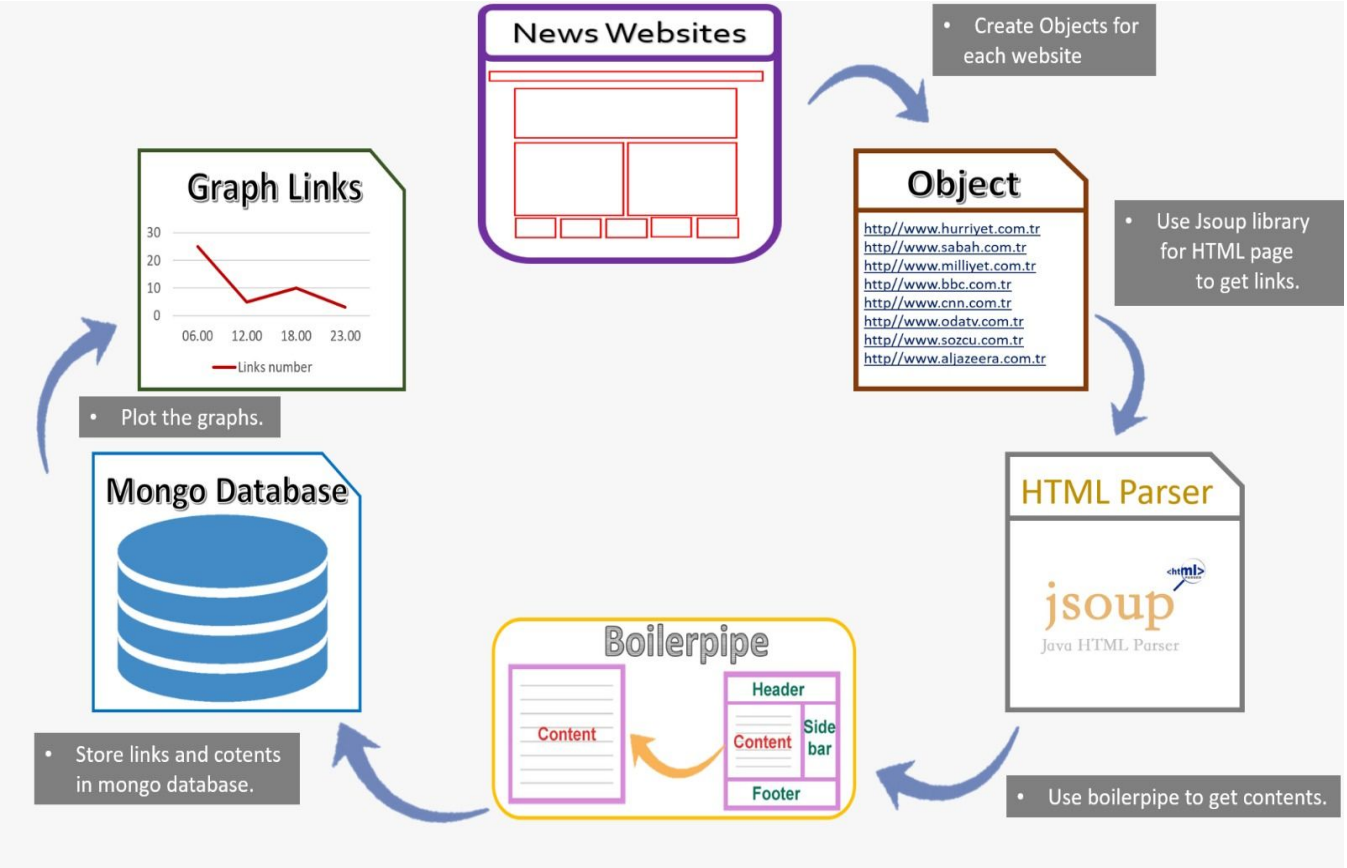
Eventually I am overwhelmed with the will to use the same concept in arabic most famous websites one day that currently no single work is done regarding such major.

References

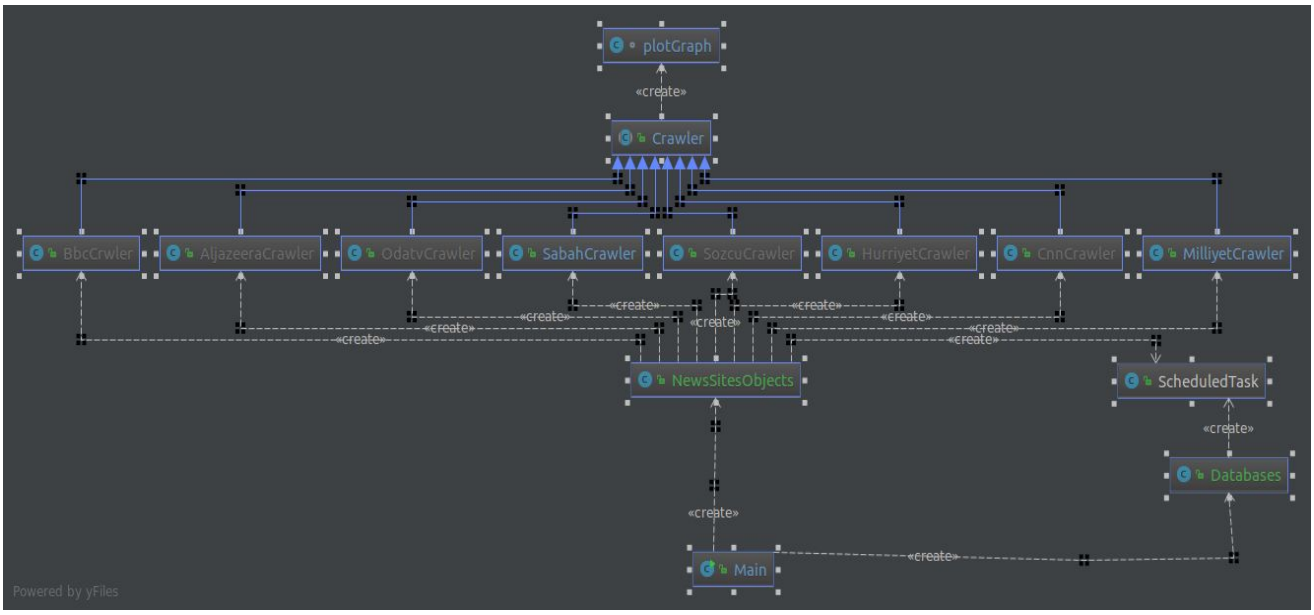
[1]https://en.0wikipedia.org/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvV2ViX3NjcmFwaW5n
[1]http://cs.hacettepe.edu.tr/
[2]https://robomongo.org/
[3]https://www.mongodb.com/
[4]http://www.milliyet.com.tr/
[5]http://www.hurriyet.com.tr/
[6]http://www.sabah.com.tr/
[7]http://odatv.com/
[8]http://www.bbc.com/turkce
[9]http://www.cnnturk.com/
[10]http://www.sozcu.com.tr/
[11]http://www.aljazeera.com.tr/front
[12]https://jsoup.org/
[13]http://www.basicsbehind.com/extract-text-webpage/
[14]https://boilerpipe-web.appspot.com/
[15]https://github.com/vanduynslagerp/boilerpipe
[16] https://www.tutorialspoint.com/jfreechart/jfreechart_overview.htm
[17]https://mvnrepository.com/artifact/jfree/jfreechart/1.0.13
[18]https://mvnrepository.com/artifact/org.codehaus.jackson/jackson-mapper-asl/1.9.13
[19]https://www.mongodb.com/compare/mongodb-mysql?jmp=docs
[20]https://www.tutorialspoint.com/jfreechart/jfreechart_overview.htm
[21]http://cs.hacettepe.edu.tr/
[22]http://web.cs.hacettepe.edu.tr/~aozsoy/index.html

Appendix

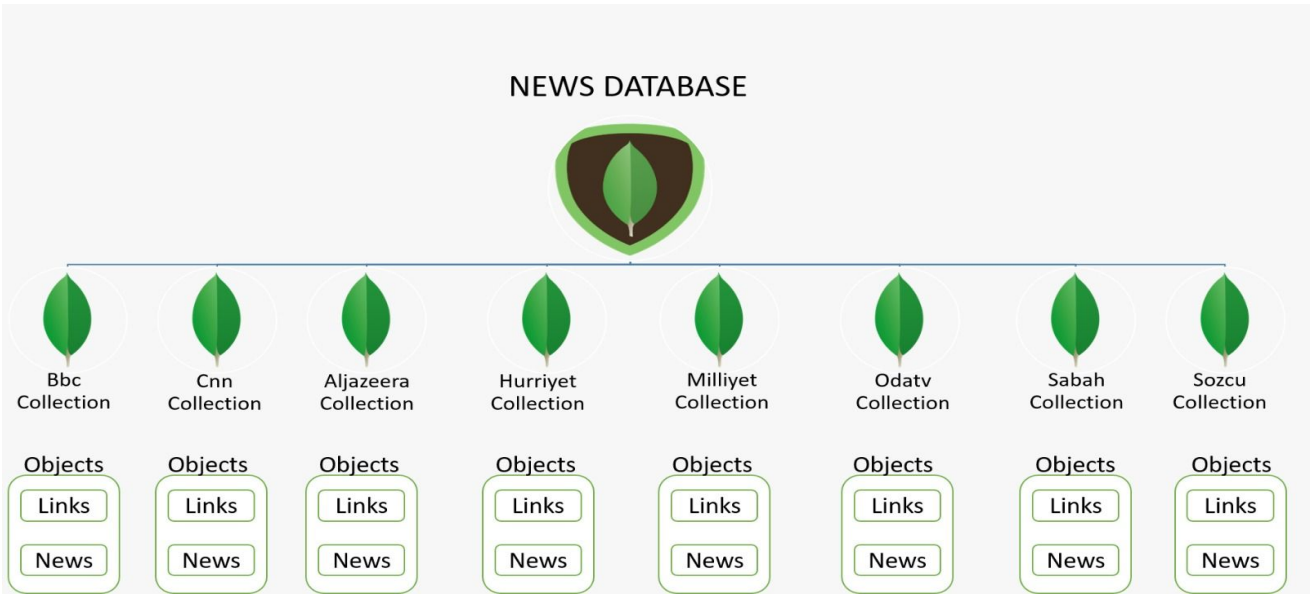
appendix A



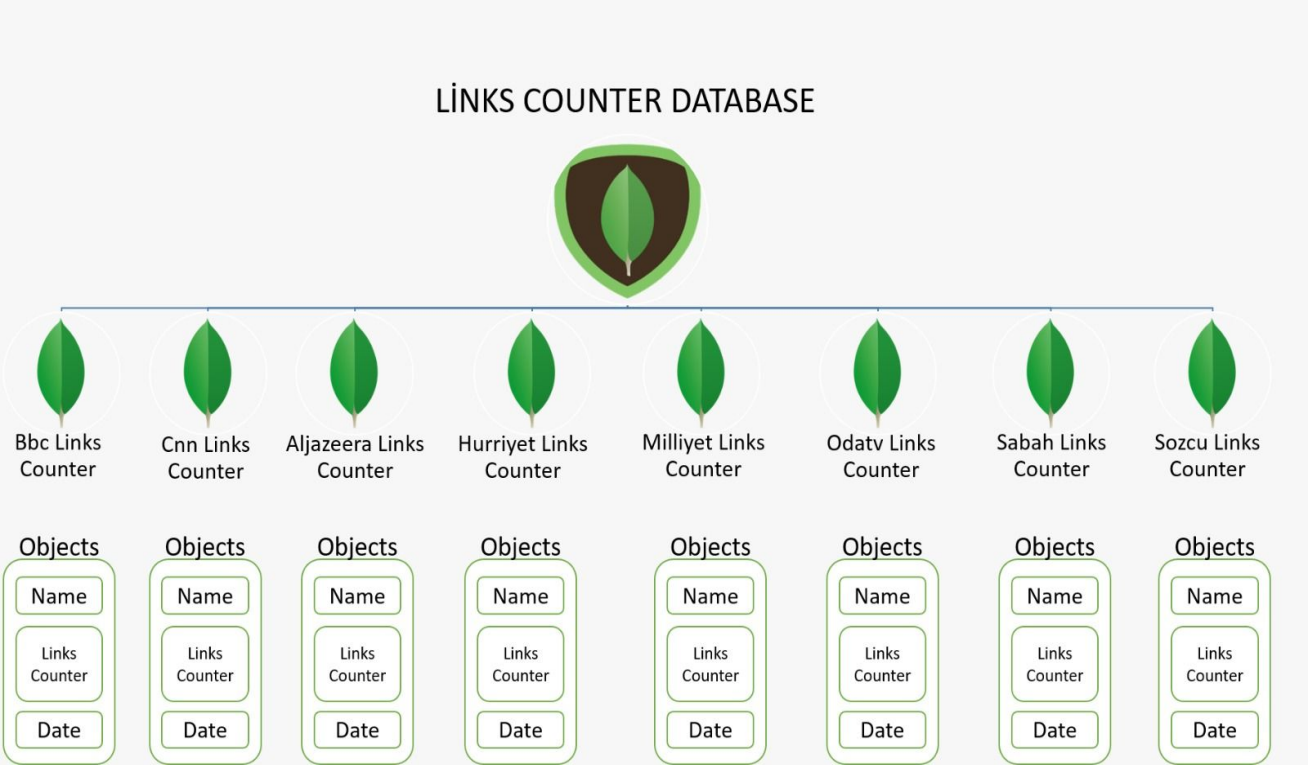
Appendix B



Appendix c



Appendix D


LİNKS COUNTER DATABASE
Bbc Links Counter
Cnn Links Counter
Aljazeera Links Counter
Hurriyet Links Counter
Milliyet Links Counter
Odatv Links Counter
Sabah Links Counter
Sozcu Links Counter
Objects
Name
Links Counter
Date