A MINOR PROJECT REPORT ON

“Web Scraping”

Submitted to

KIIT

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY (KIIT)

Deemed to be University U/S 3 of UGC Act, 1956

In Partial Fulfillment of the Requirement for the Award of

Bachelor of Technology

BY

C V B Shravya	1506119
Manvi Jha	1506139
Pritha Datta Biswas	1506156

UNDER THE GUIDANCE OF

Ms. Monideepa Roy

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

2018-2019

CERTIFICATE

It is certified that the work contained in the project report titled

“Prediction System”

by

C V B Shravya	1506119
Manvi Jha	1506139
Pritha Datta Biswas	1506156

has been carried out under our supervision and that this work has not been submitted elsewhere for a degree.

Date: 13th December 2018

(Ms. Monideepa Roy)

Project Guide

ACKNOWLEDGEMENT

We are profoundly grateful to Ms. Monideepa Roy for her expert guidance and continuous encouragement throughout to see that this project meets its target since its commencement till its completion.

C V B Shravya

Manvi Jha

Pritha Datta Biswas

Abstract

A newspaper is a periodical publication containing written information about current events. A wide variety of topics are covered, including politics, sports, and art. It may also include opinion columns or weather forecasts. With the advancement of technology the newspaper paradigm has shifted from paperback to online newspaper. A digital newspaper is an online version of a newspaper, either as a stand-alone publication or as a digital replica of a printed periodical.

News trend analysis is an approach to use computational methods to find the trending news by tallying the news headlines of various online newspapers. Several sectors such as advertising, journalism, business, and product manufacturing can benefit from it. Natural Language processing (NLP) is widely used to implement it. In this project, we gathered data from different newspaper websites such as inshorts.com and timesofindia.com. Using NLP, we analyzed the headlines of different newspapers to find the trending news that appears most frequently across different headlines.

Keywords:

Trend Analysis; News Headlines; Natural Language processing

CONTENTS

Chapter 1

Introduction

Trend analysis is the process of trying to look at current trends in order to predict future ones and is considered a form of comparative analysis. This can include attempting to determine current news trends as well as whether a trend in one market area could result in a trend in another. Though an analysis may involve a large amount of data, there is no guarantee that the results will be correct.

News trend analysis is an approach to use computational methods to find the trending news by tallying the news headlines of various online newspapers. The data is extracted from various newspapers available online. In the long run, news trend analysis will help us to identify and guess the future occurrences of various events. For analyzing the trends in the newspaper, we apply techniques like web scraping and NLP (Natural Language Processing).

1.1 Web Scraping

The need and importance of extracting data from the web is becoming increasingly loud and clear. So, we use web scraping for extracting the data. Web scraping is a computer software technique of extracting information from websites. This technique mostly focuses on the transformation of unstructured data (HTML format) on the web into structured data (database or spreadsheet). Web scraping follows a process where we need to create a mechanism to receive HTML code with a GET request. Next, inspect the DOM structure (Document Object Model) of the website to identify the nodes containing the target data. After that, create a node processor to output the data in a normalized format. The choice of format is usually based on the requirements or data processing preferences. For example, JSOUP and JAUNT.

1.1.1 Legality of Web Scraping

Web scraping is not considered to be illegal. One can scrap their own websites without any problem. The problem arises when one scrapes the website of others with their prior permission. Under many factors the owner of the site has the right to pursue a legal action against the subject who is scraping the web page. People can be sued under:

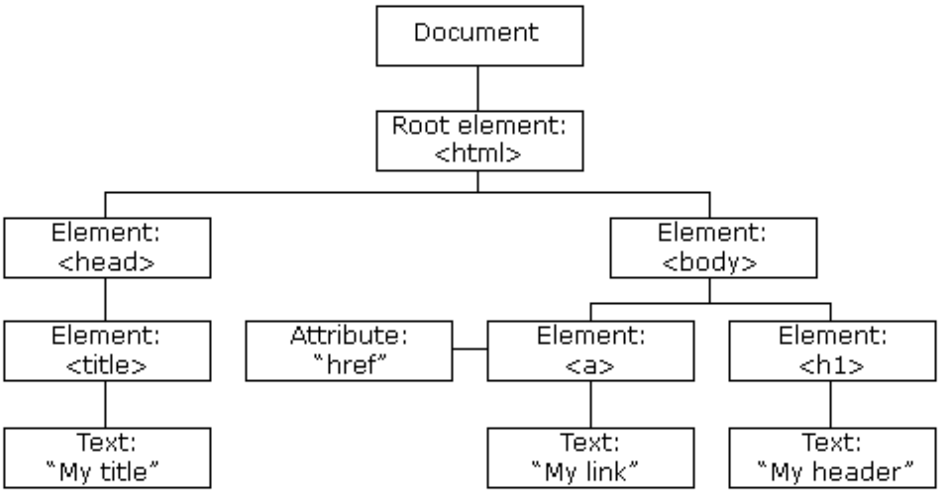
- Violation of the Computer Fraud and Abuse act
- Breach of Contract
- Trespass
- Misappropriation

1.1.2 Techniques Used for Web Scraping

Web Scraping is still considered to be a developing area. It depends mostly on the already existing applications rather than complicated breakthroughs and knowledge work. Different methods used in Web Scraping technique are-

- **Copy-Pasting** –
  It is a manual technique used by humans which provides an error free method of extracting contents. It is very handy for sites with great barriers against machine automated processes

- **DOM Parsing** –
  For inspection of a page dynamically the contents are parsed into a DOM (Document Object Model) tree. Using various codes elements and information can be retrieved from the tree



- **HTTP Programing** – Web contents can be scrapped even using HTTP requests. It can retrieve dynamic as well as static information from the web page

- **Web Scraping Software**- One can also use pre-provided software for scrapping contents of the web if he or she is not wishing to do it manually

- **Text Grepping** – One can also extract contents from the web by using Python programing or Perl along with UNIX Grep Commands

1.2 NLP (Natural Language Processing).

NLP can be used to interpret free text and make it analyzable. There is a tremendous amount of information stored in databases, like news reports of various newspapers, for example. Prior to learning-based NLP models, this information was inaccessible to computer-assisted analysis and could not be analyzed in any kind of systematic way. But NLP allows analysts to access such huge amounts of data from databases to find relevant information or trending topics.

Chapter 2

Real World Application

2.1 Web scraping Application.

2.1.1 Real Estate Listings gathering

It is a huge and growing web scraping area. This is an area where the businesses are using web scraping to gather already listed properties.

2.1.2 Email Address gathering

This is used by a lot of companies. The main purpose of this is lead generation. Once the emails are collected, bulk emails are sent.

2.1.3 Product review scrapes

The reason why many companies use it is so that they can keep an eye on their competitors.

2.1.4 Scraping to create other websites.

The purpose is to get similar data from different websites and then post all that data into one.

2.1.5 A lot of social media companies

Collecting data from different social media websites, what’s trending and what’s in. Knowing what’s trending and what’s in.

2.1.6 Getting massive amounts of data for research purposes.

This could be scraping of government websites or other websites for stats, general information and such.

2.1.7 Specific task scraping / One time scraping.

This is when you need data from a particular website for a very specific purpose just one time.

2.2 Natural Language Processor Applications

By combining the power of **artificial intelligence, computational linguistics and computer science, Natural Language Processing (NLP) helps machines read text** by simulating the human **ability to understand language**.

2.2.1 Automatic summarization

To provide an overview of a news item or blog posts, while avoiding redundancy from **multiple sources**

2.2.2 Sentiment analysis

The goal of sentiment analysis is to identify sentiment among several posts or even in the same post where emotion is not always explicitly expressed. Companies use natural language processing applications, such as sentiment analysis, to identify opinions and sentiment online to help them understand what customers think about their products and services

2.2.3 Text classification

Text classification makes it possible to assign predefined categories to a document and organize it to help you find the information you need or simplify some activities. For example, an application of text categorization is spam filtering in email.

2.2.4 Machine Translation

Machine translation helps us conquer language barriers that we often encounter by translating technical manuals, support content or catalogs at a significantly reduced cost. The challenge with **machine translation technologies is not in translating words, but in understanding the meaning** of sentences to provide a true translation.

2.2.5 Speech Recognition

Speech recognition is the problem of understanding what was said. Given an utterance of text as audio data, the model must produce human readable text.

Chapter 3

Trending News Analysis

For the given topic we have analyzed the problem statement. After analyzing we have found out the approach for this topic and the problem statement. We have found out basic codes, methods and platforms that are required for the implementation of the problem.

3.1 Java

Java is a general-purpose programming language. It is based on the concept of classes and one of its main features is that it is object oriented. The characteristics of Java Programming Language are

- Write Once Run Anytime
- Platform Independent
- Object Oriented.

Java Uses object - oriented concepts such as

- Abstraction
- Encapsulation
- Polymorphism
- Message passing

These concepts make java programming easier and very friendly. Java uses the JVM (Java Virtual Machine) for converting the source code into the byte code

3.2 Java Development Kit (JDK 8)

Java Development Kit is an implementation of Java Platform. It was introduced by the Oracle Corporation. JDK consists of a collection of programming tools. It comes with a complete Java Runtime Environment and the Java Virtual Machine.

3.3 Jsoup

Jsoup is a Java library for working with real-world HTML. It is an API used for manipulating and extracting data using DOM, CSS etc. It parses HTML to DOM. It is used for the following purposes

- Scrape and parse HTML from a URL
- Find and extract data using DOM or CSS selectors

- Manipulate HTML elements

It is used to deal with different varieties of HTML.

3.4 Java.util Package

It contains the collections framework, legacy collection classes, event model, date and time facilities, internationalization, and miscellaneous utility classes. The main usage of java.util is for the creation of the ArrayList to store the elements that have been extracted from the HTML content

3.5 Java WordNet Library

Java Wordnet Library is an API used for accessing WordNet relational dictionaries. The WordNet is majorly used for the development of NLP applications. The use of JWNL makes it easier to build a Java NLP application

3.6 Newspaper Websites

Various newspaper websites are used for extracting the headlines from the respective newspapers. For e.g. – timesofindia, inshorts etc. Analysis is performed on the extracted data of the websites

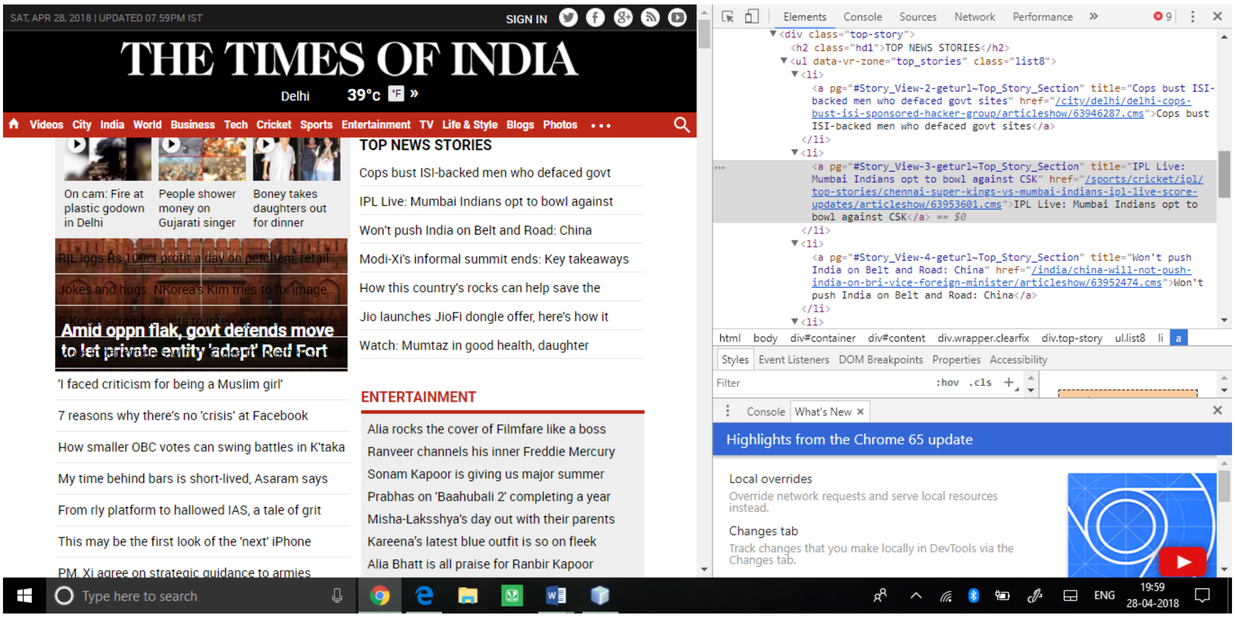
Chapter 4

Process of Analysis

The discussed project uses Java libraries and packages to analyze the news headlines. The purpose of the analysis is to find the maximum number of recurring keywords and give the conclusion. The conclusion provides us with the trending news. The following steps are needed to be followed in order to come up with a conclusion

4.1 Inspecting the web page and locating Data between HTML Source

We need to go through the developers back end HTML code in order to locate the data where we have to function.



SAT, APR 28, 2018 | UPDATED 07.59PM IST

THE TIMES OF INDIA

Delhi 39°c °F »

On cam: Fire at plastic godown in Delhi

People shower money on Gujarati singer

Boney takes daughters out for dinner

RIL logs Rs 100cr profit a day on petchem, retail

Jokes and hugs: NKorea's Kim tries to fix image

S Korea scrambles jets to intercept Chinese plane

Amid oppn flak, govt defends move to let private entity 'adopt' Red Fort

'I faced criticism for being a Muslim girl'

7 reasons why there's no 'crisis' at Facebook

How smaller OBC votes can swing battles in K'taka

My time behind bars is short-lived, Asaram says

From rly platform to hallowed IAS, a tale of grit

This may be the first look of the 'next' iPhone

PM, Xi agree on strategic guidance to armies

TOP NEWS STORIES

Cops bust ISI-backed men who defaced govt

IPL Live: Mumbai Indians opt to bowl against

Won't push India on Belt and Road: China

Modi-Xi's informal summit ends: Key takeaways

How this country's rocks can help save the

Jio launches JioFi dongle offer, here's how it

Watch: Mumtaz in good health, daughter

ENTERTAINMENT

Alia rocks the cover of Filmfare like a boss

Ranveer channels his inner Freddie Mercury

Sonam Kapoor is giving us major summer

Prabhas on 'Baahubali 2' completing a year

Misha-Laksshya's day out with their parents

Kareena's latest blue outfit is so on fleek

Alia Bhatt is all praise for Ranbir Kapoor

```
▼<div class="top-story">
   <h2 class="hd1">TOP NEWS STORIES</h2>
  ▼<ul data-vr-zone="top_stories" class="list8">
    ▼<li>
       <a pg="#Story_View-2-geturl~Top_Story_Section" title="Cops bust ISI-
       backed men who defaced govt sites" href="/city/delhi/delhi-cops-
       bust-isi-sponsored-hacker-group/articleshow/63946287.cms">Cops bust
       ISI-backed men who defaced govt sites</a>
     </li>
    ▼<li>
...    <a pg="#Story_View-3-geturl~Top_Story_Section" title="IPL Live:
       Mumbai Indians opt to bowl against CSK" href="/sports/cricket/ipl/
       top-stories/chennai-super-kings-vs-mumbai-indians-ipl-live-score-
       updates/articleshow/63953601.cms">IPL Live: Mumbai Indians opt to
       bowl against CSK</a> == $0
     </li>
    ▼<li>
       <a pg="#Story_View-4-geturl~Top_Story_Section" title="Won't push
       India on Belt and Road: China" href="/india/china-will-not-push-
       india-on-bri-vice-foreign-minister/articleshow/63952474.cms">Won't
       push India on Belt and Road: China</a>
     </li>
    ▼<li>
```

html body div#container div#content div.wrapper.clearfix div.top-story ul.list8 li a

Highlights from the Chrome 65 update

Local overrides

Override network requests and serve local resources instead.

Changes tab

Track changes that you make locally in DevTools via the Changes tab.

For extracting the web contents from the site of Times of India we need to go through the developer's tool to find the division that needs to be identified in order to extract the headlines. After identifying the division in the HTML source code, we have to carry on with the further functionalities of extracting and analysis of data. For the above case we have identified the div tag named top-story which contains the required data.

4.2 Connecting to the URL of the newspaper

In our project we use jsoup, a java library for real-world HTML. It provides an API for extracting and manipulating data using DOM, CSS and jQuery-like methods. Jsoup.connect(String URL) method is used to fetch and parse a HTML document from the web and find the data contained in it. The connect (String URL) and get () method fetches and parses a HTML file. While fetching and parsing the URL an IOException is thrown which is handled by including the keyword throwsIOException.

For e.g.:

```
Document d=Jsoup.connect("https://timesofindia.indiatimes.com/”). get ();
```

In the above code we are connecting with the site for timesofIndia using Jsoup.connect() and get () and storing the contents in an object d of interface Document

4.3 Scrapping the Data

After XML and HTML content is fetched from the site by the Jsoup.connect(String URL). get () we use a Document object to represent the contents. Document is a predefined interface in Java library whose work is to scale from very simple needs to complex needs. We use the Element interface to store the elements of the XML document. We use the select method to select the div tag whose content has to be scrapped. We further create an ArrayList of Element type to store the elements extracted.

For e.g.: -

In the below code

```
Elements ele=d.select("div.top-story");
ArrayList<Elements> Element=new ArrayList<Elements>();
for(Element element:ele)
    {
      title=element.select(".list8").text();
    }
```

ele is an object of the Element interface which is used to store the contents of an XML file in the div top-story. We are further storing the data in ele according to the sub-division in the div named .list8 into the array list Element in text format by using the method .text().

4.4 Removing stop words

Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called stop words. The general strategy for determining a stop list is to sort the terms by collection frequency (the total number of times each term appears in the document collection), and then to take the most frequent term. There is no definite list for stop words in a language. The headlines extracted from the site contain stop words such as nouns, verbs, pronouns and can cause a hindrance in the logic for calculating the trending news. Hence the stop words occurring need to be removed. In our project we will define a function where the extracted headline is passed as an argument. The content is examined and the stop words are removed.

4.5 Stemming:

Stemming is a method which reduce the words to its root. For example, two different words such as “accounts”, “account”, “acount”, “accounting”, “accounnnnt” and “accountssss” can be reduced to “account”. Stemming helps to simplify the retrieved information and assists further analyses. For stemming purpose Porter’s algorithm is used.

We further use JWNL i.e. Java WordNet Library. It is an API for accessing Wordnet Relational Library. It is used for matching the synonyms provided in different headlines for example polled and voted.

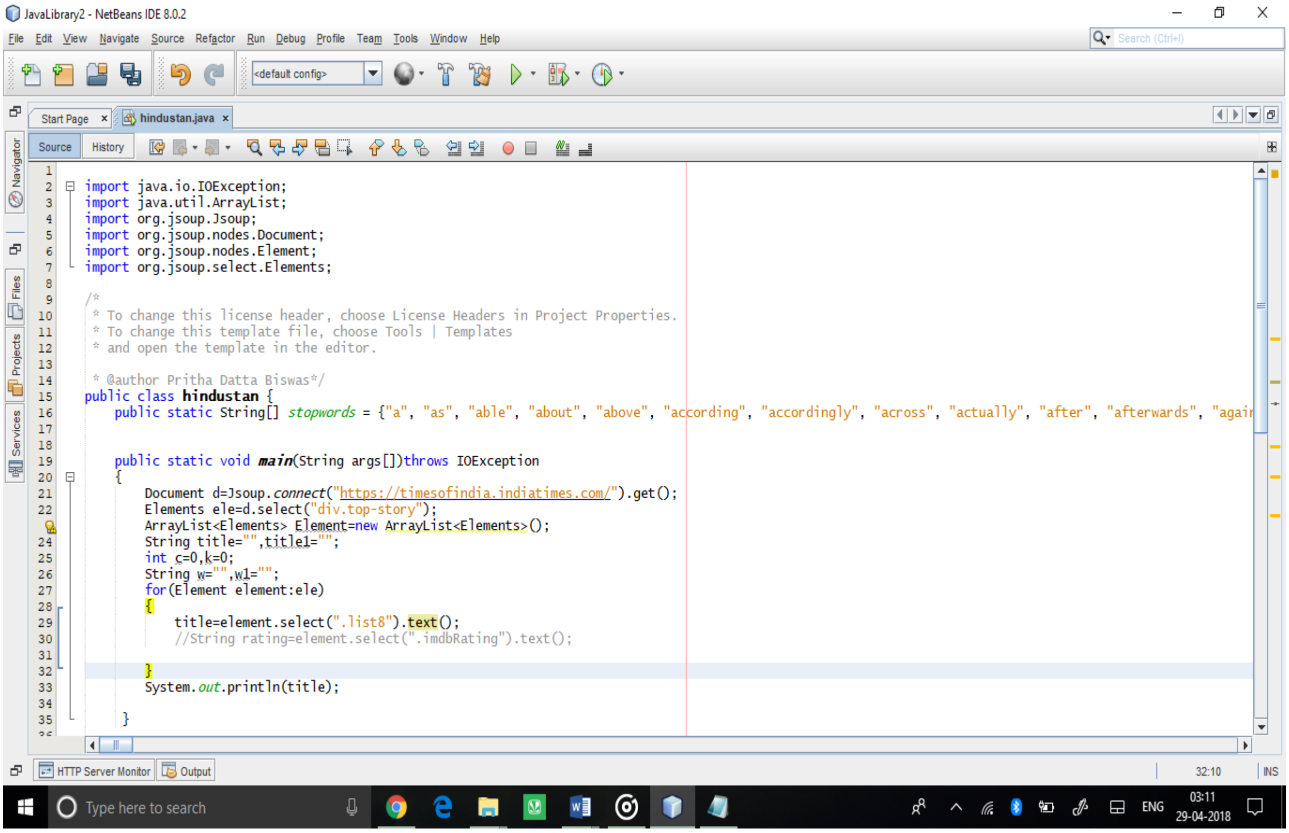
4.6 Output

The output of the following process provides us with the news that has appeared most across different headlines. In short it produces the trending news

Chapter 5

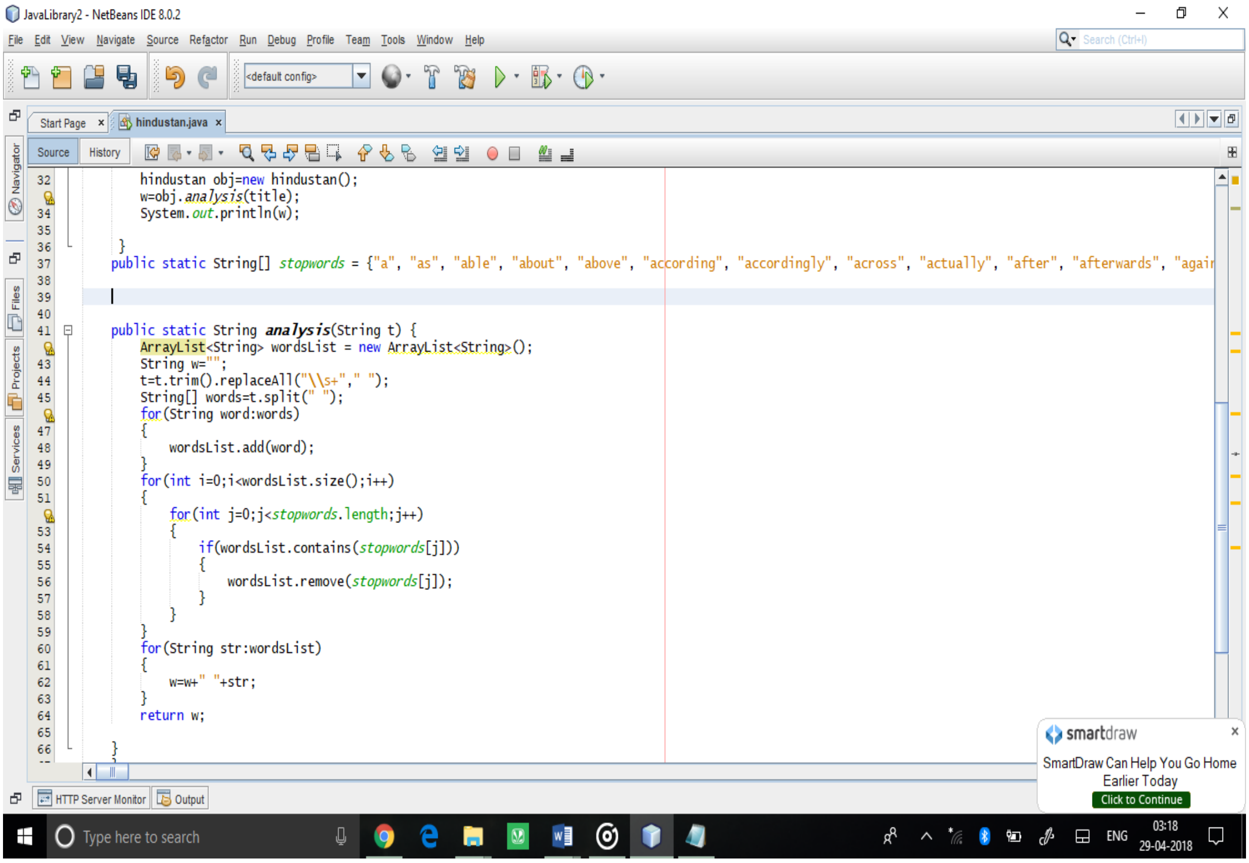
Code Snippets

Code for Extracting the contents of times of India



```java

import java.io.IOException;
import java.util.ArrayList;
import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import org.jsoup.nodes.Element;
import org.jsoup.select.Elements;

/*
 * To change this license header, choose License Headers in Project Properties.
 * To change this template file, choose Tools | Templates
 * and open the template in the editor.

 * @author Pritha Datta Biswas*/
public class hindustan {
    public static String[] stopwords = {"a", "as", "able", "about", "above", "according", "accordingly", "across", "actually", "after", "afterwards", "agair


    public static void main(String args[])throws IOException
    {
        Document d=Jsoup.connect("https://timesofindia.indiatimes.com/").get();
        Elements ele=d.select("div.top-story");
        ArrayList<Elements> Element=new ArrayList<Elements>();
        String title="",title1="";
        int c=0,k=0;
        String w="",w1="";
        for(Element element:ele)
        {
            title=element.select(".list8").text();
            //String rating=element.select(".imdbRating").text();

        }
        System.out.println(title);

    }
```

Code for removing stop words



```java
        hindustan obj=new hindustan();
        w=obj.analysis(title);
        System.out.println(w);

    }
    public static String[] stopwords = {"a", "as", "able", "about", "above", "according", "accordingly", "across", "actually", "after", "afterwards", "agai


    public static String analysis(String t) {
        ArrayList<String> wordsList = new ArrayList<String>();
        String w="";
        t=t.trim().replaceAll("\\s+"," ");
        String[] words=t.split(" ");
        for(String word:words)
        {
            wordsList.add(word);
        }
        for(int i=0;i<wordsList.size();i++)
        {
            for(int j=0;j<stopwords.length;j++)
            {
                if(wordsList.contains(stopwords[j]))
                {
                    wordsList.remove(stopwords[j]);
                }
            }
        }
        for(String str:wordsList)
        {
            w=w+" "+str;
        }
        return w;

    }
```

Chapter 6

Conclusion

Natural-language processing (NLP) is an area of computer science concerned with the interactions between computers and human (natural) languages, in particular how to program computers to fruitfully process large amounts of natural language data. For news trend analysis we have applied NLP and web scraping. Web Scraping is a technique whereby you extract data from website content. It is a form of copying, in which specific data is gathered and copied from the web. We have implemented web scraping using jsoup library. jsoup is an open-source Java library designed to parse, extract, and manipulate data stored in HTML documents. Using jsoup we have extracted headlines of different news websites and removed the stop words. This gives us the trending news.

Our heartfelt appreciation goes to **Monideepa Roy ma'am** with regards to her feedback across the course of our project from the initial phase to the conclusion and for the valuable lessons learned along the way including collaboration within a group and the challenges involved in a large-scale software development effort.

Chapter 7
Future Scope

We have covered the details of our project in the case study part. But, there are few features that we want to extend. Presently, we are doing the trend analysis on the headlines of various newspapers but in future we will try to get the trends from the complete article as well. In order to do so we will use the concept of Text Summarizer. A text summarizer provides a single line summary of an entire article. It is also an implementation of NLP (Natural Language Processing). The text summarizer analyzes the whole article and searches for the most frequently occurring word and provides a summary accordingly. We will be implementing NLP using the Java platform only. With the help of this summary it will be easier to find the trending news. This will help us get accurate news and reduce redundancy.

We will also divide the trending according to different categories i.e. political, international, entertainment, etc. The features will be implemented with the help of text summarizer.

Chapter 8

Bibliography

https://en.wikipedia.org/wiki/Web_scraping

https://www.techopedia.com/definition/5212/web-scraping

https://www.w3schools.com/js/js_htmldom.asp

https://www.udemy.com/python-master-web-scraping-course-doing-20-real-projects/

https://www.youtube.com/watch?v=CDXOcvUNBaA

https://sourceforge.net/projects/jwordnet/

http://www.ign.com/wikis/general-techniques-used-for-web-scraping

https://benbernardblog.com/web-scraping-and-crawling-are-perfectly-legal-right/

http://chianti.ucsd.edu/svn/csplugins/trunk/soc/layla/WordCloudPlugin/trunk/WordCloud/src/cytoscape/csplugins/wordcloud/Stemmer.java

https://www.investopedia.com/terms/t/trendanalysis.asp