CONTENT RECOMMENDATION FOR ARTICLES

A Report

Submitted by

Akshat Sharma (13BCE1009)

Antariksh Narain (13BCE1017)

CSE-326
Internet and Web Programming

School of Computing Science and Engineering

VIT UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
VELLORE ■ CHENNAI
www.vit.ac.in

May– 2016

Abstract

When someone writes an article he/she requires some related text, facts and figures to make the article rich and interesting to read.

Our project takes the users written article as input and processes it to find relevant keywords. These keywords are searched in the existing dataset and the summarized text is extracted. If the word is not available in the dataset a web crawler is used to search the web and get relevant summarized text.

The processed data is then presented to the user for his/her use.

Platform

The project requires Natural Language Processing Tool Kit for processing the given piece of text. It also needs Beautiful Soup, a package used for processing html pages and extracting useful text. These packages are well defined in Python, so the project is built on Python platform. As Python is platform independent language the project can run on all the available platforms.

Hardware Specification:

- AMD A10 Or Intel i5 (2.6GHz) Dual Core
- 4 GB RAM
- 1 GB Free Space in HDD

Software Specification:

- Python 2.7.4
    - Natural Language Processing Tool Kit
    - Beautiful Soup
    - Goose Parser

Analysis

Functional Requirements:

- Business Rules
    - Sentence Tokenizer
        - Tokenizes input text into sentences based on full stops. Should be able to differentiate between abbreviations, salutations and delimiters.
    - Word Tokenizer
        - Tokenizes input sentences into words.
    - Parts of Speech tagger
        - Tags each word in every sentence as a part of speech.
        - Should work with atleast 85% accuracy.
    - Summarizer
        - Should take as input entire article and return a summarized version of it.
        - The summary should contain important words and should use tf-idf weights.
    - Chunker
        - Should extract important entities from text using text relevant grammar.
    - Inverted Index
        - Should store dictionary with keys as chunked words and values as lists of files containing the word.

Non-Functional Requirements:

- Performance
    - The application should respond within a second or two.
    - The application should not lag while being used.
- Scalability
    - The application should be scalable to larger datasets.
    - The software must be modelled for ease of remodelling as a distributive system.
- Capacity
    - The application should be able to handle articles as large as 150KB (~25000 words) in less than 2 seconds.
- Availability
    - The application should be available at all times.
- Reliability
    - The output of the program should be as relevant as possible.
- Environmental
    - The program should keep continuously updated the training data.
- Data Integrity
    - The output should be in UTF-8 encoding format.
- Usability
    - The output should be well formatted and easily understandable.

- Interoperability
    - The program should provide API's for interoperability access to other platforms.

Scope

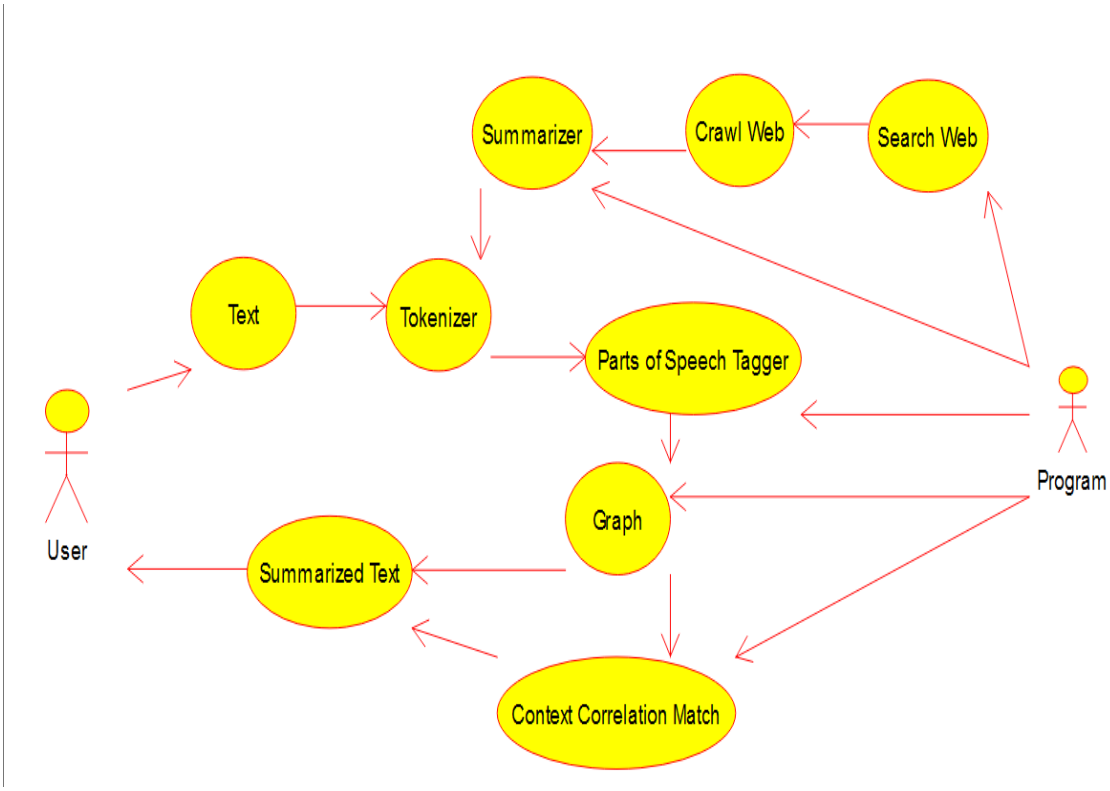
The project can easily be expanded to work on larger training sets and is ever expanding because of the presence of the crawler. By improving retrieval methods from the inverted index to retrieve specific facts and information, the data abstraction so achieved may be used as a base for a question answering machine. Also by providing some means of audio input and output the same program may be used as a speech assistant.

Feasibility

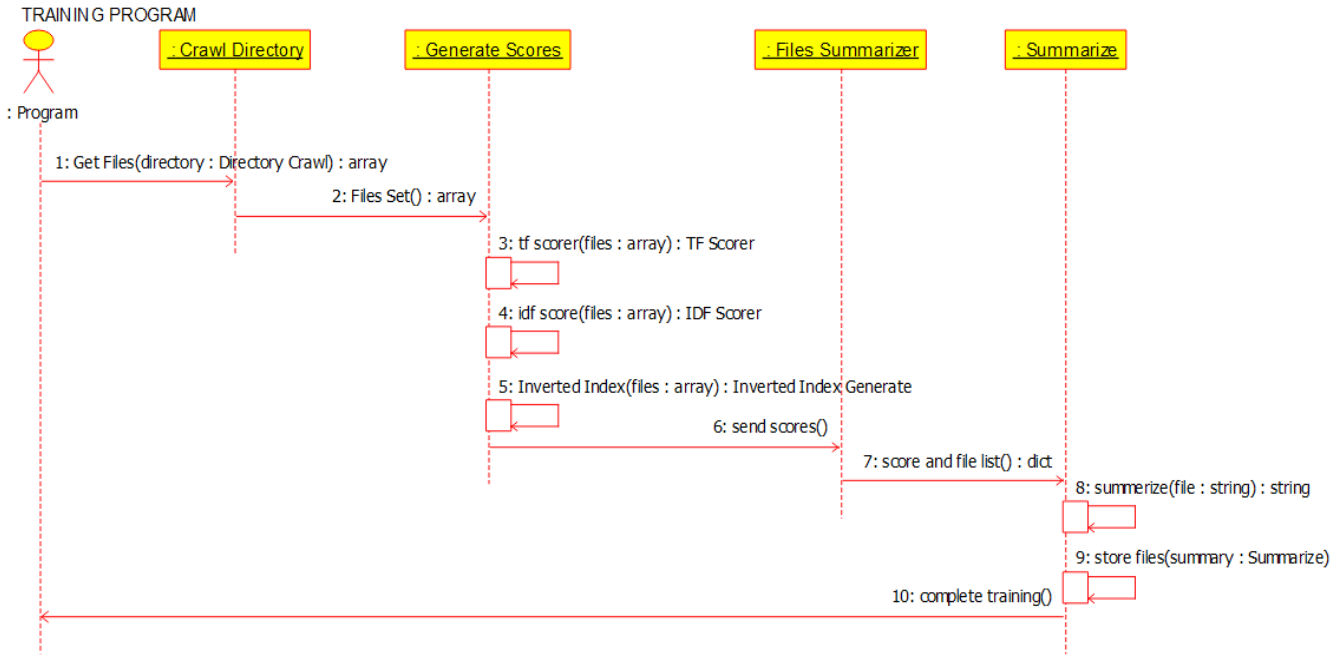
The project is feasible in all senses. It's similarity with a search engine can't be denied but it is in fact very different. First of all it works on articles and not just keywords. This is different because a search engine takes unique keys into consideration whereas our application looks at the term frequency of each word since it gives a lot of insight into the context of the article. Secondly our application is capable of working offline and makes a queue of missing keywords whose articles are fetched and stored when the system is online.

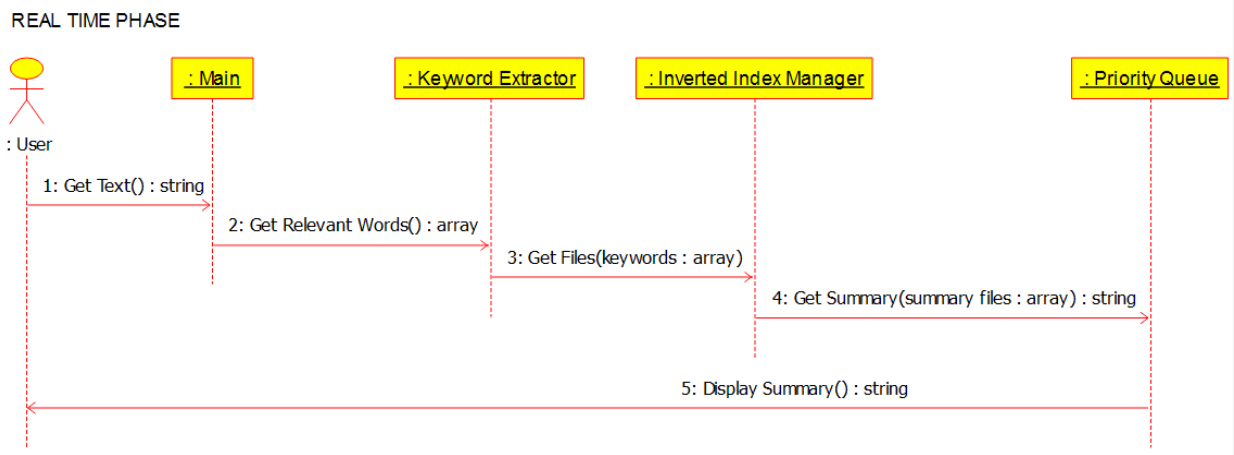
Design

UseCase Diagram

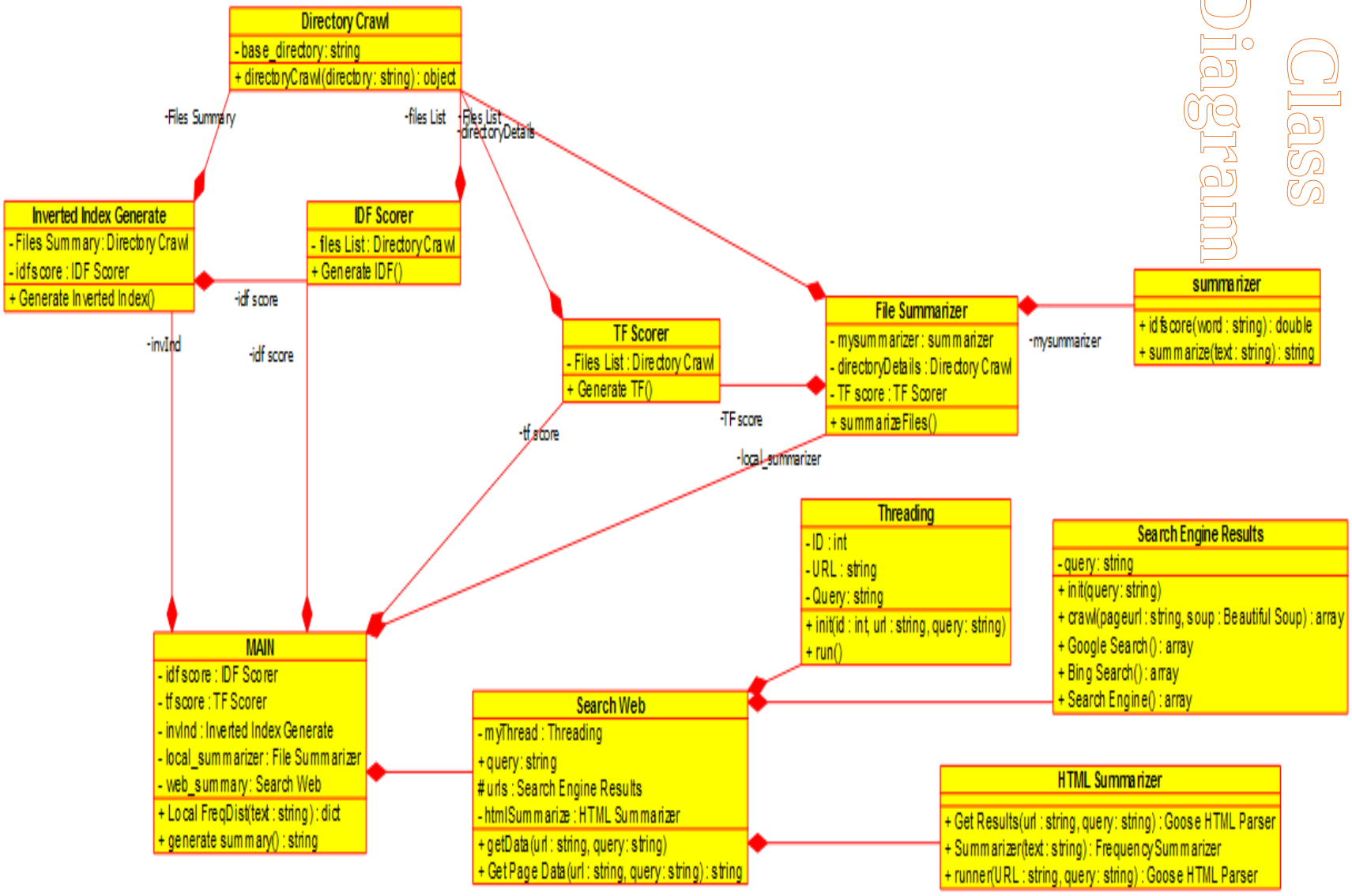


Sequence Diagram





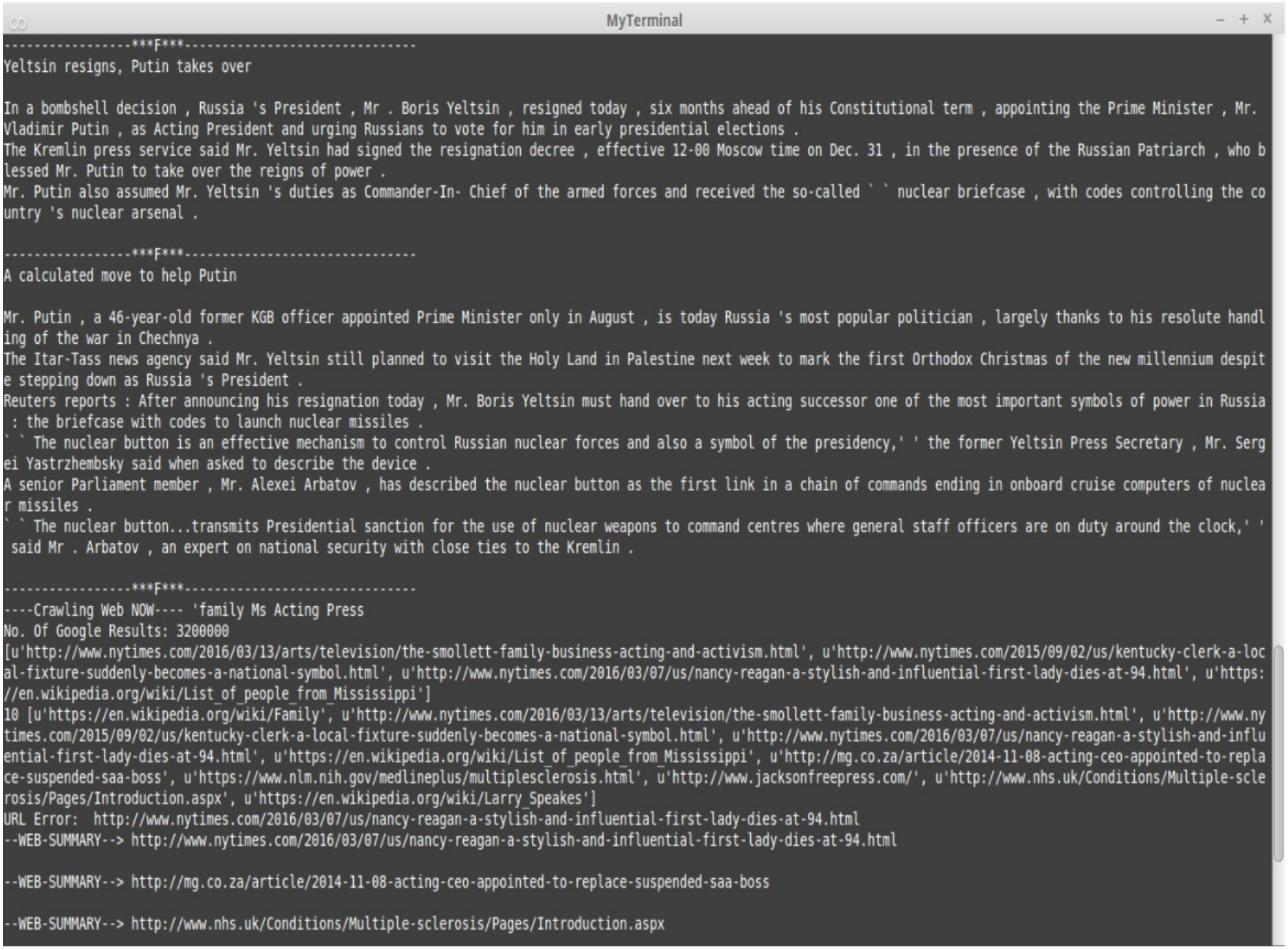
Class Diagram



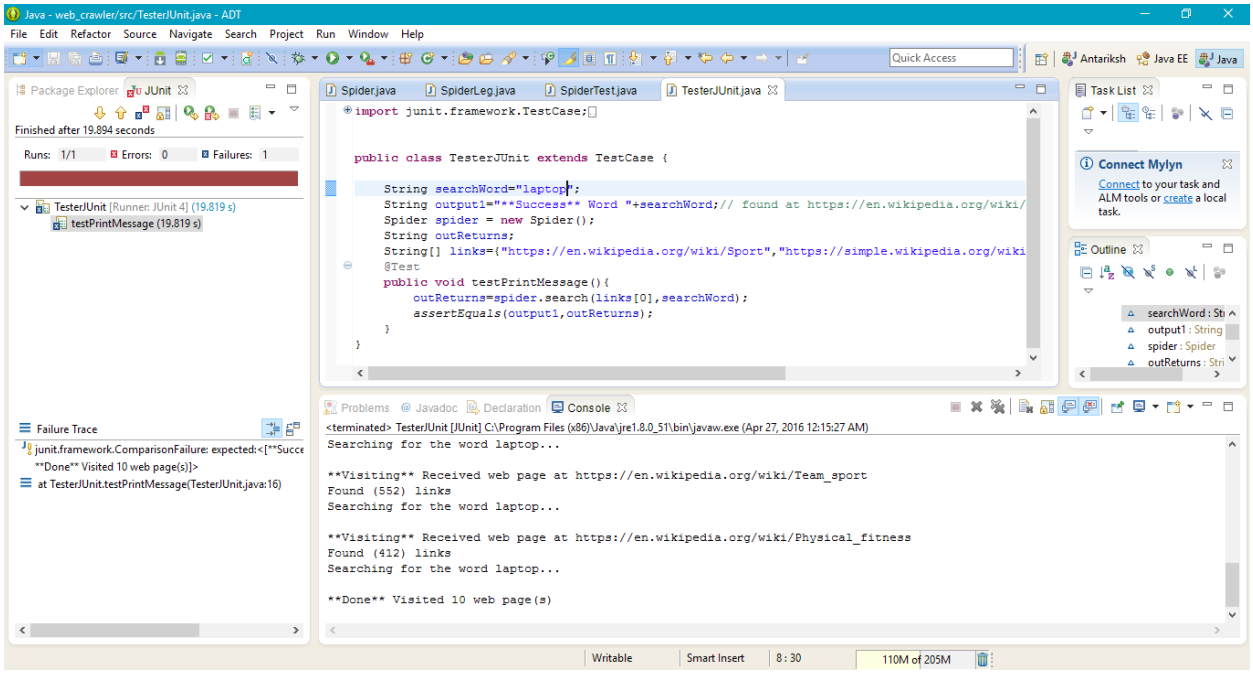
Implementation

MyTerminal

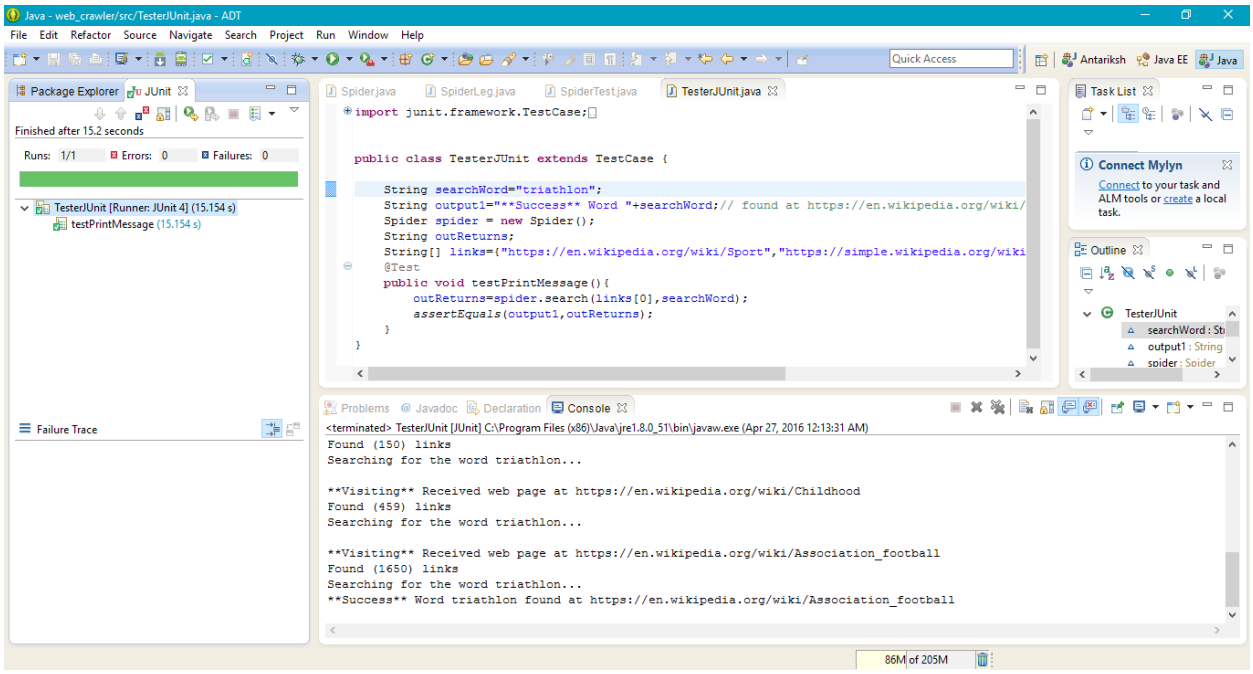
```
antariksh@antariksh-Lenovo-G580 ~/Python/Project/FINAL $ python main.py
INPUT FILE
Yeltsin's daughter sacked; Kremlin 'family' still in control

In a high-profile decision , Mr. Putin on Tuesday sacked Ms . Tatyana Dyachenko , younger daughter of the retired President , Mr . Boris Yeltsin , who had served as her
 father 's image-maker and adviser .
In a series of other changes , clearly designed to distance himself from the scandal-ridden Kremlin administration , the Acting President dismissed Mr. Dmitry Yakushkin
 , Mr. Yeltsin 's Press Secretary , Mr. Vladimir Shevchenko , chief of the Kremlin protocol sevice , and Mr. Vladimir Semenchenko , director of the President 's chancel
lery .
The two officials , together with Ms. Dyachenko , formed the core of the infamous ` ` family,' ' a shadowy clan of top Kremlin advisers and politically connected tycoon
s , who ruled Russia from behind the stage for most of Mr. Yeltsin 's second term .
Ms. Dyachenko 's main role in the Kremlin was to convey to Mr . Yeltsin recommendations of the ` ` family' ' in such a way that the fiercely independent ex-President wo
uld not reject them outright .
Mr. Mikhail Gorbachev , the last Soviet President , said in an interview on Tuesday that Ms. Dyachenko and other members of the ` ` family' ' were responsible for persu
ading Mr. Yeltsin to resign last Friday .
` ` Dyachenko does not need a Kremlin office to continue to pull strings as long as Voloshin and Yumashev stay on,' ' a Kremlin watcher remarked .

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
[('/home/antariksh/Python/Project/2000/Summarize/05/0305000j.txt', 262.6953999302655), ('/home/antariksh/Python/Project/2000/Summarize/01/01010007.txt', 89.862428682005
26), ('/home/antariksh/Python/Project/2000/Summarize/01/03010006.txt', 78.73083574858983)]
262.69539993
['Yeltsin', 'Kremlin', "'family", 'Mr.', 'Putin', 'Tuesday', 'Ms', 'Tatyana', 'Dyachenko', 'President', 'Mr', 'Boris', 'Acting', 'Dmitry', 'Yakushkin', 'Press', 'Secret
ary', 'Vladimir', 'Shevchenko', 'Semenchenko', 'Ms.', 'Russia', 'Mikhail', 'Gorbachev', 'Friday', 'Voloshin', 'Yumashev']
Yeltsin's daughter sacked; Kremlin 'family' still in control

In a high-profile decision , Mr. Putin on Tuesday sacked Ms . Tatyana Dyachenko , younger daughter of the retired President , Mr . Boris Yeltsin , who had served as her
 father 's image-maker and adviser .
In a series of other changes , clearly designed to distance himself from the scandal-ridden Kremlin administration , the Acting President dismissed Mr. Dmitry Yakushkin
 , Mr. Yeltsin 's Press Secretary , Mr. Vladimir Shevchenko , chief of the Kremlin protocol sevice , and Mr. Vladimir Semenchenko , director of the President 's chancel
lery .
The two officials , together with Ms. Dyachenko , formed the core of the infamous ` ` family,' ' a shadowy clan of top Kremlin advisers and politically connected tycoon
s , who ruled Russia from behind the stage for most of Mr. Yeltsin 's second term .
Ms. Dyachenko 's main role in the Kremlin was to convey to Mr . Yeltsin recommendations of the ` ` family' ' in such a way that the fiercely independent ex-President wo
uld not reject them outright .
Mr. Mikhail Gorbachev , the last Soviet President , said in an interview on Tuesday that Ms. Dyachenko and other members of the ` ` family' ' were responsible for persu
ading Mr. Yeltsin to resign last Friday .
` ` Dyachenko does not need a Kremlin office to continue to pull strings as long as Voloshin and Yumashev stay on,' ' a Kremlin watcher remarked .

----------***F***----------
Yeltsin resigns, Putin takes over

In a bombshell decision , Russia 's President , Mr . Boris Yeltsin , resigned today , six months ahead of his Constitutional term , appointing the Prime Minister , Mr.
Vladimir Putin , as Acting President and urging Russians to vote for him in early presidential elections .
```


MyTerminal
----------***F***----------
Yeltsin resigns, Putin takes over
In a bombshell decision , Russia 's President , Mr . Boris Yeltsin , resigned today , six months ahead of his Constitutional term , appointing the Prime Minister , Mr. Vladimir Putin , as Acting President and urging Russians to vote for him in early presidential elections .
The Kremlin press service said Mr. Yeltsin had signed the resignation decree , effective 12-00 Moscow time on Dec. 31 , in the presence of the Russian Patriarch , who blessed Mr. Putin to take over the reigns of power .
Mr. Putin also assumed Mr. Yeltsin 's duties as Commander-In- Chief of the armed forces and received the so-called ` ` nuclear briefcase , with codes controlling the country 's nuclear arsenal .
----------***F***----------
A calculated move to help Putin
Mr. Putin , a 46-year-old former KGB officer appointed Prime Minister only in August , is today Russia 's most popular politician , largely thanks to his resolute handling of the war in Chechnya .
The Itar-Tass news agency said Mr. Yeltsin still planned to visit the Holy Land in Palestine next week to mark the first Orthodox Christmas of the new millennium despite stepping down as Russia 's President .
Reuters reports : After announcing his resignation today , Mr. Boris Yeltsin must hand over to his acting successor one of the most important symbols of power in Russia : the briefcase with codes to launch nuclear missiles .
` ` The nuclear button is an effective mechanism to control Russian nuclear forces and also a symbol of the presidency,' ' the former Yeltsin Press Secretary , Mr. Sergei Yastrzhembsky said when asked to describe the device .
A senior Parliament member , Mr. Alexei Arbatov , has described the nuclear button as the first link in a chain of commands ending in onboard cruise computers of nuclear missiles .
` ` The nuclear button...transmits Presidential sanction for the use of nuclear weapons to command centres where general staff officers are on duty around the clock,' ' said Mr . Arbatov , an expert on national security with close ties to the Kremlin .
----------***F***----------
----Crawling Web NOW---- 'family Ms Acting Press
No. Of Google Results: 3200000
[u'http://www.nytimes.com/2016/03/13/arts/television/the-smollett-family-business-acting-and-activism.html', u'http://www.nytimes.com/2015/09/02/us/kentucky-clerk-a-local-fixture-suddenly-becomes-a-national-symbol.html', u'http://www.nytimes.com/2016/03/07/us/nancy-reagan-a-stylish-and-influential-first-lady-dies-at-94.html', u'https://en.wikipedia.org/wiki/List_of_people_from_Mississippi']
10 [u'https://en.wikipedia.org/wiki/Family', u'http://www.nytimes.com/2016/03/13/arts/television/the-smollett-family-business-acting-and-activism.html', u'http://www.nytimes.com/2015/09/02/us/kentucky-clerk-a-local-fixture-suddenly-becomes-a-national-symbol.html', u'http://www.nytimes.com/2016/03/07/us/nancy-reagan-a-stylish-and-influential-first-lady-dies-at-94.html', u'https://en.wikipedia.org/wiki/List_of_people_from_Mississippi', u'http://mg.co.za/article/2014-11-08-acting-ceo-appointed-to-replace-suspended-saa-boss', u'https://www.nlm.nih.gov/medlineplus/multiplesclerosis.html', u'http://www.jacksonfreepress.com/', u'http://www.nhs.uk/Conditions/Multiple-sclerosis/Pages/Introduction.aspx', u'https://en.wikipedia.org/wiki/Larry_Speakes']
URL Error: http://www.nytimes.com/2016/03/07/us/nancy-reagan-a-stylish-and-influential-first-lady-dies-at-94.html
--WEB-SUMMARY--> http://www.nytimes.com/2016/03/07/us/nancy-reagan-a-stylish-and-influential-first-lady-dies-at-94.html
--WEB-SUMMARY--> http://mg.co.za/article/2014-11-08-acting-ceo-appointed-to-replace-suspended-saa-boss
--WEB-SUMMARY--> http://www.nhs.uk/Conditions/Multiple-sclerosis/Pages/Introduction.aspx

Testing



J Unit Failure



J Unit Success

Software Testing Types

Black box testing

Internal system design is not considered in this type of testing. Tests are based on requirements and functionality.

White box testing

This testing is based on knowledge of the internal logic of an application's code. Also known as Glass box Testing. Internal software and code working should be known for this type of testing. Tests are based on coverage of code statements, branches, paths, conditions.

Unit testing

Testing of individual software components or modules. Typically done by the programmer and not by testers, as it requires detailed knowledge of the internal program design and code. may require developing test driver modules or test harnesses.

We have used Unit testing to test our web crawler as our program generates dynamic results. For a given link and a query, the program crawls through the web to search for the given word and returns positive if the word is found else negative. Using J-Unit we test the current result with the previous result and check the validity of the system.