Information Retrieval Programming Task 2

Team IR Coders :: Shikha Mehta , Rahul Kodarapu ,Chao Yan & Abdullah Al Zubaer

January 13, 2018

Abstract

The report contains the documentation of the working project i.e., Information Retrieval System Programming Task 2. It provides a brief description of all the packages in the project.

1 Overview:

- The Project crawls the seed URL, all the URL links found on the seed URL and also the links found on the subsequent links depending on the crawl depth provided by the user, while simultaneously fetching the body text from individual url and creating a local text file in the default 'public documents folder path' to store that text. Then the project performs stemming, indexing on the text files and saves the index files in the index folder provided by the user. Then the project creates a query object and compares it with the index files . Then it prints the top 10 search url links along with the other relevant information.

2 Implementation:

- Environment: Java JDK and JRE version 8.
- Libraries:
  - (1) Apache Lucene 7.1.0,
  - (2) JSoup 1.11.2.

2.1 Package: com.irpt2.main

- **StartInformationRetrieval.java:** This class contains four methods
  - (1) main: This function mainly takes the required parameters as an input from the user, checks them and then calls the methods crawlFirst and executeTask.
  - (2) crawlFirst: This function takes seed url, crawldepth, filepath, and index path as arguments. It crawls the urls, calls the nextUrl function for fetching the next url, gets links from the urls crawled and adds them to the linkedlist, finally creates the local file and adds this url to pages.txt along with its depth.
  - (3) nextUrl: It takes the linkedlist as an argument, checks for the next url, checks if it has already been visited and returns the nexturl back to the crawlFirst function.
  - (4) executeTask: This function takes indexpath, filepath, default ranking model(i.e, Vector Space) and query as arguments. Then it creates an object of the IndexingManager class , calls its startIndexing function. Then it creates an object of the SearchManager class , calls its initiateSearch function.

2.2 Package: com.irpt2.crawler

- **WebCrawler.java:** This class contains four methods namely ,
  - (1) crawl: This function takes url as an argument, connects with it using the **connection class of jsoup**, saves it into a html document, collects all the links on the page using the **Element class of Jsoup** and then adds those links to a linked list.
  - (2) TextInTheWebpage: This method returns the lowercase version of the body text in the html document.
  - (3) Title: This method returns the lowercase version of the title text in the html document.
  - (4) getLinks: This method returns the linkedlist which has all the links found on the url that has been crawled recently.

2.3 Package: com.irpt2.manager

- **IndexingManager.java:** This class contains three methods namely ,
  - (1) startIndexing: This function takes filepath(i.e., the public documents folder),index path given by the user,and default ranking model(i.e, Vector Space) as arguments. Then checks if the folder path is readable and sets up the object of the **English analyser class of Lucene** which does stemming, also sets up the object of the **Indexwriterconfig class of Lucene** which acts like an instructor to the Indexwriter, then sets up the object of the **Index writer class of Lucene** which does the indexing , calls the function IndexFiles, after that it sets up the object of the index reader class and reads the number of indexed files.
  - (2) indexFiles: This method takes the Indexwriter object and the folder path as arguments. Then it checks if it has directories or individual files and then calls the method indexFile.
  - (3) indexFile: This method takes Indexwriter object, and folder path as arguments. Then it does indexing , creates index files in the index folders and adds fields to it.

- **SearchManager.java:** This class contains two methods namely ,
  - (1) initiateSerach: This function takes index path given by the user, default ranking model(i.e, Vector Space) and the query as arguments. Then checks if the there are any special characters in the query. Then it sets up the object of the **IndexReader class of Lucene**, also sets up the object of the **IndexSearcher class of Lucene**, it creates the query object using the object of the **Queryparser class of Lucene** and the analyser and then calls the Search function.
  - (2) search: This method takes the Indexsearcher object and the query object as arguments. Then it sets up the object of the **TopDocs class of Lucene** and searches the index files in the index folder for the relevant documents. Then it sets up the object of the **ScoreDocs class of Lucene** and save the top 10 Topdocs to that. Then it prints the top 10 documents along with the other relevant information like url, title etc..

2.4 Package: com.irpt2.util

- **Utils.java** : This class contains several utility functions which are used in the other classes in several occasions.

2.5 Package: com.irpt2.constants

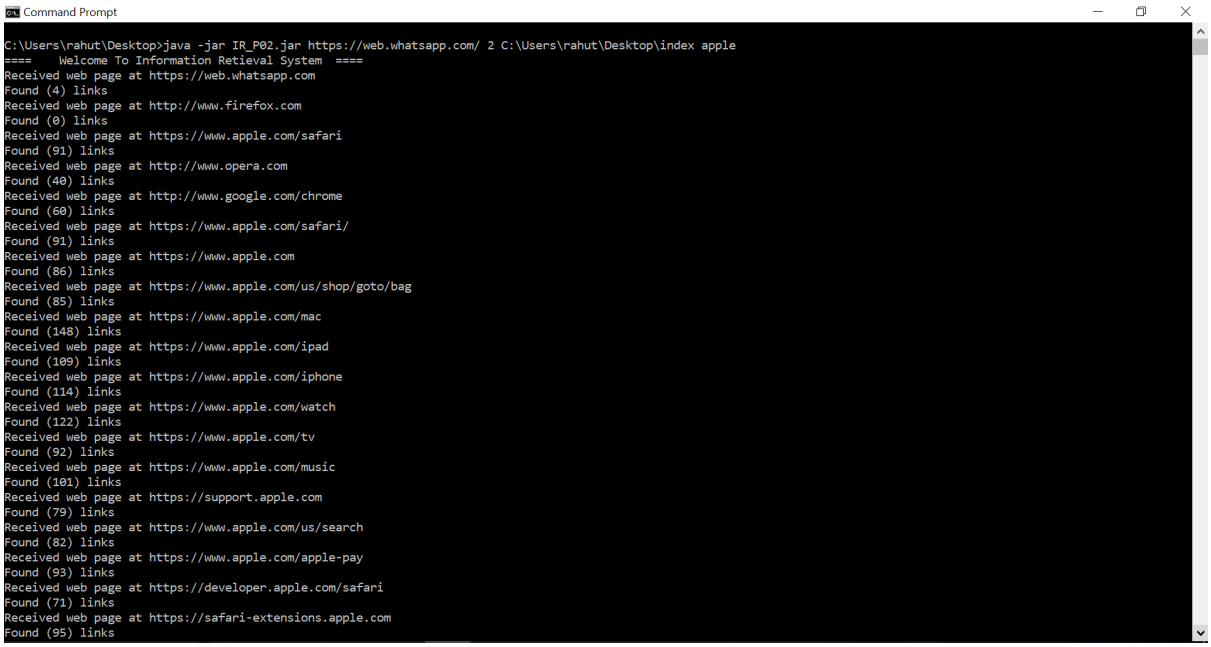
- **Constants.java** : This class contains several constant variables which are used in the other classes in several occasions.

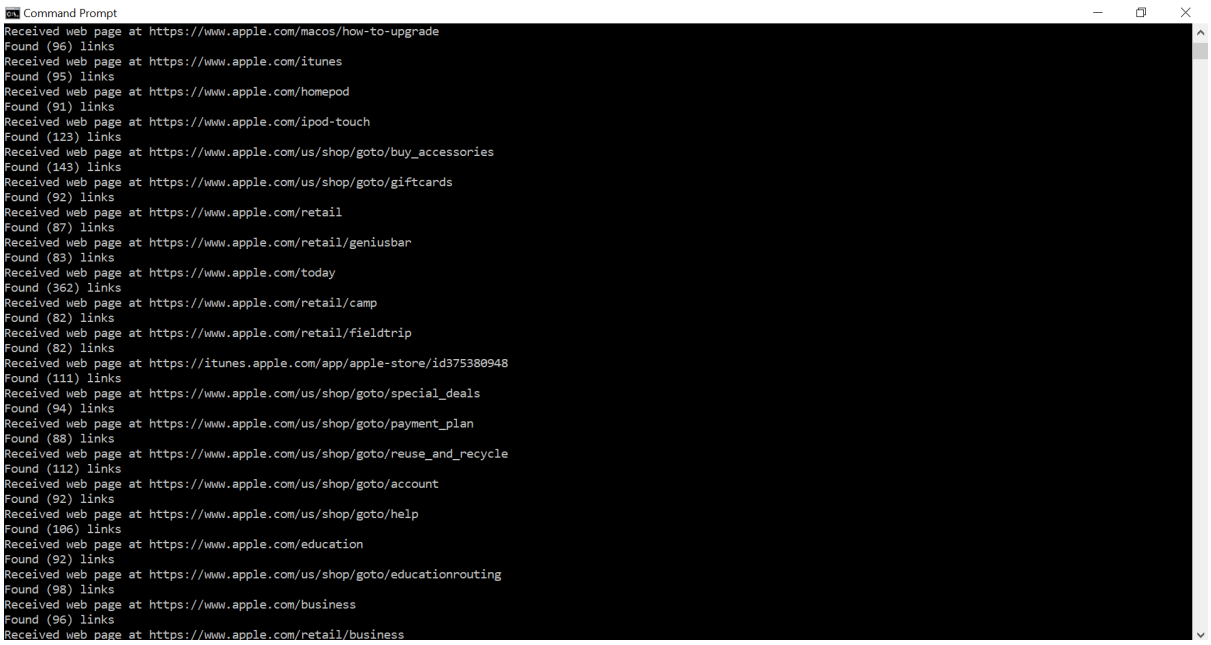
3 Snapshots of the working model :

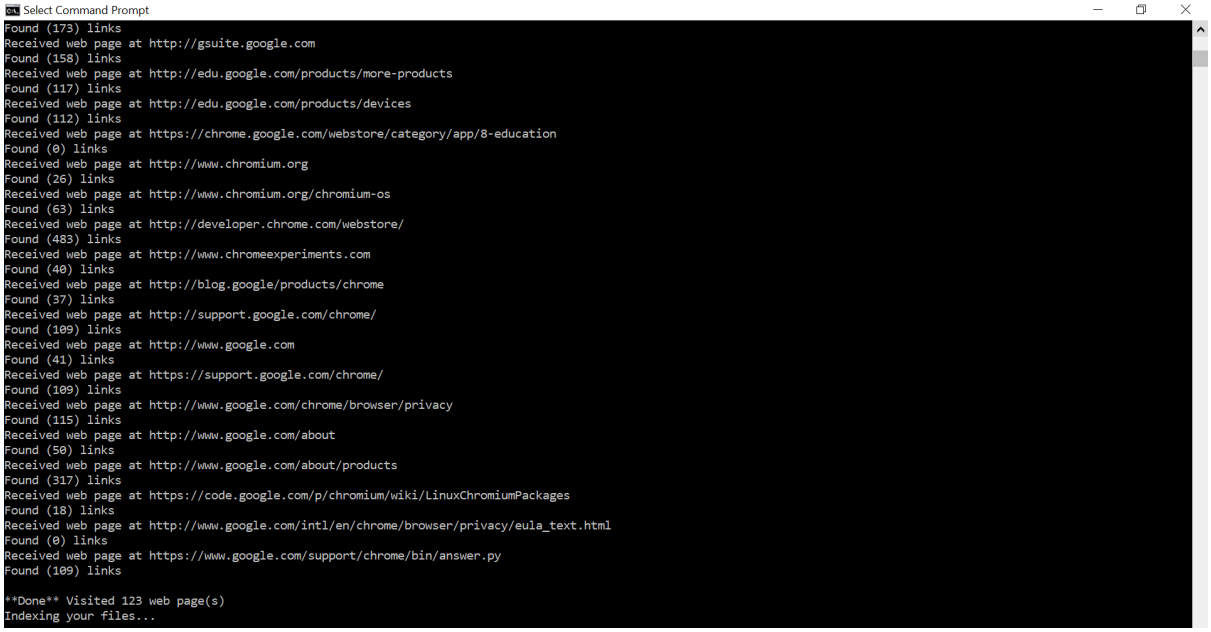
- The following command is supposed to be passed on command prompt in order to implement the program:

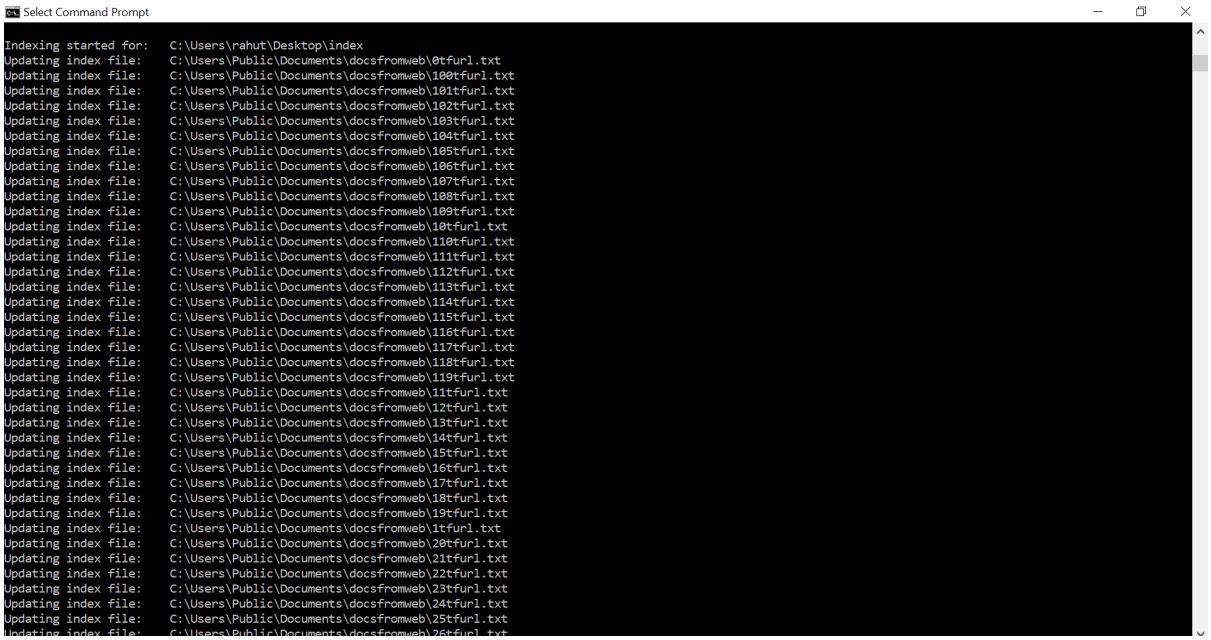
  java -jar IRP02.jar [Seed URL] [Crawl Depth] [path to index folder] [query]

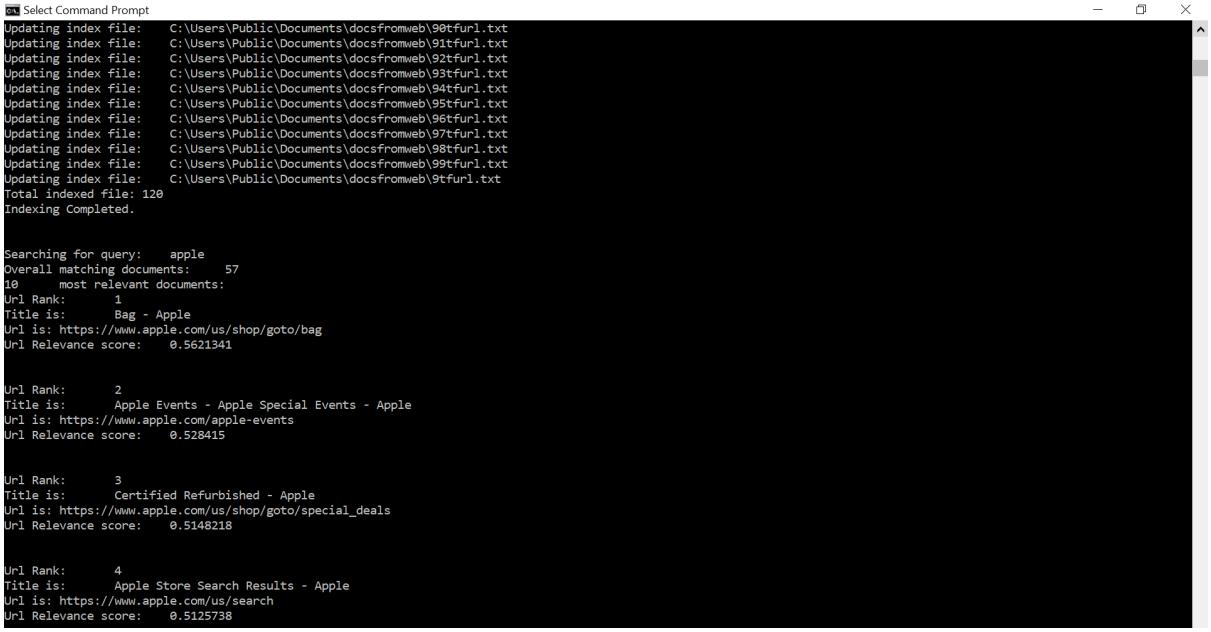
3.1 Snapshots :

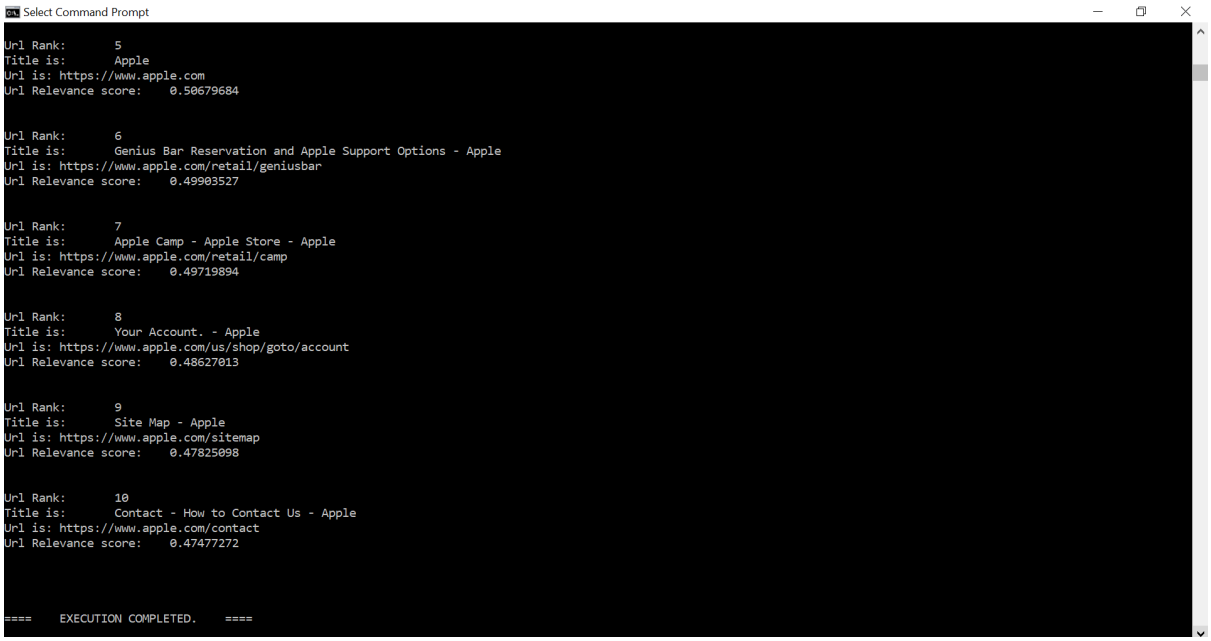





Select Command Prompt
Found (173) links
Received web page at http://gsuite.google.com
Found (158) links
Received web page at http://edu.google.com/products/more-products
Found (117) links
Received web page at http://edu.google.com/products/devices
Found (112) links
Received web page at https://chrome.google.com/webstore/category/app/8-education
Found (0) links
Received web page at http://www.chromium.org
Found (26) links
Received web page at http://www.chromium.org/chromium-os
Found (63) links
Received web page at http://developer.chrome.com/webstore/
Found (483) links
Received web page at http://www.chromeexperiments.com
Found (40) links
Received web page at http://blog.google/products/chrome
Found (37) links
Received web page at http://support.google.com/chrome/
Found (109) links
Received web page at http://www.google.com
Found (41) links
Received web page at https://support.google.com/chrome/
Found (109) links
Received web page at http://www.google.com/chrome/browser/privacy
Found (115) links
Received web page at http://www.google.com/about
Found (50) links
Received web page at http://www.google.com/about/products
Found (317) links
Received web page at https://code.google.com/p/chromium/wiki/LinuxChromiumPackages
Found (18) links
Received web page at http://www.google.com/intl/en/chrome/browser/privacy/eula_text.html
Found (0) links
Received web page at https://www.google.com/support/chrome/bin/answer.py
Found (109) links
Done Visited 123 web page(s)
Indexing your files...


Select Command Prompt
Indexing started for: C:\Users\rahut\Desktop\index
Updating index file: C:\Users\Public\Documents\docsfromweb\0tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\100tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\101tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\102tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\103tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\104tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\105tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\106tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\107tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\108tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\109tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\10tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\110tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\111tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\112tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\113tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\114tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\115tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\116tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\117tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\118tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\119tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\11tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\12tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\13tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\14tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\15tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\16tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\17tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\18tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\19tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\1tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\20tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\21tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\22tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\23tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\24tfurl.txt
Updating index file: C:\Users\Public\Documents\docsfromweb\25tfurl.txt





3.2 Drawbacks:

- This Program only works on a Windows Operating System.
- It might through NullPointerException sometimes,We request you to Please use a different Url in such cases.