**Inverted Index Report**

1. **Definition of documents:**

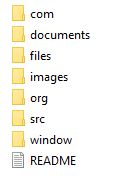
* corpus.txt: Collection of all extracted documents from Google according to queries. These queries are:
* **Query 1: what are the seven wonders of the world**
* **Query 2: how many theaters are in broadway**
* **Query 3: what is the oldest language in the world**
* **Query 4: What is the real meaning of Google in English?**
* **Query 5: where the nasa launch rockets**
* **Query 6: List of the most expensive watches in the world**
* **Query 7: What is Watson**
* **Query 8: How many pictures have the louvre museum**
* **Query 9: how many data server does Facebook have**
* **Query 10: how is mapped the internet in the ocean**

**File location:** InvertedIndex/files

There are 20 documents for each query, so there are 200 documents in total that are saved in corpus.txt file without the query, and each document is separated from other document with an empty line.

* stopWords.txt: There are 262 stop words. **File location:** InvertedIndex/files
* output.txt: prints out queries, results with snippet text provided by the application and computation of recall and precision for each query. **File location:** InvertedIndex/files

**Structure of directories**



* com: it is a directory that contains .class files according to GUI design.
* documents: list of processed documents of corpora (corpus.txt).
* files: contains corpus, stop words and output files.
* images: it contains the icon of the application.
* org: contains package org.information.retrieval. There are document and InvertedIndex classes.
* Org: it contains package org.tartarus.snowball. There are “.classes” files in order to execute the Porter’s algorithm.
* src: it contains all the source code of the application.
* window: contains package window.views. There exists the main function.
* README: Give instructions to run the application.

Therefore, the directories as com, org and windows have the Java bytecodes.

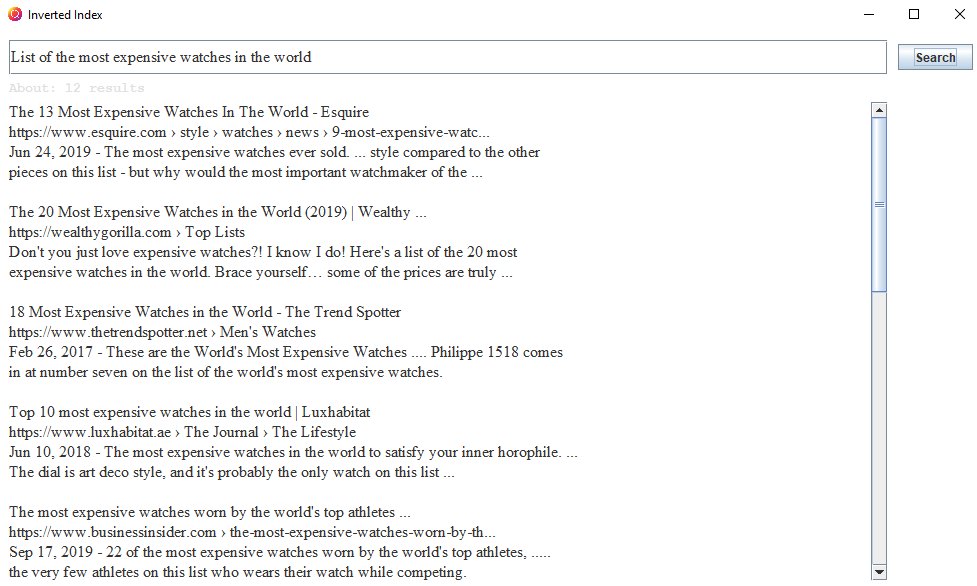
1. **Processing of Inverted Index:** *processCorpus* is the main function located in Document class (inside org.information.retrieval package). This function reads the entire corpus (corpus.txt). It processes each document in the corpus. It is important to know that it can be added more documents to the corpus file in order to test them.

Moreover, each document in corpus.txt is created in a new .txt file in “documents” directory (i.e. doc1.txt, doc2.txt, …, doc200.txt). Then, the function creates an inverted index based on each word (token) of each document, so, the structure of the inverted index is as follows:

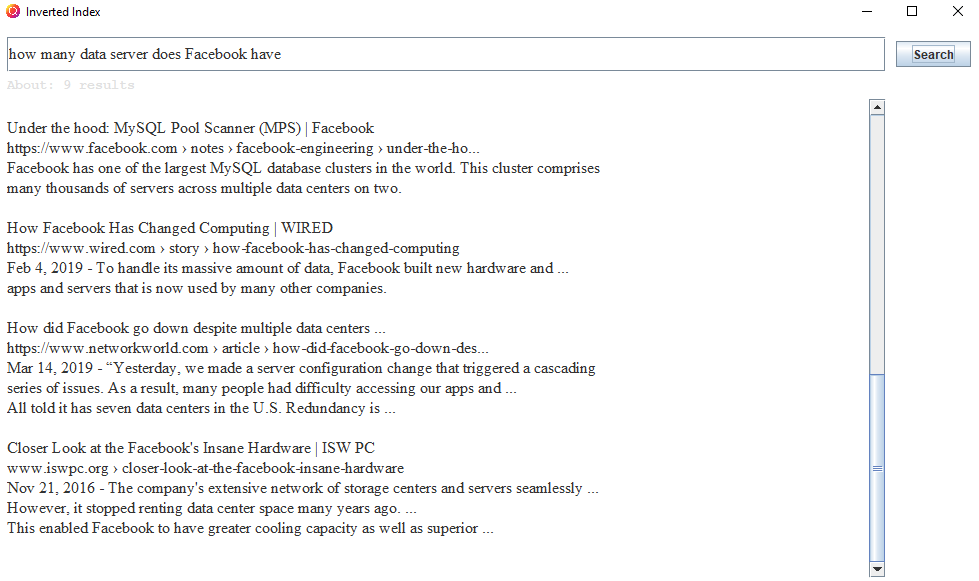
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Keywords \*** |  | **Document where appear each keyword** | | |
| **word1** |  | **1** | **2** | **3** |
| **word2** |  | **3** | **6** | **7** |
| **word3** |  | **8** | **9** |  |
| **…** |  | **…** |  |  |
| **…** |  | **…** |  |  |
| **…** |  | **…** |  |  |
| **wordn** |  | **10** | **14** |  |

**\***At first glance, it is extracted each word or token from each document (tokenization process), theneach keyword (or token) is processed in terms of lowercase, regex and Porter’s algorithm. So, in order to get a stemming word, it was used the Porter’s algorithm where its source code was downloaded from <https://snowballstem.org/download.html>. When a query is filled out in the searching text of application, there exists the same approach in order to match the query with words in the inverted index structure.

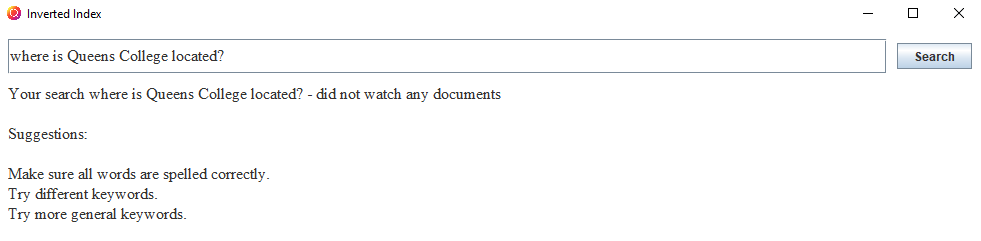
1. **Handle the application:** As Google search engine, we just fill out a query in text field and then press the Search button.

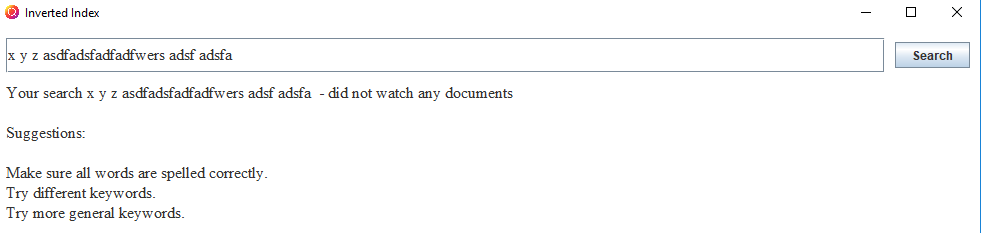


1. **Testing:**
   1. **Filling out one of the 10 queries describes in point number 1**



* 1. **Filling out unknown query documents for the application**





1. **Results:** File output.txt is created in order to show results that application prints out. For instance:

QUERY: what are the seven wonders of the world

﻿The Seven Wonders of the World - WorldAtlas.com  
https://www.worldatlas.com › articles › the-7-wonders-of-the-world  
Sep 11, 2018 - Examples of these wonders include Stonehenge, the Hagia Sophia,  
and the Leaning Tower of Pisa. Taj Mahal - India. Colosseum - Italy. Chichen Itza - Mexico.  
Machu Picchu - Peru. Christ the Redeemer - Brazil. Petra - Jordan. Great Wall of China - China.  
Great Pyramid of Giza (Honorary Candidate) - Completed c. 2560 BC.  
‎Seven Natural Wonders · ‎Seven Wonders of the Modern ... · ‎The 7 Wonders of the ...

...

PRECISION:   
RECALL:

QUERY: …

… snipped document …

PRECISION:

RECALL: