Birla Institute of Technology & Science – Pilani
Second Semester 2014-15

Date: 27.02.2015

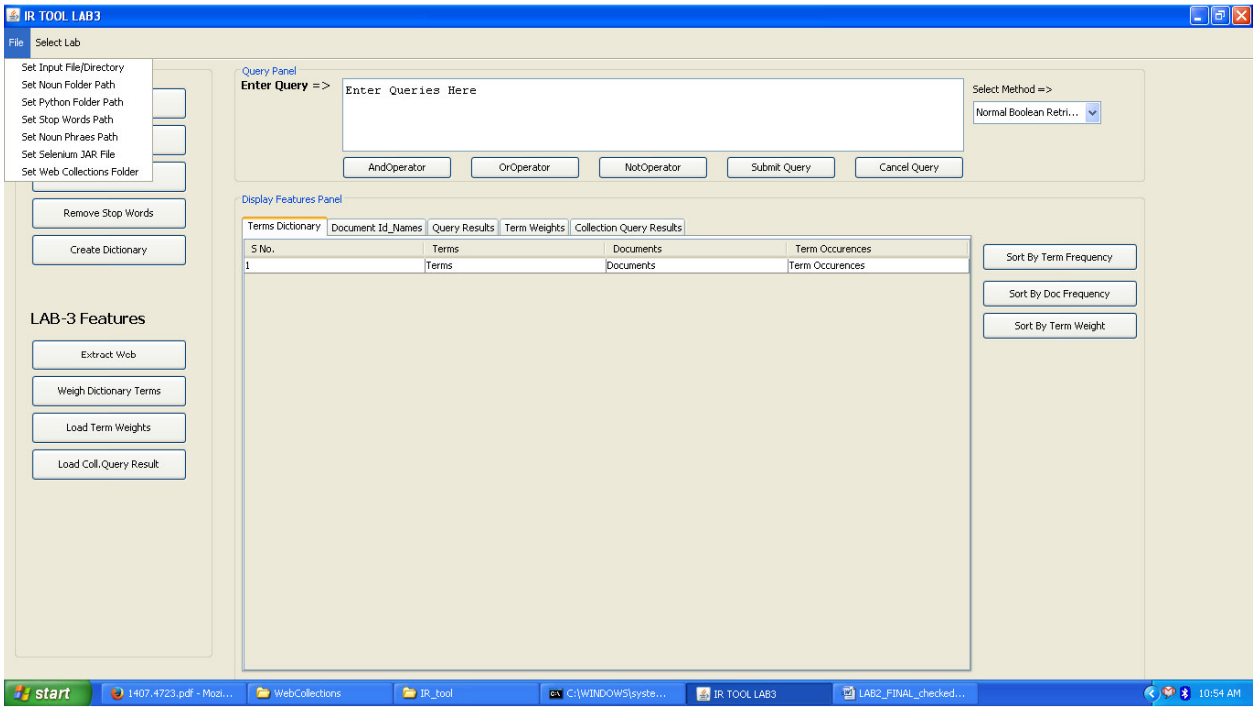
LAB-3

Course Name : Information Retrieval
Course No : CS F469
Time : 5:00 PM – 7:00 PM

The purpose of this lab is

1. To extract information from the web
2. To rank the documents based on query
3. To weigh terms/phrases in the document collection

First, click on "File" Menu and your window look like this:

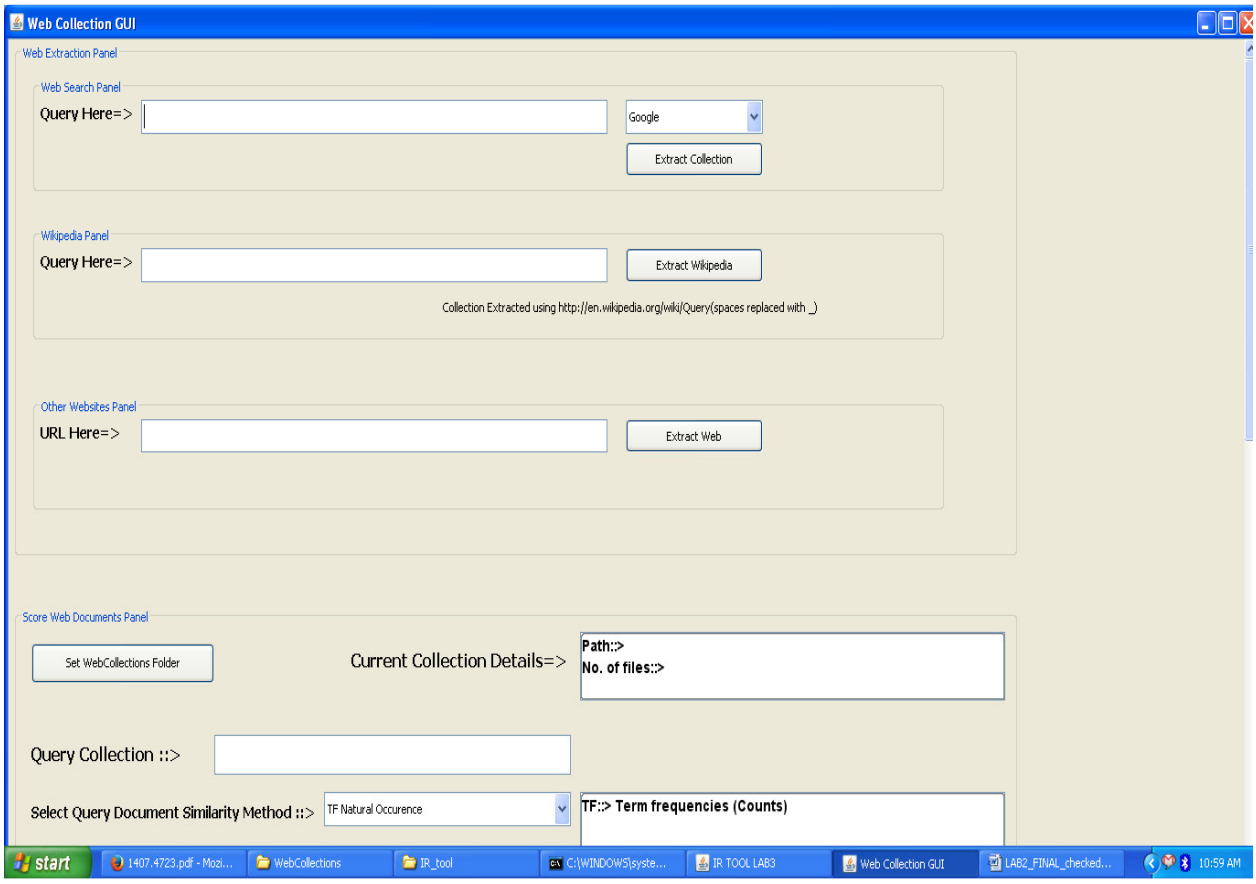


Under "File" menu, check five sub menus:

- Click on "set Input File/Directory" menu to select a single file or directory (i.e. all the files under directory) to be processed.
- Then click on "Set Noun Folder Path" menu to select the folder which contains token extraction code(s).
- Then click on "Set Python Folder Path" menu to specify the path for python
- Then click on "Set Stop Words Path" menu to select the file which contains stop words list
- Then click on "Set Noun Phrase Path" menu to select the folder which contains extracted noun phrase file(s).
- Then click on "Set selenium JAR file" menu to select the selenium JAR file for the web data extraction.
- Finally, click on "Set Web Collections Folder" to specify the path for the folder which contains extracted data from the web.

Extraction of Information from the web

- Click on "Extract Web" button in the GUI. Web Collection GUI will be displayed as shown below:



Web collection GUI has two panels: Web Extraction Panel and Score Web Documents Panel

Web Extraction Panel consists of

- **Web search panel**
  - Allows you to extract top 100 results returned by the Google search engine.
  - Enter query in the query box and click on "Extract Collection" button to extract Top-100 documents' title, Snippets and URLs. The results will be stored in the folder "Webcollections". **Note:** *Refer query_index.txt file for the verification of query name, it's id and folder name where the respective results are stored.*
- **Wikipedia panel**
  - Allows you to extract Wikipedia page related to the given query
  - Enter query in the query box and click on "Extract Wikipedia" button to extract the respective Wikipedia documents' content. The results will be stored in the folder "Webcollections".
- **Other Web sites Panel**
  - Allows you to extract the content of the page related to the given URL.

- Enter URL in the URL box and click on "Extract Web" button to extract the respective documents' content. The results will be stored in the folder "Webcollections".

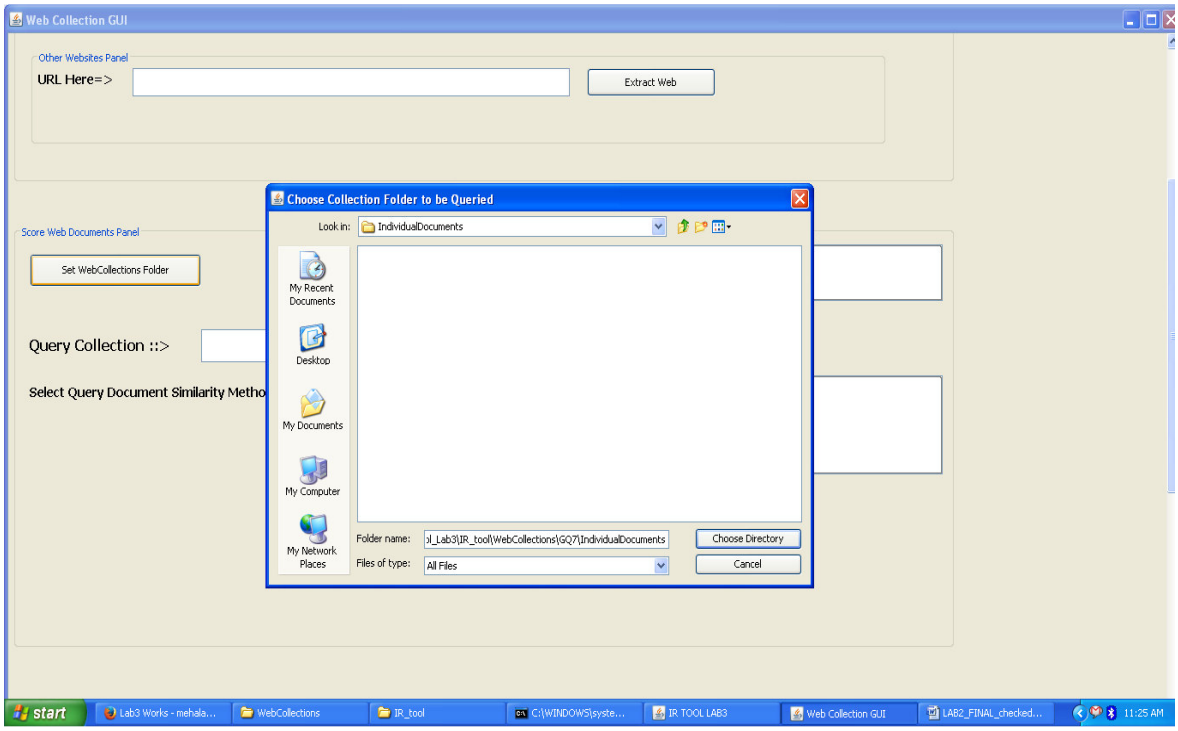
Exercise 1: Write a code to extract top 100 documents' title, snippets and URLs from Bing.

Exercise 2: Write a code to extract information about customer reviews about the product from Amazon.

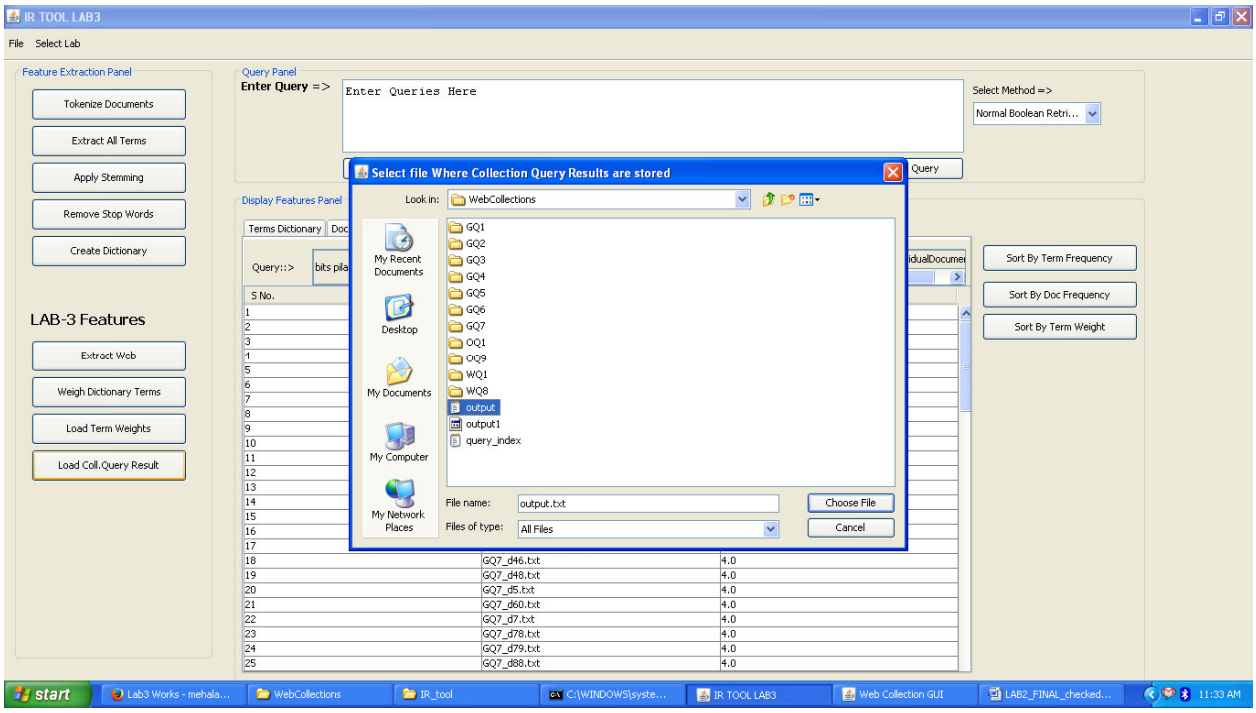
Exercise 3: Write a code to extract reviews about any smartphone from http://www.gsmarena.com/

Document Ranking

- Click on "Set Web collection Folder" to specify the folder path which contains document collections as shown below:



- Then, Query can be given in the query box.
- Then, select query document similarity method
- Then, click on "Score and Rank documents" button in the score web documents panel.
- Then store the results with .txt extension (eg. output.txt)
- To view the document relevance score, click on "Load Coll. Query result" button and select the stored file(i.e. output.txt) as shown below:



Note: *click on "Collection query Results" tab in the display features panel to view the results of the same.*

Weighing Terms/phrases importance in the collection

- Create positional/non-positional dictionary based on term/phrase/bi-word term/bi-word-phrase.
- Click on "weigh dictionary terms" button to weigh the dictionary unit term and specify the dictionary file (eg. NormalDictionary_TermIndex.txt) on which weighing method will be applied. It will create the file "WeightedNormalDictionary_TermIndex.txt" in Extractedterms folder.
- To visualize the same content in the GUI, click on "Load term weights" button in the GUI and click on "term weights" tab in the display features Panel.
  Note: Here, weight is computed using TFIDF.

Exercise 4: Implement the same using different statistical weighing measures and observe the effect with respect to different collections.