Lab Exercise 1 : Very Simple Search Engine

In this lab exercise, you will create a very simple search engine using some text analysis and dictionary based approach.

1. Create a **vocabulary/dictionary** for a given file *f1.txt*.You may use regular expressions as well.
   a. Find how many
      i. **lines**/sentences are there
      ii. **words** are there
      iii. **characters** are there (space character excluded)
   b. Create a **vocabulary** (*list of unique words*) from the text.
   c. List the words in the **vocabulary** along with **their frequency** (count).
2. Given five text files {*f1.txt, f2.txt*…..}.
   a. Get an input search query from the user. List out files in descending order of similarity.

      Hint: The dictionary/vocabulary approach implemented in Q1 can be used. One vocabulary for all text. Create a similarity measure, such as the **number of words common**, to get the search results.

Advanced Optional :

3. *If two documents have same similarity counts use frequencies of each word too.*
4. *The out-of-vocabulary (words in search query not in the vocabulary) words should be handled.*