Programming Assignment- Information Retrieval (CS F469)
Deadline: Sep 12, 2019 11:59 PM, Max Marks: 25

This assignment covers the index construction, text operations, and Boolean Retrieval model. The assignment is comprised of five tasks. First task is mandatory since it will be required to complete the other tasks.
Students are allowed to use **NLTK** and **SPACY** libraries. No other libraries such as textblob or gensim will be allowed. All the tasks in the assignment are basic IR tasks and do not require any advanced libraries.
To start with the assignment, download the following dataset: `http://tiny.cc/9wv8bz`
The corpus contains documents files named with doc_id. Each document has a set of sentences.

Task 1 [**5 Points**] Take a random sample of 100 documents with a total vocabulary size of at least 3000 words.

(a) Construct a **full-text** inverted index I_{full} using standard tokenization method.

(b) Display the size of vocabulary and dictionary.

(c) Plot the dictionary terms in the decreasing order of their document frequency. Identify the stopwords in the corpora (if any) based on the size of the posting list (not the standard lexicon of stopwords in nltk/spacy/online sources).

Task 2 [**5 Points**] Download the queries and their true annotations (relevant documents IDs) from the following links: `http://tiny.cc/8n38bz` and `http://tiny.cc/op38bz`. **query.txt** contains two columns: query id and query text. **output.txt** contains top 50 relevant documents (doc id) for each query.

(a) Take a random sample S_q of 10 queries from the text file.

(b) Compute the k^{th} level precision and recall for each query $q \in S_q$. $k = 5$, 10, and 15.

Task 3 [**5 Points**] Select one or more linguistic models (text operations) and re-construct your inverted index: a) I_P: to increase the precision and b) I_R: to increase the recall. Justify the selected linguistic model(s) and the proposed pipeline w.r.t your new inverted index.

Task 4 [**5 Points**] Report the changes in the index (I_{full} vs I_P and I_{full} vs I_R) in terms of vocabulary size, dictionary size, and posting list size after applying each linguistic model in the pipeline.

Task 5 [**5 Points**] Run the same set of queries used in **Task 2** on the new revised inverted indices (I_P and I_R) and report the k^{th} level precision and recall for each query. $k = 5$, 10, and 15. Compare the results and report

(a) Precision results of I_{full} and I_P

(b) Recall results of I_{full} and I_R

(c) Precision and Recall results of I_P and I_R

Sampling Technique:

1. **For Documents:**
   (a) Take the last two digits of the roll numbers of your group members. Compute their maximum (M).
   (b) Divide the dataset (corpus) into M parts.
   (c) Select one document (uniform at random) from each part until you get 150 documents. Now, take a sample of 100 documents with a total vocabulary size of at least 3000 words.
   (d) Example: Group 10: 2017A7PS01**37**G, 2017A7PS00**60**G, 2017A7PS00**74**G. Maximum is 74. Divide data into 74 parts. Take one document (uniform at random) from each chunk in first iteration. Take remaining 76 documents in second and third iterations (74 + 2). Now, take a random sample of 100 documents with a minimum vocabulary size of 3000.
2. **For Queries:**
   (a) Take the last two digits of the roll numbers of your group members. Compute their minimum (M).
   (b) Divide the query.txt file into M parts.
   (c) Select one query (uniform at random) from each part until you get 25 queries. Now, take a sample of 10 queries from these 25 queries.
   (d) Example: Group 16: 2019H1030**11**G, 2019H1030**12**G, and 2019H1030**10**G. Minimum is 10. Divide query.txt file into 10 parts. Take one query (uniform at random) from each chunk in first iteration. Take remaining 15 queries in second and third iterations (10 + 5). Now, take a random sample of 10 queries.

Assignment submission instructions:

1. Only one member from the group shall submit the assignment.
2. Each group needs to submit a zip file consisting of:
   (a) The set of documents taken as a random sample in the first Task.
   (b) The set of queries taken as a random sample for Tasks 3 and 5.
   (c) The jupyter notebook along with the justification and outputs.
   (d) If you are submitting a .py file and not the jupyter notebook then you must submit a report providing necessary justification and explanation asked in the respective tasks.
3. File name should be your group ID.