Assignment 1

Building Inverted Positional Index and Answering Queries

29th July 2019

This assignment is on building inverted positional indices and using them to answer different types of queries. It is highly recommended that you use python for this assignment as libraries like *nltk* will make many things easier (stop word removal and lemmatization). However, if you use any other language, you most probably have to design these modules yourselves which might not perform as good as *nltk* library in python.

Dataset:

You can find the dataset in the following link
https://drive.google.com/file/d/1sXpMTRBHNb4yTEREYlolRFZzgOoAPg_2/view?usp=sharing
The dataset contains 3253 text files each containing different English news articles on politics and religion.

Task 1 (Building Index)

1. Remove stop words, punctuation marks and perform lemmatization to generate tokens from the document. (use nltk library in python)
2. Build Inverted Positional Index (Dictionary with tokens as keys, and (file_name,positions) as postings)

Task 2 (Answering different types of queries)

Now write codes to use the built inverted positional index to answer different queries. The queries will be of different types as listed below.
(Consider five different words-- *w1*, *w2*, *w3*, *w4*, *w5*. The operators used the queries will be AND, OR, NAND, NOR, DIST. The precedence of the operators will be based on their appearance from left to right)

Query Text	What does it mean (left-> right precedence)
"*w1* AND *w2*"	*w1* AND *w2*
"*w1* OR *w2*"	*w1* OR *w2*
"*w1* NOR *w2*"	*w1* NOR *w2*
"*w1* NAND *w2*"	*w1* NAND *w2*

“*w1* AND *w2* OR *w3*”	(*w1* AND *w2*) OR *w3*
“*w1* AND *w2* AND *w3* OR *w4*”	((*w1* AND *w2*) AND *w3*) OR *w4*
“*w1* AND *w2* DIST k”	distance (*w1* AND *w2*) <= k (i.e., documents (without stop words) which have w1 and w2 within a distance of k)
“*w1* AND *w2* DIST k OR *w3*”	(distance (*w1* AND *w2*) <= k) OR *w3*
“*w1* *w2* *w3* ORDER”	If *w1*, *w2* and *w3* are in the document in the same order as in the query
“*w1* *w2* *w3* EXACT”	If there is an exact match for “*w1* *w2* *w3*” in the document

Important Instructions on How to write the code and How to submit

1. Reading the dataset: Your code should first create inverted positional index for the whole dataset. Assume the dataset to be in the path “./IR_Assignment1_Dataset”, i.e. the dataset folder is in the same path as your python code.
2. Naming the code file: The name of the code file should be in uppercase letters as below.
   ASSIGNMENT1_<ROLLNO>.py
   e.g. :- For a student with roll no 17CS92R02, the code file name should be **“ASSIGNMENT1_17CS92R02.py”**
3. Reading the queries: Write code which can take “query.txt” file as an argument as below.
   ```
   $>> python code.py query.txt
   ```
   (query.txt file will contain many queries in the above mentioned format. There will be one query in each line. This file will remain unknown to you. Your program will be evaluated based on the results (precision & recall of the results), it produces for the queries in the above file.)
4. Saving the search results: Your program should read the queries one by one and get the search results. At the end it should create a text file with results. The name of the results file should follow the below convention.
   RESULTS1_<ROLLNO>.txt
   e.g. :- **“RESULTS1_17CS92R02.txt”**
5. Python library restrictions: You can use python libraries like nltk, numpy, os, sys, collections, timeit, etc. However, you can’t use libraries like lucene, elasticsearch, or any other search api. If your code is found to use any of such libraries, you will be awarded with zero marks for this assignment without any evaluation.

6. Plagiarism Rules: If your code matches (more than 50%) with another student's code, all those students whose codes match will be awarded with zero marks without any evaluation. Therefore, it is your responsibility to ensure you neither copy anyone's code nor anyone is able to copy your code.
7. Code error: If your code doesn't run or gives error while running, you will be awarded with zero mark.