DEADLINE: 17th August, 11:59 PM

IR - Assignment 1

In this assignment, you will build an inverted index and use the index to execute search queries on a set of documents. You can use python for this assignment.
However, python will make life easier since you will be able to use nltk library for stop word removal and lemmatization.

Part 1:

#Generation of Index

1. Build an inverted index from the set of documents provided along with the assignment. Do not do any kind of pre-processing on the text. Store your index as index_raw.
2. Remove stop words from the documents and rebuild index. Store this as index_stopword.
3. In this step, remove stopwords and store only lemmatized terms in your index. Save this as index_lemmatized.
4. After removing stopwords and doing lemmatization, make Permuterm indexing. Save this as index_permuterm. *The query here will have a single wildcard (consider only *).*
5. After removing stopwords and doing lemmatization, make Bigram indexing. Save this as index_bigram. *Here the query words can have multiple wildcards (consider only *) and the characters between wildcards will be only 2. E.g. in*ma*on.*

Part 2:

Write a python script for each of the search method.

- Naive Search method: For each query, obtain the list of documents which contain all the terms present in the query.
- Taking query and a integer [1-5] as a input to load corresponding index that you constructed in part 1 of the assignment and then obtain the list of documents and store the output as output.txt in output folder.

Example:
Doc1: big kahuna burger
Doc2: chicken burger

Query1: big burger
Result: doc1

Query2: burger
Result: doc1, doc2

Deliverables:

Submit a tar.gz file containing the following:

1. Your python/java code.
2. All the code should be in src folder.
3. A document containing the following information:
    - Size of index (in KB) & number of terms in the index:
        - Without stopword removal and without lemmatization
        - For all other indexing
    - Total time taken by the naive search, and time taken by the search using inverted indexing.
    - A table containing the number of documents retrieved for each query (as shown below).

Name your file as A1_<Roll_No>.tar.gz.

Appendix

1. Data will be present in data folder. Index should be stored in index folder. Improper data paths in code will be penalized.
2. Stop word removal and lemmatization is available in the nltk library.
3. **A word of caution - please do not wait till the last moment to complete this assignment. Downloading and installing nltk takes a lot of time.**

Document Format for submission

Table1: Size of Index

		Index Size (in KB)	**Number of terms**
	Without any processing		
	Index After removing stop words		
	Lemmatized Indexing		
	Permuterm Indexing		
	Bigram Indexing		

Time taken by Naive search (in milliseconds):
Time taken using inverted index (in milliseconds):

Table2: Number of documents retrieved for each query

Query	**Number of documents**
Q1	
Q2	
Q3	