Assignment 3

Implementing Various Document Ranking Principles in Information Retrieval

20th October 2019

Submission Deadline : 11:59PM , 13th November , 2019

This assignment is on implementing various document ranking techniques (both probabilistic and non-probabilistic) and comparing their outcomes while ranking relevant documents for a given query. It is highly recommended that you use python for this assignment as in built libraries will make many things easier. However, if you use any other language, you most probably have to design these modules yourselves which might not perform as good as built-in 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. This is the same dataset you have used in assignment 1.

Task 1 (Implementing various document ranking principle)

1. Implement a ranking function that will take a query and generate ranked list of relevant documents. Ranking will be performed based on each document's RSV score (defined in slide 21 of lecture 11). Given document and query pairs model relevance using Binary Independence Model. Use Naive Bayes conditional independence assumption to decompose joint probability of document or query into term probabilities.
2. Implement a ranking function that will take a query and generate ranked list of relevant documents. Ranking must be done based on the probability that a query would be observed as a random sample from the respective document model. Use Naive Bayes conditional independence assumption to decompose joint probability of document or query into term probabilities. Consider multinomial modeling for modeling probability of a term occurrence in query conditioned over the document.

Task 2 (Comparing outcomes of various document ranking methods while ranking relevant documents for a given query)

Now write codes to use build ranking functions to retrieve a ranked list of top 10 relevant documents of a given query. The queries will be space-separated list of words without any punctuation marks or other signs. For example - Obama health plan. After processing the query, your code should generate a txt file containing the ids of relevant documents in proper order in comma separated list.

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

1. Naming the code file: The name of the code file should be in uppercase letters as below.
   ASSIGNMENT3_<ROLLNO>.py
   e.g. :- For a student with roll no 17CS92R02, the code file name should be **"ASSIGNMENT3_17CS92R02.py"**
2. 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.)
3. 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.
   RESULTS3_<ROLLNO>.txt
   e.g. :- **"RESULTS3_17CS92R02.txt"**
4. 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.
5. 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.
6. Code error: If your code doesn't run or gives error while running, you will be awarded with zero mark.