CS F469 – Information Retrieval

Assignment-1: Boolean Retrieval System

To be submitted by 2359hrs on 21-03-2022

General Instructions

1. This assignment is a coding/implementation project and is expected to be done in groups. Each group can contain at most **three** members. All members in the group should be registered for this course and try maintaining the same group for the assignments of this course.
2. This assignment is expected to be done in Python using standard libraries like NumPy, Matplotlib, NLTK and Pandas. You can use Jupyter Notebook and any other python in-built data structure or library. No other ML libraries like scikit/sklearn, TensorFlow, Torch etc., should be used.
3. All the **assignments would be run through a plagiarism detector,** and any form of plagiarism will not be tolerated and shall be brought to the notice of AUGSD/AGSRD. The final decision lies in the hand of the instructor, and only one submission per group would be allowed for one assignment.
4. All deliverable items (See below) should be put together in a single .zip file. Rename this file as FD<id-of-first-member> (preferably ID number of the student submitting) before submission.
5. Submit the zip file on CMS/Google Forms on or before the deadline, as mentioned above. The demos for this assignment will be held later, which shall be conveyed to you by the IC. All group members are expected to be present during the demo.
6. In case of any queries, please fill out the form. The responses will be shared on this doc.
7. The dataset for this assignment can be found here.
8. You should reach out to *f20180185@hyderabad.bits-pilani.ac.in* in case of any queries regarding this assignment.

Problem Statement

1. This assignment is aimed at designing and developing **Boolean Information Retrieval** S**ystem**, i.e., to return those documents (specifically their **names** from corpus/dataset given: *point 7* of General Instructions) which satisfy Boolean (**AND**, **OR** and **NOT** with their combinations).
2. The Boolean Information Retrieval System should include the following features / pre-processing steps:
    A. **Stopword Removal**: Remove the common stop words from the corpus.
    B. **Stemming or Lemmatization**: Employ either one of the techniques for normalisation.
    C. **Wildcard Query Handling:** Any one of the techniques among Permuterm or K-Gram index should be used for wildcard query management.
    D. **Spelling Correction**: **Edit Distance Method** should be employed to correct misspelled words.
3. Try to vectorise your code as much as possible to make your computations faster and more efficient. Do not hard code any parts of the implementation unless it is indispensable.

Deliverables

1. **Design Document** – This document should describe the application's architecture and the central data structures used in the project. Mention the running time for preprocessing, building index and search/retrieval.
2. **Code** – The code should be well **commented,** and all methods/classes for preprocessing, indexing and searching should be submitted.
3. **Documentation** – All the code's classes, functions, and modules must be **documented**, 'pdoc' is one such library to streamline this.
4. **README** – The README file should **describe** the **procedure** to run your code for **each function** with a **test case** of your choosing, namely:
    a. Stopword Removal
    b. Stemming/Lemmatization
    c. Building Index
    d. Querying