

BITS - Pilani, Hyderabad Campus

CS F469 IR Assignment - 1

Deadline: 02/03/2023

Task:

Retrieving a **passage/paragraph** information for the **specific section** from the listed documents.

Problem statement:

Retrieving a passage/paragraph from the sections like **exclusions/inclusions** for the documents (Any kind of documents like word/PDF/Image). Each policy documents have multiple sections like **inclusions, exclusions, conditions, definitions, extensions, covered sections, and so on**. Each document **must be extracted with its text information**. From the **policy documents, the section's** (Already mentioned above) entire passage/paragraph needs to be retrieved depending on the query.

The assignment can be done in groups of at most 4 (Four) members. All the group members are expected to contribute to all aspects of the assignment namely **design, implementation, documentation, and testing**.

Programming Languages:

The assignment can be implemented in any programming language of your choice. STL's and inbuilt packages can be used only for Normalization (C++'s Boost Library, Python's NLTK Package etc.). You are expected to code the **core functionality of the search engine**.

Additional Resources:

1. **Stemming:**
 - a. Martin Porter's '[Porter Stemmer](#)' can be used for this purpose. Implementation in multiple languages can be found in the above link.
2. **Tokenization:**
 - a. For this step you can use any standard tokenizer or inbuilt package. Following are a few sources:
 - i. Python's NLTK package.
 - ii. [Stanford Tokenizer](#).
 - iii. TM package of R.
3. **Datasets:** Attached in the Zipfile

Deliverables:

The final submission must contain the following documents:

1. **Report** – This document should contain a description of the application's architecture along with the major data structures used in the project. **Precision and Recall**, if possible, should also be calculated. Running for all the **preprocessing should be mentioned**. Also mention the **running time for search or retrieval**.
2. **Code** – The code should be well **commented**.
3. **README** – The README file should describe the **procedure to compile and run your code** for various datasets.

Submission Guidelines:

All the deliverables must be zipped and submitted in CMS latest by the **deadline**.

You are expected to demo your application and present your results as per the schedule which will be announced by end of Feb.

Evaluation Criteria for Task :

S.No.	Task	Marks
1.	Tokenization and Normalization	3
2.	Efficient usage of Data Structures with justification	3
3.	Index Construction	3
5.	Viva	3
6.	Novelty / Out-of-the-box thinking (Anything that is not covered in the lectures.)	5
	Total	17

It should be noted that 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.

