BITS Pilani Hyderabad Campus

CS F469 IR Assignment - 2

Deadline: 30 October , 2020 23:59:59 hrs

This assignment is aimed at designing a text based information retrieval system using Locality Sensitive Hashing as taught in class.

The assignment can be done in groups of 4. Please refrain from changing the groups for the assignment and the ones coming in the near future.

Programming Languages:

The assignment can be implemented in any programming language of your choice. You are expected to code the core functionality of the algorithm.

Problem Statement

In this assignment you are expected to code LSH algorithm on your own dataset or a DNA dataset given in Kaggle(Reference 3) . For example, if you work on DNA sequences we can find Duplicate or similar DNA sequences. Each DNA sequence is made up of four bases. You have to implement all the three sub-procedures for Shingling, Min-hashing and LSH step. You can use any similarity metric (Jaccard, Cosine or Euclidean) and it's respective hashing technique to find Signature matrix.

Use bigger Shingles so that your Document matrix will be long (pow(4,shingle_length)).

References

- infolab.stanford.edu/~ullman/mmds/ch3.pdf
- https://eng.uber.com/lsh/
- https://www.kaggle.com/thomasnelson/datasets

Deliverables

The final submission must contain the following documents:

1. Design Document: This document should contain the description of the application's architecture along with the major data structures used in the project. It should contain the report on the different distance measures used for the problem. Precision and Recall, if possible, should also be calculated. Running for all the preprocessing should be mentioned. Also, mention the running time of the retrieval and the search.
2. Code: The code should be well commented.
3. Documentation: All the classes, functions and modules of the code must be documented.
4. README: The README file should describe the procedure to compile and run your code for various datasets.

For any queries, contact Mr. Divakar P (f20170225@hyderabad.bits-pilani.ac.in)

Submission Guidelines:

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

You are expected to demo your application and present your results as per the schedule that will be made available.

Evaluation Criteria

- Implementation: 10 M
- Design Document and other deliverables: 5M
- Viva: 5M

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 hands of the instructor and only one submission per group would be allowed for one assignment.