BITS Pilani Hyderabad Campus  
CS F469 IR Assignment – 2

short line

# Deadline: April 9, 2023 23:59:59 hrs Max Marks:16

This assignment aims to implement a text-based information retrieval system using Locality Sensitive Hashing(LSH) as taught in class.

## 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

You must implement the LSH algorithm to build a plagiarism detector or a search engine. You have to prepare a comparative analysis for different 50, 100, and 200 hash functions. For each value, output the pairs that have an estimated similarity at least 0.5, and report the number of false positives and false negatives that you obtain. For the false positives and negatives, report the averages for 5 different runs.

Next, break up the signature table into b bands with r hash functions per band, as discussed in class, and implement Locality Sensitive Hashing. The goal is to find candidate pairs with similarity at least 0.6. Experiment with r=5, b =10 for the table with the 50 hash functions, r=5, b=20 for the table with the 100 hash functions, r = 5, b = 40 and r=10, b= 20 for the table with the 200 hash functions. Report the number of false positives and false negatives taking the average over 5 runs. How do these numbers change if we want similarity at least 0.8? What is the threshold of the sigmoid function?

You can use the datasets from the previous assignment for the same.

# References

* infolab.stanford.edu/~ullman/mmds/ch3.pdf
* https://eng.uber.com/lsh/

# 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.

## 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: 3 M
* Viva: 3 M

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.

short dash