The University of Texas at Dallas

CS 6322

Information Retrieval

Spring 2023

Class Project Proposal

Project TITLE: Search Engine for **Deserts (Sweets)**

Group:

Students:

Mehroos Ali, mxa200089@utdallas.edu

Pranjal Upadhyay, pxu210002@utdallas.edu

Abhijeetsinh Vaghela, abv210000@utdallas.edu

Utsav Adhikari, uxa200002@utdallas.edu

Karumuru Siddhartha, kvs200001@utdallas.edu

1. The Team

In this project, our team built a search engine for **Deserts (Sweets)**.

We commit to have the following distribution of work and collaboration:

Mehroos Ali shall be responsible for crawling the web to assemble our collection of web pages and its Web graph. A collection of 100,000 web pages shall be crawled. This student will deliver the results of the incremental crawl to the student responsible for indexing and relevance models. The architecture of the crawler shall be presented in our team's project reports and discussions of its functionality shall be detailed.

Pranjal Upadhyay shall be responsible for incremental indexing the crawled web pages and for the link analysis of the Web graph that was retrieved by crawling. The same student shall create two relevance models: (1) vector space relevance model as well as (2) relevance models based on Page Ranking and HITS (and their combinations with the vector-

based relevance models). This student shall collaborate with **Mehroos Ali** (to import the crawled web pages) and with **Abhijeetsinh Vaghela** to display the relevance results against any query as well as with **Utsav Adhikari** (to include clustering information in each of the relevance models) and with **Karumuru Siddhartha** to showcase the relevance results obtained for query expansion.

Abhijeetsinh Vaghela shall be responsible for enabling the queries to be entered and the results to be presented in a graphical user interface embedded in the web page that hosts your search engine. When a query is entered, the search engine should present in separate search engine web page frames of the web page the following:

- Results of your relevance models against the query – all of those that you have implemented.
- Results of the search engine relevance against the query when clustering has been incorporated – you should present the clusters that you have obtained as well.
- Results of the search engine against the query when query expansion has been enabled – you should present the expanded query as well.
- Results of Google against the query.
- Results of Bing against the query.

This student shall collaborate with **Pranjal Upadhyay** (to provide the query to the relevance models and obtain the relevance results) and with **Utsav Adhikari** to obtain the clusters of the Web collection as well as the results of using the clusters in the relevance models. In addition, collaboration with student **Karumuru Siddhartha** is required to obtain the expanded queries and the corresponding relevance results.

Utsav Adhikari shall be responsible for clustering the Web pages crawled by **Mehroos Ali**. Both flat clustering and 2 agglomerative clustering methods are required. In addition, the clusters that are obtained should be used for improving the relevance. Thus, this student needs to collaborate with **Mehroos Ali** to obtain the Web crawl, and with **Pranjal Upadhyay** and **Abhijeetsinh Vaghela** to generate experiments that showcase how the clustering information is used to enhance the relevance results. 50 queries should be considered. Student **Utsav Adhikari** will be responsible for asking 50 queries in these experiments.

Karumuru Siddhartha shall be responsible for query expansion. The Rocchio algorithm should be tested on 20 queries provided by this student. In addition, association clusters, metric clusters and scalar clusters shall be used on a different set of 50 queries to expand them and to provide new relevance results. Thus, this student needs to collaborate with **Pranjal Upadhyay** and with **Abhijeetsinh Vaghela** to obtain relevance results for each of their queries and to display both the expanded queries and their new sets of relevant results to the student **Abhijeetsinh Vaghela** for display on the search engine.

All students are responsible for the creation of the search engine –that should run correctly during the presentation of the project.