The University of Texas at Dallas
CS 6322
Information Retrieval
Spring 2016
Class Project Report

Project TITLE: Search engine for ABC

TEAM XX (i.e. 1, 2 or XX)
Students: X, Y, Z, XX and YY

1. The Problem: Generate a search engine for ABC

The entire team describes the focus of the search engine, provides a Figure with the architecture of the search engine that was developed and writes clearly which student was responsible for which part of the architecture.
In this section you also describe what you learned, what was you experience, what were the difficulties you faced and how you resolved them.

……..

2. Crawling: *written by the student that developed the crawling* module – detailing how he gathered the web pages and how he passed the collection to the index creation. Describe clearly how many web pages you were able to crawl and make use in the search engine of your project. Provide details on the web pages that were sources for your crawls. Discuss how you made sure you did not have duplication in your crawl. Elaborate on how you provided hyperlink information for the student that generated the index and relevance models. *Make sure that the name of the student appears prominently, along with the student netID on every page of the description of the crawling.*

……..

3. Indexing and relevance: written by the student that produced the index and generated the relevance methods for the search engine. Describe how you assembled the index – make a picture and included it in your description. Also elaborate on how your web graph was created – give statistics – e.g. the number of nodes, number of links, the largest number of ingoing links or outgoing links for the graph you have generated. Show how you connected the information from the graph to the index. Describe in detail the two relevance models that you created and provide

the weighting schemes that you have used. Give an example of the topic-based pageranks you have computed – showing which web pages of your search engine have the highest pageranks. Similarly, discuss the HITS scores and show which web pages have obtained the largest scores.
Elaborate on how you have collaborated with the student responsible for the user interface in generating queries to test the relevance models and to display the results of your search engine. State clearly how many queries you have used, how you have generated them and how you have judged the results of your relevance models.
Elaborate on how you have collaborated with the student responsible for the clustering for improving your relevance models.
Make sure that the name of the student responsible for this module appears prominently, along with the student netID on every page of the description of the indexing and relevance section.

……..

4. User interface and comparisons with Google and Bing: written by the student that worked on this module. Describe how you have designed the interface, how you have worked with the student that has generated the index – how you have accessed the relevance models to provide the results in you user interface. Elaborate on the number of queries you have used for testing the search engine. How many were used in collaboration with the student that built the relevance models and how many did you generate on your own. How did you collaborate with the student that produced clusters – to use the clustering information for relevance and presentation on the interface? How do you think you search engine compares to Google and Bing. Explain your judgements. If your search engine had a clustering component, how did you use the results of clustering in presenting the results of your search engine in the user interface?
   Discuss how you have decided to select the queries for the demonstration of your search engine. Provide three examples of the queries and the results produced by your search engine, as well as the results of Google and Bing.
   Make sure that the name of the student responsible for this module appears prominently, along with the student netID on every page of the description of the user interface.

……..

5. Clustering: written by the student that worked on this module. Describe how you have designed the flat clustering – how many predefined clusters did you select, and why? What did you do with the results of clustering – did you incorporate them in the relevance models – and did you provided to the user interface results that were obtained when clustering is used? Same questions – how did you use the results of agglomerative clustering – how many clusters did you obtain? How were they presented on the

user interface? How many queries did you experiment with – such that clustering could be used to improve the results of your search engine? State clearly how many queries you have used to test the impact of the results of each clustering method, how you have generated them and how you have judged the results of your relevance models. Discuss how you have decided to select the queries for the demonstration of your search engine. Provide three examples of the queries and the results produced by your search engine and the clusters that you have created.
Make sure that the name of the student responsible for this module appears prominently, along with the student netID on every page of the description of the clustering usage in your team's search engine.

……..

6. Query expansion and relevance feedback: written by the student that worked on this module. Describe how you have selected 20 queries to test the Rocchio algorithm of your search engine. List them in your reports. Give examples of the web pages that you found relevant and those that you found irrelevant – and explain your judgements. Show also the modified queries that resulted by applying Rocchio to your original queries. Discuss the 50 queries that you have used for pseudo-relevance feedback. For each of the three methods, i.e. associative clustering, metric clustering and scalar clustering show (1) examples of 3 queries, the local document set and (2) the local vocabulary and set of local stems as well as their vocabularies; (3) show the values of the correlations you computed for the queries, and discuss you selection of the clusters and (4) show the resulting expanded queries. Also show the results of the search engine on your expanded queries and discuss them. Elaborate on how you have collaborated with the student responsible for the user interface to expose the results of your expanded queries as well. Discuss which queries and their expansion you selected for the demonstration of the project.
   Make sure that the name of the student responsible for this module appears prominently, along with the student netID on every page of the description of the query expansion usage in your team's search engine.

……..

7. Discussion – all team

……..

8. Conclusion – all team