Georgetown University
COSC 488: Introduction to Information Retrieval
Spring 2018

Nazli Goharian

Project Part 3
Date Given: March 22
Submission Due: April 19, by 11:59 pm via blackboard

Grading: This assignment is 10 points out of total points (40 points) allocated for projects in the semester. It will be graded on the scale of 100.

Pick either Option-1 or Option-2:

Option-1: Query Reformulation: Expansion and Reduction

Your search engine must provide the following two functionalities:

1. Implement a query expansion approach (your choice of methodology) and compare the quality of retrieved results with the baseline, i.e., no expansion. Evaluate your engine using the short (Title) Trec quereis; report your results for various metrics.

2. Implement a query reduction technique. Compare the quality of retrieved results with the baseline which is no reduction. Consider that you can also evaluate the quality of results in the case that both reduction and expansion are done to the same query. That is, as the query is long, one can reduce the noise and at the same time add potentially relevant terms. Evaluate your engine using long (Narrative) Trec queries; report your results for various metrics

Note: You are free to utilize any methodology you want.

Option-2: Clustering

Using one of the clustering methods (hierarchical, K-means, Buckshot), your search engine must provide the following two functionalities:

1. Cluster your document collection & then perform query processing on the clustered collection. Plan to evaluate the quality of clusters. Perform query processing and evaluate results.
2. Cluster query results; use the clustered results to re-rank the result set

IMPORTANT NOTE: Part of this project is to evaluate what you have learned out of the previous projects as far as how to approach evaluating a method that you have incorporated into your engine. That is you must have and show a good **experimental plan (**what runs are needed to evaluate your approach/engine). Finally, any result without **analysis** is meaningless. Thus, as in the last two projects, you will receive points for doing a good analysis. **Use the same dataset you have been using in the semester.**

Deliverables

Cover page (1 pt): should contain the following:
a. Status of this assignment: Complete or Incomplete. If incomplete, state clearly what is incomplete.
b. Time spent on this assignment. Number of hours.
c. Things you wish you had been told regarding this assignment.

Design Document (10 pts): The design document should be written prior to coding. There should **not** be any code in your design document. No specific template is provided to you for your design. You may draw a diagram to show the architecture and the flow of the software components, and/or to provide the write-up of your design decisions.

Experimental Plan (20 pts): You should have a good experimental plan. – No need for a separate document. It will be shown from your results report

Reports & Analysis (69 pts): A working system. Results should be used to provide a good analysis of your engine. Thus, you are expected to provide a good analysis along with your results. Assumption is that you have a good experimental plan for receiving your points of reports & analysis.