Northeastern University

College of Computer and Information Science
Information Retrieval CS6200 – Fall 2016
Prof. Nada Naji najin@ccs.neu.edu

Overview: You have been introduced to the core information retrieval concepts and processes throughout the course of this semester. In this project, you will get to put these into practice by building and using your very own search engines!

Goals: Design and build your information retrieval systems, evaluate and compare their performance levels in terms of retrieval effectiveness

Dataset: CACM test-collection which is comprised of the following:

1- Textual corpus: cacm.tar.gz (3204 raw documents – except for Task 3 part B: use stemmed version cacm_stem.txt)
2- Queries (64 unprocessed cacm.query – except for Task 3 part B: use cacm_stem.query)
3- Relevance judgments (cacm.rel)
4- Stoplist: common_words.txt

Team: Teams of 2 or 3 members are to be formed. Send an email to Prof. Naji, and the TAs Dipti, Maryam, Pavitra, and Surbhi by Wednesday November 23, 2016, declaring your project team in the subject field ("we're a team!" or "IR project team" or something like that). In the message body, list ALL member names (including yourself!). Don't forget to CC your teammates. Send only one email per team. Once formed, teams cannot be altered.

Milestones:
November 20: Release of the online description for the project
November 22 by 11:59pm: Team declaration due date
~~December 6~~ December 9 by **11:00am**: Project & report (implementation & documentation) submission due date

Assessment: The project is to be graded out of 100 points and then scaled to 20% of your overall grade (see syllabus for course grade details)

Implementation: 75 points (detailed point breakdown in project task descriptions)
Documentation: 25 points. Project submissions lacking documentation (report) will NOT be accepted and hence will NOT be graded at all.
Extra credit: 20 points: All or nothing. Awarded credit applies to project & homeworks.

Academic honesty: If you get help from others you must write their names down on your submission and explain how they helped you. If you use external resources

you must mention them explicitly. You may use third party libraries but you need to cite them, too.

Project Description:

Implementation – Phase 1 :: Indexing and Retrieval

Task 1 (5 points): Build your own search engines:

A- From scratch:
   a. You may re-use your indexer from HW3
   b. Use HW4 task2 to implement each of the following models: *BM25*, *tf-idf*, and *Cosine Similarity*. You may re-use your prior implementation for *Cosine Similarity.*

B- Using Lucene: an open source library that provides indexing and searching functionalities (you may re-use your code from HW4 task1)

Task 1 Output: *Four baseline runs:* Your search engine with *BM25* as a retrieval model, your search engine with *tf-idf* as a retrieval model, your search engine with *Cosine Similarity* as a retrieval model, and Lucene's default retrieval model. Only the top 100 retrieved ranked lists (one list per run/search engine) are to be reported.

Task 2 (25 points): Pick one[1] of the four runs above and perform query expansion using *one* approach:
You may choose any of the suggested approaches below, feel free to adopt one that isn't listed but make sure to cite related literature and resources. Justify your design decisions, technical choices, and parameter setting and back them by demonstrated evidence from literature whenever applicable.

A- Inflectional and/or derivational variants
B- Pseudo relevance feedback
C- Thesauri, ontologies, etc.

Task 2 Output: *One run* using one of the base search engines with *one* query expansion technique.

Task 3 (20 points): Use the same base search engine setup (retrieval model produced in Task 1, that you chose to use in Task 2) to perform the following:

A- Stopping (using common_words.txt)
B- Index the stemmed version of the corpus (cacm_stem.txt). Retrieve results for the queries in cacm_stem.query. Perform a query-by-query analysis (see documentation) for *two* queries that you find interesting comparing stemmed and non-stemmed runs.

[1] In practice, it is advised to perform Tasks 2 and 3 using all three base search engines from Task 1 to obtain a broad view of indexing strategies and retrieval models combinations. This, however, is not mandatory for this project

Task 3 Output: *Two runs* (using the same base search engines you chose for Task 2), one run with stopping and another with the stemmed corpus and stemmed query subset.

Implementation – Phase 2 :: Evaluation

By now, you should have *six* distinct runs with results for all 64 queries. Namely, 4 baseline runs, 1 query expansion run, and one stopping run (we're not counting the stemming run here).

Produce one more (seventh) run that does *one* of the following:

1- Combines a query expansion technique with stopping
2- Uses a different base search engine than the one you chose earlier, and adopts either a query expansion technique and/or stopping

Now that you have *seven* distinct runs, it is time to assess the performance of your search engines (runs) in terms of retrieval effectiveness. Implement (from scratch) and perform the following:

1- MAP
2- MRR
3- P@K, K = 5 and 20
4- Precision & Recall (provide full tables for all queries and all runs)

Documentation:

A- ReadMe.txt: which explains in detail how to setup, compile, and run your project.
B- Report NOT to exceed 2500 words[2] in PDF format, named as follows:

firstNameInitialLastName1_firstNameInitialLastName2[_firstNameInitialLastName3].pdf

Please follow this structure:

i. First page: Project members' names, course name and semester, instructor name.
ii. Introduction: Short description of your project, detailed description of each member's contribution to the project and its documentation
iii. Literature and resources: overview of the techniques used (query expansion approaches) scholarly work and research articles to back your technique and algorithm choices, resources, third party tools that you used and referred to in your project.

[2] This document's word count is about 1000 words

iv. Implementation and discussion: More thorough description of your project and design decisions. Include query-by-query analysis in this section.
v. Results: tables reporting all results obtained for all runs and queries for all required metrics. For query level results, please provide spreadsheets, too.
vi. Conclusions and outlook: state your findings, observations and analyses of the results. Which system do you think works best? Why? For "outlook": write a few sentences stating what you would envision doing to improve your project, what other features would choose to incorporate.
vii. Bibliography: citations and links to resources

Extra credit (20 points):

This part is optional, and is all or nothing (all 20 points or none). Awarded extra credit points apply to project and homeworks.

For this part, you will choose and implement a *snippet generation* technique and *query term highlighting* within results. It is up to you to figure out which techniques to use, however, your choices are expected to be backed by the algorithm(s)/technique(s) details and citations of the respective literature.