CS380 Information Retrieval

HW 4 / Group Project 2 (may also be done solo)

Report: oral presentation only.
The presentation will be made a google hangout with a screen-shared slides. I will also accept pre-recorded presentations, but you must be available for live questions. Presentations should be in the neighborhood of 10 minutes. Each presentation should stand alone. (No coordination between groups as in group project 1.) Presentations should be similar to those done by the 4 groups prior to break.

Presentation times will be by signup in a google doc: April 27 - 29.

Due Dates:
1. Topic Selection: Send me an email by April 8 with your topic and group. If you are working in a group, one email is sufficient.
2. Presentation time signup: April 22.
3. Material supporting the presentation (slides): noon, April 27.
4. Presentations: April 27-29

Groups:
Given our virtual status, keeping the groups from project 1 is OK. However rearrangement is also fine. Similarly, solo work is fine. Some of the suggestions below are for groups only, some are solo only, some can be either.

General Instructions:
Unlike the first group project where the number of groups working on particular topics was limited, groups (individuals) are free to choose topics without restriction.

Topics:
1. Web Crawling. Write a web crawler that crawls the Bryn Mawr and/or Haverford domains. The crawler should not crawl pages outside of these domains. The crawler should retrieve essentially 100% of the documents within these domains (a part of the grade will be based on the completeness of your crawl). The crawler should retrieve any page at most twice during crawling. The crawler should also respect any robot restrictions. One result of the crawl must be a list of every page

(URL) found during the crawl along with its size and the date that the page last changed (both as given in the page response headers). This list is to be submitted electronically, NOT on paper. From all groups writing crawlers I will assemble a list of all the pages found. This merged list will be used for assessing crawl completeness.

1. Solo project: just write a crawler as above
2. Group project: Start by writing a crawler as above. Then choose either of the following extensions, or define your own extension (if you define your own, I must approve):
    1. Speed. Do everything you can to optimize the rate a which pages are crawled. Multithreading will be necessary. You will probably want to run on more than one computer. The presentation and paper should focus on how the crawler needed to be adapted make it faster.
    2. Page-rank direction. Order the crawl based on the page rank estimate for the page. This estimate should evolve while the crawl is in progress. That is, do not do a complete crawl, then calculate the page rank for each page, then recrawl with a page ranks calculated from the first, blind crawl. The presentation should focus on how the pagerank calculation can be done in parallel with a crawl and how that affects and affected the crawl. You will likely need to do two crawls, one with and one without pagerank to collect this information.
    3. Some other crawling specialzation.

2. Scatter / gather
   Group only: do scatter / gather on a queries to google. For instance
    1. get the top 300 links for a fairly general query
    2. Get text
    3. Scatter into clusters
    4. Characterize clusters (for instance, the highest centroid weighted TF-IDF token values)
    5. Select (gather) then recluster (scatter).
    6. Recharacterize new clusters.
       See “Scatter / Gather …” Hearst, Karger and Pedersen, 1995
3. Snippet finder. (Solo only) One of the things that people use to evaluate a search result is a snippet of the text of the document that suggests why the document is relevant to a query. Write a system to find snippets. For instance, a binary-search style approach would be to cut the document in half, evaluate both halves (with respect to the query) and pick the half with the higher score. Repeat until you have

an appropriately sized snippet. (This binary-search approach is flawed in several ways but can be made to work.) Other approaches might work even better. (See for instance, section 8.7 of the text.) While speed is certainly important and is a focus of the text, I am more interested in finding high quality snippets.

4. A topic of your own suggestion. If you looked at the originally posed topics, there was one additional suggestion. I removed it, but if that resonated with you, then it could reappear here. If there is something that has been discussed in class, or appears in either textbooks and particularly appeals to you, make an appointment with me to discuss your topic and how it can be reasonably sized for this project. For instance, you might do something much like the first group project … finding a research paper and trying to replicate the system described therein.