Comp 479/6791 Project 3 Final Project Fall 2018

Objectives: Experiment with web crawling, scrape and index a set of web documents, use the sentiment dictionary aFinn to associate sentiment values to the index, make document ranking reflect sentiment.

Due Date: This project has to be demonstrated during lab time December 4., 5, 6. The complete project has to be submitted by December 4, 2017. This project is a team project that must be developed in groups of 3.

Description:

- starting from pages `https://csu.qc.ca/content/student-groups-associations`, `https://www.concordia.ca/artsci/students/associations.html`, `http://www.cupfa.org`, `http://cufa.net`, crawl for links (you may use crawling tools such as Websphinx but you may also find other tools, such as NYUcrawl. To extract the text from web pages, consider Boilerpipe. Describe and attribute any tools used. Make sure you obey the standard for robot exclusion. Your crawler MUST accept as part of its input an upper bound on the total number of files to be downloaded. In developing, testing, and debugging, this number should be kept as SMALL as possible. Develop your own closed test set of HTML files for testing and debugging. The final index should index as many documents as possible. (5 pts, Attrib 5)
- create an enhanced inverted index, using tf-idf and the sentiment dictionary aFinn (`http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010`). Associate sentiment values to each term in your index vocabulary (each term should have the document frequency and the sentiment value associated in the dictionary part of the index). Note that you may use any indexer you like. Lucene is allowed, as is your classmates' code, with permission. (5 pts, Attrib 5)
- develop a simple sentiment aggregation function to associate sentiment to individual documents, to queries, or to collections of documents (3 pts, Attrib 5)
- rank retrieved documents by a partial order on (w, s), where w is tf-idf based cosine distance to the query and s is a sentiment bias. Calculate the sentiment bias as follows: if the query has overall positive sentiment, set $s_1 \leq s_2$ if s_1 is more positive than $s2$. If the query has overall negative sentiment, set $s_1 \leq s_2$ if s_1 is less positive than $s2$. (4pts, Attrib 5)

Functionality to demonstrate: for a given query, retrieve and rank documents by sentiment in the appropriate order.

You have to demonstrate system components during the final demo, such as crawling, indexing, and ranking. Note that you should ensure functionality on a subset of accessible pages first, if including all linked pages overwhelms your setup. Any limitations of the final system have to be clearly discussed in your report.

Experiment with different queries and compare their results. (1pt, Attrib 5)

Additional questions to answer on the written final report: what was the hardest step? How big is the index? How did you define what constitutes a document in your index? What observations did you make during your experiments? What did you learn from your experience? (1pt for graduate students, 3pts for undergraduate students)

Deliverable in Moodle:

- code
- the index
- the final report, documenting the design of the code and report on the experiments, the results, and answer the questions specified above. Individual contributions of each team member have to be specified at the end of the report.) (1pt, Attrib 6)