A Web Search Engine

CS 582 Term Project
Spring 2020
Total points: 100
Issued: 02/18/2020 Due: 05/07/2020

The goal of this project is to design and implement a search engine that includes components for Web crawling, webpage processing, indexing, an "intelligent" aspect of your search engine, and a friendly user interface.

1 Search engine

For the first part of the project, you will have to build a search engine for the UIC domain: you will create and deploy a Web crawler, and integrate it with an IR system that implements the vector-space model. To complete this stage, you are encouraged to (re)use the tools implemented in the assignments you completed for this class. In particular, you will probably make use of:

- the tokenizer / stopword removal / stemmer
- the vector-space model implementation

In all the steps below, make sure your Web-agent stays within the UIC domain (https://www.uic.edu/). Your agent will have to perform the following tasks:

1. Start with https://www.cs.uic.edu/
2. Perform a Web traversal using a breadth-first strategy.
3. Keep track of the traversed URLs, making sure:
   (a) they are part of the UIC domain
   (b) they were not already traversed (i.e. avoid duplicates, avoid cycles)
4. For each such URL, grab the corresponding page, process the text (eliminate SGML tags, tokenize text, stem words, remove stopwords), and index all terms using a vector-space model. Make sure you also keep track of the URL of the page.
5. Make sure your index includes at least 3000 Web pages (that is, it is ok to stop your agent after it collected 3000 Web pages, but feel free to collect more, if you want your engine to be more comprehensive).

Your search engine will take as input any user query and will return a ranked list of UIC Web pages (top 10 pages with an option to return more documents or to exit).

2 What needs to be submitted

There are two main parts for this project, each of them contributing toward the final project grade.

1. Project report, which should include:

(a) a description of your search engine, including a description of the main components (10 points)

(b) a discussion of the main challenges encountered in building the search engine (15 points)

(c) a discussion of the weighting scheme and similarity measure used, and a comparison with possible alternatives (15 points)

(d) a manual evaluation of the top 10 results for 5 sample queries of your choice (25 points)

(e) a discussion of the results, and considerations for what worked and what did not (error analysis); (5 points)

(f) a discussion of related work (2.5 points)

(g) considerations for future work (2.5 points)

Type your report using ACM style files available from this Web page
https://www.acm.org/publications/proceedings-template
Your report should be at least 4-5 pages long.

2. Software. (25 points)

3 Grading

The maximum grade for this project is 100 points.