HW Assignment 3 (Due date: Wed, Nov 20, by 10:30am)

1 Anchor Text [100 points]

Create two versions of an inverted index over the document collection created by crawling the `ohio.edu` domain, as follows:

1. The first version is the traditional index implementation.
2. In the second version, each document in the collection is augmented with the anchor text from all the links that point to it.

Create the index using the standard preprocessing steps: HTML tag removal, tokenization, case folding, stemming, stopword removal. Use VSM with standard `tf.idf` weighting for ranking documents. Run your IR engine implementation and for each version of the inverted index report the list of the top 5 documents (and their cosine similarity scores) that are returned for each of the following queries:

1. `OU Home Page`.
2. `OU Library`.
3. `ARC building`.
4. `Health Alerts`.
5. `Give to OHIO`.
6. `ADA compliance`.
7. `CATS`.
8. `University News and Updates`.
9. `HR`.
10. `Admissions Web Site`.

If neither index returns a relevant document among the top 5 documents, consider increasing the size of the crawl. Additionally, you can use a web search engine to find at least one relevant document in the `ohio.edu` domain for each query and add it to the set of seed documents to be crawled. Make sure that you do not ignore `ohiou.edu` documents, since `ohiou.edu` is equivalent with `ohio.edu`.

Compare the output that you get from the traditional index with the output that you get from the anchor-augmented index. Include a detailed analysis of the results. Furthermore, for each query and for each index version compute and report the *Precision@k* numbers, for $1 \leq k \leq 5$.

2 PageRank [100 points]

Implement the PageRank algorithm discussed in class. Your PageRank implementation should take as input 3 arguments:

1. The input graph, represented using adjacency lists.
2. The teleportation rate α.
3. The convergence threshold τ.

The output should be a vector of PageRank values, one value for each page in the graph. See the class website for an sample input graph and output file obtained using $\alpha = 0.14$.

Test your PageRank implementation on the link structure of the `ohio.edu` domain, using a teleportation rate $\alpha = 0.15$ and a convergence threshold $\tau = 0.01$ with L2 norm. Answer the following questions:

1. Rank the web pages based on the number of *in-links* and report the top 50 pages.
2. Rank the web pages based on their *PageRank* scores and report the top 50 pages.
3. Compare the 2 rankings and discuss differences between them.

Answer the same questions for the link structure of the `Wiki-Large` corpus, which is available on the course web site. This is a rather big file, so you might want to scan through it on disk rather than load it entirely in memory.

3 Submission

Please turn in a hard copy of your homework report at the beginning of class on the due date. Electronically submit a directory that has your working code, data, and images, and a concise README file describing these before class. Create a gzipped, tar ball archive of your directory, and upload it on Blackboard.

For example, if the name is John Williams, creating the archive can be done using the following commands:

> tar cvf williams_john.tar williams_john

> gzip williams_john.tar

These two steps will create the file 'williams_john.tar.gz' that you can upload on Blackboard.

Please observe the following when handing in homework:

1. Structure, indent, and format your code well.
2. Use adequate comments, both block and in-line to document your code.

3. Type and nicely format the project report, including discussion points, tables, graphs etc. so that it is presentable and easy to read. Include a description of all the design decisions that you made that may have an influence on the results.

4. Working code and/or correct answers is only one part of the assignment. The project report, including discussion of the specific issues which the assignment asks about, is also a very important part of the assignment. Take the time and space to make an adequate and clear project report.