CS6200 Information Retrieval

Homework4: Web graph computation

Objective

Compute link graph measures for each page crawled using the adjacency matrix. While you have to use the merged team index, this assignment is individual (You can compare results with your teammates)

Page Rank - crawl

Compute the PageRank of every page in your crawl (merged team index). You can use any of the methods described in class: random walks (slow), transition matrix, algebraic solution etc. List the top 500 pages by the PageRank score. You can take a look at this PageRank pseudocode (for basic iteration method) to get an idea

We have created a discussion thread in Piazza. It is meant to be a supportive space to help each other succeed in this assignment. Whether you're encountering hurdles, have discovered something interesting, or want to share your progress, this is the place!

Page Rank - wt2g graph

- Get the graph linked by the in-links in the file resources/wt2g_inlinks.txt.zip
- Compute the PageRank of every page.
- List the top 500 pages by the PageRank score.
- Display the inlink and outlink counts for each page.

The PageRank result file for both your team's merged crawl and wt2g graph should follow the format. The following represents the top 5 results for the computed page rank for the wt2g graph:

Page	PageRank	No. of Outlinks	No. of Outlinks
WT21-B37-76	0.0026944737779136915	5	2568
WT21-B37-75	0.001533178746424349	1	1704
WT25-B39-116	0.0014685026346494055	1	169
WT23-B21-53	0.0013735263702095858	1	198
WT24-B26-10	0.0012761659886317445	1	291

The results shouldn't differ if you are following the same method that has been given (it is expected that at least the first 6-8 digits after the decimal are the same).

Explain in few sentences why some pages have a higher PageRank but a smaller inlink count. In particular for finding the explanation: pick such case pages and look at other pages that point to them.

HITS- crawl

A. Compute Hubs and Authority score for the pages in the crawl (merged team index)

1. Create a root set: Obtain the root set of about 1000 documents by ranking all pages using an IR function (e.g. BM25, ES Search). You will need to use your topic as your query
2. Repeat few two or three time this expansion to get a base set of about 10,000 pages:
   - For each page in the set, add all pages that the page points to
   - For each page in the set, obtain a set of pages that pointing to the page
     - if the size of the set is less than or equal to d, add all pages in the set to the root set
     - if the size of the set is greater than d, add an RANDOM (must be random) set of d pages from the set to the root set
     - Note: The constant d can be 200. The idea of it is trying to include more possibly strong hubs into the root set while constraining the size of the root size.
3. Compute HITS. For each web page, initialize its authority and hub scores to 1. Update hub and authority scores for each page in the base set until convergence

  - Authority Score Update: Set each web page's authority score in the root set to the sum of the hub score of each web page that points to it
  - Hub Score Update: Set each web pages's hub score in the base set to the sum of the authority score of each web page that it is pointing to
  - After every iteration, it is necessary to normalize the hub and authority scores. Please see the lecture note for detail.

B. Create one file for top 500 hub webpages, and one file for top 500 authority webpages.
The format for both files should be:
[webpageurl][tab][hub/authority score]

EC1

Implement a Topical PageRank by designing categories appropriate for your crawl (merged team index)

EC2

Implement SALSA scoring on your crawl (merged team index) and compare with HITS

Rubric

Check Canvas for details