Shreesh Ayachit [sa736]
CS 4300

ASSIGNMENT 4

1A) For the first graph:

e = x	i = x/6
f = x/3	j = 5x/6
g = 5x/12	k = 11x/12
h = x/3	m = x

Ranking: Highest Rank = m and x
Lowest Rank = i

For the second graph:

e = x	i = 3x/8
f = x/4	j = 7x/8
g = 5x/16	k = 15x/16
h = x/4	m = x

Ranking: Highest Ranking = m and x
Lowest Ranking = f and h

1B)

- link is added from d_1 to d_3 : Changed, Pagerank of d3 increases
- The link from d_6 to itself is removed : Changed, Pagerank of d6 decreases
- The link from d_3 to d_4 is removed : Changed,Pagerank of d3 increases as the feed-back loop to d6 is not removed.
- The link from d_4 to d_6 is removed : Changed, Pagerank of d6 decreases
- A link is added from d_4 to d_3 : Changed, Pagerank of d3 increases

2A) Mini Search Engine:

Code Details
Implementation : Python
Required Libraries: Numpy,BeautifulSoup
Required Files: metaData.txt , pageRecord.txt [*I have not added the support of specifying file names as command-line params*]

Data structures used
I have heavily used dictionaries and lists for storing logically related elements of a webpage. For query parsing, I iterate over the anchor text stored in a dictionary with corresponding pageId and rank.

Files needed:
Th following files need to be present in the current directory

test3.txt: Reads the links to be parsed and stores them.
metaData.txt: Stores the pageId:Rank:anchor text
pageRecord.txt: Stores the pageId:Snippet

I prefetched and parsed the html files listed in text3.txt. I have extracted the main content[of each page and used part of it as the snippet.

Note: since some of the links in africana library were giving 404 error, I parsed the html files from the assignment site. However I have written the code to get the html and parse links from the net too.

How to Run
#python a3.py test3.txt metaData.txt pageRecord.txt

Test Runs

[Please refer test_runs.txt for actual code results]

1) Query: home

Results:
Page : http://www.library.cornell.edu/colldev/cdhome1.html
Rank : 0.000943

Page : http://www.library.cornell.edu/Reps/DOCS/homepage.htm
Rank : 0.000666

Shows that colldev home page is linked more than DOCS homepage.

2) Query:student
Page : http://www.library.cornell.edu/svcs/borrow/gradnewstatus
Rank : 0.000919
Page : http://www.library.cornell.edu/svcs/borrow/grad
Rank : 0.000892
Page : http://www.library.cornell.edu/svcs/borrow/undergrad
Rank : 0.000832

Clearly graduate students homepage is ranked higher

3) Query:book
Page : http://www.library.cornell.edu/svcs/borrow/renew
Rank : 0.000832
anchor : Renewing,books
Snippet : [Search , Course Help Research Help Library Services Requests , Rene

Page : http://www.library.cornell.edu/svcs/borrow/returning
Rank : 0.000832
anchor : Returning,books
Snippet : [Search , Course Help Research Help Library Services Requests , Most

Renew link is ranked higher than borrow link.

4) Query :FAQ

Page : http://www.library.cornell.edu/visualresources/faq
Rank : 0.000973
anchor : FAQ,FAQ
Snippet : [Search , Course Help Research Help Library Services Requests , Rese

Page : http://www.library.cornell.edu/annex/staff_faqs
Rank : 0.000589
anchor : Staff,FAQs
Snippet : [Search , Course Help Research Help Library Services Requests , Staf

FAQ for *visualresources* is linked more.

5) Enter Query:research

Page ID : 77
Page : http://www.library.cornell.edu/resrch/rsrchform
Rank : 0.004018
anchor :
research,consultation,appointment,Research,Consultation,Research,Consultation,Research,Consultation,Research,Consultation,Research,Consultation,Consultations,with,librarians
Snippet : [Search , Course Help Research Help Library Services Requests , To r

Only one result for research is pulled up