COMS 535 Programming Assignment 3

Hung Phang ()

Shruti Sahu (106202120)

1. **Data Structure to implement Weighted Q:**

We used LinkedHashMap to implement the WeightedQ. This allows us to order the Map in descending order of their values keeping the order in which a record was put in the map.

1. **Pseudo Code for crawling algorithm:**
2. Explore the robots.txt
3. Store all the forbidden urls in forbiddenURL hashmap
4. visited ={}, queue={}
5. root = seedUrl;
6. if root is not in forbiddenURL
7. queue.add(root)
8. download all the links in root //neighbors of roots
9. while(queue is not empty)
10. node = queue.remove();
11. for every neighbor u of node
12. if size of visited < maxNodes *// to make sure exactly the max nodes are crawled*
13. if u not in visited
14. add u to visited
15. else exit
16. if size of visited > maxNodes *// to make sure exactly the max nodes are crawled*
17. exit
18. **Output of the program WikiTennisRanker:**

***For epsilon= 0.01***

Top 10 pages with highest page rank:

[/wiki/main\_page, /wiki/spain, /wiki/world\_war\_i, /wiki/international\_standard\_book\_number, /wiki/integrated\_authority\_file, /wiki/library\_of\_congress\_control\_number, /wiki/united\_states, /wiki/digital\_object\_identifier, /wiki/biblioth%c3%a8que\_nationale\_de\_france, /wiki/wayback\_machine]

Top 10 pages with highest in-degree:

[/wiki/main\_page, /wiki/international\_standard\_book\_number, /wiki/united\_states, /wiki/france, /wiki/united\_kingdom, /wiki/tennis, /wiki/integrated\_authority\_file, /wiki/australia, /wiki/spain, /wiki/library\_of\_congress\_control\_number]

Top 10 pages with highest out-degree:

[/wiki/tennis, /wiki/olympic\_games, /wiki/us\_open\_(tennis), /wiki/career\_golden\_slam, /wiki/atp\_rankings, /wiki/grand\_slam\_(tennis), /wiki/list\_of\_grand\_slam\_related\_tennis\_records, /wiki/list\_of\_male\_singles\_tennis\_players, /wiki/list\_of\_male\_doubles\_tennis\_players, /wiki/grip\_(tennis)]

The Jaccard similarity between the top pages in ranking and top pages with highest indegree is: 0.3888888888888889

The Jaccard similarity between the top pages with highest out-degree and top pages with highest in-degree is: 0.21212121212121213

The Jaccard similarity between the top pages in ranking and top pages with highest out-degree is: 0.09859154929577464

***For epsilon= 0.005***

Top 10 pages with highest page rank:

[/wiki/main\_page, /wiki/spain, /wiki/world\_war\_i, /wiki/international\_standard\_book\_number, /wiki/integrated\_authority\_file, /wiki/library\_of\_congress\_control\_number, /wiki/united\_states, /wiki/digital\_object\_identifier, /wiki/biblioth%c3%a8que\_nationale\_de\_france, /wiki/wayback\_machine]

Top 10 pages with highest in-degree:

[/wiki/main\_page, /wiki/international\_standard\_book\_number, /wiki/united\_states, /wiki/france, /wiki/united\_kingdom, /wiki/tennis, /wiki/integrated\_authority\_file, /wiki/australia, /wiki/spain, /wiki/library\_of\_congress\_control\_number]

Top 10 pages with highest out-degree:

[/wiki/tennis, /wiki/olympic\_games, /wiki/us\_open\_(tennis), /wiki/career\_golden\_slam, /wiki/atp\_rankings, /wiki/grand\_slam\_(tennis), /wiki/list\_of\_grand\_slam\_related\_tennis\_records, /wiki/list\_of\_male\_singles\_tennis\_players, /wiki/list\_of\_male\_doubles\_tennis\_players, /wiki/grip\_(tennis)]

The Jaccard similarity between the top pages in ranking and top pages with highest in-degree is: 0.3793103448275862

The Jaccard similarity between the top pages with highest out-degree and top pages with highest in-degree is: 0.21212121212121213

The Jaccard similarity between the top pages in ranking and top pages with highest out-degree is: 0.09154929577464789

1. **Number of iterations for the page rank algorithm to converge on the graph wikiTennis.txt:**

For epsilon= 0.01 the number of iteration are: 7.0

For epsilon= 0.005, the number of iteration are: 9.0

1. **Topic chosen for MyWikiRanker and the set of topic words:**

Topic: Computer Science

Set of words: {"Data Structure", "Computer Science", "Algorithm"}

1. **Output of the MyWikiRanker:**

Number of nodes in the graph:99

For epsilon= 0.01the results are: 6.0

Top 10 pages with highest page rank:

[/wiki/ibm, /wiki/algorithm, /wiki/artificial\_intelligence, /wiki/charles\_babbage, /wiki/history\_of\_computing\_hardware, /wiki/programming\_language, /wiki/computational\_complexity\_theory, /wiki/cryptography, /wiki/punched\_card, /wiki/machine\_learning]

Number of nodes in the graph:99

For epsilon= 0.005the results are: 8.0

Top 10 pages with highest page rank:

[/wiki/ibm, /wiki/algorithm, /wiki/artificial\_intelligence, /wiki/charles\_babbage, /wiki/history\_of\_computing\_hardware, /wiki/programming\_language, /wiki/computational\_complexity\_theory, /wiki/cryptography, /wiki/punched\_card, /wiki/machine\_learning]