Com S 535x
Project Report #3

Yunxi Guo, Qingqin Hou

1. Description of (pseudo code) of your crawler

There are six methods in our crawler.

private boolean validURL(String target)
Method used to check whether target URL is a valid URL which could be visited. First, check whether target URL is included in blocking list. Second, check whether target URL contains "#" or ":", if contains, it will be out of domain. Finally, check whether the target URL start with "/wiki". If not, it is out of domain.

private ArrayList<String> findAllLinks(String content, String url)
Method used to find all links on current page. First, find all links on the page. For each link, if the link url is valid and the link url is not current url, then add this url to result.

private void initialBlockList()
Method to get blocking list. First open robots.txt. For each line in robots.txt, split the line by ":", if the first element of array equal to "Disallow", then add the url followed by "Disallow" to block page list.

private String fatchPage(String target, boolean filter)
Method used to fetch page content. If we have finished 100 requests, wait for 5 seconds. If filter is true, then add each line contains in . If filter is false, just add each line in the document.

private boolean hasKeywords(String target)

Method used to check whether target contains all keywords, save pages contains all keywords into HashMap. First, fetch raw text page of target link, since it is raw text page, we don't need to filter this page. If the raw text page contains each keyword, then store the page and return true, otherwise return false.

public void crawl()

Method used to crawl from seedURL. First do BFS to crawl nodes. Then find all edges between those nodes.

Algorithm 1 validURL(*url*)

```
1: if blockinglist contains url then return false
2: if url contains # or url contains : then return false
3: if url is not start with "/wiki" then return false
return true
```

Algorithm 2 findAllLinks(*content*,*url*)

```
1: result ← new list
2: for each link in content do
3:     if link is valid & link is not in result & link is not url then
4:         add link to result
   return result
```

Algorithm 3 initialBlockList()

```
1: stream ← robots.txt
2: for each line l of stream do
3:     if l start with "Disallow" then
4:         link ← content after "Disallow"
5:         add link to blockPages
```

Algorithm 4 fatchPage(*target*,*filter*)

```
result ← new string
if 100 fetching request has been finished then
    sleep for 5 second
if filter then
    for each line l do
        if m is inside of < p > and < /p > then
            append m to result
else
    for each line l do
        append l to result
return result
```

Algorithm 5 hasKeywords(target)

```
content ← raw text page of target url
for each keyword k do
    if content contains k then
        return false
Add content to storage
return true
```

Algorithm 6 crawl()

```
1: q ← new Queue
2: q.add(seed URL)
3: visited ← new HashSet
4: visited.add(seed URL)
5: file ← new File
6: while q is not empty do
7:     first ← q.next element
8:     content ← get content of first
9:     outpage ← all links in content
10:    for each page in outpage do
11:        if visited does not contain page & page contains keywords then
12:            if visited is not full then
13:                visited.add(page)
14:                q.add(page)
15: for each url in visited do
16:    list ← all links in url
17:    for each link out in list do
18:        if visited contains out then
19:            Write edge into file
```

2. For each epsilon (0.01, 0.005): List webpages with top 15 page rank. How does the list change as epsilon changes?

Using PavanWikiTennis.txt and 0.01 as epsilon.

/wiki/Grand_Slam_(tennis)
/wiki/French_Open
/wiki/The_Championships,_Wimbledon

/wiki/US_Open_(tennis)
/wiki/Australian_Open
/wiki/France
/wiki/International_Tennis_Federation
/wiki/Association_of_Tennis_Professionals
/wiki/Roger_Federer
/wiki/Women%27s_Tennis_Association
/wiki/Rafael_Nadal
/wiki/Serena_Williams
/wiki/Fed_Cup
/wiki/Fred_Perry
/wiki/Rod_Laver

Using 0.005 as epsilon

/wiki/Grand_Slam_(tennis)
/wiki/French_Open
/wiki/The_Championships,_Wimbledon
/wiki/US_Open_(tennis)
/wiki/Australian_Open
/wiki/France
/wiki/International_Tennis_Federation
/wiki/Association_of_Tennis_Professionals
/wiki/Roger_Federer
/wiki/Women%27s_Tennis_Association
/wiki/Rafael_Nadal
/wiki/Serena_Williams
/wiki/Fed_Cup

/wiki/Fred_Perry
/wiki/Rod_Laver

The lists are the same using these two epsilons.

3. For each epsilon (0.01, 0.005): Number of steps that your page rank algorithm took to converge (within epsilon)

When using 0.01 as epsilon it took 4 steps to converge. When using 0.005 it took 5 steps.

4. For wikiTennis: Pages with top 15 page rank, top 15 indegree, and top 15 outdegree and Jaccard Similarities among these sets. Do this for epsilon 0.005

Top 15 rank:

/wiki/France
/wiki/Australia
/wiki/United_Kingdom
/wiki/Grand_Slam_(tennis)
/wiki/Spain
/wiki/Switzerland
/wiki/Czech_Republic
/wiki/Tennis
/wiki/Romania
/wiki/Belarus
/wiki/South_Africa
/wiki/Serbia

/wiki/French_Open
/wiki/Australian_Open
/wiki/The_Championships,_Wimbledon

Top 15 indegree:

/wiki/Grand_Slam_(tennis);
/wiki/Australia;
/wiki/United_Kingdom;
/wiki/France;
/wiki/Tennis;
/wiki/The_Championships,_Wimbledon;
/wiki/Spain;
/wiki/French_Open;
/wiki/South_Africa;
/wiki/Australian_Open;
/wiki/Switzerland;
/wiki/US_Open_(tennis);
/wiki/Czech_Republic;
/wiki/Romania;
/wiki/Belarus;

Top 15 outdegree:

/wiki/The_Championships,_Wimbledon;
/wiki/Grand_Slam_in_tennis;
/wiki/Career_Golden_Slam;
/wiki/List_of_Grand_Slam_related_tennis_records;

/wiki/Grand_Slam_(tennis);
/wiki/French_Open;
/wiki/Roger_Federer;
/wiki/List_of_Wimbledon_Open_Era_champions;
/wiki/List_of_Wimbledon_gentlemen%27s_singles_champions;
/wiki/Serena_Williams;
/wiki/Rafael_Nadal;
/wiki/Margaret_Court;
/wiki/List_of_Wimbledon_ladies%27_singles_champions;
/wiki/List_of_Wimbledon_mixed_doubles_champions;
/wiki/Australian_Open;

Sets	Jaccard Similarity
pagerank vs indegree	0.875
indegree vs outdegree	0.15384615384615385
pagerank vs outdegree	0.15384615384615385

5. For your favourite topic: Web pages with top 15 page rank, top 15 indegree, ad top 15 outdegree. Report Jaccard similarities among these sets. Do this for epsilon = 0.005

Using /wiki/basketball as seed url. The key words are “national basketball association”

Top 15 rank:

/wiki/National_Basketball_Association
/wiki/United_States
/wiki/Basketball
/wiki/Women%27s_National_Basketball_Association
/wiki/Los_Angeles_Lakers
/wiki/Major_League_Baseball
/wiki/Boston_Celtics
/wiki/National_Hockey_League
/wiki/New_York_Knicks
/wiki/Detroit_Pistons
/wiki/Chicago_Bulls
/wiki/Golden_State_Warriors
/wiki/Eastern_Conference_(NBA)
/wiki/Western_Conference_(NBA)
/wiki/Sacramento_King

Top 15 indegree:

/wiki/National_Basketball_Association
/wiki/Los_Angeles_Lakers
/wiki/Boston_Celtics
/wiki/New_York_Knicks
/wiki/Detroit_Pistons
/wiki/Western_Conference_(NBA)
/wiki/Eastern_Conference_(NBA)
/wiki/Basketball
/wiki/Golden_State_Warriors
/wiki/Chicago_Bulls

/wiki/Sacramento_Kings
/wiki/Philadelphia_76ers
/wiki/Atlanta_Hawks
/wiki/United_States
/wiki/Phoenix_Suns

Top 15 outdegree:

/wiki/NBA.com
/wiki/NBA
/wiki/National_Basketball_Association
/wiki/Western_Conference_(NBA)
/wiki/Eastern_Conference_(NBA)
/wiki/Magic_Johnson
/wiki/50_Greatest_Players_in_NBA_History
/wiki/National_Basketball_Association_Christmas_games
/wiki/2015%E2%80%9316_NBA_season
/wiki/List_of_National_Basketball_Association_seasons
/wiki/Bill_Russell
/wiki/List_of_National_Basketball_Association_head_coaches
/wiki/List_of_National_Basketball_Association_head_coaches_with_400_games_coached
/wiki/NBA_Finals
/wiki/Scottie_Pippen

Sets	Jaccard Similarity
pagerank vs indegree	0.6666666666666666
indegree vs outdegree	0.1111111111111111
pagerank vs outdegree	0.1111111111111111