Com S 435/535: Large Scale Dataset
Web Crawler, Page Rank and Spam Farm
Authors: Yijia Huang and Mohammad Wardat

1-Weighted Q:

Weighted Q is consist of

- **Tuple(String, double)**
    - A class with two attributes edge, weight.
- **HashMap with <String, Tuple>**
    - Take edge as key and the corresponding Tuple as value.
- **TreeSet<Tuple>**
    - Build a new TreeSet with a comparator
        - (t1, t2) -> t2.getWeight() >= t1.getWeight() ? 1 : -1
            - It is used to sort tuples in descending order
            - Keeping the new element in the FIFO order in the sequence if it already contains an element with the same weight.
- **add()** {
    - Use HashMap to check the existence edge.
        - Yes. Replace the old one by comparing the weight.
        - No. Add a new Tuple into the TreeSet.
- }
- **extract()** {
    - Remove and output the first element in the Treeset.
- }

2-Pseudo code of crawling Algorithm

int max

- Maximum number of pages wanted to crawl.

HashMap<String, Integer> visited

- It is used to check the visited urls and assign a new index with it.

```
crawling() {
    ●  Initialize weighted Q with seedUrl.
    ●  While (Q is not empty) {
        ○  Pop the most weighted tuple t of Q and extract the edge(u, v) in it.
        ○  If (The size of visited is not over the max and v is not visited.) {
            ■  Visit v.
                ●  Add new links to Q.
        ○  }
        ○  If (v is visited) {
            ■  Add edge(u, v) to the output list.
        ○  }
    ●  }
    ●  Write the output list to file.
}
```

In the algorithm, we apply BFS and we made the graph has maximal 500 vertices by stop adding links into the Weight Q after visiting the first 500 links (max). Therefore, the graph wihh have at most 500 vertices. Also, Each time When I add an edge I check if edge’s source and destination are visited. Thus, ensure it has to be both sides in the map.

3- Output of WikiTennisRanker

SeedUrl: /wiki/Tennis, Max Pages: 500
Topics: [racket, court, game]

Top 20 page rank: [442, 1, 42, 5, 449, 450, 448, 40, 39, 451, 38, 238, 41, 180, 202, 478, 77, 431, 311, 80]
Top 20 in degree: [5, 42, 39, 38, 40, 41, 442, 222, 232, 82, 123, 56, 238, 105, 180, 151, 155, 62, 57, 237]
Top 20 out degree: [5, 42, 38, 39, 40, 82, 222, 232, 16, 442, 41, 123, 105, 139, 115, 144, 156, 162, 217, 124]

Pair	Jaccard Similarities
A, B	0.42857142857142855
A, C	0.21212121212121213
A, D	0.25
A, E	0.25
B, C	0.2903225806451613
B, D	0.3793103448275862
B, E	0.42857142857142855
C, D	0.8181818181818182
C, E	0.7391304347826086
D, E	0.9047619047619048

4~5- Number of iterations

Epsilon	Beta	No. of Iterations
0.01	0.85	7
0.005	0.85	10
0.001	0.85	15
0.01	0.25	2
0.005	0.25	3
0.001	0.25	3

6- Spam Farm

Epsilon: 0.001, Beta: 0.85, Num of Iterations: 15
Before Spam Farm
Lowest Ranked Node: 53 with Rank: 3.490127268623332E-4

Epsilon: 0.001, Beta: 0.85, Num of Iterations: 15
After Spam Farm
Node: 53 with Rank: 0.015445353465935247
Difference: 0.015096340739072914