PAGERANK FOR IDENTIFYING CENTRAL PEOPLE IN NEWS ARTICLES

Karahan Sarıtaş - 2018400174
Department of Computer Engineering, Boğaziçi University
karahan.saritas@boun.edu.tr

ABSTRACT

In this assignment, we are expected to implement a PageRank-based method to identify the most important people occurring in news articles. The graph has been constructed from a subset of 3000 news articles from the Reuters-21578 corpus by identifying the person names. The vertices of the graph are defined as distinct people. An edge is constructed between two people if their names appear in the same news article. The resulting social network consists of 459 nodes and 1422 edges.

1 PageRank Calculation

We first construct the probability matrix P where P_{ij} stands for the probability of visiting page j from page i. In order to avoid getting stuck at dead ends, we jump to a random web page. At any dead-end, jumping to a random web page is called *teleporting*. In our case, we set the teleportation rate to 0.1. Initially P_{ij} is calculated as follows:

$$P_{i,j} = \begin{cases} \frac{1}{C(j)}, & \text{if there is a link from page } i \text{ to } j \\ 0, & \text{if no link from page } i \text{ to } j \end{cases}$$

$$C(j) = \text{number of out-going edges from page } j$$

Then we add the teleportation rate:

$$P_{ij} = P_{ij} * (1 - \text{teleportation rate}) + \frac{\text{teleportation rate}}{\text{total number of pages}}$$

To determine the most central people in the co-occurrence graph, we implement the power iteration method. We should multiply the probability vector by increasing powers of P until the product looks stable. Basically, what we have to do is to repeat the following operation until the stopping condition holds for an arbitrary value ϵ:

$$r^{(t+1)} = r^{(t)} * P$$

where $r^{(t)}$ is the probability row vector of size $R^{1 \text{ x (\# of pages)}}$.

$$\text{Stopping condition: } \|r^{(t+1)} - r^{(t)}\|_1 \leq \epsilon$$

where

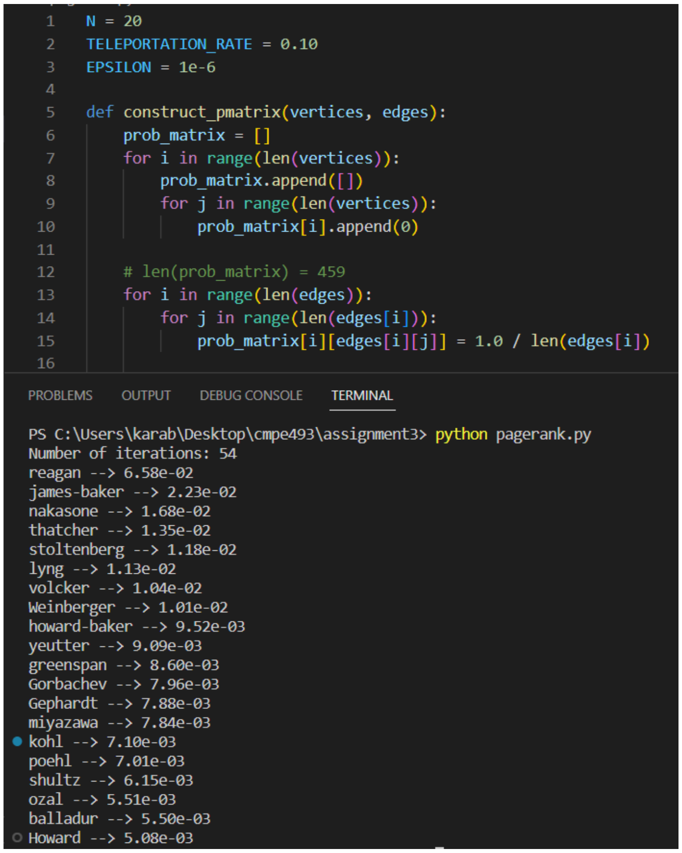
$$\|r^{(t+1)} - r^{(t)}\|_1 = \sum |r_i^{(t+1)} - r_i^{(t)}|$$

2 Results

Here you can find the list of the top 20 people along with their PageRank scores for $\epsilon = 10^{-6}$. As it can be seen, a majority of the people are prominent politicians of ’80s. Reagan was an American politician and actor who served as the 40th president of the United States from 1981 to 1989. James Baker worked as 67th United States Secretary of the Treasury under President Ronald Reagan. Nakasone was a Japanese politician who served as Prime Minister of Japan from 1982 to 1987. Nakasone was best known for his close relationship with U.S. President Ronald Reagan, popularly called the "Ron-Yasu" friendship - which explains why he is listed as one of the most central people. Since our dataset consists of texts from financial newswires, it is quite reasonable why we get famous politicians and not singers/actors/other prominent figures of that time.

```
reagan --> 6.58e-02
james-baker --> 2.23e-02
nakasone --> 1.68e-02
thatcher --> 1.35e-02
stoltenberg --> 1.18e-02
lyng --> 1.13e-02
volcker --> 1.04e-02
Weinberger --> 1.01e-02
howard-baker --> 9.52e-03
yeutter --> 9.09e-03
greenspan --> 8.60e-03
Gorbachev --> 7.96e-03
Gephardt --> 7.88e-03
miyazawa --> 7.84e-03
kohl --> 7.10e-03
poehl --> 7.01e-03
shultz --> 6.15e-03
ozal --> 5.51e-03
balladur --> 5.50e-03
Howard --> 5.08e-03
```

3 Screenshot



```python
N = 20
TELEPORTATION_RATE = 0.10
EPSILON = 1e-6

def construct_pmatrix(vertices, edges):
    prob_matrix = []
    for i in range(len(vertices)):
        prob_matrix.append([])
        for j in range(len(vertices)):
            prob_matrix[i].append(0)

    # len(prob_matrix) = 459
    for i in range(len(edges)):
        for j in range(len(edges[i])):
            prob_matrix[i][edges[i][j]] = 1.0 / len(edges[i])

```

```
PS C:\Users\karab\Desktop\cmpe493\assignment3> python pagerank.py
Number of iterations: 54
reagan --> 6.58e-02
james-baker --> 2.23e-02
nakasone --> 1.68e-02
thatcher --> 1.35e-02
stoltenberg --> 1.18e-02
lyng --> 1.13e-02
volcker --> 1.04e-02
Weinberger --> 1.01e-02
howard-baker --> 9.52e-03
yeutter --> 9.09e-03
greenspan --> 8.60e-03
Gorbachev --> 7.96e-03
Gephardt --> 7.88e-03
miyazawa --> 7.84e-03
kohl --> 7.10e-03
poehl --> 7.01e-03
shultz --> 6.15e-03
ozal --> 5.51e-03
balladur --> 5.50e-03
Howard --> 5.08e-03
```