Cmpe 493 Introduction to Information Retrieval, Fall 2015
Assignment 3 - PageRank for Identifying Central People in News Articles
Due: 11/01/2016 (Monday), 23:00

In this assignment you will develop a PageRank-based method to identify the most important people occuring in news articles. The *data.txt* file is a plain text file containing an undirected and unweighted graph of social network of co-occurrence in news articles. The graph has been constructed by reading 3000 news articles from the Reuters-21578 corpus and 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. The format of the data.txt file is as follows.

∗Vertices <number of vertices>
1 "label1"
2 "label2"
...
∗Edges
vertex1 vertex2
vertex3 vertex4
...

Implement and run the Pagerank algorithm (the power iteration method) to determine the most central people in the co-occurrence graph. Note that the provided network is undirected. Therefore, before applying the PageRank algorithm you should first convert it to a directed network as follows. For each edge $vertex1$ $vertex2$, include an edge in the opposite direction, i.e., $vertex2$ $vertex1$. Set the teleportation rate to 0.15.

You may use any programming language of your choice. Your program should take as input a text file in the same format as the provided datat.txt file and output the names of the top 50 people.

Submission: You should submit a *".zip"* file named as YourNameSurname.zip containing the following files using the Moodle system:

1. Report: List the names of the top 50 people. Discuss whether they make sense given that the data set is from the 1987 newswire.
2. Source code and executable: Commented source code and executables.
3. Readme: Describing how to run your program. I should be able to run your program using a different data set (in the same format as the data.txt file).

Late Submission: You are allowed a total of 5 late days on homeworks with no late penalties applied. You can use these 5 days as you wish. For example, you can submit the first homework 2 days late, then the second homework 3 days late. In that case you will have to submit the remaining homeworks on time. After using these 5 extra days, 10 points will be deducted for each late day.