Cmpe 493 Introduction to Information Retrieval, Spring 2023
Assignment 3 - PageRank for Identifying Central People in News Articles
Due: 05/06/2023 23:59 o'clock

In this assignment you will develop a PageRank-based method to identify the most important people occurring 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 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. 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 0.10.

You should use Python to implement your algorithms. We should be able to run your program by following the instructions in your readme file. You should NOT use any third party libraries, except the ones available in the Python Standard Library.

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

1. Report:
   (i) List the names of the top 20 people as well as their PageRank scores. Discuss whether they make sense given that the data set is from the 1987 newswire.

   (ii) Provide screenshots of running your program.

2. Source code: Commented source code.

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).

Honor Code: You should work individually on this assignment and all the source code should be written by you. You are NOT allowed to use any available libraries or any code written by other

people. Violation of the Honor Code will be strictly penalised, not only by a zero grade from the homework, but also by filing a petition to the Disciplinary Committee.

Late Submission: You are allowed 3 late days (until 08 June 23:59 o'clock) for this assignment with no late penalty. After that, 1 point will be deducted for each late hour (unless you have a serious excuse).