CSEN 272 Web Search and Information Retrieval

Project 1 (100 points)

Due: 5:40pm (submit your code and report to Camino), Tuesday, Feb. 18, 2025

Introduction

In this assignment you will implement the PageRank algorithm used for ranking webpages. You will then analyze the results of the algorithm as you vary its parameters.

You must complete this assignment on your own, not in a group. You are highly encouraged to use Python.

- Your code must obey the specified input-output format, accepting input from a text file and writing output to another text file.
- Other than numerical computing libraries like NumPy or SciPy, and built-in functions such as a random number generator, you cannot use any libraries.
- You need to turn in source code and a report (PDF and 1 page).

Input:

Your PageRank program will receive a text file as input and a teleporting parameter (denoted as *d* in the slides). The text file will encode a graph. In the text file, each line starts with a vertex number *i*, followed by a colon, followed by a sorted comma-separated list of numbers, where each number corresponds to a vertex that vertex *i* has edges to. For example, the text file might look like this:

0:2
1:2
2:0,1,3
4:0

In this example, vertex 0 has an edge to vertex 2, vertex 1 has an edge to vertex 2, vertex 2 has edges to vertices 0, 1, and 3, vertex 3 has no outgoing edges, and vertex 4 has an edge to vertex 0.

Output:

Your code will compute the PageRank vector, which is a length-*n* vector corresponding to the PageRank values of the *n* webpages. The output should be *n* lines, one for each vertex, where the *i*-th line outputted contains the value of the PageRank vector for the *i*-th vertex (where we index the first line outputted with 0)

For example if your PageRank vector is the length-*n* vector *p*, where the *i*-th entry is *p(i)*, then the output is:

p(0)
p(1)
p(2)
..
p(n)

Since each *p(i)* will be some decimal, we further require that each line be formatted in scientific notation to 10 decimal places.

Running your code:

Your code should run as follows:

python PageRank.py input.txt d > output.txt

If you want to use other languages than Python, please specify in the report how to run your code.

Report: Analyze different values of *d*

Create a web graph with 10,000 webpages. The value of *d* changes the PageRank vector. You will vary *d* between .75 and .95 with 0.05 increment to see how the PageRank vector changes.

- What convergence criterion did you use? How long did your PageRank algorithm converge in one run on average?
- How different are the top sites for each *d*? How different are the PageRanks of the top sites? How does the PageRank vector change as a whole? Write any observations you find.