BITS - Pilani,Hyderabad Campus

CS F469 IR Assignment - 2

Deadline: 21/10/2017

This assignment is aimed at designing and implementing the PageRank algorithm or the HITS Algorithm for ranking pages in a network.

Kindly continue with the same team for this assignment.

Programming Languages:

The assignment can be implemented in any programming language of your choice. STL's and inbuilt packages can be used for tasks like sparse representation of the matrix, visual representation and matrix multiplication. You are expected to code the core functionality of the model that you choose.

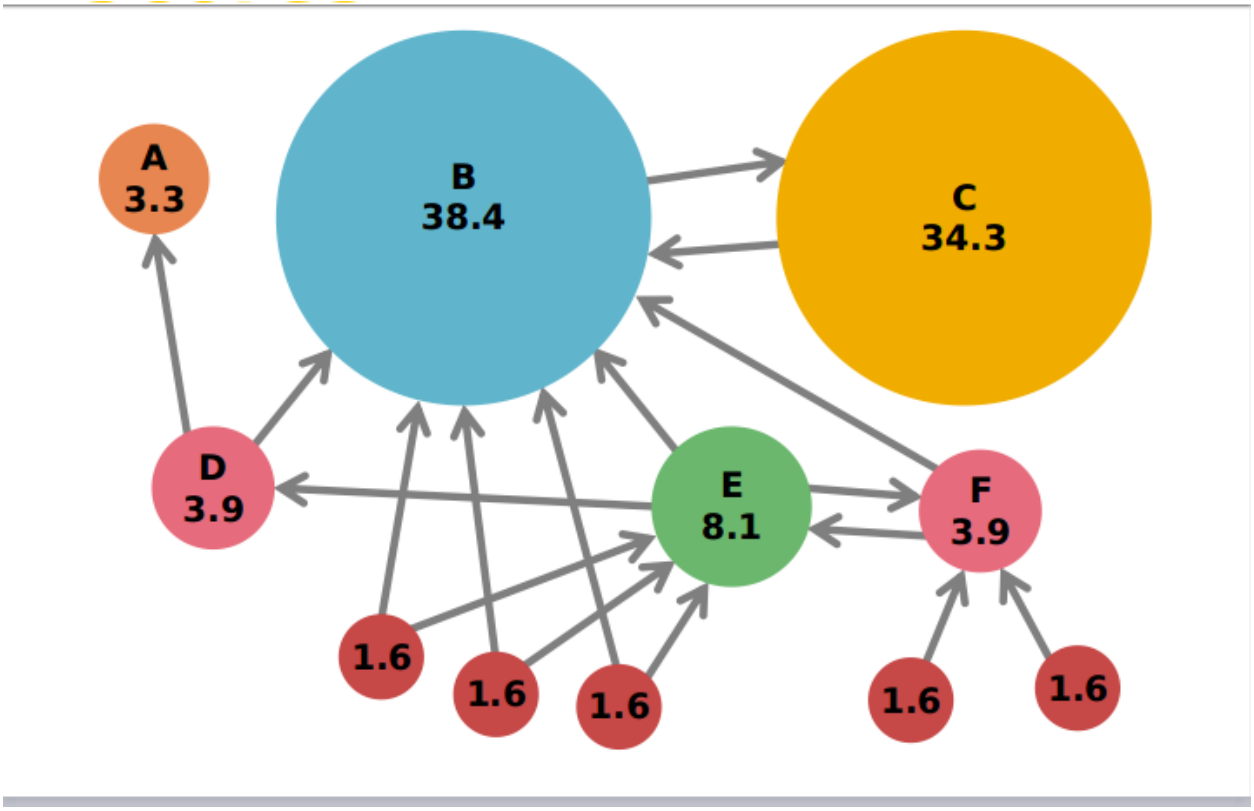
Task:

The task is to implement either the PageRank algorithm or the HITS - Hubs and Authorities algorithm taught in the class. The aim of implementing these algorithms is to rank the pages in the corpus by considering the inlinks and outlinks. These algorithms have varied applications like finding the most valuable research paper on a particular topic, highly talked about news articles, influential people in the social media network etc.

You are free to use your own dataset, scrape the internet, or use any of the datasets mentioned below.

Expectations:

1. Successful implementation of the algorithm on a reasonably sized dataset.
2. Handling Spider traps and Dead ends.
3. A design document which clearly mentions the formulas used and clear explanation of the results and insights that it gives.
4. After the entire algorithm has run, it should display a network of interconnected circles whose size varies as their importance. The edges are nothing but the inlinks and outlinks. For Eg:



Here the circles represent a page and the number inside them is some score which ranks them relatively with respect to each other. If the network has >100 Nodes you can display 30 - 40 nodes randomly.

5. Apart from the traditional implementation, students are expected to add some enhancement to the algorithm. Some possible enhancements have been mentioned in the *instructions* section

Additional Resources:

1. PageRank Algorithm:
    a. Mining in Massive Datasets - Stanford Module 2 and 15
    b. PageRank implementation
    c. PageRank on wikipedia data
    d. PageRank on twitter Memes Dataset
2. HITS Algorithm - Hubs and Authorities:
    a. Hubs and Authorities - Stanford
    b. Mining in Massive Datasets - Stanford Module 15 (video #91)
    c. University of Michigan - Coursera Video

3. Datasets:
    a. Stanford SNAP Datasets
    b. PR Data Code
4. For Visualization
    a. Graph - tool
    b. Plotly
    c. Boost - C++ libraries
    d. iGraph
    e. For JAVA

Instructions:

1. Take data of reasonable size. Make sure the data can be loaded entirely into the RAM at once. As the algorithms involve matrix multiplications and multiple iterations ensure that the data isn't too large.
2. Few enhancements one can implement:
    a. Topic specific PageRank
    b. Comparing the results of search with and without implementing PageRank. (Use the IR Model made in the first assignment.)
    c. Trying to implement it on very large graphs using sparse matrix representation (Make sure the data is large enough in this case)
    d. Incorporating Trust rank into your model
    e. Visualising each iteration and change in each page's score graphically

Deliverables:

The final submission must contain the following documents:

1. **Design Document** –This document should contain all the formulas and packages used along with a brief description. All the assumptions, pros & cons of your model, running time etc. should be well documented. Make sure names and ID numbers of all the team members are mentioned on the first page itself.
2. **Code** – The code should be well commented.
3. **Documentation** – All the classes, functions and modules of the code must be documented. Software that automatically generate such documents can be used – pydoc for Python, Eclipse for Java etc.
4. **README** – The README file should describe the procedure to compile and run your code for various datasets.

Submission Guidelines:

All the deliverables must be zipped and submitted to **bphc.ir@gmail.com** latest by **deadline.**

You are expected to demo your application and present your results as per the schedule that will be made available.

Evaluation Criteria for Task :

S.No.	Task	Marks
1.	Implementation of the Algorithm	15
2.	Design Document	5
3.	Visual Representation	10
4.	Handling Spider traps and Dead Ends	5
5.	Viva	5
6.	Enhancement	10
	Total	**50**

It should be noted that all the assignments would be run through a plagiarism detector and based on the results, the marks would be altered. The final decision lies in the hand of the instructor and only one submission per group would be allowed for one assignment.