Spring 2009	**CSC 466:** Knowledge Discovery from Data	Alexander Dekhtyar

Lab 7: PageRank and Link Analysis

Due date: Friday, June 5, midnight.

Overview

In this assignment you will implement PageRank ranking algorithm and run it on two datasets: the 2009 NCAA Football season data and the list of state borders within the USA.

Assignment Preparation

This is a pair programming assignment. Each student teams up with a partner. Each team submits only one copy of the assignment deliverables.

Data

You will be using two datasets provided to you by the instructor. The first dataset, `NCAA-FOOTBALL` dataset, available at

> http://wiki.csc.calpoly.edu/datasets/wiki/NCAAFootball

contains information about **every single game** played by Division I teams in the 2009 NCAA regular football season (before bowls and championships started). A total of 1537 games was played, their results are documented in the `NCAA_football.txt` file.

The second dataset, `STATES`, (or `StateNeighbors`) contains information about the 48 mainland U.S. states plus the District of Columbia and the borders that these states share. (Alaska and Hawaii are excluded from the dataset since they do not share borders with other states). The dataset is available at

> http://wiki.csc.calpoly.edu/datasets/wiki/StateNeighbors

There are a total of 214 borders reported in `stateborders.csv` file. This dataset is symmetric (if a state X shares a border with a state Y, then a state Y shares a border with state X), therefore only 107 actual shared borders exist in the US[1].

Data Format. To simplify parsing, both datasets are available to you in a uniform data format. This format makes sense for the football games, but has some redundant information in the state borders data file.

The format of the `NCAA_football.txt` file from the `NCAA-FOOTBALL` dataset is:

```
<Team1> <Team1_Score>, <Team2> <Team2_Score> [(OT)]
```

For example, `NCAA_football.txt` file contains the following lines:

```
Ball State 48, Northeastern 14
Buffalo 42, UTEP 17
Central Michigan 31, Eastern Illinois 12
Southeast Missouri State 35, Southwest Baptist 28 (OT)
```

The interpretation of this data is straightforward. For example, the first line above states that `Ball State` and `Northeastern` played a game and `Ball State` won with the score of 48 to 14. Games that went into overtime (e.g., the game between `Sotheast Missouri State` and `Southwest Baptist`) are marked with an `(OT)` note at the end of the line.

The format of the `stateborders.csv` file from the `STATES` dataset is:

```
<State1> 0, <State2> 0
```

Here, `<State1>` and `<State2>` are two-letter codes for state (e.g., AZ, CO, DC). Zeroes are redundant, they are included for compatibility with the format of the `NCAA_football.txt` file. `stateborders.csv` file represents an undirected graph: each state border is listed twice. (you can treat the .csv file as the list of directed edges).

Here are a few sample lines from the `stateborders.csv` file:

```
DE 0, MD 0
  FL 0, GA 0
  FL 0, AL 0
  GA 0, FL 0
```

[1]The dataset does not treat the Four Corners intersection of Utah, Colorado, New Mexico and Arizona as a four-way border, i.e., New Mexico DOES NOT share a border with Utah and Colorado DOES NOT share a border with Arizona.

These four lines indicate that Delaware shares a border with Maryland, and that Florida shares borders with Alabama and Georgia. Note that the latter shared border is listed twice: once in a `FL, GA` pair and once in a `GA, FL` pair.

The `STATES` dataset was adapted from a one of the data files for Hartigan's book on clustering (see Lab 4 data). The link to the original file is available on the `STATES` dataset wiki page, and is also provided here:

http://people.sc.fsu.edu/~burkardt/datasets/hartigan/file28.txt

The data format is explained inside the file. The data is presented in the form of adjacency lists.

Lab Assignment

Write a program that takes as input a data file formatted in the way described above (the format of `NCAA_football.txt` and `stateborders.csv` files), runs the PageRank analysis on the graph extracted from the input file and outputs the individual items (football teams, states) ranked in descending order of their computed PageRank together with the PageRank score and the rank.

The program, `pageRank.java` shall take as input the file name. It may also (if you want) take as input a flag specifying whether the dataset you are reading in represents a directed or undirected graph (some of you may find it useful, but it is not absolutely necessary so it is left up to you. Please specify in the README file if you have the flag and if it is mandatory for your implementation).

It should parse the input, create a graph structure from it in main memory and run a version of PageRank ranking algorithm to rank the nodes in the graph.

You may implement the version of PageRank that was discussed in the class. You may also use extra information available to you (in case of the football season data: the score of the game and whether there was an overtime) to adapt your PageRank computation.

Your program should output an ordered list of **all** nodes in the graph. It should print the rank and the PageRank score of each of them. For example, your program can output:

```
1       obj: Mississippi with pagerank: 0.030110446720353286
2  obj: Florida with pagerank: 0.02406493620983336
3  obj: Utah with pagerank: 0.01609201202237497
4  obj: Oklahoma with pagerank: 0.01531666186988849
```

as the first four items for the `NCAA_Football.txt` input file.

(note, the actual results may vary from the one above)

Extra Credit

Experimentation with PageReank is subject to extra credit in the amount of 15% – 20%. To be eligible for the extra credit, your submission must have the following:

- Implement an extension of PageRank that tries to take in to account the score of the game (i.e., the numeric values used in the input) or any other information.
- Make this extension customizable by a input parameter (collection of input parameters).
- Describe your extension in the README file.

To get the extra 5% of the credit you may also do the following:

- Try to find the optimal values of the custom input parameters that provide for the best, in your opinion, ranking of the (at least) top 25 - 40 football teams. Feel free to compare the rankings you obtained with the final regular season rankings by AP, Coaches, BCS standings, etc. Links to these rankings will be made available to you.
- Report your findings either in the README file, or in a separate micro-report file that you submit.

Deliverables and submission instructions

This lab has only electronic deliverables. Submit the following:

- Source code for your program.
- `README` file.
- Any extra credit files.

Submit all electronic deliverables as a single zip of gzipped tar archive (`lab07.zip` or `lab07.tar.gz`). Use the following command

```
$ handin dekhtyar-grader lab07 lab07.<ext>
```