Home Assignment 2

Version 1.0

Organization

Build groups of five people and complete the provided assignment with the help of the Map/Reduce framework *Hadoop*. Nominate a team captain who registers the group by creating a subversion repository named `WSWT10_<GROUPNAME>` (replace <GROUPNAME>with your chosen groupname). Add the study assistants as readers and the group members as members.

If there are problems with the assignment please use the **newsgroup** as a primary method for communication. Others might have the same/similar problems.

Task

Your task is to implement the MapReduce steps for the Twitter Ranking algorithm TunkRank. TunkRank is a ranking measure developed by Daniel Tunkelang to measure the influence of users of Twitter. The approach is to some extent similar to *PageRank*, for details about the TunkRank algorithm see the Tutorial slides. For simplicity, you only need to implement the first iteration of the TunkRank algorithm in this assignment and set the p value to 0.02 globally. The influence of each node at start is 1.0.

Implementation

Use *Python* to implement the MapReduce step(s) for TunkRank. Use *Bash* scripts if you decide to wrap your code. Every file you submit must contain the names and matr. numbers of all group members.

You do not have to solve the assignment in one MapReduce step or one program. Just make sure you explain how to use your code in the readme file.

More information on how to set up and configure Hadoop can be found in the corresponding tutorial slides: `http://kmi.tugraz.at/staff/markus/courses/SS2010/707.000_web-science/slides/hadoop-tutorial.pdf`

Provided Files

For the assignment you are given a sample of a dataset. The archive contains a single file which is tab separated and each line represents a user (first column) and one of her followers (second column).

`ftp://ftp.tugraz.at/pub/WSWT10/a2/twitter_rv_5533517.tar.gz` [530MB]

Structure of your repository

- `report.pdf` (one page explaining your approach – keep it short! max. 1 page)
- bash script (optional)
- `python/`
  - `mapper_1.py` (mapper)
  - `reducer_1.py` (reducer)
  - `...` (additional mappers and reducers - if needed)
  - `readme.txt` (UTF-8-encoded)

  `results/`

- `tunkrank_run_1.txt` (top 10000 twitterers in descending order + their tunkrank score)
- `...` (additional iterations for bonus points)

Submission

Home Assignment 2 is due **June 18, 2010 12:00** (high noon).

The due date is a *soft deadline*. That is, your score on the assignment will be rated 100% if you hand in the assignment before 12:00. The following 12 hours are suitable for a submission as well, *but* your points will be rated 66%. Read: 1/3 of your points will be subtracted if you hand in your assignment between 12:00 and 23:59. 24:00 is the *hard deadline*; if you hand in anything after 24:00 you will not receive any points.

Submission is done using the SVN version control system. (See instructions on the course website.)

The "Abgabegespräche" will be on Tuesday, June 22nd 2010. The exact time slots for each meeting will be announced in the newsgroup for this course. **It is mandatory for each team member to participate**. Students who are not present at the "Abgabegespräche" will get no points on the assignment.

Policies

- By submitting your code and report you agree that your submission will be checked for plagiarism.

Resources

- `http://www.michael-noll.com/wiki/Running_Hadoop_On_Ubuntu_Linux_%28Single-Node_Cluster%29`
- `http://www.michael-noll.com/wiki/Running_Hadoop_On_Ubuntu_Linux_%28Multi-Node_Cluster%29`
- `http://www.michael-noll.com/wiki/Writing_An_Hadoop_MapReduce_Program_In_Python`