Task: Compute the PageRank scores on the Wikipedia dataset.

Dataset:

- [links-simple-sorted.zip](http://users.on.net/~henry/pagerank/links-simple-sorted.zip) (323MB)
- [titles-sorted.zip](http://users.on.net/~henry/pagerank/titles-sorted.zip) (28MB)

The above files contain all links between Wikipedia pages (English page only). In links-simple-sorted.zip, there is one line for each page which contains the destination pages of the out-links from one specific page. The format of the lines is as follows:

```
from1: to11  to12  to13 ...
from2: to21  to22  to23 ...
...
```

Here from1 is an integer denoting the source page, and to11 to12… are integers denoting the destination pages of the out-links from the corresponding source page. The n-th line in the file titles-sorted.zip is the page title of page n.

In this project, you need to install your Hadoop environment and use MapReduce to complete the PageRank calculation on the given datasets. You need to report the titles of Top 100 pages and the corresponding PageRand scores. You can choose different parameters, such as the teleport parameter, of PageRank algorithm and do comparison on different results. One result you must report is the result when the teleport parameter (beta) is set to 0.85. In addition, you need to compare the PageRank results with the top pages ranked by the number of in-links and give some analysis.

Of course, you can do more interesting work on the dataset. You are greatly encouraged to implement your own ideas on the datasets. Teams with novel ideas or more trials will get a 20% bonus in the final score.

You need to write a report about this project. The report should include but not limited to the following contents:

1. Background, understanding and analysis of the task;
2. Details of the algorithm and platform;
3. Experimental results and findings;
4. The contribution of each member of the group;
5. Reference.

The deadline for Project 2 is July 02, 2013 (23:59pm).

Package your clean code, final report and other relevant files, name the package as MMDS-PRO2-[成员姓名].[rar/zip].

Some notes:

1. You may find useful information from:
   - Course slides: [http://www.cs.sjtu.edu.cn/~liwujun/course/mmds.html](http://www.cs.sjtu.edu.cn/~liwujun/course/mmds.html)

2. Final package should be sent to TA (xiaoyu199175@gmail.com) by the team leader.

3. Only cluster mode will be accepted. It is strongly recommended to use real Ethernet instead of virtual machines. If you do so, please mention it in your report.

All your submitted assignments must be entirely your own (or your own group's). Any student found cheating or performing plagiarism will receive a final score of zero for this course.