Computational Social Science
APAM E4990
Spring 2013

Homework 2

1. Word count

In this problem you'll use Amazon's Elastic MapReduce to count word occurences throughout the English version of Wikipedia. A recent dump of Wikipedia[1] is available on Amazon's S3 at `s3://css2013/enwiki-text`, where each line represents one article in the following format:

```
id    url                                          text
--    ---                                          ----
673   http://en.wikipedia.org/wiki/Atomic_number   Atomic number.\nIn chemistry and physics, ...
```

Write and execute a Pig job that splits the text in each line to individual words using Pig's tokenize function and then counts the occurence of each word across all articles. The output of your job should be a two-column tab-separated file with words and their counts for all words that appear at least 10 times.

See Peter Norvig's "How to Write a Spelling Corrector"[2] if you'd like to use these counts to build a simple spell checker.

2. Page popularity

Hyperlinks between these Wikipedia articles are available on Amazon's S3 at `s3://css2013/enwiki-edges`, where each line represents one link in the following format:

```
source_id    source_url                                   target
---------    ----------                                   ------
673          http://en.wikipedia.org/wiki/Atomic_number   Physics
673          http://en.wikipedia.org/wiki/Atomic_number   Chemistry
673          http://en.wikipedia.org/wiki/Atomic_number   Moseley%27s_law
```

Write and execute a Pig job that calculates page popularity, as measured by the number of distinct incoming links to each article. The output of your job should be a three-column tab-separated file with source ids, source URLs and their popularity, ordered by descending popularity. There is no need to submit the complete output file—instead provide a plot of the degree distribution for incoming links across all articles and include the counts for the top 10 articles in your report.

3. Tie strength

In this exercise you'll investigate tie strength between co-authors as measured by number of common collaborators. In particular, you'll examine the network of authors who have published at least one paper with the prolific mathematician Paul Erdos[3].

The provided file represents an undirected co-authorship network where an edge exists between two individuals if they have jointly authored at least one paper together. For each pair of co-authors, calculate the tie strength between them as the number of co-authors they have in common divided by the total number of distinct co-authors between the pair (i.e., the Jaccard Index). Provide a plot of the distribution of these tie strengths along with a tab-separated file that lists each author, their total number of collaborators, and their top three collaborators ordered by descending tie strength.

In contrast to the first two problems, there's no need to use Hadoop for these calculations.

[1] `http://en.wikipedia.org/wiki/Wikipedia:Database_download`
[2] `http://norvig.com/spell-correct.html`
[3] `http://vlado.fmf.uni-lj.si/pub/networks/data/Erdos/Erdos02.net`