Cloud Computing: Lab Exercise 4

1. Repeat the word counting exercise for the literary texts, making sure to eliminate punctuation marks such as ”!'.?;:, etc.

   Hint: take a look at the `sed` and `tr` Unix commands.

2. Use the Spark infrastructure to generate a Monte Carlo estimate for the value of π; you can take the sample code `pi.py` from BlackBoard (adapted from `https://github.com/apache/spark/blob/branch-1.6/examples/src/main/python/pi.py`. How does it work? Modify it to compute:

   (a) The value of π from the volume of a sphere.

   (b) The value of the natural logarithm of 2.

3. Complete the `parse_edges,py` example to parse a graph from an input file; you will need to filter out comment lines (starting with `#`);

4. Write a Spark to compute the outdegree (indegree) of each node in a graph, i.e. the number of edges that have that particular node as the first (second) in the edge;

5. Complete the pagerank example, and apply it to the graph data available.

6. Modify the pagerank example to allow for a weight vector.

7. Use the Spark infrastructure to study the data provided for the DEBS 2015 Grand Challenge `http://www.debs2015.org/call-grand-challenge.html`. The dataset records a years worth of taxi trips for the city of New York. The website contains links to subsets of the data for testing purposes. To get the data for this example first go to https://stackoverflow.com/a/39225039 and copy that code into a file dlChromeLink.py. Then run

   ```
   python dlChromeLink.py 0B0TBL8JNn3JgTGNJTEJaQmFMbk0 taxi.csv.gz
   sudo yum install gzip.x86_64
   gunzip taxi.csv.gz
   ```

   In particular:

   (a) Compute the extremals for the longitutde/latitude location data.

   (b) Divide the pickup (dropoff) locations into a grid with approximately 100 (1000) square cells, and count how many trips start (end) in each cell.

How to reduce the verbosity of `spark-submit`:

- Go to directory `/etc/spark/conf`
- Execute `sudo cp log4j.properties.template log4j.properties`
- Edit `log4j.properties` and change the appropriate line from
  `log4j.rootCategory=INFO, console`
  to
  `log4j.rootCategory=ERROR, console`