Homework 3

January 19, 2018

- Find the top 100 words of the file quixote.txt with map-reduce.
- Find the number of words that start with each letter of the English alphabet in the file quixote.txt.
- Write a pySpark program that implements the social network friendship recommendation algorithm, which was discussed in class, using the data soc-data.txt. You may need to use –driver-memory 8G to set the runtime memory to 8GB.
- Transform the rows of the first two trade files of the course to a human-readable version, and then create two tables with those two files using Spark.
  - See https://github.com/rxin/spark/blob/master/examples/src/main/python/sql.py
- Explain the algorithm sortByKey of Spark.
  - See https://github.com/apache/spark/blob/master/examples/src/main/python/sort.py
- Implement multinomial logistic regression using Spark.
  - See https://github.com/apache/spark/blob/master/examples/src/main/python/logistic_regression.py
  - Train your algorithm using the MNIST dataset