CENG 686 Distributed Data Processing and Analysis
Erdogan Dogdu

Assignment 2

Subject: Spark

1. Download (The Complete Works of William Shakespeare) http://www.gutenberg.org/cache/epub/100/pg100.txt
2. Develop a Spark program to compute the term co-occurrence matrix for the downloaded text using the "**Pairs**" approach as explained in the MapReduce-2 lectures notes (pg. 26-29). Consider "sentences" only for the term co-occurrence context. Give the first 50 terms' results.
3. Compute the term co-occurrence matrix for the downloaded text using the "**Stripes**" approach as explained in the MapReduce-2 lectures notes (pg. 30-36). Give the first 50 terms' results.
4. Make sure the results are the same for both solutions in the above two steps. Now report the computation time difference (in seconds) between the two approaches. Which one is better and how much better?

Resources:

- Section 3.2 "Pairs and Stripes" of the text book "Data-Intensive Text Processing with MapReduce" by Jimmy Lin and Chris Dyer (http://lintool.github.io/MapReduceAlgorithms/).
- Spark Programming Guide https://spark.apache.org/docs/2.2.0/rdd-programming-guide.html
- Spark Tutorial: https://spark.apache.org/docs/latest/quick-start.html
- You can run the programs online at: https://community.cloud.databricks.com/
- Spark Tutorial at Cloudera https://www.cloudera.com/documentation/enterprise/5-5-x/topics/spark_develop_run.html
- To install Spark http://spark.apache.org/downloads.html