CS 6350

Assignment 1

## Name (NetID): Reg Gonzalez (rdg170330)

## Number of free late days used: 0

## Sources/References:

* “Topic 3 - Hadoop MapReduce” Slides & Notes
* “Topic 4 - Information Retrieval using MapReduce” Slides & Notes
* “Topic 4 - Apache Spark” Slides & Notes
* PySpark Lab 2
* PySpark Lab 3
* Natural Language Toolkit (NLTK) Official Documentation (<https://www.nltk.org/>)
* “The Discoverie of Witchcraft” book from Project Gutenberg (<https://www.gutenberg.org/ebooks/60766>)
* CMU Movie Summary Corpus (<https://www.cs.cmu.edu/~ark/personas/>)
* Documents Similarity (<https://text2vec.org/similarity.html>)
* Tf-Idf and Cosine Similarity (<https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/>)