CS 6350

ASSIGNMENT \_\_\_\_\_\_1\_\_\_\_

Names of students in your group:

Rui Yang

Number of free late days used: \_\_\_\_\_\_\_\_\_\_\_\_\_1\_\_\_\_\_\_\_   
Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

<https://mallikarjuna_g.gitbooks.io/spark/content/spark-mllib/spark-mllib-transformers.html#StopWordsRemover>

<https://mallikarjuna_g.gitbooks.io/spark/content/spark-mllib/spark-mllib-transformers.html>

<https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/>

<https://spark.apache.org/docs/latest/rdd-programming-guide.html>

<http://bigdatatn.blogspot.com/2017/05/joining-spark-rdds.html>

<https://johnsnowlabs.github.io/spark-nlp-workshop/databricks/index.html#scala/annotation/2-%20Pre-trained%20Pipelines%20-%20onto_recognize_entities_sm.html>

<https://johnsnowlabs.github.io/spark-nlp-workshop/databricks/index.html#scala/annotation/2-%20Pre-trained%20Pipelines%20-%20onto_recognize_entities_sm.html>

https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/

https://stackoverflow.com/questions/51890678/how-to-convert-key-arrayvalue-to-key-value-in-spark

Part1

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/8750190078783311/1293224766374745/5504697096027154/latest.html>

Part2

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/8750190078783311/3860953457484921/5504697096027154/latest.html>