Files Submitted

* Read Me
* Source Code folder containing the source code for both pyspark and neo4j
* Datamanipulation folder containg the bash sscript for processing the raw data create a better format of ratings.csv files
* Project Report

HOW TO EXECUTE

**DATA**

* Please download the formatted dataset for execution from the following link-

<https://drive.google.com/file/d/0B2yCALEhyyVVVHVVUFdGZUdCZE0/view?usp=sharing>

* Data is too big to attach in the original submission

1. **PySpark** - Instructions for executing the scala code in Databricks-

* -Create a Spark Cluster version - Spark 1.6.1 (Hadoop 2)
* create a python notebook and attach the code to the cluster
* open the CL.ipynb in notepad++ and copy the code to this created notebook.
* create a path for dataset by uploading the dataset in the link (

<https://drive.google.com/file/d/0B2yCALEhyyVVVHVVUFdGZUdCZE0/view?usp=sharing>

) given above to Databricks. This is the new path that is generated

* Load the python code to this notebook and Replace the path of the ratings.csv file with the new path
* Click Run-all to run the code.

1. **Neo4j**

Note – please upload only a sample of dataset as uploading all the data will take very large amount of time in neo4j

* Install the community edition of Neo4j from their official website also listed below:
* <https://neo4j.com/download/>
* Run Neo4j Community Edition, set the database path and click on start.
* The windows shows the default localhost to open in order to access Neo4j i.e http://localhost:7474/. Open it in browser
* Set a new password as a first time user and type :server connect in the editor.
* Once connected, create a new directory called as import inside the database path and copy the dataset to that folder
* Run the code provided inside jokeRecommenderNeo4j.txt and view the output.

1. Data Manipulation code located in the data manipulation folder

* This code manipulates the raw data to create new raw data in a triple format.
* download the original raw data from http://eigentaste.berkeley.edu/dataset/

jester\_dataset\_1\_1.zip: (3.9MB

jester\_dataset\_1\_2.zip: (3.6MB

jester\_dataset\_1\_3.zip: (2.1MB)

* Unizip the file and store it in a location
* In the same location store the files ‘scriptname.sh’ and ‘DataManipulation.py’ . These files can be downloaded from the folder Data Manipulation
* Run the following shell script in Linux command bash ( Ubuntu 14.04 Trusty) ; package requirement - Package Apt-get, Python
* R un the shell script using the command : bash scriptname.sh
* This will generate the ratings.csv file which can be fed as input path to the pyspark and neo4j code