**Environment Setup: All instructions are for windows environment only**

**Main Attachment: Project Prototype.zip contains all files necessary to run the project**

**All coding is done using jupyter notebook and files are in ipynb format. Please use jupyter notebook for code execution.**

Software to install:

1. Python – Version 3.7.9
2. Mongo DB – Version 4.4.1 -> <https://docs.mongodb.com/manual/tutorial/install-mongodb-on-windows/>
3. Apache Spark – Version 3.0.0 -> <https://spark.apache.org/downloads.html>
4. Dependency jars for spark-mongo db connection: All jar files are enclosed with the main attachment.

mongo-spark-connector\_2.12-3.0.0.jar

mongodb-driver-core-4.0.5.jar

mongodb-driver-sync-4.0.5.jar

scala-library-2.12.11.jar

bson-4.0.5.jar

Copy the above jar files from jar\_files folder to spark-3.0.1-bin-hadoop2.7/jars folder

Make sure that path variables are set in environment variables for JAVA\_HOME, SPARK HOME.

In our case, below are the user variables

|  |  |
| --- | --- |
| Variable name | Value |
| JAVA\_HOME | C:\Java\jre1.8.0\_261 |
| SPARK\_HOME | C:\Spark\spark-3.0.0-bin-hadoop2.7\spark-3.0.0-bin-hadoop2.7 |

Please update the environment variables accordingly.

1. Configuring pyspark in jupyter notebook: Please follow ‘Method 2 — FindSpark package’ in the below link

<https://www.sicara.ai/blog/2017-05-02-get-started-pyspark-jupyter-notebook-3-minutes>

**Instructions to load the data in mongo DB:**

We have created ‘Twitter’ database and total eight collections for our data while the data is loaded directly into mongo database using REST API with different search criteria for each genre. To completely load the data, please follow the below steps,

* Copy the data dump (File name: Full\_json file) into the mongo db installed bin folder.
* Go the mongo db bin directory using the command terminal and import the data using the below commands,

mongoimport --db Twitter --collection twitter\_action\_netflix --file Full\_json/twitter\_action\_netflix.json --jsonArray

mongoimport --db Twitter --collection twitter\_action\_rakhi --file Full\_json/twitter\_action\_rakhi.json --jsonArray

mongoimport --db Twitter --collection twitter\_comedy\_netflix --file Full\_json/twitter\_comedy\_netflix.json --jsonArray

mongoimport --db Twitter --collection twitter\_action\_netflix --file Full\_json/twitter\_action\_netflix.json --jsonArray

mongoimport --db Twitter --collection twitter\_comedy\_rakhi --file Full\_json/twitter\_comedy\_rakhi.json --jsonArray

mongoimport --db Twitter --collection twitter\_horror\_netflix --file Full\_json/twitter\_horror\_netflix.json --jsonArray

mongoimport --db Twitter --collection twitter\_romance\_netflix --file Full\_json/twitter\_romance\_netflix.json --jsonArray

mongoimport --db Twitter --collection twitter\_romance\_rakhi --file Full\_json/twitter\_romance\_rakhi.json --jsonArray

Note: If there is any trouble in setting up MongoDB or loading data to it, Please start with Recommendation System\_Data Preprocessing.ipynb and run cell 1 and 2 which installs and loads libraries and then start the execution from cell 8(convert the cell type to code). Ignore all cells in between.

This will load the data from csv file (df\_show.csv) which has data read from Mongo DB to the pyspark dataframe (df\_show) and continue the execution.

**Notebooks**

Please run the notebooks in below order

1. Recommendation System\_Data Preprocessing.ipynb
2. Recommendation System\_Sentiment\_Analysis\_Model.ipynb

**Instructions to run the jupyter notebook file 1: Data extraction and processing in Spark**

1. First stage is to read the data from mongodb collections into pyspark dataframe, preprocess, extract show names, match with actual show names and write the final dataframe to a csv file.

Use Recommendation System\_Data Preprocessing.ipynb

Please place Netflix\_updated.csv file in the home directory of the notebook so that no changes are required in the code.

Input csv – Netflix\_updated.csv

Output csv – mytweetmodel\_.csv, mynetflixcsv.csv

Please note that the complete execution of this file will take at most 20 minutes with the below system configuration:

Operating system: Windows

RAM size: 8 GB

Processor: Intel Core 8265U CPU @ 1.80 GHz

**Instructions to run the jupyter notebook file 2: Sentiment Analysis and Model creation in Python**

1. Here we calculated the sentiment score for each user and create a deep learning keras model to recommend Netflix shows based on tweet sentiment. We tried one conventional matrix factorization approach (SVD) as well.

* Use Recommendation System\_Sentiment\_Analysis\_Model.ipynb
* Please place all csv files (mytweetmodel\_.csv, mynetflixcsv.csv, Netflix\_updated.csv, IMDB\_full\_rating\_.csv) in the home directory of the notebook so that no changes are required in the code. The output file mytweetmodel\_.csv created in preprocessing is to be used as input here. Please find the backup mytweetmodel\_.csv file enclosed with the attachment.

Input csv – mytweetmodel\_.csv, Netflix\_updated.csv, IMDB\_full\_rating\_.csv