20181117_Batch 44_CSE 9099c_PhD_Instructions

Source of Data

Data is available in two of MySQL table, details of that given below:
Database: - insofe_b44_phd_data
Table name: b44_phd_train (For training data)
b44_phd_test (For test data)

Data Description

Data set contains two fields, review and rating. Reviews are about various eateries and ratings provided by different customers.

Rating	Review
3	Not sure why there are such bad reviews for
5	This is Jersey Boys as in Frankie Valli and the
1	I am curious know of how much they have pa
3	Wynn oh how I want to love you so... with sp
2	I took my kid in for wash/deep cond, she has
5	There is not a single thing about my experier
2	Not that authentic. Taqueria Guadalajara is m

The range for rating is from 1 to 5, 1 being the lowest and 5 is the highest rating.
Review field of the dataset may be empty or contains only single digit number, please treat them as bad records and you can ignore them.

Data Ingestion

Prepare sqoop jobs to take the training and test data to your respective hdfs location.
Format for HDFS path: "/user/<your user id>/B44/PHD_DATASET/ "

Data Pre-processing

Once data is available in your respective HDFS location, as part of the next step you need to pre-process the data.
You are supposed to follow all the pre-processing steps which you are familiar of, like tokenization, removal of noisy words like URLs, numbers, words containing special characters which do not add much value to the meaning of the word, stop word removal etc.
All this has to be done using Spark-Data Frame. Refrain from using pandas data frame and later on converting it to Spark Data Frame.

Model Building - Classification

You need to build a classification model using Spark ML You are free to use any model of your choice and metric to evaluate the model's performance is **Accuracy**. Perform experiments which will help in increasing the accuracy of your model.

20181117_Batch 44_CSE 9099c_PhD_Instructions

Note: Please process train and test data separately, you should not read both the data at the same time and do your pre-processing, then later you split to train and test. We have given separate train and test data.

Model Building – Clustering

Apart from classification you are required to build a clustering model. Please take the reviews from the train data-set and try to segregate them into different cluster, you can use clustering algorithm like K-means. Justify your approach for number of clusters chosen.

Once you are done with the clustering model, predict on train data and try to see the distribution of labels/ratings in each cluster. Please create simple bar-plots (using python), showing the number of different kinds of rating in each cluster. (For example, cluster 1 may have 50 numbers of reviews of rating 1, 40 number of reviews of rating 5 so on)

Use the model to predict on test data and do the same bar-plots to see the rating distribution in each cluster.

Submission of results

You are required to submit **.ipynb files & .html file of the same**, format for the naming convention will be "<your user id>_First Name_Last Name_b44_phdsolution.ipynb.

Please use proper comments for each section of code, also as you are supposed to submit few visualization plots, please have proper label on them.