CSE 487/587 Assignment 3: Predictive Analytics with Spark

The objective of this assignment is to get started with predictive Analytics with Apache Spark. The goals of the assignment are to use Spark Libraries to implement an end to end Predictive Analytics Pipeline and introduce you to the data science competition platform Kaggle.

Movie Genre Prediction

- The objective of the assignment is to implement a movie genre prediction model using Apache Spark
- The dataset provided to you contains information about movies.
- Kaggle competition has been launched: signup link
- Create an account for your team(1-3 members).
- You can download the train.csv, test.csv, sample.csv, and mapping.csv from the Kaggle website
- train.csv has movie summaries of around 31K movies along with their genres. You will use this to train your predictive analytics model
- test.csv has just plot summaries. You will be predicting the genre of these movies
- The task of predicting the genre is essentially a multi-label classification problem. A movie can have multiple genres associated with it. Your model should be able to predict all the genre associated with the movie
- sample.csv is a sample submission file. This is the format in which you will upload the predictions to the Kaggle website. The format of the submission file is movie_id (string), predictions(string). The predictions are expected to be a string of 1's and 0's corresponding to the presence and absence of a particular genre.
- The mapping of the genre to the string index is given by the mapping.csv. For example presence of genre 'Drama' is indicated by a '1' in the first position of the prediction string and an absence of this genre is indicated by '0 in the first position,

PART 1 - Basic Model - 5 points

- Analyze the data and preprocess it if needed
- Create a machine learning model (use any algorithm) in spark to use the information provided in the train set to predict the genres associated with a movie.
- You should create a term-document matrix from the plots and use these as feature vectors for the machine learning model.
- Generate predictions for the test set and upload to Kaggle website
- Report macro F1 score obtained for your submission from the Kaggle website

PART 2 - Use TF-IDF to improve the model - 7 points

- Focussing on the summary of the movie, implement Term Frequency-Inverse Document Frequency (TF-IDF) based feature engineering technique to improve the performance of the model
- Similar to part 1, generate predictions and upload to the Kaggle website

- Ideally, your model should improve performance from the previous step

PART 3 - Custom Feature Engineering - 8 Points

- Implement any one of the modern text-based feature engineering methodology to improve the performance of the model

- Some of the methods to consider are (But not limited to)
    - Word2vec
    - Glove
    - Doc2vec
    - Topic Modelling
    - etc…
- This is an open-ended part of the assignment, where you are free to explore any new text processing methodology to create custom features
- Custom feature engineering would be deemed successful only if the model performs better than the model of part 2

BONUS:- 5 Points

- Kaggle maintains a leaderboard of all the submissions
- There are two leaderboards, public, and a private leaderboard
- You would only be able to see the public leaderboard.
- Generally, the ranking in the public leaderboard also reflects in the private leaderboard
- If the performance of your model ranks in the top 10 of the entire class in the private leaderboard, you will get the bonus 5 points

Submission Instructions

You will submit Assignment3.zip or Assignment3.tar.gz, a compressed archive file containing the following files:

- PySpark code for each part of the assignment.

- The code should follow a spark based programming model. If your code does not use PySpark to process the data and create models, you will not get any points.

- A video (or link to a video) of you running the assignment and explaining what you did for each part

- A report explaining the logic that you have used to implement each part of the assignment and the F1 score you got from Kaggle. (make sure that you write names of your team members on the report)

Submission is due 05/16/2020, Saturday, 11:59 PM EST. Please use the submit_cse487 or submit_cse587 script in Timberlake to submit your assignment.

Ps: Lecture on 04/28/2020 would have more details and discussion on Assignment3