CSE 487/587 Assignment 3: Predictive Analytics with Spark

Christopher Sam Roy
(50320374)
Department of Computer Science
State University of New York at Buffalo
Buffalo, NY 14260
croy2@buffalo.edu

Keshav Jethaliya
(50317073)
Department of Computer Science
State University of New York at Buffalo
Buffalo, NY 14260
keshavje@buffalo.edu

Vishal Singh
(50317996)
Department of Computer Science
State University of New York at Buffalo
Buffalo, NY 14260
vsingh27@buffalo.edu

PART 1: Basic Model

1. Loading of data (train.csv, test.csv and mapping.csv)
2. First, we convert genre column in list as it was in string format.
3. Then we did one hot encoding on list of genres which gave us labels.
4. For plot column, we removed all special character, white spaces and left with only alphabets in lower case.
5. Then we tokenized the plot column based on white spaces.
6. Now making Term document matrix for plot with only 10000 most frequently words.
7. For each label we created a separate Logistic Regression model i.e. for 20 models.
8. Test data is passed to these 20 models, so each model predict for one of the 20 genres.
9. Combining all prediction to form a string of length 20.
10. Converting it to csv with movie id and 20 long string.

F1 Score: 0.89306

PART 2: Using TF-IDF

1. First 4 steps are same as part 1.
2. We created countVectorizer for plot column.
3. Then passed above output to IDF to get TF-IDF.
4. Remaining steps are same.

F1 Score: 0.89743

PART 3: Custom Feature Engineering

1. First 4 steps are same as part 1.
2. For plot we used word2VecEstimator feature extraction.
3. And using above feature we trained the model.

F1 Score: 0.91321

Video Link:

DRIVE: https://drive.google.com/drive/folders/1MQdffA77KxPFhkJvpe-bdtYyl_Q8iUNx?usp=sharing

BOX: https://buffalo.box.com/s/r6iqu723m6czxycpxacc1dpjshw60cxw