Harman Singh

Student Id - 301326898

CMPT 353 Project

Wikidata, Movies, and Success

Problem

In this project I worked on movie analysis project particularly how a movie's success or failure can be classified using plot or genre or both.

Data Collection

All the data used was provided by Greg Baker on the sfu cloud. Data was given in .json.gz format.

Data Preprocessing

Given data included lots of 'nan' and categorial values. To work with genre, plot and movie audience ratings I combined two different datasets. I merged rotten tomatoes and omdb to get plot, genre and ratings in same dataset. I converted the genre list to string by using join method so that it can be passed to the model. I created a new data frame with columns 'Genre', 'Plot', 'Ratings' and 'Plot/Genre'(which is result of genre+plot columns) from rotten tomatoes data frame. Resulted data frame has 6595 rows of clean data.

Classification

By looking at the data I thought it would a good idea and to classify success of movie based on plot or genre or both. To do so I converted the given number ratings to 0 and 1(fail/success) by assuming movie with rating equal to or more than 50 is successful whereas movie with rating less than 50 is flop movie. To use plot/genre in model, they need to be processed before.

For that I used 'Natural Language Processing' tool 'CountVector'. At first I applied countvector on plot and used it as independent variable and movie success/fail as dependent variable. Second time I used genre as independent variable and third time I combined genre and plot into column and used them as independent variable. For classification I used SVM, Logistic Regression, KNN, Naive Bayes and Random Forest models. I got following accuracy scores:

Using Plot as Independent Variable	
Classifier	Accuracy Score

SVM(rbf kernel)	0.74
Logistic Regression	0.68
KNN(with 5 neighbors)	0.74
Naive Bayes(Multinomial)	0.67
Random Forest	0.70

Using Genre as Independent Variable	
Classifier	Accuracy Score
SVM(rbf kernel)	0.73
Logistic Regression	0.73
KNN(with 5 neighbors)	0.70
Naive Bayes(Multinomial)	0.70
Random Forest	0.72

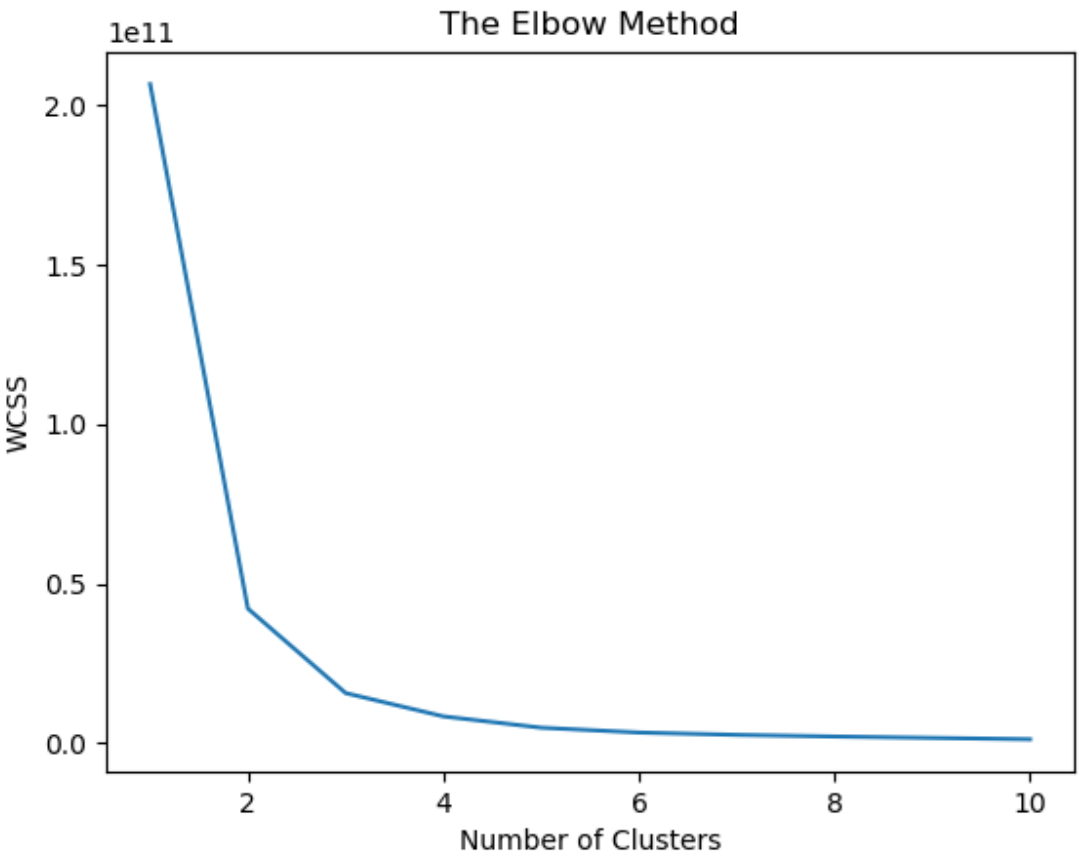
Using Plot+Genre as Independent Variable	
Classifier	Accuracy Score
SVM(rbf kernel)	0.71
Logistic Regression	0.70
KNN(with 5 neighbors)	0.74
Naive Bayes(Multinomial)	0.67
Random Forest	0.71

Clustering

Given the data about movie genre and publication date, I thought it would a good idea to cluster movie genre. To do that, I created a new data frame with column ‘Genre’ and ‘Date’ from wikidata data frame. I converted the dates given from string to float by first slicing the string to just get the year and then casting it to float. To use genre in the model it need to be encoded first. I used LabelEncoder form sklearn.preprocessing to do that.

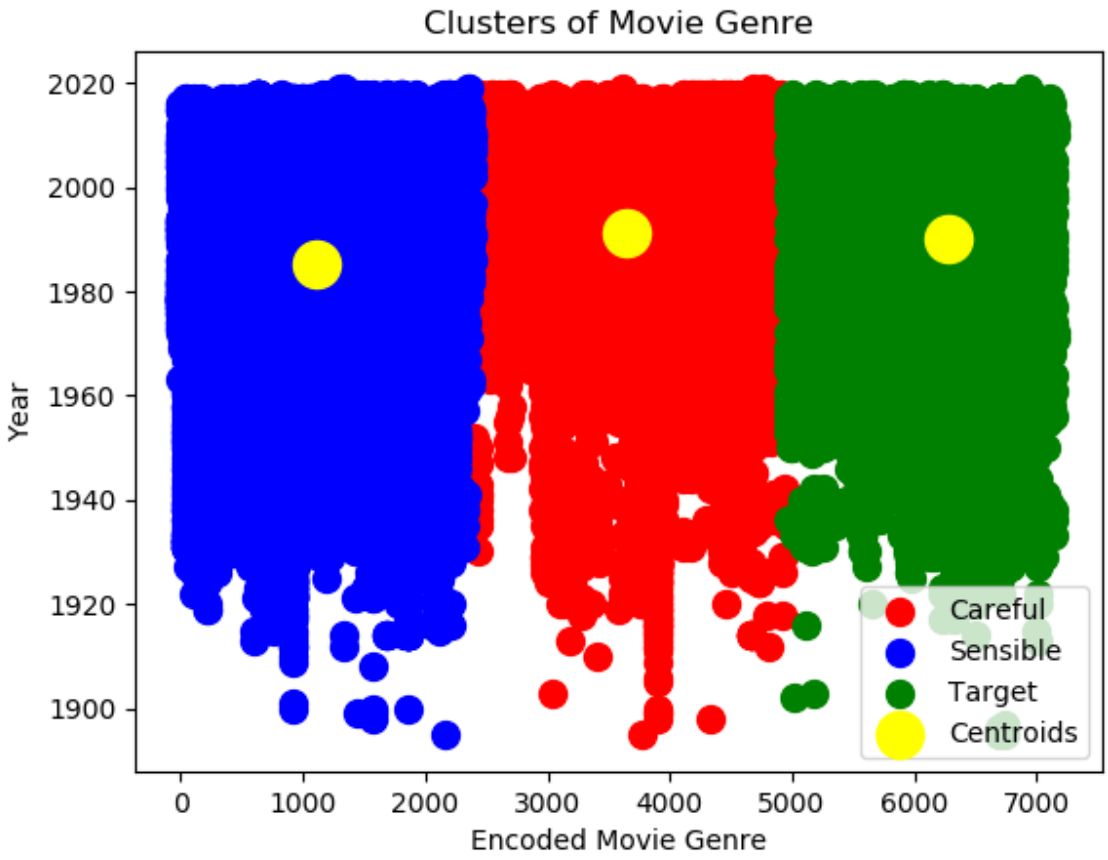
After processing the data and before making cluster, I found optimal number of clusters by using the ‘elbow’ method which involved fitting different number of clusters and visualizing the result to get the right number.

WCSS means Within cluster sums of squares.



WCSS changes drastically for first few numbers after which there not much change. Thats how I get the right number of clusters which is 3.

I used Kmeans clustering from sklearn.cluster library with number of clusters = 3. I got the following result



Results

On average I got 70% accuracy from classification results with max being 74%. But one thing that should be noted here that the score highly depend on rating classification. I chose above 49% as success whereas someone else can choose different number for in place of this and end up getting different results. On average SVM performed best among others.

Accomplishment Statement:

Natural Language Processing

- Classified movies as successful/failed using genre/plot data
- Used CountVector as NPL tool to convert text to token, and pandas and multiple classifiers for analysis
- Achieve a accuracy score of maximum 74% on average through different classifiers.