

International Conference on

Big Data Analytics and Data Mining

November 25-26, 2019 | Frankfurt, Germany

Insights regarding IMDB ratings and reviews through lexicon comparison


Bryant Hwang

Korea International School, South Korea

In this research, sentiment analysis has been used on a data set found in a website called Kaggle, an online community of data scientists. The data set chosen for this study contains 50,000 IMDB review data. The 50,000 IMDB reviews are evenly split into 25,000 training and 25,000 testing data sets. There is an additional 50,000 unsupervised reviews data set. In this study, however, only 25,000 training data sets are used to analyze the sentiments of reviews. The training and test data sets contain folders labeled "pos" and "neg," which represents positive and negative. The data set has sorted which review received a IMDB score above 6 and below 5. Each category contains IMDB review files in a form of individual .txt file. Each .txt file is numbered from 0 to 25,000 and the title of the file contains the IMDB ratings score that respective reviewer gave.

The data set chosen to perform lexicon comparison in this study was the positive and negative training sets. Using RStudio, reviews were put into a data frame and every word from each review were tokenized according to its document id. The sentiment of each review was acquired by summing up the sentiment value found through comparing each tokenized word to four different lexicons (Syuzhet, Bing, Afinn, NRC), which are authoritative dictionaries that contains a plethora of vocabulary and their respective sentiment values. Then, the IMDB ratings of each review was acquired. The score of each review was acquired simply by removing the string of the document id before the underscore and after the period. The two variables were compared to each other using ggplot and other variations of plotting method. Also, using the lexicon NRC, acquiring the distribution of emotions was possible.

e: bhwang0712@gmail.com

 Notes: