Movie Review Sentiment Analysis

TEAM : HYZZ

Han Huang, Yiming Tang, Zixuan Yang, Zhuangfuli Shen

1. Data Preprocessing

After observing the data files in Kaggle, we choose to use *read.delim* to read in the movie review data. Since this time, we mainly focus on text analysis, we remove any punctuation, numeric character and blank space that have nothing to do with our analysis. Besides, we transfer all the capital letters into lowercase to make sure the same word in different forms will be treated as the same. We do not remove the stop words when fitting our model since some stop words like *not* may affect the sentiment of the review and removing them may cause inaccuracy of our model.

2. Using *text2vec* Package

Next we convert data frame to data.tables by reference using *setDT()* and create iterator over train data and test data. Then we create a vocabulary of unique terms and we prune vocabularies by deleting terms that have very high or very low rates or show in only 0.1% of reviews. What's more, we connect the words which can make more sense when they show together. For example, *pretty_good* means differently from *pretty* and *good*. Last, we create a document-term matrix before fitting our model.

3. Building LASSO Model

Since we get a large document matrix, adding a penalization when fitting a model is a reasonable choice. We use 6 folds cross validation and plot the AUC graph with respect to log of lambda. When lambda is -6 the AUC value is the largest and we use this value when predict on test data. We upload the data set on Kaggle and the result is **0.95852**.

4. Data Visualization

(1) Word Cloud

We draw two word clouds graph for positive words and negative words in the reviews from test data based on values of their coefficients and choose top 50 to show in our clouds. As we can see, in positive word cloud, words like *finely*, *refreshing*, *excellently* have high influence on reviews to be positive. In negative word cloud, words like *uneducated*, *waste* and *disappointment* have high influence on reviews to be negative.

Figure 1 Positive Word Cloud

Figure 2 Negative Word Cloud

(2) Highlight Sentimental Words

In order to present a concise highlighting graph, when doing visualization, we remove the connection of two words and also remove the stop words. Then we fit a new model to do following visualization. After fitting the model, we extract coefficients of all words and words themselves which were used in the model to build a new bag of words. We choose words with positive coefficients as positive bag of words and words with negative coefficients as negative bag of words. Then match words of each movie reviews in test data with these two bag of words to find words we want to highlight. Finally, we write a loop to highlight each word contained in our positive and negative bag of words in test review.

We highlight first 50 reviews in test data and select 3 to show here. Here, yellow means positive and pink means negative.

- **"12311_10"** sentiment = 1

  "Naturally in a film who's main themes are of mortality, nostalgia, and loss of innocence it is perhaps not surprising that it is rated more highly by older viewers than younger ones. However there is a craftsmanship and completeness to the film which anyone can enjoy. The pace is steady and constant, the characters full and engaging, the relationships and interactions natural showing that you do not need floods of tears to show emotion, screams to show fear, shouting to show dispute or violence to show anger. Naturally Joyce's short story lends the film a ready made structure as perfect as a polished diamond, but the small changes Huston makes such as the inclusion of the poem fit in neatly. It is truly a masterpiece of tact, subtlety and overwhelming beauty."

- **"8348_2"** sentiment = 0

  "This movie is a disaster within a disaster film. It is full of great action scenes, which are only meaningful if you throw away all sense of reality. Let's see, word to the wise, lava burns you; steam burns you. You can't stand next to lava. Diverting a minor lava flow is difficult, let alone a significant one. Scares me to think that some might actually believe what they saw in this movie. Even worse is the significant amount of talent that went into making this film. I mean the acting is actually very good. The effects are above average. Hard to believe somebody read the scripts for this and allowed all this talent to be wasted. I guess my suggestion would be that if this movie is about to start on TV ... look away! It is like a train wreck: it is so awful that once you know what is coming, you just have to watch. Look away and spend your time on more meaningful content."

- **"5828_4"** sentiment = 1

  "All in all, this is a movie for kids. We saw it tonight and my child loved it. At one point my kid's excitement was so great that sitting was impossible. However, I am a great fan of A.A. Milne's books which are very subtle and hide a wry intelligence behind the childlike quality of its leading characters. This film was not subtle. It seems a shame that Disney cannot see the benefit of making movies from more of the stories contained in those pages, although perhaps, it doesn't have the permission to use them. I found myself wishing the theater was replaying "Winnie-the-Pooh and Tigger too ", instead. The characters voices were very good. I was only really bothered by Kanga. The music, however, was twice as loud in parts than the dialog, and incongruous to the film. As for the story, it was a bit preachy and militant in tone. Overall, I was disappointed, but I would go again just to see the same excitement on my child's face. I liked Lumpy's laugh...."

Figure 3

Reference:

Dmitriy S. (2016) Analyzing Texts with the text2vec package. Retrieved from https://cran.r-project.org/web/packages/text2vec/vignettes/text-vectorization.html#feature_hashing