Sentiment Analysis on Movie Reviews

Yuhui Mou
Xi'an Shiyou University
Xi'an, China
905925264@qq.com

ABSTRACT

Each of us has seen the film,We might write down what we think of it on a movie review website
Film reviews can reflect the quality of films to a certain extent, which depends on the comments of the film critics and their emotions. People who like the film will have good comments, while those who don't like the film will have bad comments.Therefore, we can reflect a person's feelings based on film reviews.

KEYWORDS

Film reviews,BoW,Logistic Regression,LaTeX, CNN

ACM Reference Format:
Yuhui Mou. 2021. Sentiment Analysis on Movie Reviews. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Summary:Classify the sentiment of sentences from the Rotten Tomatoes dataset
Every years,there are many movies appear on the screen.We can read com- ments to know a movie is good or not
Movie Reviews come from varying people.Some may say the positive re- views,others may say the negative comments of the movies.We can classify a movie by it's comments.But the reviews is a large dataset,people can't read every comments.
Here it is,we can use computer to classify the dataset. According to the reviews to distinguish sentiments.

2 DATA

This event provides two data sets that can be used: train.tsv test.tsv
train.tsv: contains the phrases and their associated sentiment labels. We have additionally provided a SentenceId so that you can track which phrases belong to a single sentence.
test.tsv:contains just phrases. You must assign a sentiment label to eachphrase.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

3 DATA PROCESSING

- Read Data
  - Read train.tsv
  - Read test.tsv
- Processing Model:Bag-of-words model (BoW model)
  - BoW early in Natural Language Processing and Information Retrieval This model ignored the grammar and word order elements such as text, just as it is a collection of several words, the emergence of each word in the document are independent of each otherBoW to use an unordered list of words to express a text or a document.
  - CountVectorizer is a characteristic class of common numerical calculation, Is a text feature extraction method.For each training text, it only considers each of these words in the frequency of the training in the text. CountVectorizer Converts text of the words in the word frequency matrix. It does this by fit_transform function calculating the number of occurrences of all words.
  - TF-IDF model
    TF - IDF (term frequency, inverse document frequency) is a kind of commonly used for information retrieval and data mining weighted technique, often used for digging the key words in the article, and the algorithm is simple and efficient, has often been industry for the first text data cleaning A word in the article the TF - the larger the IDF, so in general the word in this article the importance of the higher, so each word in the article by calculation of the TF - IDF, from big to small order, the top of a few words, is the key of the article
- Logistic Regression
  Used to estimate the likelihood of something, and also to classify.
  Logical regression is such a process: in the face of a regression or classification problem, the cost function is established, and then the optimal model parameters are solved iteratively through the optimization method, and then the quality of the model we solve is tested and verified.
  Although Logistic regression has "regression" in its name, it is actually a classification method, mainly used for two classification problems (that is, there are only two outputs, representing two categories respectively).

4 OTHER WAY

CNN NLP
The CNN model was initially applied in the field of image recognition. Text is different from image, and pixel matrix points of image

are dense, but text does not have these characteristics. Each word in the sentence is represented by a vector, and the vector of each word is arranged together to form a "graph", which is then processed by CNN.

5 CONCLUSIONS

I learned the steps of natural language processing.
Know how to deal with large text class data set—-BoW model.
Learn a classifier-LR:can predict the test data and classify test data.
The CNN network model can also be used to deal with NLP problems.

ACKNOWLEDGMENT

We want to thank Kaggle for providing this unique dataset. Kaggle is hosting this playground competition for fun and practice. The authors would like to thank …