Sentiment Analysis On Movie Reviews

MingZhu Kang

Xi'an Shiyou University

Chinese Academy of Sciences

November 22, 2020

Introduction

Project Introduction

Defn

- Project Introduction

  This competition presents a chance to benchmark your sentiment-analysis ideas on the Rotten Tomatoes dataset. You are asked to label phrases on a scale of five values: negative, somewhat negative, neutral, somewhat positive, positive. Obstacles like sentence negation, sarcasm, terseness, language ambiguity. The Rotten Tomatoes movie review dataset is a corpus of movie reviews used for sentiment analysis, This project needs to classify the sentiment of sentences from the Rotten Tomatoes dataset.

Import Data Set

Preprocessing Of Project Data Sets

- Data Collection

  Download dataset from the kaggle project.

  train.tsv contains the phrases and their associated sentiment labels.

  Official website also have additionally provided a SentenceId so that you can track which phrases belong to a single sentence.

	PhraseId	SentenceId	Phrase	Sentiment
0	1	1	A series of escapades demonstrating the adage ...	1
1	2	1	A series of escapades demonstrating the adage ...	2
2	3	1	A series	2
3	4	1	A	2
4	5	1	series	2

Preprocessing Of Project Data Sets

- Data Collection

  test.tsv contains just phrases. You must assign a sentiment label to each phrase.

	PhraseId	SentenceId	Phrase
0	156061	8545	An intermittently pleasing but mostly routine ...
1	156062	8545	An intermittently pleasing but mostly routine ...
2	156063	8545	An
3	156064	8545	intermittently pleasing but mostly routine effort
4	156065	8545	intermittently pleasing but mostly routine

Build A Corpus

Build A Corpus

- To extract the text contents of the training set and the test set, a corpus was constructed,and the text contents of the training set and the test set were combined together to extract the emotional tags of the training set through concat function.

```
0        A series of escapades demonstrating the adage ...
1        A series of escapades demonstrating the adage ...
2                                                A series
3                                                       A
4                                                  series
                              ...
66287               A long-winded , predictable scenario .
66288                 A long-winded , predictable scenario
66289                                   A long-winded ,
66290                                     A long-winded
66291                                  predictable scenario
Name: Phrase, Length: 222352, dtype: object
```

Import The Stop Lexicon

- You need to determine how to deal with frequent words that don’t make sense. A stop-word database is not helpful for emotion analysis. These words are called "stop-words". In English, they include words like "a", "and", "is" and "the".

```
[″\ufeffain′′,
 'happy',
 'isn',
 'ain',
 'al',
 'couldn',
 'didn',
 'doesn',
 'hadn',
 'hasn',
 'haven',
 'sn',
 'll',
 'mon',
```

Characteristics Of The Engineering

Build The Model

Throw these words directly to the computer, the computer can’t calculate them, so we need to convert the text into vectors and use the word bag model for text feature engineering.

There are several common text vector processing methods, such as: word bag model,TF-IDF model, WORD2VEC model for text feature engineering.

Bag-of-words Model

- The word bag model learns vocabulary from all documents and then models each document by counting the number of occurrences of each word.
  The corpus is used to construct the word bag model,
  and the constructed word bag model is used to carry out feature engineering
  for each word in the training set and the verification set and turn it into a vector

TF-IDF Model

- TF-IDF is a statistical method used to assess the importance of a word to one of the documents in a document set or corpus.
  The importance of a word increases proportionally with the frequency of its occurrence in the document, but decreases inversely with the frequency of its occurrence in the corpus.
  Word frequency (TF) represents the frequency of entries (keywords) in the text.

Construct The Classifier Algorithm

Classifier Algorithm

- Machine learning and data mining are carried out for the text processed by the word bag model.
  Use the logistic regression activation function to convert the output label of the category variable to a numeric variable.
  The word bag method was used for text feature engineering, and sklearn default logistic regression classifier was used to verify the prediction accuracy on the set.

Predict Test Set Data

Predict Test Set Data

- The text in the test set was predicted using the LG_FINAL logistic regression classifier,
  the predicted results were viewed, the test results were added to the test set, each film comment in the test set was tagged, and submitted in the phrase ID-emotion tag format.