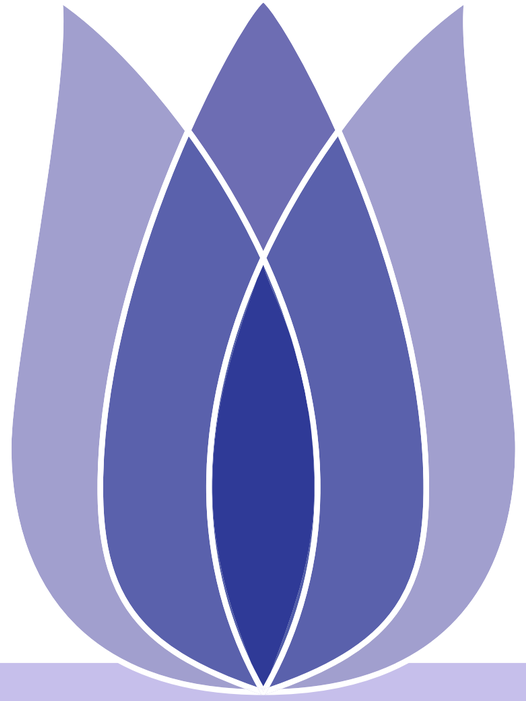
Sentiment Analysis on Movie Reviews

Yuhui Mou

Xi'an Shiyou University

January 15, 2021



Overview

Introduction

Data

Data Processing

- Step One
- Step Two
- Step Three

Output

Other Way

Conclusion

Introduction

Introduction

sentiments

Summary:Classify the sentiment of sentences from the Rotten Tomatoes dataset

Every years,there are many movies appear on the screen.We can read comments to know a movie is good or not

Movie Reviews come from varying people.Some may say the positive reviews,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.

Data

Data

Data

The dataset is comprised of tab-separated files with phrases from the Rotten Tomatoes dataset.

Each Sentence has been parsed into many phrases by the Stanford parser.Each phrase has a PhraseId. Each sentence has a SentenceId.Phrases that are repeated (such as short/common words) are only included once in the data. The date is stored in tsv format.

- 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 each phrase.

In the test file test.tsv, the format of a Sentence is the same as train.tsv, only the sentiment type is removed, because it is the target variable you are going to predict.

	PhraseId	SentenceId	Phrase	Sentiment
1	1	1	A series of escapades demonstratin	
2	2	1	A series of escapades demonstratin	
3	3	1	A series	2
4	4	1	A	2
5	5	1	series	2
6	6	1	of escapades demonstrating the ada	
7	7	1	of	2
8	8	1	escapades demonstrating the adage	
9	9	1	escapades	2
10	10	1	demonstrating the adage that what	
11	11	1	demonstrating the adage	2
12	12	1	demonstrating	2
13	13	1	the adage	2
14	14	1	the	2
15	15	1	adage	2
16	16	1	that what is good for the goose	2
17	17	1	that	2
18	18	1	what is good for the goose	2
19	19	1	what	2

Figure 1: train-data

	PhraseId	SentenceId	Phrase
1	156061	8545	An intermittently pleasing
2	156062	8545	An intermittently pleasing
3	156063	8545	An
4	156064	8545	intermittently pleasing bu
5	156065	8545	intermittently pleasing bu
6	156066	8545	intermittently pleasing bu
7	156067	8545	intermittently pleasing
8	156068	8545	intermittently
9	156069	8545	pleasing
10	156070	8545	but
11	156071	8545	mostly routine
12	156072	8545	mostly
13	156073	8545	routine
14	156074	8545	effort
15	156075	8545	.
16	156076	8546	Kidman is really the only
17	156077	8546	Kidman
18	156078	8546	is really the only thing t
19	156079	8546	is really the only thing t

Figure 2: test-data

Each training data has four parts:PhraseId SentenceId Phrase Sentiment. The features of the reviews come form phrase, and the training goal is Sentiment.

Data Processing

Step One

- Read Data
  - Read train.tsv
  - Read test.tsv
- Build a corpus
  - Build a corpus.
    The text contents of the training set and test set are merged together using the contcat function in Pandas
  - Import the stop word library
    We use words like a, an, and, or, of, at, the and so on. The information about these words is extremely limited.Therefore, what we need to do is to remove the pause words in the text in the NLP analysis process. The advantage of doing this is that we reduce the vocabulary and then reduce the dimension of our feature vector

Step Two

Characteristics of the engineering

- Bag-Of-Words 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.
- 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

Step Three

Build a classifier

- Logistic Regression
  Used to estimate the likelihood of something, and also to classify.

Forecast the data in the test set

- Feature engineering the text in the test set.
- Logical regression classifier is used to predict the text in the test set.
- Format it and save it as a.csv file

Output

	A	B
1	PhraseId	Sentiment
2	156061	3
3	156062	3
4	156063	2
5	156064	3
6	156065	2
7	156066	3
8	156067	3
9	156068	2
10	156069	3
11	156070	2
12	156071	2
13	156072	2
14	156073	2
15	156074	2
16	156075	2

Figure 3: outp-data

This is a screenshot of the output file with the sentence ID in the first column and the emotion ID in the second column.
0 - negative; 1 - somewhat negative; 2 - neutral; 3 - somewhat positive; 4 - positive

Other Way

CNN-NLP

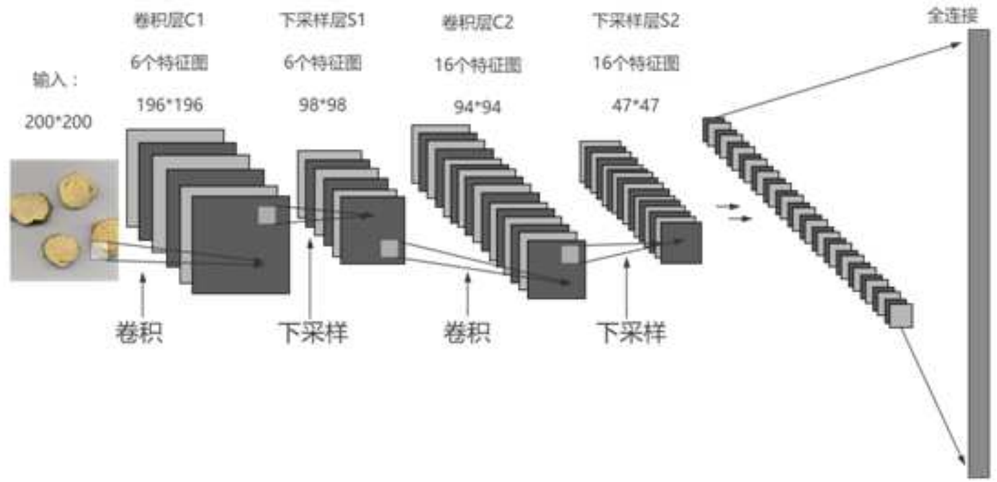


Figure 4: CNN

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.

Conclusion

Conclusion

- I have learned the steps of natural language processing.
- I have knew how to deal with large text class data set—-BoW model.
- I have learned 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.

Contact Information

Thanks for watching!

Yuhui Mou
Xi'an Shiyou University