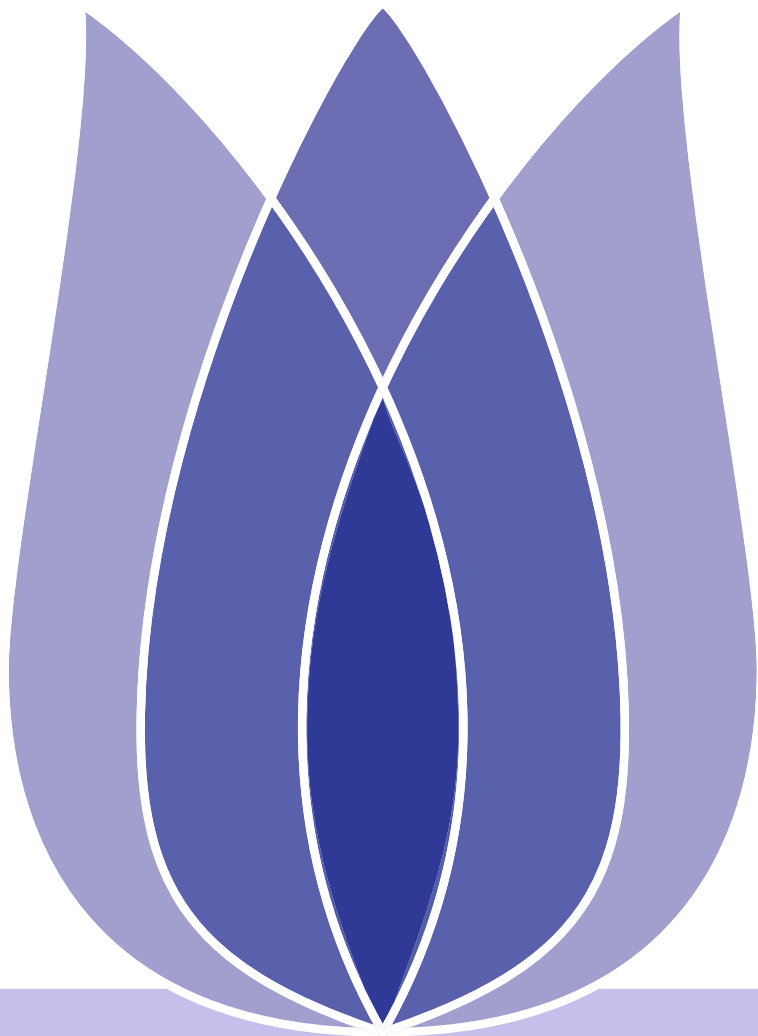




# FLIP01 MIDTERM PRESENTATION

Zebin Ju  
Xi'an Shiyou University

January 13, 2020





# Overview

[Problem Description](#)

[Problem Analysis](#)

[Text Feature Extraction](#)

[Methods](#)

[The Ending](#)

## Problem Description

## Problem Analysis

## Text Feature Extraction

## Methods

## The Ending



Problem Description

Description

Problem Analysis

Text Feature Extraction

Methods

The Ending

# Problem Description



# Description

Problem Description
Description
Problem Analysis
Text Feature Extraction
Methods
The Ending

Rotten tomato movie review data set is a movie review corpus for emotional analysis. It is an opportunity for us to build your idea of Emotional Analysis on rotten tomato data set. It is required to mark phrases with five levels of values: negative, some negative, neutral, some positive, positive. Negative sentences, satire, conciseness, language ambiguity and other obstacles make this task very challenging.



- [Problem Description](#)
- [Problem Analysis](#)**
- [Data Analysis](#)
- [Data Glance](#)
- [Type Of Review](#)
- [Text Feature Extraction](#)
- [Methods](#)
- [The Ending](#)

# Problem Analysis



- [Problem Description](#)
- [Problem Analysis](#)
- [Data Analysis](#)
- [Data Glance](#)
- [Type Of Review](#)
- [Text Feature Extraction](#)
- [Methods](#)
- [The Ending](#)

The dataset consists of tab delimited files and phrases in the rotten tomatoes dataset. Each phrase has a phraseid. Each sentence has a sentenceid.

- 0 for negative
- 1 for somewhat negative
- 2 for neutral
- 3 for somewhat positive
- 4 for positive



- [Problem Description](#)
- [Problem Analysis](#)
- [Data Analysis](#)
- [Data Glance](#)**
- [Type Of Review](#)
- [Text Feature Extraction](#)
- [Methods](#)
- [The Ending](#)

The following table shows the data characteristics.

	Phraseld	Sentenceld	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

Figure 1: data characteristics table





# Type Of Review

- [Problem Description](#)
- [Problem Analysis](#)
- [Data Analysis](#)
- [Data Glance](#)
- [Type Of Review](#)**
- [Text Feature Extraction](#)
- [Methods](#)
- [The Ending](#)

Next, we can see the category distribution of comments. From the figure, we can see that emotion tag 2, that is, the most neutral comments.

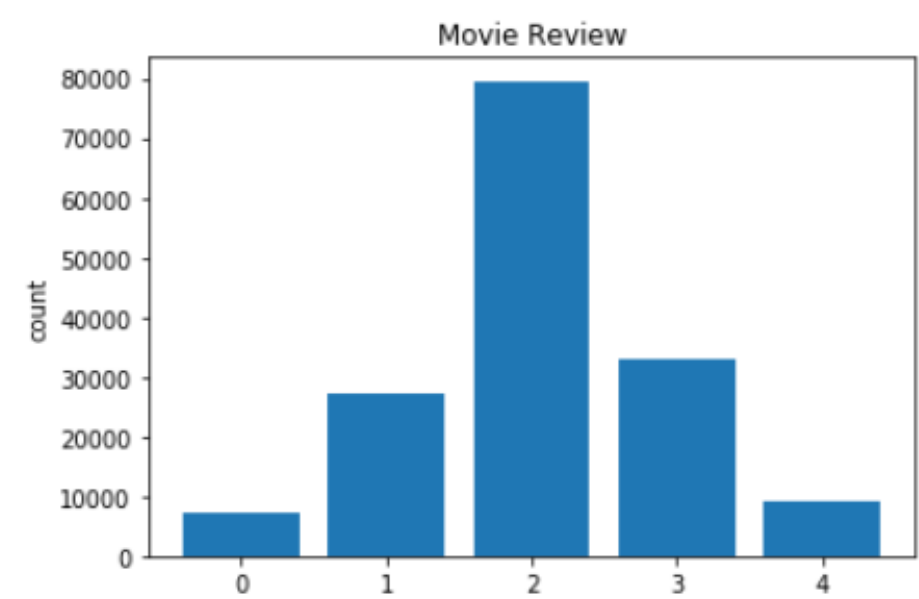


Figure 2: type of review



[Problem Description](#)

[Problem Analysis](#)

[Text Feature Extraction](#)

[Methods](#)

[The Ending](#)

# Text Feature Extraction



# Text Feature Extraction

<a href="#">Problem Description</a>
<a href="#">Problem Analysis</a>
<a href="#">Text Feature Extraction</a>
<a href="#">Methods</a>
<a href="#">The Ending</a>

It is the process of transforming text data into feature vector. Machine learning algorithm often can't directly process text data, so it needs to transform text data into numerical data.

- CountVectorizer
- TfidfVectorizer





[Problem Description](#)

[Problem Analysis](#)

[Text Feature Extraction](#)

[Methods](#)

[The Ending](#)

# Methods



# Naive Bayes

[Problem Description](#)

[Problem Analysis](#)

[Text Feature Extraction](#)

[Methods](#)

[The Ending](#)

Naive Bayes is a classification algorithm based on probability theory, which can predict classification by considering feature probability.

- CountVectorizer:0.6715045495322312
- TfidfVectorizer:0.6308070613866461

Phraseld Sentiment			Phraseld Sentiment		
66287	222348	1	66287	222348	1
66288	222349	1	66288	222349	1
66289	222350	1	66289	222350	1
66290	222351	1	66290	222351	1
66291	222352	1	66291	222352	2

Figure 3: naive bayes



# Logistic Regression

- [Problem Description](#)
- [Problem Analysis](#)
- [Text Feature Extraction](#)
- [Methods](#)
- [The Ending](#)

Logical regression can solve the problem of two categories, but it can also solve the problem of multiple categories.

- CountVectorizer:0.7002354863514033
- TfidfVectorizer:0.6662581699346405

	Phraseld	Sentiment		Phraseld	Sentiment
66287	222348	1	66287	222348	1
66288	222349	1	66288	222349	1
66289	222350	1	66289	222350	2
66290	222351	1	66290	222351	2
66291	222352	2	66291	222352	1

Figure 4: logistic regression



[Problem Description](#)

[Problem Analysis](#)

[Text Feature Extraction](#)

[Methods](#)

[The Ending](#)

# The Ending