**哈尔滨工业大学理学院**

**小组实训项目总结报告**

|  |  |
| --- | --- |
| 实训项目名称： | 基于电影评论的情感分析 |
| 姓 名： |  |
| 系： |  |
| 专 业： |  |
| 年 级： |  |
| 学 号： |  |
| 实训地点： |  |
| 指导教师： |  |
| 完成起止日期： |  |

年 月 日

基于电影评论的情感分析

摘要

关键词：

Emotional analysis of film review

Abstract

Key words:

目录

1. 挖掘目标
2. 分析方法与过程
3. 结果与讨论
4. 总结
5. 参考文献

**7.附件**

**附录1标签数据集**

id sentiment review

0 5814\_8 1 With all this stuff going down at the moment w...

1 2381\_9 1 \The Classic War of the Worlds\" by Timothy Hi...

2 7759\_3 0 The film starts with a manager (Nicholas Bell)...

3 3630\_4 0 It must be assumed that those who praised this...

4 9495\_8 1 Superbly trashy and wondrously unpretentious 8...

5 8196\_8 1 I dont know why people think this is such a ba...

6 7166\_2 0 This movie could have been very good, but come...

7 10633\_1 0 I watched this video at a friend's house. I'm ...

8 319\_1 0 A friend of mine bought this film for £1, and ...

9 8713\_10 1 <br /><br />This movie is full of references. ...

10 2486\_3 0 What happens when an army of wetbacks, towelhe...

11 6811\_10 1 Although I generally do not like remakes belie...

12 11744\_9 1 \Mr. Harvey Lights a Candle\" is anchored by a...

13 7369\_1 0 I had a feeling that after \Submerged\", this ...

14 12081\_1 0 note to George Litman, and others: the Mystery...

15 3561\_4 0 Stephen King adaptation (scripted by King hims...

16 4489\_1 0 `The Matrix' was an exciting summer blockbuste...

17 3951\_2 0 Ulli Lommel's 1980 film 'The Boogey Man' is no...

18 3304\_10 1 This movie is one among the very few Indian mo...

19 9352\_10 1 Most people, especially young people, may not ...

20 3374\_7 1 \Soylent Green\" is one of the best and most d...

21 10782\_7 1 Michael Stearns plays Mike, a sexually frustra...

22 5414\_10 1 This happy-go-luck 1939 military swashbuckler,...

23 10492\_1 0 I would love to have that two hours of my life...

24 3350\_3 0 The script for this movie was probably found i...

25 6581\_7 1 Looking for Quo Vadis at my local video store,...

26 2203\_3 0 Note to all mad scientists everywhere: if you'...

27 689\_1 0 What the ........... is this ? This must, with...

28 9152\_1 0 Intrigued by the synopsis (every gay video the...

29 6077\_1 0 Would anyone really watch this RUBBISH if it d...

... ... ... ...

24970 9389\_7 1 Red Rock West (1993)<br /><br />Nicolas Cage g...

24971 9251\_9 1 what can i say?, ms Erika Eleniak is my favori...

24972 1422\_10 1 The spoiler warning is for those people who wa...

24973 7415\_2 0 What do you call a horror story without horror...

24974 7492\_7 1 Though not a horror film in the traditional se...

24975 7689\_10 1 This was what black society was like before th...

24976 12370\_4 0 They probably should have called this movie Th...

24977 5625\_8 1 Attractive Marjorie(Farrah Fawcett)lives in fe...

24978 9397\_9 1 Vaguely reminiscent of great 1940's westerns, ...

24979 5992\_7 1 I admit I had no idea what to expect before vi...

24980 2488\_10 1 To me, the final scene, in which Harris respon...

24981 9627\_10 1 This is by far the funniest short made by the ...

24982 3822\_2 0 To be a Buster Keaton fan is to have your hear...

24983 5983\_4 0 I was one of those \few Americans\" that grew ...

24984 8021\_2 0 Visually disjointed and full of itself, the di...

24985 3471\_3 0 this movie had more holes than a piece of swis...

24986 6034\_10 1 Last November, I had a chance to see this film...

24987 1988\_9 1 First off, I'd like to make a correction on an...

24988 7623\_9 1 While originally reluctant to jump on the band...

24989 5974\_7 1 I heard about this movie when watching VH1's \...

24990 2034\_9 1 I've never been huge on IMAX films. They're co...

24991 9416\_3 0 Steve McQueen has certainly a lot of loyal fan...

24992 10994\_1 0 Sometimes you wonder how some people get fundi...

24993 10957\_3 0 I am a student of film, and have been for seve...

24994 2372\_1 0 Unimaginably stupid, redundant and humiliating...

24995 3453\_3 0 It seems like more consideration has gone into...

24996 5064\_1 0 I don't believe they made this film. Completel...

24997 10905\_3 0 Guy is a loser. Can't get girls, needs to buil...

24998 10194\_3 0 This 30 minute documentary Buñuel made in the ...

24999 8478\_8 1 I saw this movie as a child and it broke my he...

25000 rows × 3 columns

**附录2 测试数据集**

id review

0 12311\_10 Naturally in a film who's main themes are of m...

1 8348\_2 This movie is a disaster within a disaster fil...

2 5828\_4 All in all, this is a movie for kids. We saw i...

3 7186\_2 Afraid of the Dark left me with the impression...

4 12128\_7 A very accurate depiction of small time mob li...

5 2913\_8 ...as valuable as King Tut's tomb! (OK, maybe ...

6 4396\_1 This has to be one of the biggest misfires eve...

7 395\_2 This is one of those movies I watched, and won...

8 10616\_1 The worst movie i've seen in years (and i've s...

9 9074\_9 Five medical students (Kevin Bacon, David Labr...

10 9252\_3 'The Mill on the Floss' was one of the lesser ...

11 9896\_9 I just saw this film at the phoenix film festi...

12 574\_4 \The Love Letter\" is one of those movies that...

13 11182\_8 Another fantastic offering from the Monkey Isl...

14 11656\_4 This was included on the disk \Shorts: Volume ...

15 2322\_4 I'm not really much of an Abbott & Costello fa...

16 8703\_1 This movie was dreadful. Biblically very inacc...

17 7483\_1 I don't think I've ever gave something a 1/10 ...

18 6007\_10 Excellent story-telling and cinematography. Po...

19 12424\_4 I completely forgot that I'd seen this within ...

20 4672\_1 I like action movies. I have a softspot for \B...

21 10841\_3 This is one of the worst Sandra Bullock movie ...

22 8954\_7 Watched this flick on Saturday afternoon cable...

23 7392\_1 I went to see \TKIA\" with high expectations, ...

24 10288\_8 All credit to writer/director Gilles Mimouni w...

25 5343\_4 As a writing teacher, there are two ending I n...

26 4950\_1 I don't know why this has gotten any decent re...

27 9257\_4 This film was released in the UK under the nam...

28 8689\_3 Uncle Fred Olen Ray once again gives us a litt...

29 4480\_2 OK, it's watchable if you are sick in bed or h...

... ... ...

24970 6857\_10 With \Anatomy\" the german film producers have...

24971 11091\_8 This movie is one of my all-time favorites. I ...

24972 4167\_2 I found Code 46 very disappointing. I thought ...

24973 679\_4 Tamara Anderson and her family are moving once...

24974 10147\_1 Now I've seen it all. Just when I thought it c...

24975 6875\_1 In this movie everything possible was wrong an...

24976 923\_10 Well every scene so perfectly presented. Never...

24977 6200\_8 Sleeper Cell is what 24 should have been. 24 i...

24978 7208\_8 Not for everyone, but I really like it. Nice e...

24979 5363\_8 Set just before the Second World War, this is ...

24980 4067\_8 Contains Spoiler The movie is a good action/co...

24981 1773\_7 This is one of several period sea-faring yarns...

24982 1498\_10 Hearkening back to those \Good Old Days\" of 1...

24983 10497\_10 I thought this to be a pretty good example of ...

24984 3444\_10 Seeing this film, or rather set of films, in m...

24985 588\_2 I didn't like this movie for many reasons - VE...

24987 1983\_9 eXistenZ combines director David Cronenberg's ...

24988 5012\_3 this movie is allegedly a comedy.so where did ...

24989 12240\_2 The Comebacks is a spoof on inspirational spor...

24990 5071\_2 I'd love to write a little summary of this mov...

24986 9678\_9 I absolutely love this show!!!!!!!, Its basica...

24991 5078\_2 Obvious tailored vehicle for Ryan Philippe. It...

24992 10069\_3 <br /><br />JURASSIC PARK III \*\_\_\_ Adventure <...

24993 7407\_8 If you're even mildly interested in the War be...

24994 7207\_1 It used to be that video distributors like Sub...

24995 2155\_10 Sony Pictures Classics, I'm looking at you! So...

24996 59\_10 I always felt that Ms. Merkerson had never got...

24997 2531\_1 I was so disappointed in this movie. I am very...

24998 7772\_8 From the opening sequence, filled with black a...

24999 11465\_10 This is a great horror film for people who don...

25000 rows × 2 columns

**附录3 预测数据集**

id review

0 "9999\_0" "Watching Time Chasers, it obvious that it was...

1 "45057\_0" "I saw this film about 20 years ago and rememb...

2 "15561\_0" "Minor Spoilers<br /><br />In New York, Joan B...

3 "7161\_0" "I went to see this film with a great deal of ...

4 "43971\_0" "Yes, I agree with everyone on this site this ...

5 "36495\_0" "Jennifer Ehle was sparkling in \"Pride and Pr...

6 "49472\_0" "Amy Poehler is a terrific comedian on Saturda...

7 "36693\_0" "A plane carrying employees of a large biotech...

8 "316\_0" "A well made, gritty science fiction movie, it...

9 "32454\_0" "Incredibly dumb and utterly predictable story...

10 "37128\_0" "After reading the comments for this movie, I ...

11 "19439\_0" "It's hard to describe Elfen Lied to someone w...

12 "10760\_0" "Of all the bile-inducing vomitoriums to squee...

13 "15073\_0" "This is quite an underrated Hitchcock movie. ...

14 "33119\_0" "Being a huge Gary Oldman fan, I had high hope...

15 "38735\_0" "For the most part the acting was poorly done ...

16 "12041\_0" "Ram Gopal Varma does it again yet another mov...

17 "41565\_0" "I gave it 2 for some attempt at historical de...

18 "48612\_0" "I wanted to watch this one sense I am a fan o...

19 "17525\_0" "Che is a good film for chronicling the events...

20 "32062\_0" "This film is very disturbing. It reminded me ...

21 "48625\_0" "Let me put it this way. I love bad movies. I ...

22 "48176\_0" "From the combined ministries of Jack and Rexe...

23 "16232\_0" "Ghost World is one of the better teen or comm...

24 "12528\_0" "Outstanding film dealing with a 16 year old w...

25 "39976\_0" "This movie did play like a dream or maybe eve...

26 "25857\_0" "This film is supposedly based on the true acc...

27 "10552\_0" "I know that cartoons aimed towards adults gen...

28 "26437\_0" "...but what a disappointment. The acting was ...

29 "16262\_0" "As a long-time Holmes fan, and particularly o...

... ... ...

49970 "25941\_0" "Wouldn't suit all tastes, but if you like it ...

49971 "36778\_0" "As a great fan of Dwight Yoakam's music as we...

49972 "37114\_0" "I'm a huge fan of Whoopi and have seen all of...

49973 "11962\_0" "It not very often that I go to the same movie...

49974 "32293\_0" "I, like most, worried that it would degrade \...

49975 "7005\_0" "This movie makes no bones about what it is an...

49976 "112\_0" "The eyebrow-wagging Lionel Q. Deveraux (Grouc...

49977 "46123\_0" "There maybe spoilers in the following; if you...

49978 "10628\_0" "This is one of my favorite films. <br /><br /...

49979 "22572\_0" "A novelist who could find a funny side of a d...

49980 "40601\_0" "This is the worst drivel I've ever seen come ...

49981 "31509\_0" "\"Jaguar Lives\" is very slow and plodding en...

49982 "33703\_0" "This should have been better, much better. A ...

49983 "7379\_0" "Look at the number of film festival awards \"...

49984 "29370\_0" "This movie is one of the best of all time.Fir...

49985 "27604\_0" "The earliest episodes with John Amos as the c...

49986 "9197\_0" "I've been looking for this movie since I beca...

49987 "34018\_0" "OK, this isn't a good film by any stretch of ...

49988 "15625\_0" "It is arguably the most boring film I've ever...

49989 "35740\_0" "Oh joy,, another foreign torture/bloody movie...

49990 "20947\_0" "In 1994, a year that gave us Forrest Gump, Pu...

49991 "7432\_0" "http://eattheblinds.blogspot.com/<br /><br />...

49992 "35236\_0" "Taste the Blood of Dracula marks a first for ...

49993 "12075\_0" "There were some aspects of this movie I reall...

49994 "23455\_0" "Look, I don't know what people have against E...

49995 "18984\_0" "The original Man Eater by Joe D'Amato is some...

49996 "16433\_0" "When Home Box Office was in it's early days m...

49997 "16006\_0" "Griffin Dunne was born into a cultural family...

49998 "40155\_0" "Not a bad story, but the low budget rears its...

49999 "35270\_0" "This not-very-good mummy-alien flick does fea...

50000 rows × 2 columns

**附录4用Keras搭建LSTM并预测测试集程序**

**数据集分析**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import re

def dataCleaning(totalText):

total=[]

for i in totalText:

temp=i.lower()

temp=re.sub('[^a-zA-Z]',' ',temp)

tempArr=[j for j in temp.strip().split('\t') if isinstance(j,str)]

tstr=' '.join(tempArr)

total.append(tstr)

return total

def loadData(name):

data=pd.read\_csv(name,delimiter='\t')

totalText=data['review']

totalText=dataCleaning(totalText)

totalLabel=data['sentiment']

return totalText,totalLabel

def getTest(name):

data=pd.read\_csv(name,delimiter='\t')

totalText=data['review']

totalText=dataCleaning(totalText)

return totalText

**加载数据集**

import os

os.chdir(r'C:\Users\lenovo\Desktop\电影评论情绪分析')

data=pd.read\_csv('labeledTrainData.tsv',sep='\t')

data2=pd.read\_csv('testData.tsv',sep='\t')

sub = pd.read\_csv('unlabeledTrainData.tsv', header=0, delimiter="\t", quoting=3 )

#totalText,totalLabel=loadData('labeledTrainData.tsv')

#testText=getTest('testData.tsv')

totalText,totalLabel=loadData('labeledTrainData.tsv')

testText=getTest('testData.tsv')

**数据查看**

print(len(totalText))

print(len(totalLabel))

print(len(testText))

**查看评分的类别分布**

uniqueLabel=set(totalLabel)

x=[]

y=[]

for i in uniqueLabel:

x.append(i)

y.append(totalLabel[totalLabel==i].size)

plt.figure(111)

plt.bar(x,y)

plt.xlabel('type of review ')

plt.ylabel('count')

plt.title('Movie Review')

plt.show()

**Tokenizer**

from keras.preprocessing.text import Tokenizer

train\_tokenizer = Tokenizer()

train\_tokenizer.fit\_on\_texts(totalText)

train\_sequences = train\_tokenizer.texts\_to\_sequences(totalText)

test\_sequences = train\_tokenizer.texts\_to\_sequences(testText)

# 获得所有tokens的长度

num\_tokens = [ len(tokens) for tokens in train\_sequences ]

num\_tokens = np.array(num\_tokens)

print(len(num\_tokens))

print(T.one\_hot(text2,10))

import keras.preprocessing.text as T

# 平均tokens的长度

print('mean',np.mean(num\_tokens))

# 最长的评价tokens的长度

print('max',np.max(num\_tokens))

# 最长的评价tokens的长度

print('min',np.min(num\_tokens))

**查看训练数据的长度**

plt.hist((num\_tokens), bins = 50)

plt.ylabel('number of tokens')

plt.xlabel('length of tokens')

plt.title('Distribution of tokens length')

plt.show()

max\_tokens = np.mean(num\_tokens) + 2 \* np.std(num\_tokens)

max\_tokens = int(max\_tokens)

max\_tokens

# 取tokens的长度为599时，大约94%的样本被涵盖

np.sum( num\_tokens < max\_tokens ) / len(num\_tokens)

**对向量化后的文本进行padding**

# from keras.preprocessing sequence import sequence

train\_Data=sequence.pad\_sequences(train\_sequences,maxlen=20)

test\_Data=sequence.pad\_sequences(test\_sequences,maxlen=20)

#这里的长度可以是前面取到的max\_tokens，也可以是他的最大长度 前者可以节约计算时间

print(train\_Data.shape)

print(test\_Data.shape)

train\_Data

**转换标签**

from keras.utils.np\_utils import to\_categorical#转换标签

train\_label=to\_categorical(totalLabel,5)

print(train\_label.shape)

print(train\_label)

**在测试样本上切分数据集和测试集**

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(train\_Data , train\_label, test\_size=0.25, random\_state=42)

print(X\_train.shape)

print(X\_test.shape)

print(y\_train.shape)

print(y\_test.shape)

**使用Keras来搭建LSTM**

#max\_features为最多单词数

max\_features=len(train\_tokenizer.index\_word)

max\_len=48 #这个是要和前面padding时的长度一致

epochs = 5 #训练次数

emb\_dim = 128 #128代表embedding层的向量维度

batch\_size=80 #这是指定批量的大小

**LSTM搭建**

from keras.models import Sequential

from keras.layers import Dense ,Embedding,Activation

from keras.layers import LSTM

from keras.preprocessing import sequence

from keras.preprocessing.sequence import pad\_sequences

model = Sequential()

model.add(Embedding(max\_features,emb\_dim,mask\_zero=True))

model.add(LSTM(64,dropout=0.4, recurrent\_dropout=0.4,return\_sequences=True))

model.add(LSTM(32,dropout=0.5, recurrent\_dropout=0.5,return\_sequences=False))

model.add(Dense(5, activation='softmax'))

model.compile(loss='categorical\_crossentropy', optimizer='adam', metrics=['accuracy'])

print(model.summary())

model.fit(X\_train,y\_train,batch\_size=batch\_size,epochs=epochs,validation\_data=(X\_test,y\_test))

**验证**

model.evaluate(X\_test,y\_test)

**最后预测测试集**

X\_train

data2

predict=model.predict\_classes(test\_Data)

sub = pd.read\_csv('sampleSubmission.csv')

sub['sentiment'] = predict

sub.to\_csv('sub\_lstm.csv', index=False)

import os

os.chdir(r'C:\Users\lenovo\Desktop\电影评论情绪分析')

sub = pd.read\_csv('unlabeledTrainData.tsv', header=0, delimiter="\t", quoting=3 )