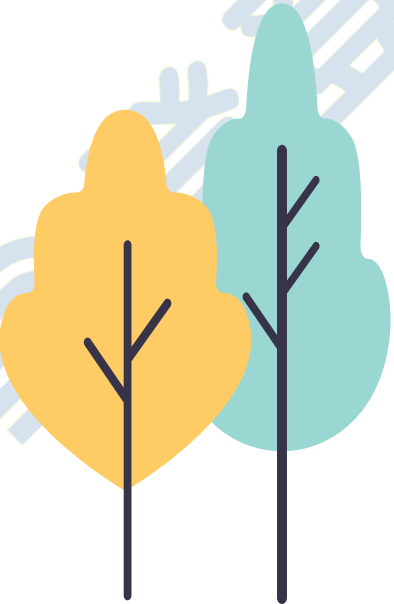
COMP 9414

ass2 情感分类

主讲人：Tara

2020.7

一 Assignment 要求

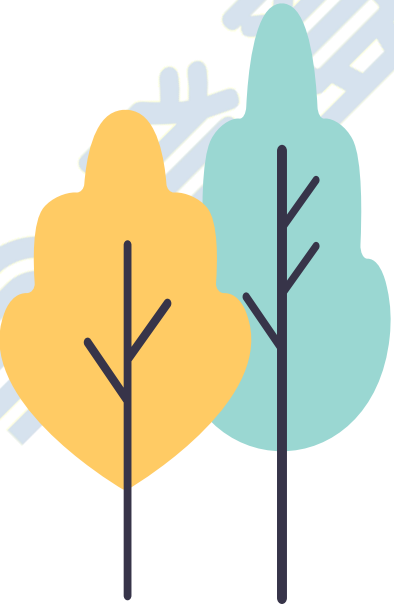


Assignment 描述

有一个数据集，里面是Twitter 的反馈，来确定客户对你的公司及其竞争对手的看法

我们的任务：

- 用这个给定的数据做训练集，对数据进行处理后，采用三种方法Decision Trees(DT), Bernoulli Naive Bayes(BNB) 和 Multinomial Naive Bayes (MNB)建立三个基础的模型，对给定的测试集进行测试，输出测试结果。(DT_sentiment.py, BNB_sentiment.py, MNB_sentiment.py)
- 对以上三个模型进行优化，挑选一个最好的，作为自己的模型。（sentiment.py）
- 完成report。（report.pdf）



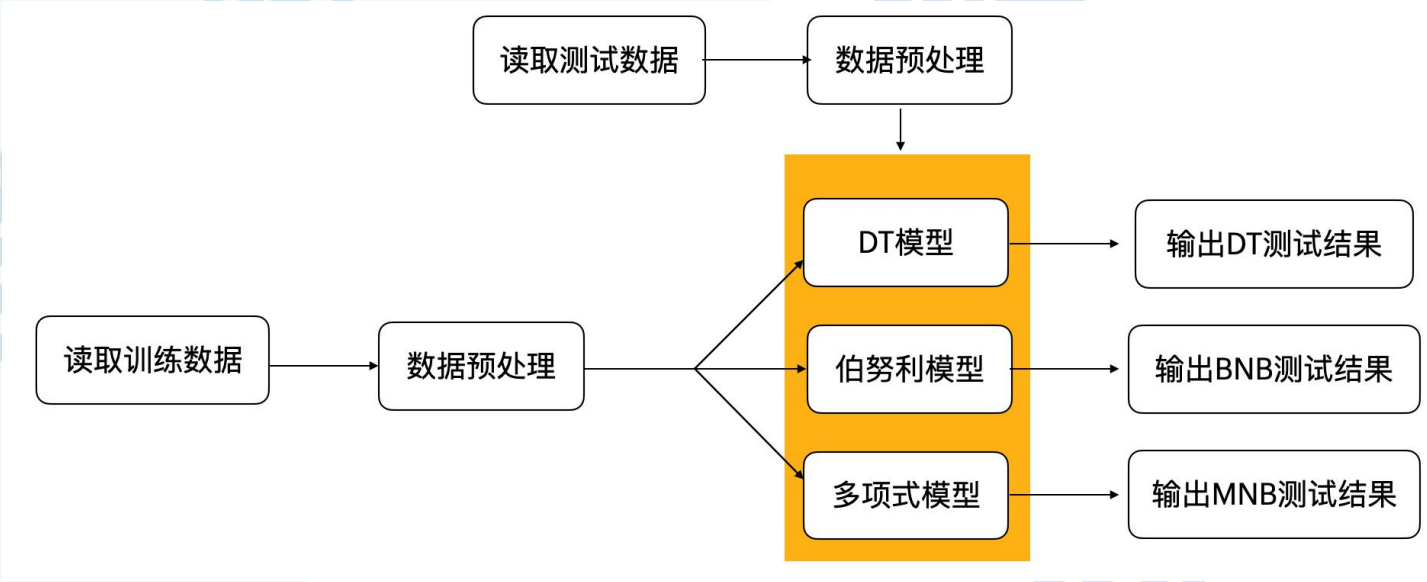
二

DT_sentiment.py,

BNB_sentiment.py,

MNB_sentiment.py

用这个给定的数据做训练集，对数据进行处理后，采用三种方法Decision Trees(DT), Bernoulli Naive Bayes(BNB) 和 Multinomial Naive Bayes (MNB)建立三个基础的模型，对给定的测试集进行测试，输出测试结果。



1 读取数据

```
python3 DT_sentiment.py training.tsv test.tsv > output.txt
```

import sys
data_set_file = sys.argv[1]
test_set_file = sys.argv[2]

1	**@JetBlue thank you for incredible customer svc from gate to flight. Mint experience is magic.**	**positive**
2	@united I was well taken care of, thanks. I've already been sent a survey request & I'll share my positive experience (despite delay)	positive
3	@united Too Late Flight, damage has been done. Easily the worst airline experience of my life. Missed two connecting flights & days of work. #UA49	negative
4	@USAirways of course! "Yeah, travel has gotten harder. Ask other passengers if they will switch" it's not the fact, it's the attitude!	negative
5	@united I'm very frustrated and have wasted 2 days now due to your equipment failures.	negative
6	@united please update what is going to happen to passengers now that flt ua14 has been Cancelled Flightled	neutral

```python
import pandas as pd
import numpy as np
# read files
train = pd.read_csv(data_set_file, sep = '\t', header = None)
test = pd.read_csv(test_set_file, sep = '\t', header = None)

# get sentences
train_sentence = np.array(train[1])
test_sentence = np.array(test[1])
test_id = np.array(test[0])
```

```python
import csv
id = []
sentence = []
y = []
with open(file_name,'r') as f:
    csv_reader = csv.reader(f,delimiter = '\t')
    for line in csv_reader:
        id.append(line[0])
        sentence.append(line[1])
        y.append(line[2])
```

2 数据预处理

For all models except VADER, consider a tweet to be a collection of words, where a *word* is a string of at least two letters, numbers or the symbols #, @, _, $ or %, delimited by a space, after removing all other characters (two characters is the default minimum word length for CountVectorizer in scikit-learn). URLs should be treated as a space, so delimit words. Note that deleting "junk" characters may create longer words that were previously separated by those characters.

- 去掉url，替换成空格
- 去掉 junk 字符，替换成‘’
- 设置一个单词至少有两个字符

➢去掉url，替换成空格

利用正则表达式
url = re.compile(.....)
re.sub(url,' ',sentence)

➢去掉junk 字符，替换成‘’

junk_character = 除了字母，数字，#,@, _,$or%
利用正则表达式
junk = re.compile(.....)
re.sub(junk,'',sentence)

正则表达式语法：https://www.runoob.com/regexp/regexp-syntax.html
正则表达式教程：https://www.runoob.com/python/python-reg-expressions.html
去掉url教程：https://stackoverflow.com/questions/11331982/how-to-remove-any-url-within-a-string-in-python

- 设置一个单词至少有两个字符

CountVectorizer()

见example.py

3 模型构建

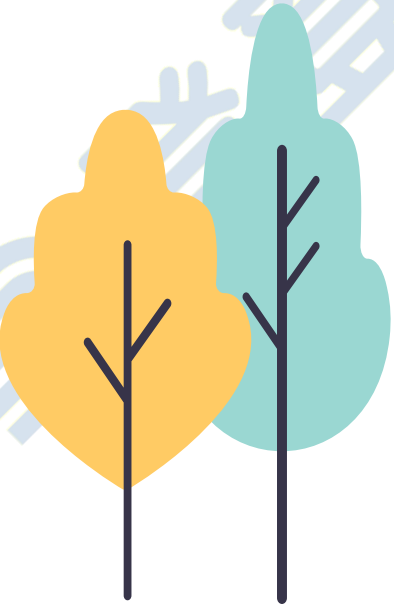
➢决策树模型

Train the three standard models on the supplied dataset of 5000 tweets (the whole of `dataset.tsv`). For Decision Trees, use scikit-learn's Decision Tree method with criterion set to 'entropy' and with random_state=0. Scikit-learn's Decision Tree method does not implement pruning, rather you should make sure Decision Tree construction stops when a node covers fewer than 50 examples (1% of the training set). Decision Trees are likely to lead to fragmentation, so to avoid overfitting and reduce computation time, for all Decision Tree models use as features only the 1000 most frequent words from the vocabulary (after preprocessing to remove "junk" characters as described above). Write code to train and test a Decision Tree model in `DT_sentiment.py`.

➢BNB模型和MNB模型

For both BNB and MNB, use scikit-learn's implementations, but use *all* of the words in the vocabulary as features. Write two Pythons programs for training and testing Naive Bayes models, one a BNB model and one an MNB model, in `BNB_sentiment.py` and `MNB_sentiment.py`.

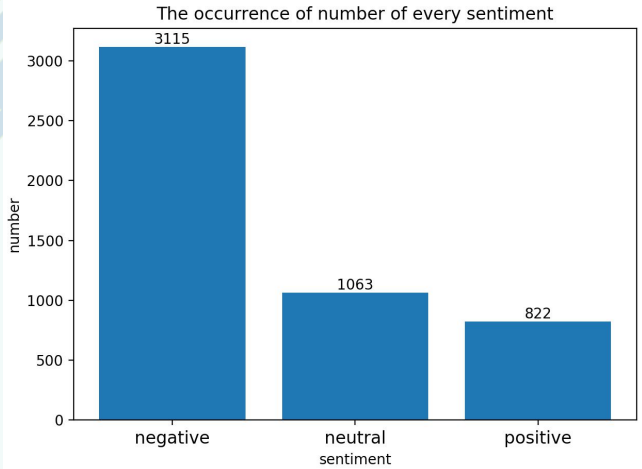
三 report + sentiment.py



In the report, you will first evaluate the standard models, then present your own model. For evaluating all models, report the results of training on the first 4000 tweets in `dataset.tsv` (the "training set") and testing on the remaining 1000 tweets (the "test set"), rather than using the full dataset of 5000 tweets for training, so stopping the Decision Tree classifiers when nodes cover less than 40 tweets rather than 50. Use the metrics (micro- and macro-accuracy, precision, recall and F1) and classification reports from scikit-learn. Show the results in either tables or plots, and write a *short* paragraph in your response to each item below. The answer to each question should be self contained. **Your report should be at most 10 pages.** Do not include appendices.

使用图表，拒绝直接截图分类报告!

1. (1 mark) Give simple descriptive statistics showing the frequency distribution for the sentiment classes for the whole dataset of 5000 tweets. What do you notice about the distribution?



2. (2 marks) Develop BNB and MNB models from the training set using (a) the whole vocabulary, and (b) the most frequent 1000 words from the vocabulary (as defined using CountVectorizer, after preprocessing by removing "junk" characters). Show all metrics on the test set comparing the two approaches for each method. Explain any similarities and differences in results.

For this setting (multi-class but each instance is contained in, and predicted to be in, one and only one class), micro-F1 = micro-precision = micro-recall = accuracy.

bnb all

	precision	recall	f1-score	support
negative	0.71	0.99	0.82	628
neutral	0.82	0.30	0.44	210
positive	0.95	0.25	0.40	162
accuracy			0.73	1000
macro avg	0.83	0.51	0.55	1000
weighted avg	0.77	0.73	0.68	1000

bnb 1000

	precision	recall	f1-score	support
negative	0.90	0.85	0.87	628
neutral	0.61	0.68	0.65	210
positive	0.63	0.68	0.65	162
accuracy			0.78	1000
macro avg	0.71	0.74	0.72	1000
weighted avg	0.79	0.78	0.79	1000

mnb all

	precision	recall	f1-score	support
negative	0.75	0.98	0.85	628
neutral	0.80	0.32	0.46	210
positive	0.84	0.47	0.60	162
accuracy			0.76	1000
macro avg	0.80	0.59	0.64	1000
weighted avg	0.78	0.76	0.73	1000

mnb 1000

	precision	recall	f1-score	support
negative	0.85	0.90	0.87	628
neutral	0.67	0.55	0.60	210
positive	0.68	0.69	0.69	162
accuracy			0.79	1000
macro avg	0.73	0.71	0.72	1000
weighted avg	0.78	0.79	0.79	1000

3. (2 marks) Evaluate the three standard models with respect to the VADER baseline. Show all metrics on the test set and comment on the performance of the baseline and of the models relative to the baseline.

NLTK includes a hand-crafted (crowdsourced) sentiment analyser, VADER,[1] which may perform well in this domain because of the way it uses emojis and other features of social media text to intensify sentiment, however the accuracy of VADER is difficult to anticipate because: (i) crowd-sourcing is in general highly unreliable, and (ii) this dataset might not include much use of emojis and other markers of sentiment.

	precision	recall	f1-score	support
negative	0.91	0.48	0.63	628
neutral	0.36	0.43	0.39	210
positive	0.34	0.89	0.49	162
accuracy			0.54	1000
macro avg	0.54	0.60	0.51	1000
weighted avg	0.70	0.54	0.56	1000

4. (2 marks) Evaluate the effect of preprocessing the input features by applying NLTK English stop word removal then NLTK Porter stemming on classifier performance for the three standard models. Show all metrics with and without preprocessing on the test set and explain the results.

Stop Words

一类是人类语言中包含的功能词，这些功能词极其普遍，与其他词相比，功能词没有什么实际含义，比如'the'、'is'、'at'、'which'、'on'等

另一类词包括词汇词，比如'want'等，这些词应用十分广泛，但是对这样的词搜索引擎无法保证能够给出真正相关的搜索结果，难以帮助缩小搜索范围。

stemming 词干提取

通过将单词的不同形式转换为基本形式来减少单词量。spatial--spat

bnb remove stop words and stemming

	precision	recall	f1-score	support
negative	0.69	0.98	0.81	628
neutral	0.75	0.24	0.36	210
positive	0.88	0.22	0.35	162
accuracy			0.70	1000
macro avg	0.77	0.48	0.51	1000
weighted avg	0.73	0.70	0.64	1000

dt remove stop words and stemming

	precision	recall	f1-score	support
negative	0.78	0.85	0.81	628
neutral	0.43	0.39	0.41	210
positive	0.70	0.56	0.62	162
accuracy			0.70	1000
macro avg	0.64	0.60	0.61	1000
weighted avg	0.69	0.70	0.69	1000

mnb remove stop words and stemming

	precision	recall	f1-score	support
negative	0.75	0.97	0.85	628
neutral	0.77	0.31	0.44	210
positive	0.79	0.51	0.62	162
accuracy			0.76	1000
macro avg	0.77	0.60	0.64	1000
weighted avg	0.76	0.76	0.73	1000

5. (2 marks) Evaluate the effect that converting all letters to lower case has on classifier performance for the three standard models. Show all metrics with and without conversion to lower case on the test set and explain the results.

dt lower case = False

	precision	recall	f1-score	support
negative	0.73	0.91	0.81	628
neutral	0.46	0.26	0.33	210
positive	0.70	0.46	0.55	162
accuracy			0.70	1000
macro avg	0.63	0.54	0.57	1000
weighted avg	0.67	0.70	0.67	1000

6. (6 marks) Describe your best method for sentiment analysis and justify your decision. Give some experimental results for your method trained on the training set of 4000 tweets and tested on the test set of 1000 tweets. Provide a brief comparison of your model to the standard models and the baseline (use the results from the previous questions).

mnb
max_feature = 1000
stemming

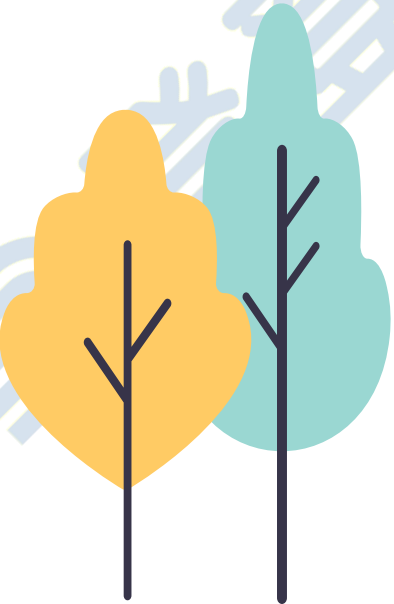
sentiment.py

跟三个标准模型+Vader对比

词形归并(效果一般)：https://www.jianshu.com/p/32258d3b02f6

mnb max features = 1000, stemming

	precision	recall	f1-score	support
negative	0.85	0.89	0.87	628
neutral	0.65	0.56	0.60	210
positive	0.71	0.71	0.71	162
accuracy			0.79	1000
macro avg	0.74	0.72	0.73	1000
weighted avg	0.79	0.79	0.79	1000



有问题发群里

需要1v1 debug的咨询艾米酱

登登教育

THANKS

更多UNSW CSE课程，请咨询微信