使用NLTK训练情感分析器

介绍

今天我们使用NLTK工具包，对`movie_reviews`语料库进行训练。训练过程中使用NLTK自带的贝叶斯分类器模型。最终结果可以对一段英文的电影评价进行积极/负面的情感预测。

准备

- python3
- nltk（使用pip安装，`pip install nltk`）
- nltk.corpus.movie_reviews （在python程序内使用`nltk.download('movie_reviews')`安装）

具体过程

1. 新建一个setupNLTK.py文件，导入nltk包并下载所需语料库

```
# setupNLTK.py
import nltk
if __name__ == '__main__':
    nltk.download('movie_reviews')
```

创建完成后执行，命令行会提示下载完成

2. 新建一个main.py文件，导入下列python包

```
import nltk.classify.util
from nltk.classify import NaiveBayesClassifier
from nltk.corpus import movie_reviews
```

3. 导入movie review数据

```
if __name__ == '__main__':
    positive_fileids = movie_reviews.fileids('pos')
    negative_fileids = movie_reviews.fileids('neg')
```

4. 定义一个函数，用来提取特征数据

```
def extract_features(word_list):
    return dict([(word,True) for word in word_list])
```

5. 将语料库中的数据通过刚才定义的函数提取出来

```
features_positive =
[(extract_features(movie_reviews.words(fileids=[f])),
'Positive') for f in positive_fileids]
features_negative =
[(extract_features(movie_reviews.words(fileids=[f])),
'Negative') for f in negative_fileids]
```

6. 将数据分成训练数据和测试数据

```
# Split the data into train and test (80/20)
threshold_factor = 0.8
threshold_positive =
int(threshold_factor*len(features_positive))
threshold_negative =
int(threshold_factor*len(features_negative))
features_train = features_positive[:threshold_positive]
+ features_negative[:threshold_negative]
features_test = features_positive[threshold_positive:] +
features_negative[threshold_negative:]
```

7. 使用朴素贝叶斯分类器训练

```
# Train a Naive Bayes classifier
classifier = NaiveBayesClassifier.train(features_train)
print ("\nAccuracy of the classifier:",
nltk.classify.util.accuracy(classifier, features_test))
```

8. 分类器对象中存有从训练数据中获取的对语义最有影响的单词，我们将它们输出

```
print ("\nTop 10 most informative words:")
for item in classifier.most_informative_features()[:10]:
    print (item[0])
```

9. 给一些输入文本

```
# Sample input reviews
input_reviews = [
    "It is an amazing movie",
    "This is a dull movie. I would never recommend it to
anyone.",
    "A complete and utter destruction of one of the most
iconic superheroes. 0.1 effort and thought put into the
storyline. A coming of age awkward teenage movie with a
'spiderman' stamp put on it. Bad jokes aimed at
teenagers (at best). A complete caricature of a villain,
a complete caricature of a Spiderman. Just please stop
making this garbage Put some god damn effort! A total
waste of time",
    "Just staving off some negative reviews. Fits well
into the Marvel movies to date and is an excellent
follow up to Avengers: Endgame."
]
```

```
### 10. 用我们之前训练出的分类器预测这些文本的分类
```

```
print ("\nPredictions:")
for review in input_reviews:
    print ("\nReview:", review)
    probdist =
classifier.prob_classify(extract_features(review.split()
))
    pred_sentiment = probdist.max()
    print ("Predicted sentiment:", pred_sentiment )
    print ("Probability:",
round(probdist.prob(pred_sentiment), 2))
```

结果

```
[nltk_data] Downloading package movie_reviews to
```

```
[nltk_data]     /home/xinrui/nltk_data...
[nltk_data]   Package movie_reviews is already up-to-
date!
['films', 'adapted', 'from', 'comic', 'books', 'have',
...]

Number of training datapoints: 1600
Number of test datapoints: 400

Accuracy of the classifier: 0.735

Top 10 most informative words:
outstanding
insulting
vulnerable
ludicrous
uninvolving
avoids
astounding
fascination
seagal
anna

Predictions:

Review: It is an amazing movie
Predicted sentiment: Positive
Probability: 0.61

Review: This is a dull movie. I would never recommend it
to anyone.
Predicted sentiment: Negative
Probability: 0.77

Review: A complete and utter destruction of one of the
most iconic superheroes. 0.1 effort and thought put into
the storyline. A coming of age awkward teenage movie
with a 'spiderman' stamp put on it. Bad jokes aimed at
teenagers (at best). A complete caricature of a villain,
a complete caricature of a Spiderman. Just please stop
making this garbage Put some god damn effort! A total
waste of time
Predicted sentiment: Negative
Probability: 0.99
```

```
Review: Just staving off some negative reviews. Fits
well into the Marvel movies to date and is an excellent
follow up to Avengers: Endgame.
Predicted sentiment: Positive
Probability: 0.92
```