复旦大学大数据学院
School of Data Science, Fudan University

Assignment 2
Stock Market Prediction

Description

- In this assignment, you are required to predict whether the price of a stock will go up or down based on financial news.
- You need to implement the Naïve Bayes model that maximizes the jointly likelihood of word features and target labels for text classification.
- At least three other classification models provided by NLTK should be tested and reported.

Formal Definition

Input

A group of financial news about a target stock that publish in the same day.

Output

Predict whether the price of the target stock will go up or down.

1. ‘+1’: price will go up
2. ‘-1’: price will go down

Save your answer in “result.txt”. Each line should contain one prediction inline with instances provided in “test.txt”.

Provided Files

- news.txt
  - Each line contains a single instance of news sample which is stored as data type of dictionary.
  - Each sample of news has three fields: *ID*, *title* and *content*

```
In [9]: import codecs
   ...: f = codecs.open('news.txt','r','utf8')
   ...: news = [eval(i) for i in f.readlines()]
   ...: f.close()

In [10]: news[3]
Out[10]:
{'content': '【中信证券2014年净利同比增长115%】28日讯，中信证券晚间发布2014年度业绩快报。2014年公
司实现营业收入295.11亿元,同比增长83.13%;归属于母公司股东的净利润112.95亿元,同比增长115.39%;每股收
益1.03元。',
 'id': 30886,
 'title': '中信证券2014年净利同比增长115%'}
```

Provided Files

- train.txt & test.txt
  - These two files contain 7,954 and 3,000 instances for training and testing, respectively.
  - Each instance contains several pieces of financial news related to a company from a same day and the trend of the stock price resulted.

```
In [13]: f = open('train.txt','r')
    ...: train = [i.split() for i in f.readlines()]
    ...: f.close()

In [14]: train[0]
Out[14]: ['-1', '30819,30826,30876']
```

  - Obviously, you can not use the label in the testing file for prediction. Labels for testing instances are given to support your report write-up.
- scorer.py
  - This file can help you evaluate the performance of your model (recall, accuracy, precision and F1-score).

Submission

- Generate a zip file and name it as "sid_homework-2.zip".
- It should include a python file named "prediction.py", an output file named "result.txt" and a written report named "stock market prediction.pdf".
- Program: codes should be written in python.
- Report: the report needs to be written in English with no more than 4 pages.

Evaluation

- We will mark your homework based on the three criteria:
  - Final accuracy (20%)
  - Program (30%)
  - Report (50%)

Due

- Submit your homework via E-learning system.
- Deadline: Mid-night at Nov. 17th, 2017

- If you have any questions about this homework, send email to TA or our course mailbox.

- TA in Charge
  - 陈婷(17210980029@fudan.edu.cn)