

项目配置说明：

一、爬虫模块：

1、文件结构：包含Weibo和dianping两个爬虫文件，分别对应微博数据和大众点评爬虫。



2、安装Python2.7：从官网上下载Python2.7。官方网址：<https://www.python.org/download>

3、安装Python包管理工具pip。

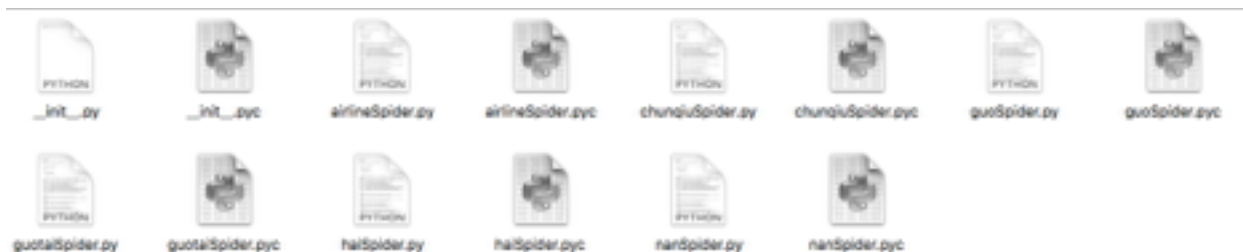
从网址：<https://bootstrap.pypa.io/get-pip.py> 上得到get-pip.py文件，使用命令安装pip：

```
[~] python get-pip.py
```

4、使用pip命令安装Scrapy。

```
[~] pip install Scrapy
```

5、Weibo文件中有spiders文件夹，其中包含着所有的爬虫。



6、使用命令进入Weibo文件夹中，并通过scrapy crawl +爬虫名爬取数据。其中chunqiuSpider为爬虫名。

```
[Spider] cd Weibo 16:5
[Weibo] scrapy crawl chunqiuSpider 16:5
2017-06-18 17:03:06 [scrapy] INFO: Scrapy 1.2.0 started
2017-06-18 17:03:06 [scrapy] INFO: Overridden settings:
{'botname': 'tutorial.spiders', 'ROBOTSTXT_OBEY': True, 'SPIDER_MODULES':
['BOT_NAME': 'tutorial', 'COOKIES_ENABLED': False, 'DEFAULT_ITEMS':
['tutorial.items.DmozItem', 'DOWNLOAD_DELAY': 1]}
```

二、数据分析模块：

1、使用pip安装ipython和notebook

```
[~] pip install ipython
```

```
[~] pip install notebook
```

2、开启交互式IPython

```
[~] ipython notebook
```

3、打开数据分析源代码中的Untitled.ipynb即可运行项目。

Select items to perform actions on them.

☐
/ Desktop / 郭泰成_社交网络数据分析与可视化工具的设计与实现 / 项目源代码 / 数据分析代码

<input type="checkbox"/>	..
<input type="checkbox"/>	randomcolor
<input type="checkbox"/>	Untitled.ipynb
<input type="checkbox"/>	1.png
<input type="checkbox"/>	neg.txt
<input type="checkbox"/>	official
<input type="checkbox"/>	pos.txt
<input type="checkbox"/>	sentiment_marshall
<input type="checkbox"/>	sentiment_df
<input type="checkbox"/>	stop_words
<input type="checkbox"/>	tag_cloud_negative_东航.png
<input type="checkbox"/>	tag_cloud_negative_国航.png
<input type="checkbox"/>	tag_cloud_negative_春秋航空.png
<input type="checkbox"/>	tag_cloud_negative_海航.png

项目使用说明：

针对每一条语句，点击图中红色标记的按钮或者按住shift+enter即可运行。

jupyter
Untitled (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help
Python [default]

主题：通过微博数据对各大航空公司的公众情感倾向以及服务质量进行分析

一、加载原始数据

1、从mongodb导出的csv文件中加载数据

```

In [1]: import pandas as pd
import numpy as np
from snowlp import SnowLP
from snowlp import sentiment
import matplotlib.pyplot as plt
import matplotlib

df1 = pd.read_csv('weibo_posts.csv')
df2 = pd.read_csv('weibo_posts_4.6.csv')
df = df1.append(df2)
    
```

2、统计爬取的原始数据的数量