说明文档

代码分成两个部分

（1） 短文本内容爬取。爬取 twitter 官方日报的数据-> 对数据进行情感分析 -> 结果可视化

（2） 长文本内容爬取。爬取官方网站的官网数据-> 网页解析，获取正文内容 -> 对正文进行停用词处理 -> 获取主题词 -> 结果可视化

需要额外依赖的包：

tweepy python twitter api 工具包

matplotlib 画图工具包

scipy 函数库

numpy 数值计算包

pandas 基于 Numpy 构建的含有更高级数据结构和工具的数据分析包

dataset 数据库管理工具

textblob 进行文本数据处理的工具包，用到它情感分析的部分

plotly 可视化工具，生成地图 wordcloud 可视化工具，生成词云

jieba 分词工具，这里用它来提取关键词

安装方法：

```
pip install -r requirements.txt
```

第一部分：

设置：

在 `private.py` 中

TWITTER_APP_KEY =

TWITTER_APP_SECRET =

TWITTER_KEY =

TWITTER_SECRET =

是用来配置 twitter api 参数的，具体获得方法：

（1）注册一个 twitter

https://dev.twitter.com

（2）创建一个 app 应用

https://apps.twitter.com

（3） 获得授权和密码

在 settings.py 中

TRACK_TERMS = ["THAAD"]

定义了查询的关键词

CONNECTION_STRING = "sqlite:///tweets.db"

定义了数据库的连接名字

OFFICIAL_LIST=['PDChina','XHNews','YonhapNews', 'cnn','WashTimes']

定义了要查询新闻的名称

OFFICIAL_LOC =['CHN','CHN','KOR','USA','USA']

媒体所在国家

CSV_NAME_OFFICIAL='twitter_message_official.csv'

定义输出的文件名称

TABLE_NAME_OFFICIAL = "THAAD_OFFICIAL"

定义了数据库中表的名字

配置好后，执行顺序为

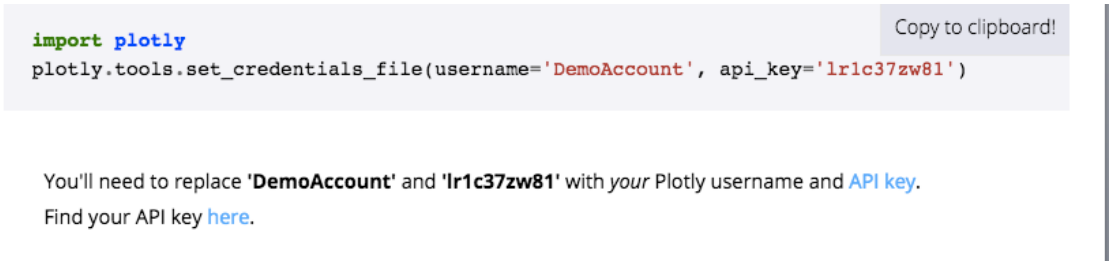
scraper_official.py -> dump.py -> translationData.py -> layout.py

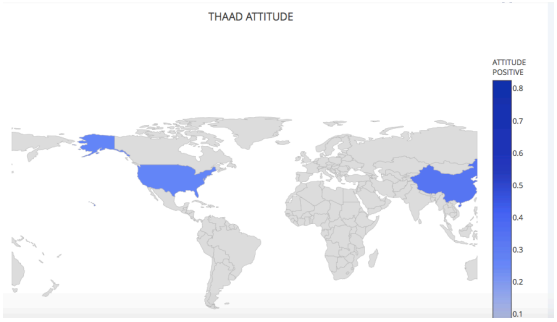
爬取数据，情感分析 -> 存储数据 -> 转换数据 -> 数据可视化

备注：plotly 是一个需要注册使用的工具包，注册方法

https://plot.ly/python/getting-started/

在使用之前要在电脑前初始化下





第二部分：

设置：

`setting.conf` 包含输出文件路径，和特征参数设置

`getnews.py -> getdetails.py -> transfer.py -> lda.py ->`

获取网站 `url` -> 获取网页内容 -> 去除常用词 -> 主题词生成 ->

`extract.py`

生成词云