**本文档为项目说明文档，请按顺序浏览文档并执行程序！**

**一、开发环境：**

Python3.7 + django 2.1.7 + numpy 1.16.3 + pandas 0.24.2 + jieba 0.39 + wordcloud 1.5.0 + matplotlib 3.0.3 + snownlp 0.12.3

**二、编辑器：**

**使用pycharm或其它可运行 python 项目的编辑器；推荐使用pycharm；**

**三、安装必须的包：**

**首先，必须先安装Python3.7；**

**然后，在项目根目录下，打开终端（Terminal），输入命令：**

**pip install -r requirements.txt**

**稍等片刻，即可安装完必须的包；**

**四、运行项目：**

**1. 当安装完必须的包（执行完上一步）之后，继续在终端（Terminal）下，输入命令：**

**Python manage.py makemigrations**

**然后，继续执行：**

**Python manage.py migrate**

**以上命令执行完成后，即可完成数据库的迁移；**

1. **执行爬虫程序（必须保证网络通畅）：**

**在项目的microblog文件夹下（Consensus/microblog/）：**

**有一个文件 wb\_spider.py ，在pycharm中打开此文件，直接运行该文件，即可运行爬虫程序；稍等一会，爬虫程序即可完成；**

1. **运行web项目：**

**在终端下执行命令：**

**Python manage.py runserver**

**然后，在浏览器中输入：**<http://127.0.0.1:8000/>

**即可打开web界面；**

1. **在主页可看见共有导航栏共有三个功能按钮和一个搜索按钮：**
2. **主页、知乎、登录和搜索功能；**
3. **主页可直接点击一键保存本页数据（必须保证网络通畅）；**
4. **在主页可浏览最新微博、可查看详细内容、微博原文详情、查看专题内容（蓝色标题）、 可点击搜索功能（搜索之后可查看舆情分析情况）；在专题页面也可一键保存该专题所有微博；**
5. **点击登录，会跳转到登录注册页面，即可登录注册账号；**
6. **使用技术**

**微博：**

1. **爬虫：**

**在爬取新浪微博时，使用了requests 库进行数据的抓取，为了节省爬取数据的时间，使用了python的threading模块，利用多线程技术进行数据的爬取；同时，在爬虫脚本文件中配置了django的环境，然后可以使用django的ORM技术进行数据库的存储；**

1. **保存数据库：**

**在保存数据库时，如果同时将大量的数据写入数据库，数据库会自动上锁，此时，可以使用threading模块的threading.Lock()技术进行对存储数据库时进行上锁，当一个线程执行完保存数据库后，释放锁，交给第二个线程，这样，就解决了多线程保存数据库时数据库锁死的问题；**

1. **从数据库读取数据并进行数据分析：**

**从数据库读取数据，同样使用了django的ORM（在分析脚本文件中配置了django环境），读取数据后；**

**数据预处理：**

**①将数据读取为pandas的DataFrame格式；**

**②对DataFrame数据进行去空值；**

**③对DataFrame数据进行去重；**

**④利用自定义的函数对每条数据进行机械压缩去词；**

**数据分析：**

**①使用snownlp库的sentiments方法进行情感分析；**

**②使用jieba库进行分词；**

**③去除停用词；**

**画图：**

**①使用wordcloud画出分析后的词云图；**

**②使用matplotlib的bar方法画出高频词柱状图；**

**③使用matplotlib的pie方法画出正负面百分比饼图；**

1. **页面渲染**

**使用django的模板渲染，对数据进行渲染；**

1. **搜索**

**使用django的Q表达式进行对数据库的多个表达式查询；**

**知乎：**

**一、知乎主页展示：**

上面的四个轮播图，是随便在网上找的，文字也是网上随便找的，就是让好看点。（实时刷新的）

下面的图片也是固定的10个图片轮着，不是爬取的图片，图片下面的标题是知乎热榜前12（默认12，修改方式：zhihu/views.py中第18行 修改12）个。点击图片或文字都可查看详细信息。 技术难点：因为知乎主页必须登录才能查看（内容页不需要），我就使用第三方网站：AnyKnew - 效率资讯 - 高效读新闻, 5分钟遍历全网热点（网址：<https://www.anyknew.com/#/> ）。 这个第三方网站，里面有知乎的热榜50个（问题）。通过这个网站爬取知乎热榜问题的链接和问题（通过接口爬，接口网址：<https://www.anyknew.com/api/v1/sites/zhihu> ），然后在主页展示。

爬虫技术：要携带useragent 的请求头。

**二、详情页：**

展示10个（默认10个，修改方式：zhihu/views.py中第19行 修改10）回答。通过知乎详情页的接口爬取内容，（接口网址： https://www.zhihu.com/api/v4/questions/271861610/answers?include=data[\*].is\_normal,admin\_closed\_comment,reward\_info,is\_collapsed,annotation\_action,annotation\_detail,collapse\_reason,is\_sticky,collapsed\_by,suggest\_edit,comment\_count,can\_comment,content,editable\_content,voteup\_count,reshipment\_settings,comment\_permission,created\_time,updated\_time,review\_info,relevant\_info,question,excerpt,relationship.is\_authorized,is\_author,voting,is\_thanked,is\_nothelp,is\_labeled,is\_recognized,paid\_info;data[\*].mark\_infos[\*].url;data[\*].author.follower\_count,badge[\*].topics&limit=5&offset=5&platform=desktop&sort\_by=default ）

三、点击原链接，可跳转到知乎的原链接。

点击保存，可将此问题及默认300回答（修改方式：zhihu/views.py中第58行 修改300）保存到数据库。

四、点击搜索，搜索的是数据库内保存过的问题数据。

搜索的结果是数据库内的问题，图片也是固定的图片。点击问题后，进入这个对问题的回答的舆情分析结果的图片展示。（展示的有情感分析词云图，分喜欢情感和不喜欢情感2个图；热词频率统计柱状图，分喜欢情感和不喜欢情感2个图，一共4个展示图）

**五、数据舆情分析技术：**

先用snownlp库进行舆情分析，再用jieba库进行分词，在用matplotlib库画图并展示。

画图有：1.词云图，用wordcloud画词云图，再用matplotlib画图；2.高频词柱状图，先用collections库进行统计jieba分词，再用matplotlib画图。