**一 微博爬虫程序**

1. **程序主体：**

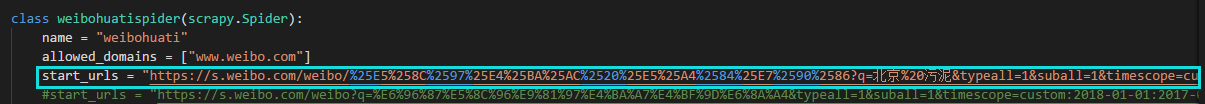
微博爬虫程序->weibohuati->函数文件main.py

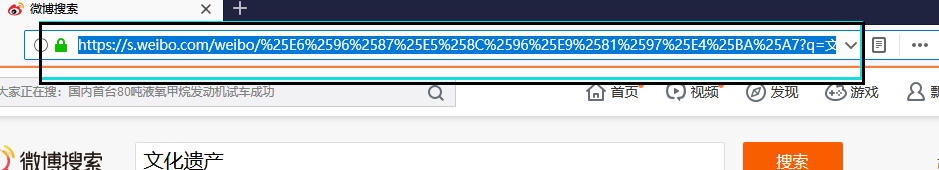
微博爬虫程序->weibohuati->spiders->函数文件weibohuati.py和xpath.py。

1. **程序执行步骤：**
2. 在微博的高级搜索中输入关键词和时间，爬取给定关键词和时间段的数据，如下图：

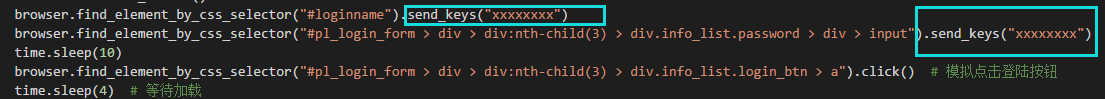


1. 在weibohuati.py函数文件的对应位置输入搜索出来的内容对应的网址

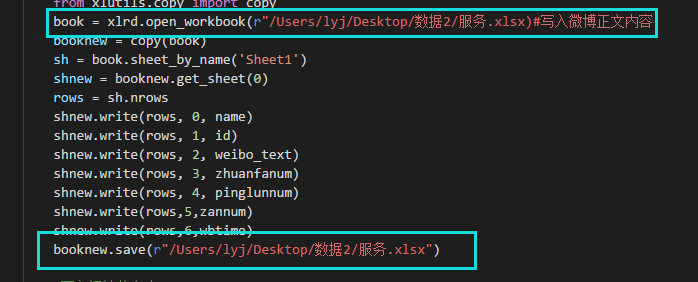




3)在weibohuati函数文件中的下图中的位置输入微博账号和密码



4)在xpath.py函数文件的对应位置修改微博保存的文件路径



5)执行main.py函数运行程序

**二 热点主题发现和关键词提取**

**提取关键词文件中的程序实现了热点主题发现和关键词提取，数据文件是程序运行所需的数据和生成的数据。**

**1、聚类word2vec(TF-IDF).ipynb这个python函数文件实现了政协提案的热点主题发现，并实现了一种关键词提取方法，程序直接顺序运行即可。**

**程序执行所需的文件的路径分别是：**

1. “百度停用词列表（去除英文）+tycb\_symbol.txt”的路径：项目\停用词\数据
2. “提案数据”的路径：项目\提取关键词\数据\提案数据\提案数据
3. “sogou.txt”和“标题补充词典.txt”的路径：项目\补充词典\数据

**2、key\_words.ipynb这个python函数文件实现了另一种关键词提取方法，其中需要用到聚类word2vec(TF-IDF).ipynb中的运行结果。**

**程序执行所需的文件的路径是：**

1. 所需文件1的路径：项目\提取关键词\数据\候选关键词
2. 所需文件2的路径：项目\提取关键词\数据\聚类结果\word2veccluster

**3、 key\_words.ipynb这个python函数文件对两种关键词提取算法进行了对比。所需的文件的路径分别是：**

1. 项目\提取关键词\数据\关键词文档
2. 项目\提取关键词\数据\关键词文档1
3. 项目\提取关键词\数据\关键词文档2

**三 情感分析**

**1、 create\_tycb.ipynb这个python函数文件用于对百度停用词表进行改进。**

**2、 情感分析模型LSTM.ipynb这个python函数文件训练了一个基于双向LSTM的情感分类模型。**

**所需的文件的路径是：**

1) 项目\情感分析\数据\所有微博数据

**生成数据的保存路径是：**

1. 项目\情感分析\数据\所有打标数据

**这个文件夹下保存了所有的微博数据和每个微博数据对应的预测标签和得分。**

**3、情感强度.ipynb这个python函数文件将所有微博数据的强度以主题为文件进行保存。**

**程序执行所需的文件的路径是：**

1) 项目\情感分析\数据\所有打标数据

**生成数据的保存路径是：**

1. 项目\情感分析\数据\强度集合

**4、情感极性.ipynb这个python函数文件统计每个主题在不同时间段的不同情感极性的微博数、每个主题在不同时间段的不同情感极性的微博数占该时间段微博总数的比例、每个主题的微博数随时间的变化曲线。**

**程序执行所需的文件的路径是：**

1) 项目\情感分析\数据\强度集合