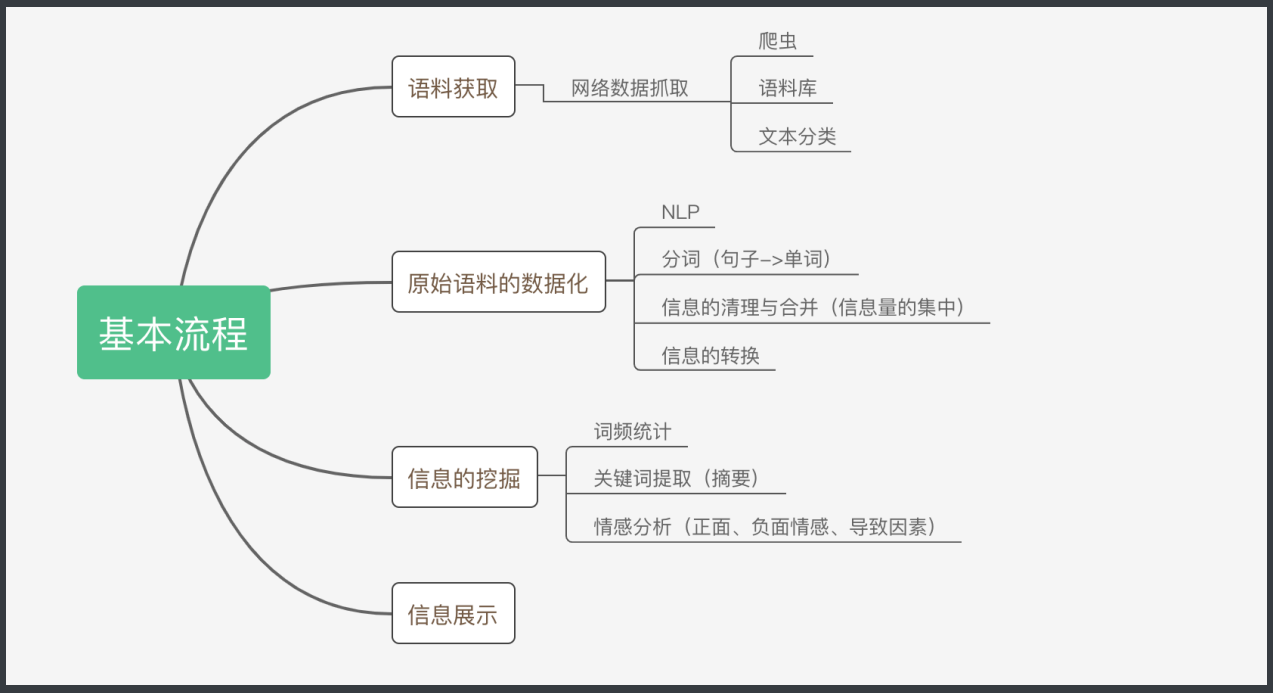
数据科学基础大作业

基本流程



STEP1.语料获取：

主要流程：

1.将目标数据源、获取对象进行分类：

- **主流媒体**新浪、百度、天涯的**新闻标题**
- 新闻的**评论**(新浪微博等)
- 重点新闻的**重点内容**

2.语料的获取：

按照时间发展对疫情划分为4个阶段

利用爬虫技术以及百度、新浪等提供的API**分别获取这4个阶段的语料对象**

3.对获取的语料进行归类和整理

- 将4个阶段的语料数据按照第一步所划分的三个部分进行分类
- 存储到**数据库**中，如果不调用第三方中文NLP接口（**如snownlp和jieba**），而是直接使用类似nltk库，那么还要对原始语料进行标记，以便nltk的处理。

使用工具、技术

- 爬虫技术：

> 1.请求阶段：引入requests包实现本地对网络的请求，得到网页的源码
>
> 2.解析阶段：BeautifulSoup + lxml解析器
>
> 3.数据保存：保存数据有两种方式：
>
> 一种是直接保存到文本，并生成文本文件到指定路径，但如果数据量大，不易进行数据的分类
>
> 另一种是保存到数据库，比如mysql等

参考资料

1.爬虫简介：https://blog.csdn.net/aaronjny/article/details/77945329

2.

STEP2.数据分析+结果呈现

主要流程：

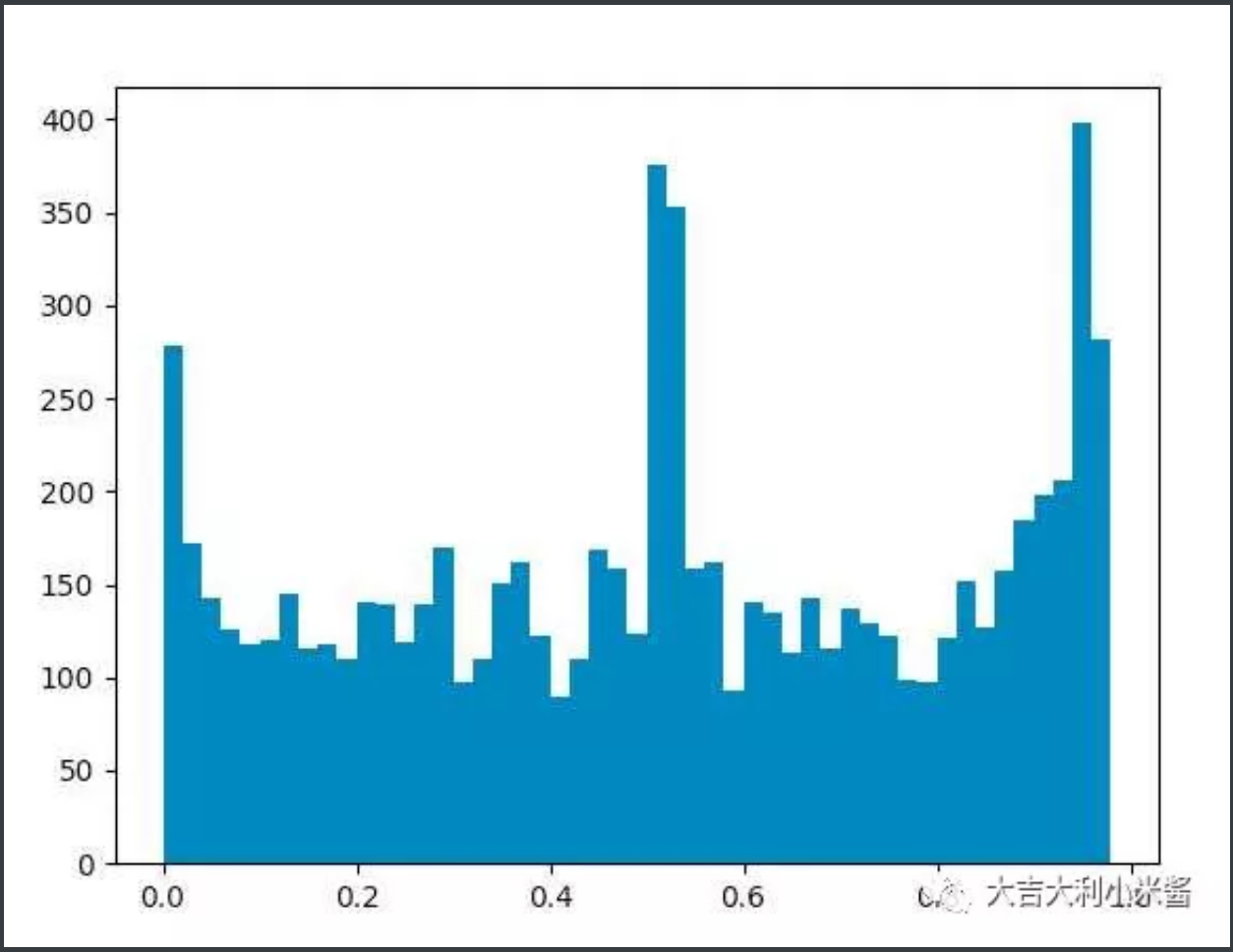
> 我觉得如果要自己实现**分词**、**同义词合并**、**关键词提取**、**词频统计**等工作操作难度较大
>
> 详细的数据分析步骤可参考：https://blog.csdn.net/yawei_liu1688/article/details/79011697
>
> 如果直接采用**jieba**这种第三方预训练模型，会大大简化实验的难度

使用的工具、技术

> 1.NLP库，比如NLTK，下载方式比较麻烦，下载完nltk库相当于只有数据分析的算法实现，还要下载nltk包含的语料库，直接百度。
>
> 2.第三方中文自然语言处理模块：比如snowNLP，jieba等，推荐jirba，下载地址 https://github.com/fxsjy/jieba
>
> 3.如果要生成词云，还要install **wordcloud** （用来生成词云）还有辅助作图工具：pandas,PIL
>
> 4.Excel也不妨一试

数据分析的不同角度（自己想了一部分，还可以补充）

1.感情值分布统计图（感情极差分析）



2.表情统计图


鹿晗微博评论表情统计
总计 53
[跟]
[牛郎]
[撒手]
[all微博益起来]
[呵呵]
[太開心]
[机智]
[悲傷]
[嘿哈]
[加油]
[ok]
关晓彤微博评论表情统计
总计 20
[ok]
[太開心]
[挖鼻屎]
[bm投诉]
[别烦我]
[加油]
大吉大利小米当

3.词云

4.感情变化趋势统计图

这一个是一定需要的，在ppt的第10页已经提供了这个思路

参考资料

1. NLTK的系统教程：https://www.bilibili.com/video/BV1dK4y1t7zD?p=1&share_medium=ipad&share_plat=ios&share_source=QQ&share_tag=s_i&timestamp=1606205413&unique_k=tstMJ2

STEP3：结果展示

> 最后的结果展示阶段应该要做一份报告，建议每个人做自己负责的部分，作为大作业的亲历者才能呈现出本次作业最为关键的点。

案例参考

1. 用python对鹿晗、关晓彤微博进行情感分析 https://mp.weixin.qq.com/s/a0904t-7Yvhi0VO_n-6CEw
2. 疫情期间的微博评论分析 https://github.com/NULL-JC/Text_Analysis
3. 京东商城某一红酒评论，情感分析https://blog.csdn.net/yawei_liu1688/article/details/79011697