**高考分数线的数据采集和分析**

**——计科1603班凌晓**

1. **数据爬取对象和范围**

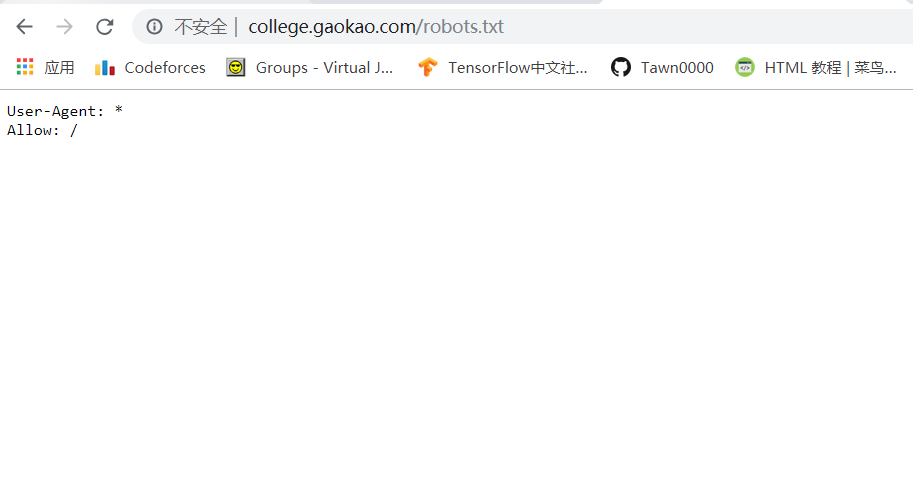
无论何时，高考一直都是一个受到大家广泛关注的话题。如何取得一个好的高考成绩很重要，但是如何在已知成绩下选择适合自己的学校和专业也同样重要。于是，本项目利用网络爬虫技术从展示了高考分数线的网站：<http://college.gaokao.com/spepoint> 爬取了从2010年到2017年高考分数线数据，包括大学，地区，专业，录取平均分，最高分等信息。

网页界面：



1. **爬取对象的限制和约束**

众所周知，爬虫技术存在着很大的风险，因此，在网络爬虫领域有一个不成文的道德约束，即robots.txt，所有爬取的行为需要遵循robots.txt的规定。首先我们来看一下该网站的robots.txt文件的内容。



如上图，我们可以看到该网站允许了所有爬虫的任何爬取行为。因此，我们可以放心的爬取该网站数据作为项目数据。

1. **爬取流程和数据存储**
2. 爬取URL的确定

首先我们可以观察到该网址的url查询存在着一定的规律：[http://college.gaokao.com/spepoint/a{}/s{}/y{}/p{}](http://college.gaokao.com/spepoint/a%7b%7d/s%7b%7d/y%7b%7d/p%7b%7d)

a 代表着地区编号如湖南，江苏等，s代表着理科，文科等学科大类，y代表着考试年份。所以我们利用这个规律来枚举url进行数据的爬取。

1. 使用的爬虫框架

此次爬取使用的爬虫框架是Scrapy: Scrapy（/ˈskreɪpi/) 是一个Python编写的开源网络爬虫框架。它是一个被设计用于爬取网络数据、提取结构性数据的程序框架。

使用Scrapy框架可以很方便地爬取网站。

Scrapy的具体使用和编写可见该教程：<https://tawn0000.github.io/2019/10/03/scrapy-introduction/>

1. 爬取过程中的注意事项

虽然该网站允许任何爬虫的任意行为，但是为了不必要的麻烦，我进行了以下几个操作，1）设置了IP 代理和UA代理, 防止其能反向识别和追踪我的爬虫。2）限制了并发线程的数量，使其不至于影响对方网站服务器的正常运行。

1. 爬取数据的提取和存储

利用Scrapy爬下html 页面，同时使用reponse.xpath 和regrex正则进行网页上的数据字段的提取并定义为item。

将得到的数据利用自行编写的scrapy pipeline组件将item存入本地的MongoDB。为后面的数据分析和文本分析提供数据支持。同时将数据.json的格式存在硬盘，以供提交查验。

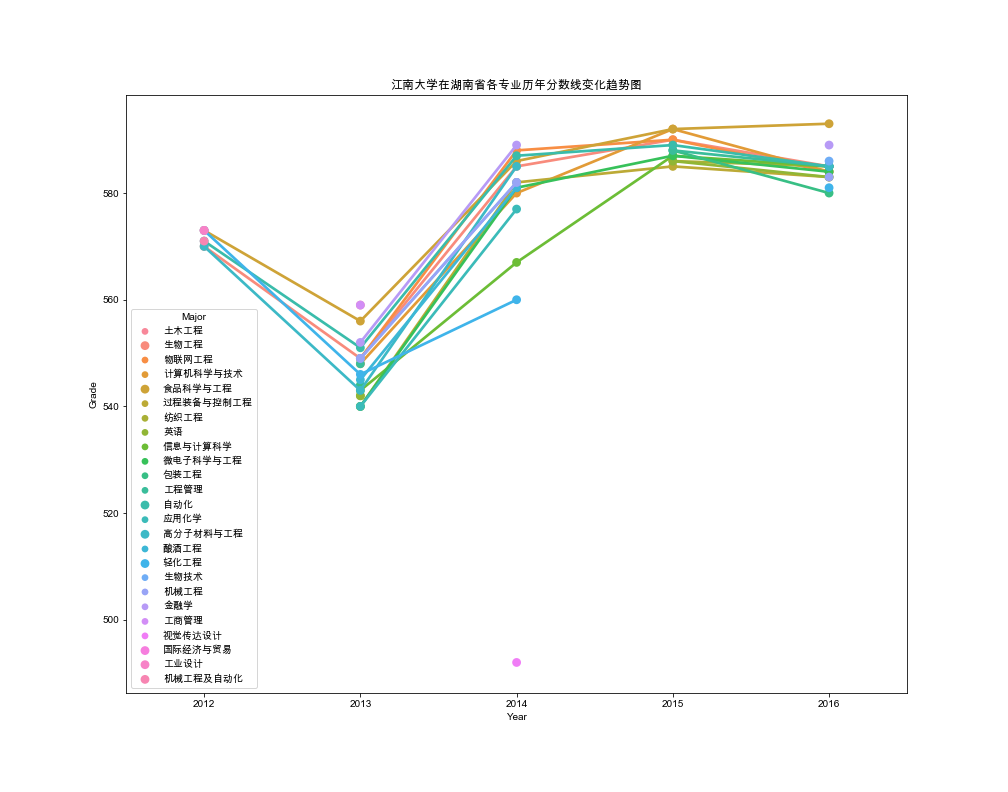
部分数据：



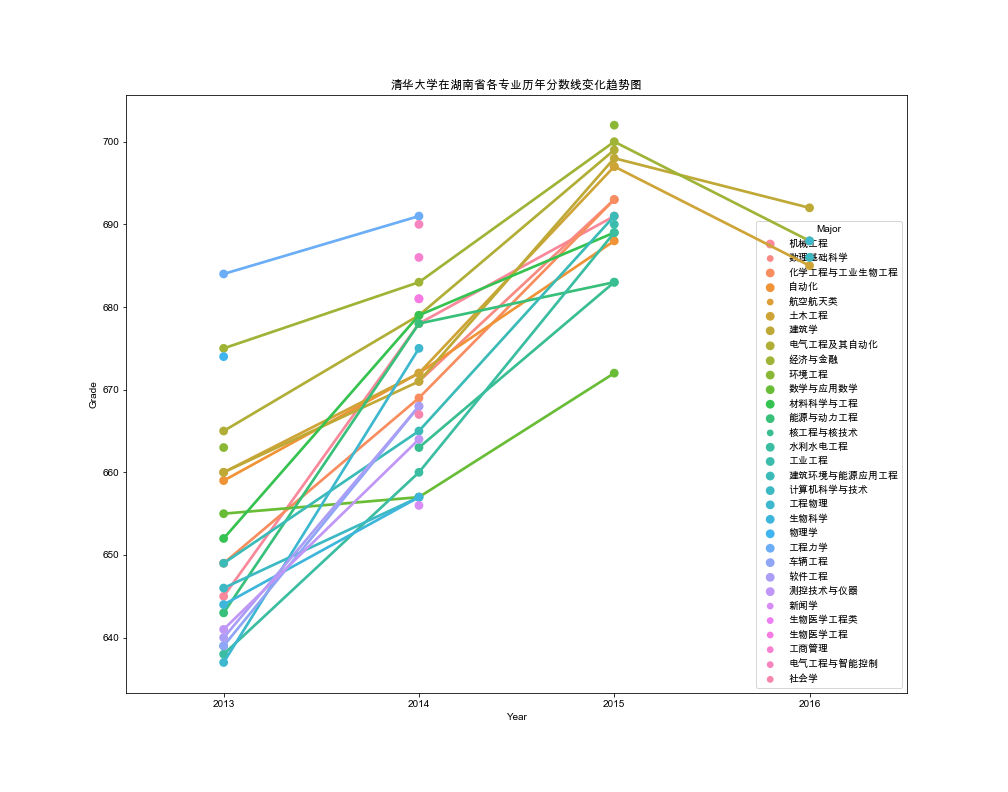
1. 爬虫代码

见附件或<https://paste.ubuntu.com/p/sNx2VTbvY9/>

1. **数据分析和文本分析**
2. 数据分析，将爬回的高考分数线数据利用seaborn画图包进行了构建了历年各大学对各省份的录取平均分数线变化趋势。如下图：江南大学在湖南各专业的历年分数线变化



清华大学在湖南各专业历年的分数线变化：



2、文本分析

将所有大学的所开设的理科可报考的专业都统计出来，并利用Wordcloud 生成了词云图。



从上图我们可以发现计算机科学与技术专业和土木工程开设院校较多，一定程度上也证明了这两个专业的受欢迎程度。

1. 代码

见附件或<https://paste.ubuntu.com/p/J5PVzwQYNG/>