**题目： 大数据告诉你为什么要逃离北上广**

**小组成员**：董文伟 软件工程

郑宝英 语言学及应用语言学

匡燕芳 语言学及应用语言学

**主要内容及技术路线：**

1. 数据来源：<http://www.tianqihoubao.com/lishi/beijing.html>

获取方法：通过python的scrapy包对网站按广度优先进行爬取。

1. 初步分析：爬取了北京2011-2017年每天的天气数据，经过初步分析2015和2016年雾霾比较严重，又以2016年为例，对全国的省会天气数据进行爬取，来用可视化展示全国雾霾严重的几个城市。
2. 统计与展示：用正则对网页结构进行解析，获取有用信息，首先将对北京2011-2016年雾霾天数进行统计，来观察这几年北京雾霾天数的变化，通过折线图展示，然后以2016年为例观察北京雾霾各月的变化，然后以2016年为例，来对全国省会雾霾的严重度在地图上标明，以及通过分析与雾霾相关联词的词频，来做一个词云，结合以上再根据近几年的国家治理雾霾办法来分析雾霾的成因. 通过flask+echarts在网页上进行展示。

**可能遇到的问题：**

1. 爬虫可能会不断爬取，得到太多冗余信息，不容易找到想要的，解决方法，设置爬虫爬取的深度，并且在爬取时，限制域名. 提取数据时。
2. 爬下来的数据网页结构不统一，造成用正则提取重要信息时有难度，解决办法：对多种不同的数据格式进行分析，分别写对应的正则，提取尽可能对的有用信息。
3. 数据显示可能会遇到一些技术问题，解决办法：查阅相关文档，来学习api的调用格式。

**进度安排：**

12.10——12.15：数据处理；

12.16——12.21：对所得数据进行统计与分析；

12.22——12.28：数据可视化；

12.28——1.6：搭建可视化网站，整理所有相关文档，完成项目报告。