## 计算机网络数据爬取作业方案

为进一步结合网络课程理论学习内容，进行网络实践操作，现公布本课程的数据爬取作业方案：

1. 每人自行**调研并选取相关的数据网站**，完成爬虫设计、编写爬虫脚本和相关文档，给出运行测试结果（具体的爬虫编写说明见**附件1**，爬虫编写样例见**附件2**）。

注：附件2给出的样例为编写下限，仅供参考和入门学习。

1. 计划安排：
   1. **需求阶段**：在腾讯文档“计算机网络数据爬取作业-设计文档”上申报自己的爬虫设计方案。

（链接：https://docs.qq.com/sheet/DV1JvRERKaHFDUkdQ?tab=BB08J2）。第一行为填写样例，每个人选择爬取的**网页（URL）不能重复**，可为同一网站的不同网页，先填先得，不要改别人填写的内容。

注：助教每日会在腾讯文档最后一栏**给出可行性评估结果**，各位同学自行查阅，评估结果为“可行”的同学可继续进入下一阶段的设计开发工作。

最晚提交填写时间：2020年11月28日（周六）晚21：00。

* 1. **设计开发阶段**：待评估结果为“可行”之后，遵循**爬虫编写规范**编写爬虫，并爬取和清洗数据，以相应的存储格式存储数据。

1. 使用python语言，推荐Scrapy框架（简单易上手），选择会定期更新的数据（即处于动态持续爬数据的状态）。
2. 爬取的数据范围：
   1. 全国各主要城市的交管局、公交公司、地铁、交通委和交通厅的官网，寻找相关动态数据。
   2. 全国各主要机场、航空公司的关于航班的动态信息，延误、天气等事件信息。
   3. 其他**交通相关**的主题。
3. 作业提交：
   1. 交付物：从爬虫项目根目录打包的zip压缩包、爬虫说明文档（包含爬虫使用说明、数据介绍、爬取结果截图等，文档旨在体现爬虫的可用性）。
   2. 提交方式与命名规范：从爬虫项目**根目录**打包的**zip**格式压缩包命名为“学号\_姓名\_爬虫项目”、爬虫说明文档命名为“学号\_姓名\_爬虫说明文档”。将**项目和文档**再打成一个zip压缩包（命名为“学号\_姓名\_爬虫作业”）发送至**各班学委**处。各班学委收齐后发送至邮箱：[15801189928@163.com](mailto:15801189928@163.com)。
   3. 评分标准：分为优、良、中三档进行评分，根据数据质量、爬虫编写规范性、运行情况、数据清洗工作等标准给分。请严格按照交付物命名规范命名，不符合命名规范的最高评级为良。
   4. 邮件提交截止时间：2020年12月11日（周五）晚21：00。