数据采集和分析课程考核

## 一、课程考核内容

考核总分100分，考核分为两部分：(1)数据采集（50分）；(2)数据分析（50分）。撰写《课程考核报告》，提交能正常运行的源代码。

1. 数据采集

根据某一应用场景需求，采用Scrapy框架设计一个爬虫系统。

1. 数据分析

根据某一数据统计特征，抽取其中有价值的统计指标或挖掘一些隐含信息或预测发现一些新现象。

## 二、考核要求

1. 总要求

（1）、杜绝抄袭，若发现抄袭，比如直接网络下载或下载后改动极少，少于20%，没有体现个人工作量，评定为0分。

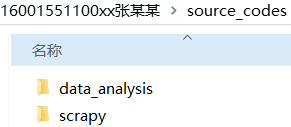
（2）、同一班级或不同班级间，一经发现两份高度雷同（相似度80%以上），评定0分。

（3）、不能直接采用书上或教师上课提供了的源代码，作为考核内容可以在这基础上改动，要求改动工作量在50%以上，体现设计工作量。

（4）、《课程考核报告》撰写规范，按照大家毕业设计任务撰写规范性标准提交。

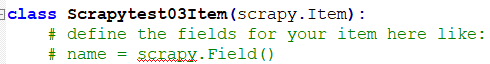
（5）、提交文档命名必须按提交模板来写，替换必要的实际内容即可（标注红色部分）

（6）、源码和数据放到已经建好的对应目录，不能随意放置。



（7）、提交源码必须能重现《课程考核报告》中所写的内容，第三方组件必须在Readme.md文件中说明清楚。

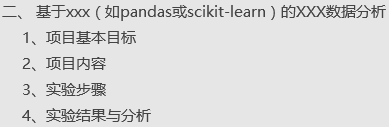
1. 数据采集
2. 必须基于Scrapy平台实现数据爬取，不能是其它，否则不给分。
3. 必须做items.py设计



1. 必做pipeline.py中的 process\_item函数



1. Spider设计中要做必要的爬去内容分解，用json或response.xpath等解析器。
2. 数据分析
3. 项目中必须使用Matplotlib实现绘图，项目中必须有数据绘图；
4. 项目中必须使用Pandas或Scikit-learn可以二选一。
5. 项目必须连贯，确保是一个整体，有一定的实际应用价值。



1. Readme.md文件中写清楚环境配置和重现实验结果的代码。

## 三、各环节评分比例

1、文档规范和内容充实，40%；

2、代码完整、能重现《课程考核报告》的内容50%；

3、资料按要求认真规范提交，10%。

## 四、提交方式和时间

1、提交压缩目录文件给学习委员



2、提交截至时间

本学期第10周星期五（2019年11月8号）