# 《面向数据科学的编程语言》结业考核大作业及评分要求

2024年春季学期 网安、大数据、计科2022级

本次期末考试为数据获取分析任务，考察学生应用PYTHON基础知识进行数据获取、预处理以及综合分析的能力。学生在规定时间内完成任务，并在截止时间前在指定平台（学习通）提交总结报告、源代码（含基本注释）和程序运行结果视频(不超过 2 分钟，只需要关键帧部分)。

**一、评分标准**

1. 提交材料包括实验报告、源代码和运行视频（时长2分钟以内），若未提交源代码，扣10分，未提交运行视频，扣10分；
2. 题目1和题目2任选其一。每一题目分值均为100分。请按照每一题的考核要求，综合本课程的知识点进行作答；
3. 对题目要求步骤有缺失或者错误，酌情扣分；
4. 未对代码和运行结果未解释，酌情扣分；
5. 大报告排版格式不佳、缺少代码、缺少代码运行视频，也应酌情扣分；
6. 程序界面截图和运行截图，有缺失或错误扣10分，描述不完整，酌情扣分；
7. 报告总结，有缺失或错误扣10分，描述不完整，酌情扣分。

**二、大作业要求**

**题目1：亚马逊手机商品文本对比分析**

爬取亚马逊指定手机商品的评论内容并进行文本分析（至少爬取前 3000 条评论，不足3000条，则爬取全部内容，不得使用selenium）。

一）任务链接地址

1. 苹果 iPhone 11 Pro,64GB,金色:

[https://www.amazon.com/-/zh/dp/B08BHJ5XH9/ref=sr\_1\_7?crid=3PR2HOFK5T4AB&keywords=Mobile%2Bphone&qid=1680827474&refinements=p\_89%3AApple&rnid=2528832011&s=wireless&sprefix=%E6%89%8B%E6%9C%BA%2Caps%2C445&sr=1-7&th=1#customerReviews](https://www.amazon.com/-/zh/dp/B08BHJ5XH9/ref=sr_1_7?crid=3PR2HOFK5T4AB&keywords=Mobile%2Bphone&qid=1680827474&refinements=p_89%3AApple&rnid=2528832011&s=wireless&sprefix=%E6%89%8B%E6%9C%BA%2Caps%2C445&sr=1-7&th=1" \l "customerReviews)

1. 三星 Galaxy S21 FE 5G 手机

[https://www.amazon.com/-/zh/dp/B09BFTMQH9/ref=sr\_1\_4?crid=3PR2HOFK5T4AB&keywords=Mobile+phone&qid=1680827671&refinements=p\_89%3ASAMSUNG&rnid=2528832011&s=wireless&sprefix=%E6%89%8B%E6%9C%BA%2Caps%2C445&sr=1-4#customerReviews](https://www.amazon.com/-/zh/dp/B09BFTMQH9/ref=sr_1_4?crid=3PR2HOFK5T4AB&keywords=Mobile+phone&qid=1680827671&refinements=p_89%3ASAMSUNG&rnid=2528832011&s=wireless&sprefix=%E6%89%8B%E6%9C%BA%2Caps%2C445&sr=1-4" \l "customerReviews)

1. 小米红米 Note 11 4G 全球版

<https://www.amazon.com/-/zh/dp/B09QSB4N2C/ref=sr_1_13?crid=3PR2HOFK5T4AB&keywords=Mobile%2Bphone&qid=1680827758&refinements=p_89%3AXiaomi&rnid=2528832011&s=wireless&sprefix=%E6%89%8B%E6%9C%BA%2Caps%2C445&sr=1-13&th=1>

1. 诺基亚 225

<https://www.amazon.com/-/zh/dp/B08Q1YWQQ9/ref=sr_1_2?crid=3PR2HOFK5T4AB&keywords=Mobile%2Bphone&qid=1680828024&refinements=p_89%3ANokia&rnid=2528832011&s=wireless&sprefix=%E6%89%8B%E6%9C%BA%2Caps%2C445&sr=1-2&th=1>

1. TCL 10L

<https://www.amazon.com/-/zh/dp/B0951TM7JS/ref=sr_1_4?crid=3PR2HOFK5T4AB&keywords=Mobile%2Bphone&qid=1680828155&refinements=p_89%3ATCL&rnid=2528832011&s=wireless&sprefix=%E6%89%8B%E6%9C%BA%2Caps%2C445&sr=1-4&th=1>

二）任务要求

1. 学号末位数为 1、2 的同学爬取链接地址a和b的数据
2. 学号末位数为 3、4 的同学爬取链接地址a和c的数据
3. 学号末位数为 5、6 的同学爬取链接地址b和c的数据
4. 学号末位数为 7、8 的同学爬取链接地址c和d的数据
5. 学号末位数为 9、0 的同学爬取链接地址c和e的数据

并按照下文任务实现步骤开展数据分析。

三）任务实现步骤：

步骤1:

分析并爬取手机评论信息: 包括评分、日期、是否已购买、评论标题、评论内容、评论有用人数。



步骤 2:

爬取的内容存入 Excel 文件（每位同学有两个xls文件，文件名称为：学号\_姓名\_手机品牌.xls）；

步骤3 :

分别读取爬取的数据文件数据，提取有效评论（已确定购买，点击评论有用人数大于10），分析评论内容的词性、高频词，画出词云；

步骤 4:

撰写总结报告（内容见总结报告模板），根据获取的高频词信息，对相应手机给出简短评价，然后对比分析两款手机，选出自己觉得相对更好的手机，并给出选择原因。

附加任务（非必须完成任务，仅针对有能力完成的同学）: 对获取的评论内容做文本情感分析，可以是命名实体识别，也可以是文本分类等等。

**题目2：超市顾客分类**

对Mall Customers 的200\*4数据集提供的信息，如下图，使用K-means算法对顾客进行分类。

表格

描述已自动生成

注意：

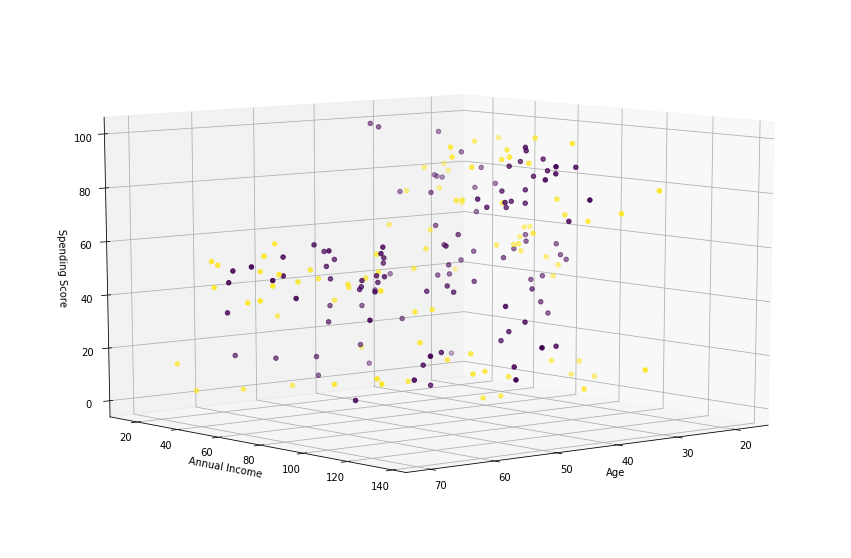
学号最后一位数为1、2、3的同学使用KMeans聚类分析

学号最后一位数为4、5、6的同学使用均值漂移聚类分析

学号最后一位数为7、8、9、0的同学使用DBSCAN聚类分析

具体步骤如下，以使用KMeans为例：

步骤1：把csv转换为dataframe格式，把所有样本通过3D方式展示，如下图:



步骤2：

通过K-means聚类clustering：

(1)使用from sklearn.cluster import KMeans 实现。

(2)不使用sklearn等机器学习包，自定义函数实现。

(3)用可视化包或库(如，matplotlib )显示分类的结果，如图所示：

图表, 散点图

描述已自动生成

步骤3

撰写总结报告，包含代码基本注释、实验流程，简短总结。