课程考察方法

1. 形式要求

1. 请使用`研究生课程作业答题本电子版.doc`作为文件模板（打印封面、评分页即可，作业内容用空白纸打印附后）。
2. 提交**纸本作业、作业（pdf）及所处理语料`样本`电子版**。
3. **2026.01.01**前：纸本和电子版交***李梦茹***。

2. 作业内容

选择一

> **人文学院、外语学院同学**

1. 基础工作

1. 完成`康熙字典.txt`中指定内容（见任务表，或全文）的数据处理工作，构建`康熙字典数据表`，其结构如下：

字头	篇	集	集部	部	字	正文
一	正集/补遗/备考/考证	子	上	一	一	〔古文〕弋【唐韻】【韻會】於悉切.....

> 参考原文
> 一　【子集上】【一字部】　一　〔古文〕弋【唐韻】【韻會】於悉切【集韻】【正韻】益悉切

2. [可选] 构建`康熙字典关系数据库`：使用Access、SQLite、MySQL等数据库软件（引擎）构建包括`篇`、`集`、`部`、`字`在内的4张子表及表间关系的关系数据库。
3. 根据需求设计数据库（数据表）“应用程序”。如基于Excel构建`康熙字典`字典小工具，使用VBA开发嵌入Word的康熙字典小工具，使用Python、Html+js等技术开发针对不同平台、不同应用场景的小工具或网站。请重点思考词典之外的其他需求（如音韵学需求）和应用方式。

2. 附加题目[必做]

请以自己的数据部分为材料，检索、统计所有的唐韵反切，请列举检索表达式和具体操作步骤。

> 一　【子集上】【一字部】 一　〔古文〕弋【`唐韻`】【韻會】`於悉切`【集韻】【正韻】益悉切，丛漪入聲
>
> 七　【子集上】【一字部】 七　【`唐韻`】`親吉切`【集韻】【韻會】
>
> 丂　【子集上】【一字部】 丂　【`唐韻`】【集韻】丛`苦浩切`，音考。

提取结果：

```
一　　於悉切
七　　親吉切
丐　　苦浩切
```

数据分配表

请按下表从`康熙字典.txt`中提取数据作为个人的作业数据基础：

姓名	院系名称	数据范围
赵婧淼	外国语学院	0-1000行
赵亮	外国语学院	1000-2000行
崔雨晴	外国语学院	2000-3000行
苏佼阳	外国语学院	3000-4000行
许若涵	外国语学院	4000-5000行
张蓉	外国语学院	5000-6000行
李洁琛	外国语学院	6000-7000行
丁雨萱	人文学院	7000-8000行
冉文婷	人文学院	8000-9000行
王嘉雯	人文学院	9000-10000行
张羽雯	人文学院	10000-11000行
李芊荷	人文学院	11000-12000行
王昊	人文学院	12000-13000行
李明柔	人文学院	13000-14000行
周莲	人文学院	14000-15000行
王佳琪	人文学院	15000-16000行
李梦茹	人文学院	16000-17000行
王美文	人文学院	17000-18000行
段奕鑫	人文学院	18000-19000行
杨其铮	人文学院	19000-20000行
周子涵	人文学院	20000-21000行
李景姗	人文学院	21000-22000行
甘丹彤	人文学院	22000-23000行
龚成龙	人文学院	23000-24000行
李碧轩	人文学院	24000-25000行
唐静茹	人文学院	25000-26000行
孙子焜	人文学院	26000-27000行
初同飞	人工智能与自动化学院	27000-28000行

姓名	院系名称	数据范围
张国豪	人工智能与自动化学院	28000-29000行
吴佳乐	能源与动力工程学院	29000-30000行
陈昊	计算机科学与技术学院	30000-31000行
刘黄海	计算机科学与技术学院	31000-32000行
刘弈成	计算机科学与技术学院	32000-33000行
邱启航	计算机科学与技术学院	33000-34000行
张翔	计算机科学与技术学院	34000-35000行
张云开	计算机科学与技术学院	35000-36000行

技术与内容要求

技术要求

- 必须体现技术：`Word高级查找与替换`或`正则表达式`；
- 可选使用技术：**Excel VBA**，`Excel公式、筛选等`，`Access、Sqlite数据库开发`，`各类编程技术`，`网页开发技术`，等等。

内容要求

1. 原始语料分析
2. 工作目的、处理思路和流程总体说明
3. 每步处理方案，如Word检索与替换，说明`查找模式`、`替换模式`，以10条有代表性的数据为例列举处理前后效果

   电子版样本分步骤示例，

   1. 原始数据；
   3. Word（txt）处理完成；
   4. Excel格式化完成；
   5. 数据库（可选）
   6. 最终成品（可选）

选择二：基于大语言模型的国际音标OCR工具设计与开发

> 非人文学院、外语学院同学
>
> 可自由组队完成：最多3人一队，每个人的作业中请详细阐述自己完成的工作内容。

选择合适的技术开发一个国际音标OCR工具

- 无标记图像-文本数据，仅有文本型IPA数据集（`ipa词表.csv`，8万余行，见附件）；
- 基于`ipa词表.csv`中的国际音标词库数据，以常用的国际音标字体（`Times New Roman、IPAPANNEW、`

`CharisSIL`，见附件）渲染、生成图像，对图像进行数据增强（旋转、扭曲、模糊、添加噪点等），得到带标签的训练图像数据集；

- 选择、设计合适的OCR模型；
- 模型训练与评估；
- 提交成品

形式要求

作业文本：

1. 说明工具开发思路、技术路线；
2. 简要说明具体开发过程；
3. 展示模型训练效果评估结果；

电子材料

1. 提交可供验证的成果。

选择三：自选

- 人文学院、外语学院同学：

  根据自己的研究课题需要，选择一个具体的语料处理和语料库建设案例，完成相关处理工作，说明工作目的、思路和流程，提交数据样本（参考选择一）。自选作业需使用技术与选择一相同。

- 非人文学院、外语学院同学

  设计一个你感兴趣的与语言、文字相关的计算机技术（特别是AI）项目，完成项目开发工作。说明项目目的、思路和流程，提交项目成果。

  - 参考题目：
    - 甲骨文等古文字识别
    - 基于 Transformer 的古籍文字识别与版面结构分析
    - 文本到知识图谱的自动构建系统
    - ……