## 期末大作业：三校择业意向分析

1. 抓取北京邮电大学、西安电子科技大学、电子科技大学，自2021年9月1日以来的就业帖子。西电、成电，获取字段包括：招聘主题、用人单位、发布日期、浏览次数。北邮除了这四个字段，再增加1个字段：职位数量。各存储为1个csv文件，名字分别为：BEIYOU\_1、XIDIAN\_1、CHENGDIAN\_1。
2. 编写程序，遍历3个文件，对3个文件的数据做如下处理。
   1. 去掉招聘主题两端的空格等无关字符（如空格、tab、换行等）（有可能有，有可能没有）。
   2. 统一把日期处理为：YYYY-MM-DD的格式。
   3. 统一把访问次数为空的置为0，字符串统一转为整数格式。职位数量如果为空，置为1，统一转为整数。
3. 预处理完成的结果，存入一个Excel文件中，Excel包括4个标签页，分别为：北邮、西电、成电、分类，**或者**分别存入四个新的csv文件BY\_D、XD\_D、CD\_D、CAT\_A。
   1. 西电、成电标签页，表头字段一致，包括：序号（数字序号）、招聘主题、用人单位、发布日期、浏览次数。选中列设定对应列的数据类型：文本、文本、日期（YYYY-MM-DD）、数字（整数）。北邮标签页，前面四个之后，还有第5个字段：职位个数；对应的Excel类型为数字（整数）。
   2. 第4个标签页，表头字段为：序号（数字序号）、招聘主题、雇主类型。其中，招聘主题来自于前三个标签页去重之后的结果，雇主类型由完成作业的人自行设定，如中国移动雇主类型可以设为“电信运营商”，BAT可以设为“互联网平台企业”等，自己总结，自己设定，存入汉字即可。
4. 编写程序，对于前述Excel进行分析处理。分析得到：
   1. 最受北邮学生关注的招聘TOP20、最受西电学生关注的招聘TOP20、最受成电学生关注的招聘TOP20（以浏览次数决定排序）。
   2. 最受北邮学生关注的雇主类型TOP10、最受西电学生关注的雇主类型TOP10、最受成电学生关注的雇主类型TOP10。
   3. 北邮招聘职位总数、招聘职位数量TOP10（简单按帖子主题就行，不需要解析企业名称）、招聘职位所属雇主类型TOP10（即前述的自己建立的分类）。
   4. 最关注ICT行业的招聘主题TOP10（在三个高校都出现，并且按浏览次数、职位个数，两重排序；主要考虑帖子同名，可以不使用用人单位也可以结合，或者自定义规则并说明）。
5. 编写报告，总结本次作业。内容主要包括：
   1. 需求分析。写自己对于本次作业的理解和分析。
   2. 模型设计。介绍自己是如何设计雇主分类，包括哪些分类，每个分类里面包括哪些招聘主题。
   3. 系统设计。介绍自己是如何设计各个程序的。
   4. 数据分析。按照前页要求的指标，详细列出自己分析的最终数据，并配以饼图或柱状图等合适形式（可视化图表可以通过matplotlib这样的可视化工具生成，也可以用Excel制作）。如果自己有前述指标之外的个人观点，也可以在此描述，比如就业形势、行业情况等。
   5. 作业小结。对作业的工作进行小结。

**最后总体要求要言之有物，语言精练，不用凑字数，不要凑字数！（不要超过15页）**

**（截止：1月10日晚上24点）**