**高考志愿智能推荐系统开发需求**

**1. 概述**

目前市场上有很多高考志愿填报推荐系统，考生只需要输入自己所在省份以及自己的分数，系统就会给出推荐的若干所学校，一般会按照“冲一冲、稳一稳、保一保”的策略各给出三到四所学校名单，做得好的系统还会给出推荐的专业，以及可能录取的概率（从低到高排列）。

**2. 实现原理**

这种推荐系统实现的原理比较简单，它并不是根据考生的分数来推荐学校的，因为每年的分数变化很大，并不具有多大的参考意义。它是查当年所在省份（因为中国高考是分省录取的）的“一分一段表”，然后根据考生的分数计算出他当年的省内排名，再根据过去3~5年每所大学在本省录取线换算出来的排名给出推荐学校。

比如今年湖南省理科某考生570分，当他来查询推荐学校时，系统会先查今年湖南省的一分一段表，570分的排名在36000名左右。然后再查过去5年各所学校在湖南招生的最低录取排名的平均值计算出来，比如湖南科技大学是42000名左右，湘潭大学是32000名左右，长沙理工是30000名左右，于是系统按照录取概率的大小，分别将长沙理工、湘潭大学和湖南科技大学归入到“冲一冲”、“稳一稳”和“保一保”的档次。

如果考虑到各个专业录取的最低分数排名不同，还需要根据专业给出推荐。不过目前各个系统做到分专业推荐的还很少，因为专业录取分数线搜集很困难。

**3. 开发需求**

本系统分为两个相对独立的部分：一个是信息采集（爬虫）部分，一个是面向用户的业务系统，两个部分之间无直接调用关系，只是会访问同一个数据库。

**3.1 信息采集需求**

**（1）学校基础信息采集**

这部分包括两类学校：本科院校和高职专科学校。

本科院校采集的信息包括学校的名称、招生代码，学校类别（一本、二本、三本，公办、民办，985、211、双一流）学校排名（至少包括武书连和网大排名，以近5年的为准，计算平均值），学校专业排名（主要是第4轮学科排名、武书连排名）

高职专科学校采集的信息主要是学校名称、招生代码和民间的排名。

**（2）招生数据采集**

这部分包括各省（全国有31个省份）近5年（2016-2020）的招生数据。包括：历年的各科最低控制分数线（理科、文科必须采集，艺考、体育生等可暂缓），一分一段线数据，各个学校在本省招生的最低投档线（能采集到平均分和最高分更好）。如果可能的话，采集专业投档线。以上分数线均要转换成排名。

今年各学校在本省的招生名额（分专业人数和总人数）。

**3.2 业务系统需求**

本系统在采集到的高考数据基础上实现高考志愿推荐功能。采用Java Web技术实现。

**（1）管理员角色功能需求**

系统预置至少一个管理员，拥有以下功能：

* 账户管理功能：登录、修改密码
* 其他用户管理功能：包括查看、禁用、启用普通用户，
* 统计功能：查看、统计用户充值缴费情况
* 数据管理功能：搜索、查看、统计、修改、删除、增加采集到的各类数据
* 配置功能：增、删、改、查各个阈值。这里主要是指推荐策略需要用到的数值，包括推荐的学校数目、排名变动的范围

**（2）普通用户（考生）角色功能需求**

* 注册功能：手机号码注册
* 账户管理功能：登录、修改密码、密码找回
* 充值功能：支付宝或者微信充值（暂时做假的，有界面即可）
* 查看推荐学校：输入自己的分数，查看推荐结果，如不满意，可以换一批
* 高校浏览功能：查看各个高校的基本数据和在本省的招生情况
* 评价功能：对推荐结果进行评价
* 交流功能：类似BBS板块，考生之间可以发帖互相交流

补充：可设计成充值用户拥有更多的推荐功能，例如不充值只能使用一次推荐，或者推荐的学校较少。

**（3）推荐功能需求**

设计合适的且灵活可配置的推荐策略，可以提供多组备选推荐（可加入随机权重），或者按照“保守”、“激进”、“稳重”等选择偏好来进行推荐。

**4. 技术选型**

* 开发和运行的操作系统是centOS 7.x
* 爬虫采用python 3.7实现，推荐使用Beautiful Soup + Xpath，爬虫调度框架推荐使用Scrapy。
* 采集过程中可能会要用到OCR技术，源码我将上传到群共享中。
* 数据库统一使用MySQL 8.0。
* 业务系统采用Java Web技术，JDK 11，前后台分离。前端使用Vue 框架，后端使用Spring Boot，Web容器采用Nginx。前后端通信采用restful接口。
* 使用gitee（码云）管理本项目所有源代码。

**5. 人员安排**

所有还未确定题目的16位同学均需参与此课题开发，不再提供其他课题选择。

（1）大组长一名，负责部署维护服务器、设计数据库、撰写详细的开发接口文档和说明，协调各组之间的开发进度，维护gitee项目。

（2）信息采集组长一名，负责管理爬虫开发事宜，负责调度框架的开发与维护，负责爬虫写入数据库的API开发。所有的爬虫均不能直接访问数据库，需调用数据库接口，需设计好爬虫部分的接口规范。要保证各省数据入库格式的一致性。采集信息数据库的设计要与大组长协商一致。

（3）业务系统组长一名，负责管理业务系统前后端的开发事宜，需要设计好前后端的调用接口；并负责考生角色的后端功能开发。业务数据库的设计要与大组长协商一致。

（4）管理员角色的后端功能开发人员一名。

（5）管理员角色的前端开发人员一名。

（6）考生角色的前端开发人员一名。

（7）推荐功能开发人员一名。

（8）普通本科院校基础信息采集人员一名。

（9）高职专科院校基础信息采集人员一名。

（10）各省招生信息采集人员7名，各负责3~4个省份的数据采集，注意不要重复，优先选择东部和中部省份（即此次暂不采集新疆、内蒙、西藏、宁夏、青海、甘肃信息）。

其他未尽事宜，请在群中留言。