命题赛三：求职者智能分析系统答题指南

第四届全国高校云计算应用创新大赛

杭州华三通信技术有限公司

资料版本：1.0

产品版本：0.4.1

http://www.h3c.com.cn

前言

本指南主要介绍了求职者智能分析系统赛题的情况及答题提纲，用于指导报名该赛题的同学可以按照提纲开发出优秀的作品。

读者对象

本手册主要适用于报名第四届全国高校云计算应用创新大赛“命题赛三：求职者智能分析系统”的人员。此外，读者还需具备一定的大数据知识背景、爬虫、数据挖掘以及 ETL 基础知识。

第四届全国高校云计算应用创新大赛

目录

第四届全国高校云计算应用创新大赛

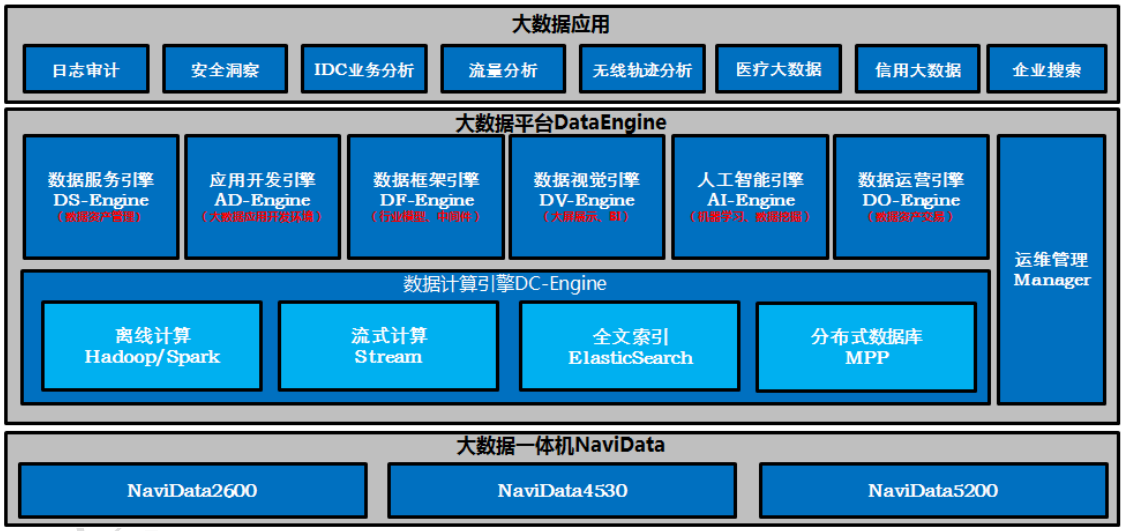
1 概述

1.1 赛题简介

本赛题题名为求职者智能分析系统，目标是完成“求职智能分析系统”的设计、开发、部署工作。通过网络爬虫爬取智联招聘、51job等招聘网站上，大数据相关职位的招聘信息，提取出其中的关键数据，包括但不限于职位名称、职位待遇、职位描述、公司介绍、公司规模、公司性质等信息。通过对这些信息的挖掘分析，可以更加精准、清晰的指导求职者所在行业的待遇水平、自身可能的待遇、以及对公司、行业的选择。

本赛题重点考察参赛选手如下知识的掌握使用情况：1、爬虫工具；2、大数据平台；3、数据ETL工具；4、数据挖掘；5、应用开发；请参赛选手根据赛题要求，使用赛题要求的工具，在规定的时间内高质量完成，达到学以致用的目的。

1.2 H3C产品概述



H3C DataEngine 系列产品

1. 大数据平台产品

- 离线计算

  产品定位：提供企业增强 Hadoop &Spark 版本。
- 流式计算

  产品定位：提供流式计算框架，支持 Storm、Spark Streaming。
- 全文检索

  产品定位：提供企业版 Elastic Search 产品。
- 分布式数据库

  产品定位：提供大规模并行计算数据库 MPP 产品。

2. 大数据中间件产品

- 数据服务引擎
  产品定位：数据全生命周期管理。
- 应用开发引擎
  产品定位：大数据应用开发。
- 数据框架引擎
  产品定位：企业数据仓库，数据模型管理。
- 数据视觉引擎
  产品定位：数据可视化。
- 人工智能引擎
  产品定位：数据挖掘，人工智能。
- 数据运营引擎
  产品定位：城市数据运营。

1.3 产品推荐

1. 本次赛题可能用到的产品主要有：

- 大数据平台产品：提供 Hadoop & Spark 计算框架、NoSQL 数据库 HBase、分布式文件系统 HDFS 等。

- 数据服务引擎：数据分布式 ETL、爬虫工具。
- 人工智能引擎：提供相应算法库。
- 应用开发引擎：模型训练，大数据应用开发。

（备注：用到的产品会提供相应的产品说明文档和使用手册。）

第四届全国高校云计算应用创新大赛

2 答题步骤

2.1 网络数据爬取

1. 目标

解决招聘网站公布的招聘信息的爬取问题。

2. 工具

H3C Datahunter （推荐）或 其他开源爬虫及基于开源爬虫框架编写的爬虫

3. 网站

移动版前程无忧 m.51job.com （推荐） 或 其他招聘网站

4. 格式

企业信息

推荐包含以下信息：

- 企业名称
- 企业形式
- 企业所属行业
- 企业规模
- 企业介绍
- 企业页面地址
- 数据添加时间
- 数据来源网站

职位信息

推荐包含以下信息：

- 职位名称
- 职位所属企业名称
- 工作性质
- 最低月薪
- 最高月薪
- 工作地点
- 职位发布时间
- 学历
- 招聘人数
- 最低工作经验
- 最高工作经验
- 职位类型
- 性别要求
- 最低年龄要求
- 最高年龄要求

第四届全国高校云计算应用创新大赛

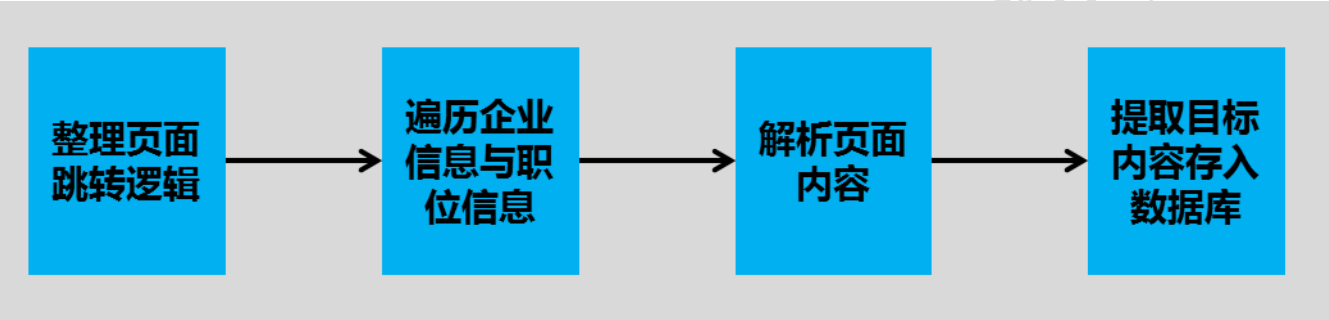
- 语言要求
- 其他要求
- 数据添加时间
- 数据来源网站

5. 存储

MySQL（推荐），或其他数据库

6. 流程概述

爬虫的主体流程应如下图所述，但不限于该流程，最终以采集到的数据量与精细程度评分。



2.2 网络数据ETL

1. 目标

解决爬取下来的数据抽取、转换、加载到大数据平台的问题。

2. 工具

H3C DataEngine DI（推荐），或其他工具

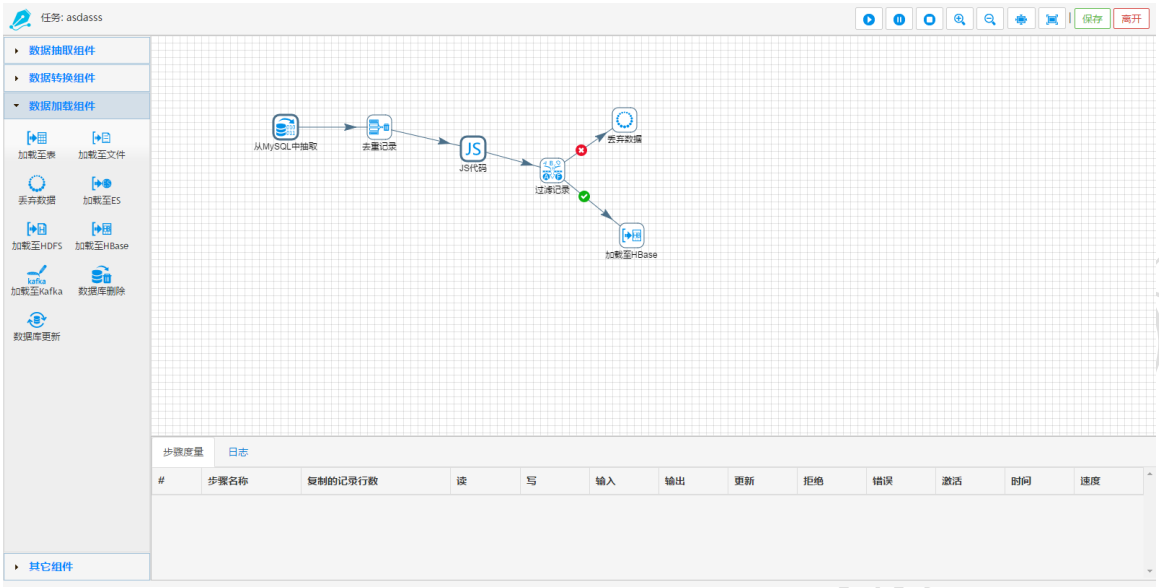
3. 存储

限定使用 HBase

4. 流程概述

使用工具将 MySQL（或其他）中数据进行抽取，在抽取过程中需要去重数据、校验数据的格式等。

第四届全国高校云计算应用创新大赛



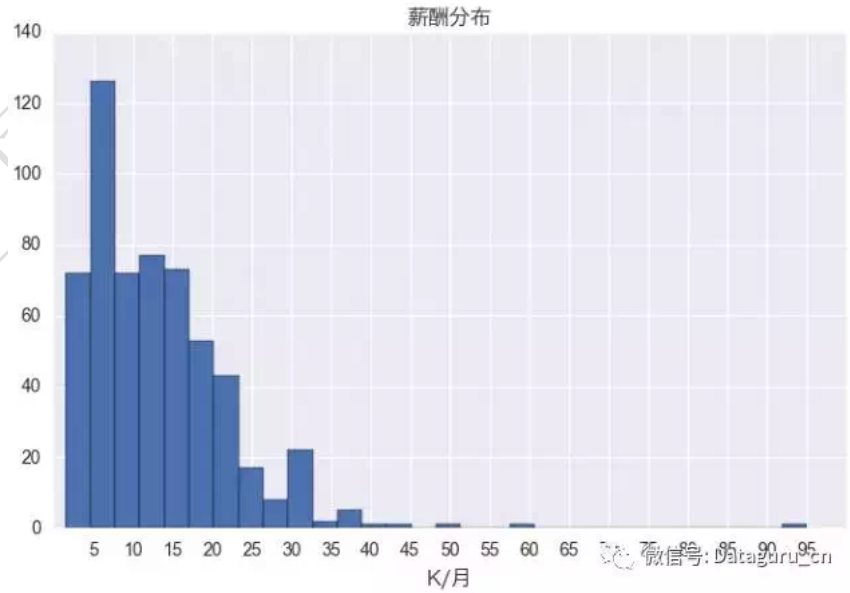
2.3 网络数据分析

1. 目标

从爬取的数据中挖掘出有价值的结果，指导求职者找工作。参赛者结合数据情况，尽可能寻找挖掘点。现提供以下几个参考主题：

- 主题一：大数据职位描述性统计

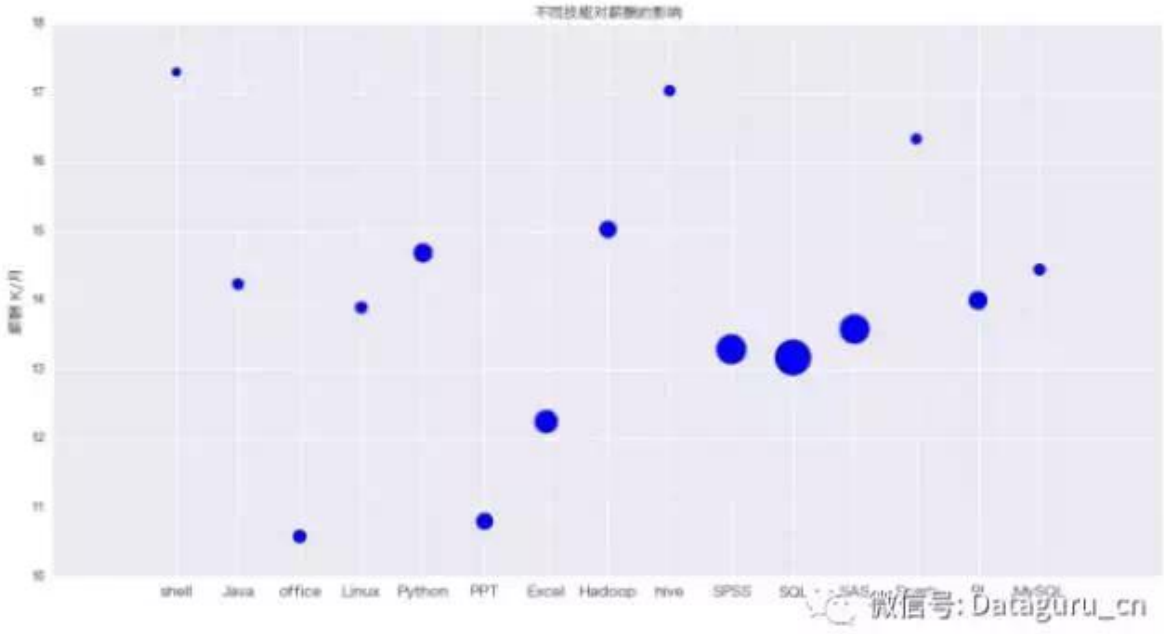
使用相关数据分析软件（jupyter、python），分析大数据职位的地域、薪酬、行业等分布情况。如图：



- 主题二：岗位待遇的影响因素。

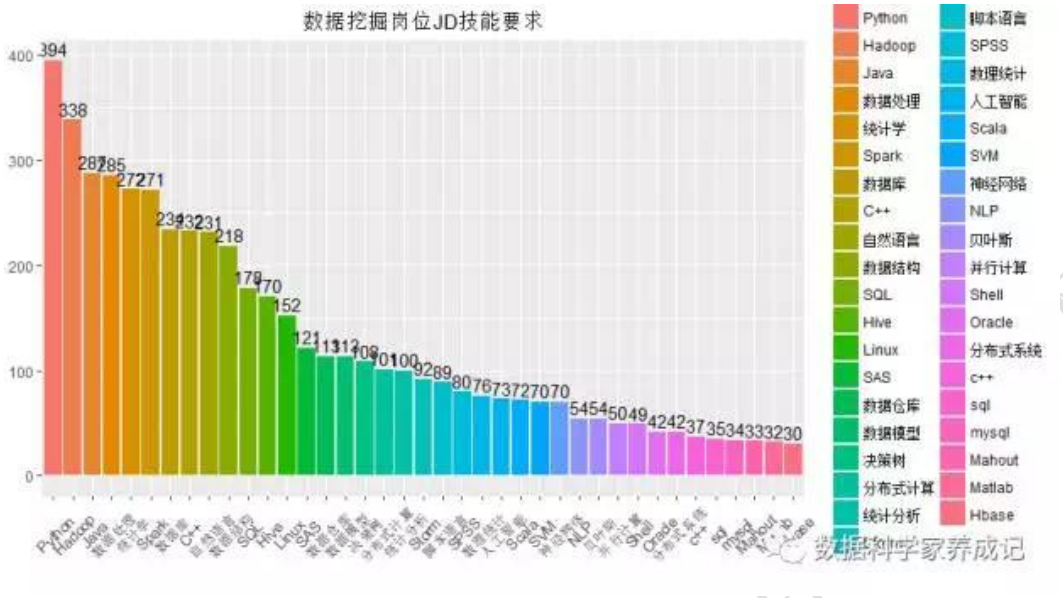
参赛选手可从有工资信息的记录里抽取样本进行研究，提取出有价值的信息。比如，“震惊！英语 6 级能够提高工资 4000 元！”。

选手从样本数据中提取学历、工作年限、地域、性别、行业、技能种类数、项目个数等特征来分析岗位待遇的影响因素。使用逻辑回归、多元回归识别出影响待遇水平的重要因素。如图：



- 主题三：大数据职位技能需求图谱

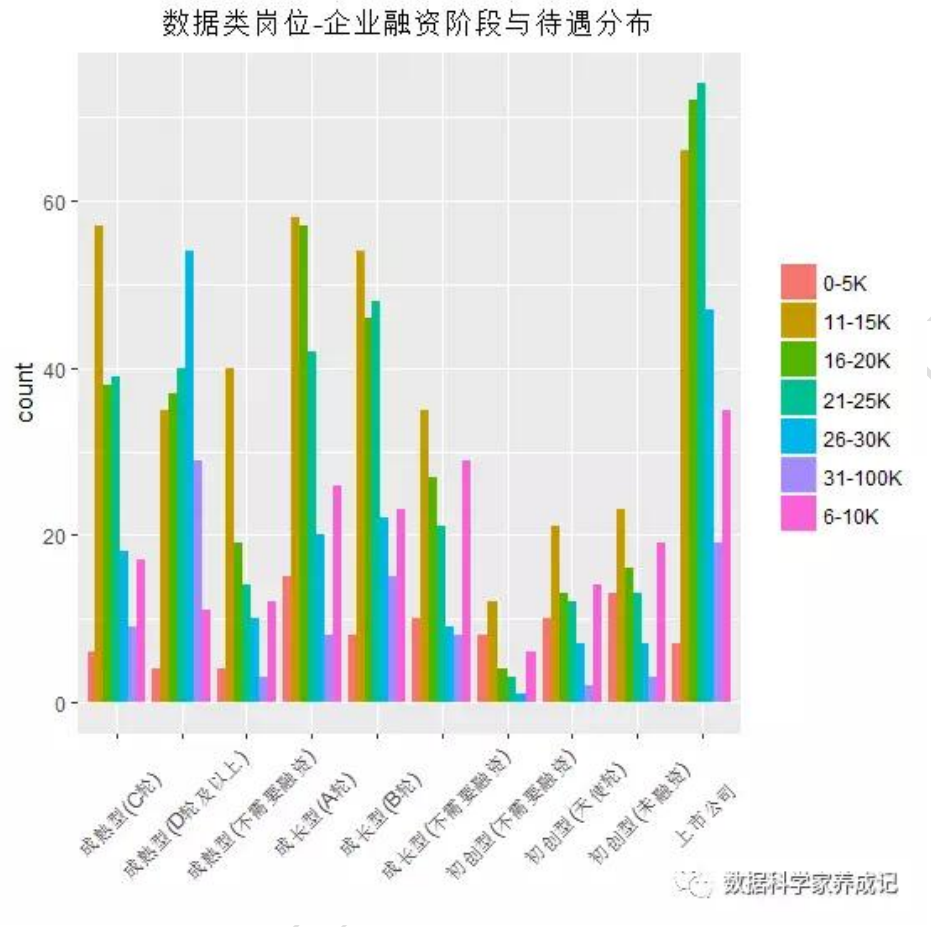
参赛者需要将招聘信息进行解析（jupyter、python）、分词（python）、提取关键字（python）等方式，得到每个大数据职位的需求图谱，并展示词云（python）。如图：



注：为了得到更好的效果，参赛者也可以通过文本信息提取（条件随机场、crf++、深度学习、tensorflow）等相关算法和工具进行命名实体识别，从而提取出更详尽、准确的文本信息。

- 主题四：招聘企业画像

招聘大数据人才的企业有什么特点，比如处于什么发展阶段、什么行业。总而言之，就是给招聘企业贴上各种标签。如图：



最重要的两步：

a.文本数据结构化，提取有意义的信息。比如，技能要求。

b.数据预处理，了解数据的基本统计分布情况，是否有异常值、缺失值，并从中抽取重要特征。

2. 工具

jupyter、python、spark、crf++

3. 相关算法

逻辑回归、概率密度分布、分词、关键词提取、词云、文本信息提取（条件随机场、深度学习，有能力者可以试试）

2.4 智能检索和推荐

1. 目标

解决求职者输入个人信息后，智能推荐出适合自己的职位的问题，以及其它更多元化的信息。如图：

a.分析求职者输入岗位名称、工作地点、行业、公司名称等信息，系统将跟根据这些信息进行初步检索，向使用者展示出合适的职位，这个过程要实现关键词匹配（工具不限）或者词相似性（jupyter、python、spark、tensorflow）等功能；

b..结合用户前期的浏览记录，向其推荐更多的相关职位。由于用户信息偏少，这里的推荐算法建议使用基于内容的推荐。

c.基于内容的推荐算法（spark）需要计算招聘信息的相似性。选手可从招聘信息中提取重要特征，比如词向量特征（python）、词频特征（python），实现招聘信息的相似性计算。

d.根据计算出的相似信息列表和主索引，对用户进行相关职位推荐。同时展示需求 2 关于大数据职位的分析挖掘结果。

2. 工具

jupyter、python、spark、tensorflow

3. 相关算法：

基于内容的推荐、词相似性（词嵌入、word2vec）、文本相似性(余弦相似性)

注：该部分提到工具和方法仅供参考，选手可根据自己的习惯自由选择。

2.5 应用开发和部署

1. 目标

开发可与用户进行交互的推荐系统，有快速和精准的推荐和良好的用户体验效果。如图：

1) 提供简洁的交互式界面供用户录入个人信息包括：原从事行业、原职位、岗位、学历、地区、年龄、性别以及预期待遇，有详细校验和友善提示信息；
2) b..将用户录入的个人信息调用步骤 4 中建立的数据模型进行相关性匹配，匹配最接近用户需求的招聘信息，要求调用数据模型能够及时响应；。
3) 根据数据模型相关性分析，将相关度最高的前 Top10 进行列表展示，招聘信息展示要求包含：行业、地区、公司名称、招聘需求、待遇信息。
4) 开发的应用进行实地部署，可通过浏览器进行访问

第四届全国高校云计算应用创新大赛

3 备注说明

1. 参赛队伍要求

面向但不仅限于各高校计算机学院、微电子学院、数字艺术学院等对云计算、人工智能和其开发感兴趣的选手。在云计算大赛官网（https://cloud.seu.edu.cn）完成注册之后，报名时选择“命题赛三”即可完成报名。按照大赛要求，每个队伍需由 1 名指导老师带队，至少包含 1 名队长和 3 名队员。本题推荐 3-4 人组队完成。

2. 作品评审周期

命题赛三时间节点在大赛整体流程框架内有如下细分：

a) 初赛：2017 年 9 月 1 日～2017 年 11 月 30 日
提交《项目设计方案》（包括作品内容（功能描述、页面设计），技术路线（所用框架，模块设计，语言，开发环境等）。

b) 复赛：2017 年 12 月 15 日～2018 年 2 月
原型系统开发，提交可运行的系统，并根据每个晋级队伍的特点提出进一步的指导和完善意见。

c) 决赛：2018 年 3 月-4 月
进一步完善作品后决赛展示。

3. 环境说明

本次赛题涉及到华三系列产品，除了爬虫工具可以自由选择外，其他产品如：大数据平台，应用开发平台，数据ETL平台必须选用华三产品。

4. 产品配套材料

H3C 大数据引擎，H3C Data Engine 介绍：
http://www.h3c.com/cn/Products___Technology/Products/Big_Data/Catalog/DataEngine/DataEngine/

本赛题提供软件的安装包、技术手册和技术支持，请在报名后联系本命题联系人索取相关资料。

5. 命题联系人

杨东东 yangdongdong.09262@h3c.com 0571-86762656

第四届全国高校云计算应用创新大赛