**大数据分析与挖掘**

**（该命题有两个方向，分别有北京以利天诚科技有限公司和山东交通学院担任命题单位）**

方向一 基于互联网数据的大数据分析

**命题教师：崔云飞 韩帅**

**欢迎加入该命题的QQ群：699204653**

**命题企业：北京以利天诚科技有限公司**

1. 项目背景

信息科技经过 60 余年的发展，已经从高端科技普及到社会运转的每一个角落中。随着信息技术在国家治理、经济运行的方方面面的运转，大量的数据随之产生。而互联网技术的爆发式发展使得近两年来产生的数据总量超过了人类有史以来的数据总和。如何利用这些数据，挖掘数据的深层价值，是未来信息科技发展的趋势，也是大数据技术产生的背景。

本赛项的设置，是为提高大数据专业的学生的综合实践能力，适应大数据开发应用产业对高素质技术技能型人才的需求，重点检验学生对数据获取，数据处理，数据分析，数据可视化等操作流程的实际动手能力。激发学生对大数据知识和技术的学习兴趣，提升学生能力和素质，努力为中国大数据应用产业发展储备及输送新鲜血液。

通过本大赛，提升高校电子信息类及相关专业毕业生能力素质，满足社会用人需求，实现行业资源、企事业资源与教学资源的有机融合，使高校在专业建设、课程建设、人才培养方案和人才培养模式等方面，跟踪社会发展的最新需要，缩小人才培养与行业需求差距，促进专业教学建设和教学改革。

希望通过本次大赛，实现以下实践意义：

1. 可以极大的提高学生的大数据挖掘、分析实践经验，有助于大学生对大数据挖掘、分析工具、语言的熟练应用，有助于学校学习的理论知识和实践相结合。
2. 掌握大数据分析的方法和思路，切实了解社会对大数据技能的需求。
3. 可以让学生在校学习时了解企业对人才需求的动向，综合开率专业，知识技能，收入，地域等因素。既进行操作实践，又可以对未来的学习有所规划。
4. 可以对各个高校在设置专业的时候了解企业对人才的要求，对学生人才培养方向上有比较实际的参考价值。

二. 项目要求

* **项目介绍**

本次项目的的主要目的是通过编写网络爬虫程序爬取互联网中2018年与计算机相关专业的招聘信息，以及收集各专业学科学习课程内容。通过数据清洗，转化，可用信息提取等技术手段获得可用的原始数据，并进行可视化的展示。利用数据分析技术，把项目要求的最终结果通过可视化的手段展现出来。其中抓取网站包括不限于：智联招聘、前程无忧、拉勾网、齐鲁人才网等等、在数据源收集上学生可以自由发挥，不违反相关法律法规即可，数量和质量都是最后考核的标准之一。数据的处理和分析流程，可以通过代码完成，也可以通过封装、可视化操作等进行优化，作为评分标准的内容。

以下是本项目的数据处理流程图：



* **开发学习准备阶段**

根据项目内容，制定需求方案，技术方案。进行自主或其他形式的学习培训，有针对性有目标的进行大数据技能的学习准备。

学习阶段建议使用“以利天诚大数据教学实训平台”。

* **原始数据准备**

可使用自己熟悉的语言编写网络爬虫，也可以使用网上的开源或商业工具对互联网招聘网站进行抓取，抓取网站可选择知名度较高的招聘网站，抓取的数据量要足够多。具体抓取信息内容可根据最终分析内容进行选取。

此阶段要求完成计算机相关专业的就业前景分析的原始数据准备工作即可，若能保证易用性，通用性可成为加分项。例：以客户端或B/S架构等形式提供，可自由输入目标网站和目标数据，使爬虫程序可根据现实需求快速改变原始数据的内容。

* **数据存储**

上一步所抓取下来的数据，可存储于Apache Hadoop平台的HDFS中，为后续的数据处理和可视化做准备，对存储的类型不做限制，也可以用其他的平台和容器存储。

* **数据处理**

数据处理手段可自主选择；

通过对原始数据半结构化数据的清洗、转换和汇总形成结构化数据后可以为后续数据分析提供直接可用的数据源。

本阶段的项目目标是为后续数据分析提供直接可用的资源即可。若满足以下几点优化，可作为加分项：

大数据平台性能调优，可利用调优前后的数据对比形成分析分档，说明调优目的，调优方法。

易用性和通用性：以客户端或B/S架构等形式提供操作界面，可对目标数据进行相应的可选择处理操作，最终在制定的位置或目标存储系统中生成优质数据。

* **数据分析**

数据分析的技术手段可自主选择；

通过对原始半结构化数据的清洗、转换和汇总形成结构化数据后，可实现以下分析功能（举例）：

* 计算机专业薪水最高的前10名招聘职位
* 大数据职位需求量最高的前10名城市
* 大数据职位需求量最高的前10大行业（如互联网、金融、电子商务等）
* 计算机相关专业技能需求前10名。
* 企业对哪类大数据人才需求最为迫切（大数据分析师、大数据架构师等等）

本阶段项目目标为根据处理后的数据进行可用价值的提取和分析。若实现以下功能为加分项：

录入学生所学专业课程，可以得到该学生的就业前景和就业分析。如适合岗位前5名排行，在各地区能拿到的薪水，有多少家企业可供选择；

学生学业成绩和就业前景的相关性分析，如某一或几门课程取得的成绩可以在未来的就业中取得多少薪水的概率。若无法取得相关成绩与薪水的历史数据，可利用企业开出薪水和要求技能掌握程度进行预测。

易用性和通用性：以客户端或B/S架构等形式提供操作界面，可对目标数据进行相应的指定统计和分析操作。

* **数据可视化**

对以上分析结果进行可视化展示，可自主编写程序或使用开源的可视化工具来做展示。

**作品提交规范**

每个参赛队的作品必须以压缩包形式上交，压缩包内需包含以下内容：

在目录下建立四个文件夹，分别命名为“源程序”、“数据”、“文档”和“演示”，必要时自己根据情况建立二级文件夹并清晰命名。

1. 将作品的源程序放入“源程序”文件夹中；
2. 将作品的数据放入“数据”文件夹中，训练数据与测试数据分开，还应包括必要的数据导入脚本；
3. 文档需包括分析报告与系统说明文档，都放到“文档”文件夹中。分析报告采用论文的形式说明系统的背景、意义、数据处理过程、数据分析与挖掘方法、结论等内容。系统说明文档应包括系统的运行环境、使用方法、测试方法、数据导入方法，应能够根据文档在全新环境中还原系统。

**四、作品上交时间：请关注大赛网站通知**[www.softqilu.com](http://www.softqilu.com) **及QQ群通知。**

**五、评比方式**

* 1. 提交的材料完整、齐全、可读性好。如果所提交的文档不完整，仍可以参加评比，但要适当扣分。
  2. 评比首先要看程序是否能完全还原，并正常运行，即运行中出现的错误不能使程序退出、陷入死循环或造成死机。如果出现上述情况，可以重新运行，但总尝试次数不能超过三次。
  3. 作品展示效果占比分的20%，功能及数据分析为80%。在正常运行基础上，以功能评比为主，在功能相近的情况下，考察程序的其他特色，重点看数据分析的思维，以鼓励学生运用大数据分析技术解决实际问题。
  4. 综合作品、文档、数据等多方面得出最终评比结果

**六、参赛注意事项**

* 多关注大赛网站 <http://www.softqilu.com>
* 多关注大赛论坛
* 加入本命题QQ群
* 加入相关专业开发群
* 参加巡回报告会，与命题老师和往届获奖学生面对面交流
* 参加暑期训练营，与命题老师和其他参赛同学直接深入交流
* 多与往届获奖选手交流，知道一等奖应该具备什么水平，是如何炼就的
* 主动与指导老师交流
* 组建一支最有战斗力最合适的团队
* 与相关专业结合
* 多买些专业书籍看
* 多上网看一些高水平论文
* 多上网查询一些相关资料
* 认真审题，分析得分点
* 注意细节，细节决定成败
* 注意参赛的各个时间点和参赛的所有信息

**七、参考资料**

Hadoop官网<http://hadoop.apache.org/>

Spark官网<http://spark.apache.org/>

R语言官网<https://www.r-project.org/>

WEKA官网<http://www.cs.waikato.ac.nz/ml/weka/>

**方向二：工业与交通大数据挖掘**

**命题单位：山东交通学院 命题负责人: 司冠南 qq群号：427674697**

**一、项目介绍**

“大数据”一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资产。

大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换而言之，如果把大数据比作一种产业，那么这种产业实现盈利的关键，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台的计算机进行处理，必须采用分布式架构。它的特色在于对海量数据进行分布式数据挖掘。但它必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术。

随着云时代的来临，大数据也吸引了越来越多的关注。大数据需要特殊的技术，以有效地处理大量的容忍经过时间内的数据。适用于大数据的技术，包括大规模并行处理数据库、数据挖掘、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统。

**二、项目要求**

出题背景：近几年来，随着计算机和信息技术的迅猛发展和普及应用，行业应用系统的规模迅速扩大，所产生的数据呈爆炸性增长，因此，寻求有效的大数据处理技术、方法和手段已经成为现实世界的迫切需求。大数据分析挖掘将能为企业带来巨大的商业价值，实现各种高附加值的增值服务，进一步提升企业的经济效益和社会效益。在未来，一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分，对数据的占有、控制和运用也将成为国家间和企业间新的争夺焦点。

每团队3-5人，请使用目前流行的大数据框架与分析挖掘工具，如：Hadoop、Spark、R语言等，完成一个实用的软件系统，为行业、企业或个人生活提供有益的支持。选题方向可结合以下方面：

（1）工业系统数据分析：通过分析真实的工业应用系统数据，为该系统的功能或性能优化提出可行的建议。

（2）网页文本信息的获取与处理：通过对真实网页数据的爬取与搜集，为政府、企业或个人提供有价值的信息，如：舆情信息、某一方面的流行趋势、热门话题等信息。

（3）公共交通数据的分析与挖掘：通过分析真实的公共交通流量数据，如：道路交通数据、铁路客运信息、民航客票信息等，提出线路优化或个人出行建议。

系统要求如下：

1. 系统采用Hadoop、Spark等大数据平台构建；
2. 使用或改进常用的数据分析与挖掘方法完成数据分析功能，可使用WEKA、Python、R等工具或语言提供的接口，或使用SAP水晶易表进行可视化分析；
3. 系统完成后需提供训练数据与测试数据，以检验系统的有效性；
4. 系统数据来源需真实，重要数据可进行脱敏处理，不可自行生成随机数据，数据来源的真实性需可验证，数据量不少于100万条；
5. 根据数据分析与挖掘结果给出报告。