任务6-2 大数据技术

《信息技术基础》

01

任务描述

将实际工作中的信息处理任务设计为相应的课堂学习行为，形成以培养信息处理能力为核心、学习能力和社会能力为两翼的课堂教学任务。

任务6-2 大数据技术

大数据时代已经来临，生活在信息时代，每个人都要能正确认识大数据，并掌握常用的大数据技术，对数据进行专业化处理，实现大数据的价值增长。本任务将为大家介绍大数据的诞生、常用技术及典型应用。

02

技术分析

梳理分析实现操作任务中需要掌握的知识点、技能点，明确学习目标，实现方法，确定学习重难点。

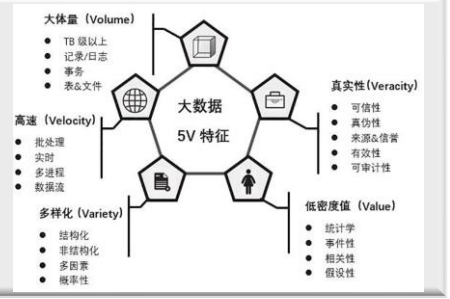
一、知识点

大数据的诞生

2008年9月，美国《自然》杂志，正式提出“大数据”概念。
2009年，大数据正式成为互联网技术行业的热门词汇。

大数据的特征

1、数据量大（Volume）
2、输入和处理速度快（Velocity）
3、数据多样性（Variety）
4、价值密度低（Value）
5、真实性（Veracity）



大数据相关技术

1、大数据采集
2、大数据预处理
3、大数据存储及管理
4、大数据挖掘和分析
5、数据可视化
6、大数据的相关软件平台

大数据的影响

1、大数据对科学研究的影响
2、大数据对思维方式的影响
3.大数据对社会发展的影响

二、技能点

NLPIR是中科院张华平博士开发中文分词系统，被誉为自然语言处理奠基之作，目前国际、国内测评双第一。NLPIR分词系统前身为2000年发布的ICTCLAS词法分析系统，从2009年开始，调整命名为NLPIR分词系统，推广NLPIR自然语言处理与信息检索共享。现在的NLPIR大数据语义分析系统能够全方位多角度完成对大数据文本的处理需求，包括大数据完整的技术链条：网络抓取、正文提取、中英文分词、词性标注、实体抽取、词频统计、关键词提取、语义信息抽取、文本分类、情感分析、语义深度扩展、繁简编码转换、自动注音、文本聚类等。

03

示例演示

主讲教师进行操作示范，讲解操作方法、要点和技巧，突破重难点。

一、启动分词系统

操作演示1

在浏览器网页输入网址（http://ictclas.nlpir.org/）即可进入分词系统

二、抓取数据

操作演示2

验证通过后，在“验证成功”信息的下端 “网页URL”处输入待抓取数据的URL（http://world.people.com.cn/n1/2021/0307/c1002-32045025.html），点击“抓取”按钮开始抓取数据，也可以粘贴数据到文本编辑区域，作为分析基础数据。

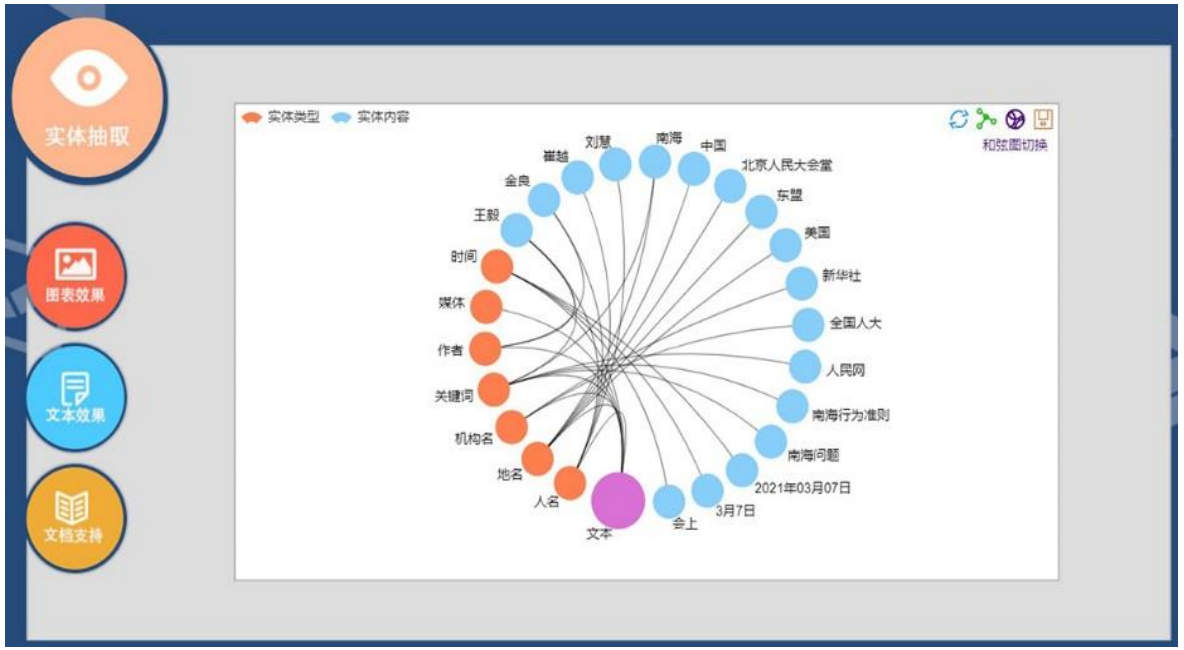
三、分词标注

操作演示3

在数据抓取成后，点击“分词标注”按钮，在分词标注页面可以看到：“王毅”、“南海”、“问题”、“国际”等名词是橘黄色标记，“谈”、“排除”、“干扰”等动词是蓝色标记，剩下少量修饰的介词、标点符号都有不同的颜色标注。新词发现了“南海行为准则”、“南海问题”、“南海和平”，这都是系统自动学习的结果。

四、实体抽取

操作演示4

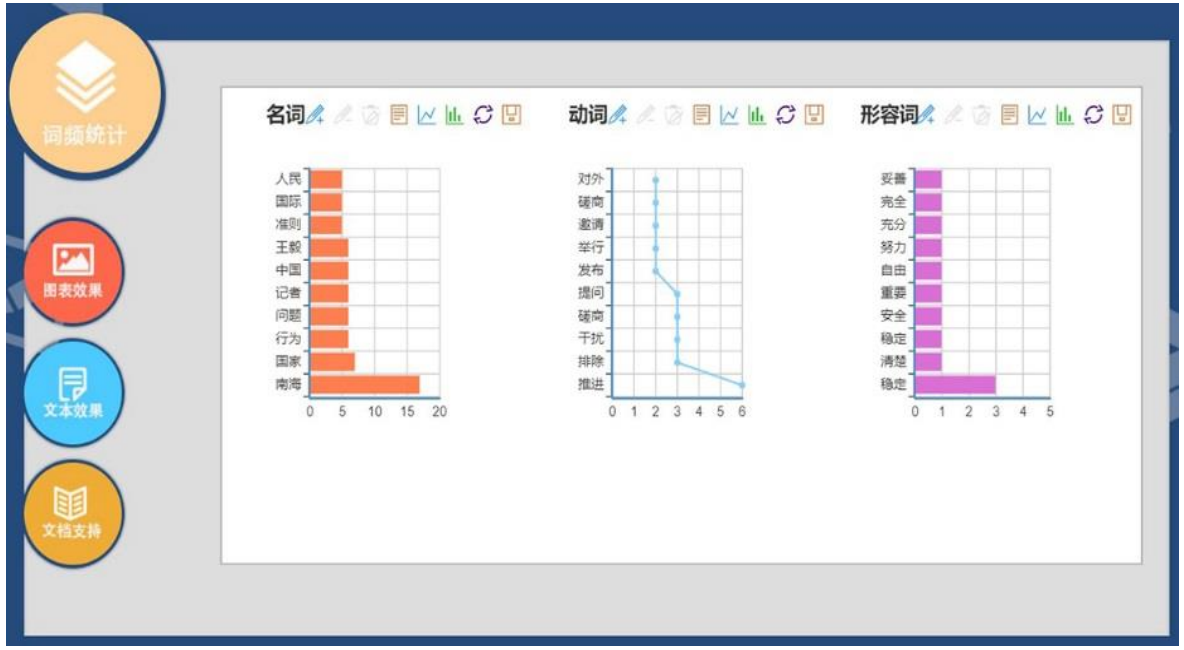


点击“实体抽取”按钮，可以看到实体抽取的蓝色内容有“王毅”、“南海”、“北京人民大会堂”等非常正式的新闻发言稿内容实体，由此推断，这是一篇关于南海问题的非常正式的新闻。

五、词频统计

操作演示5

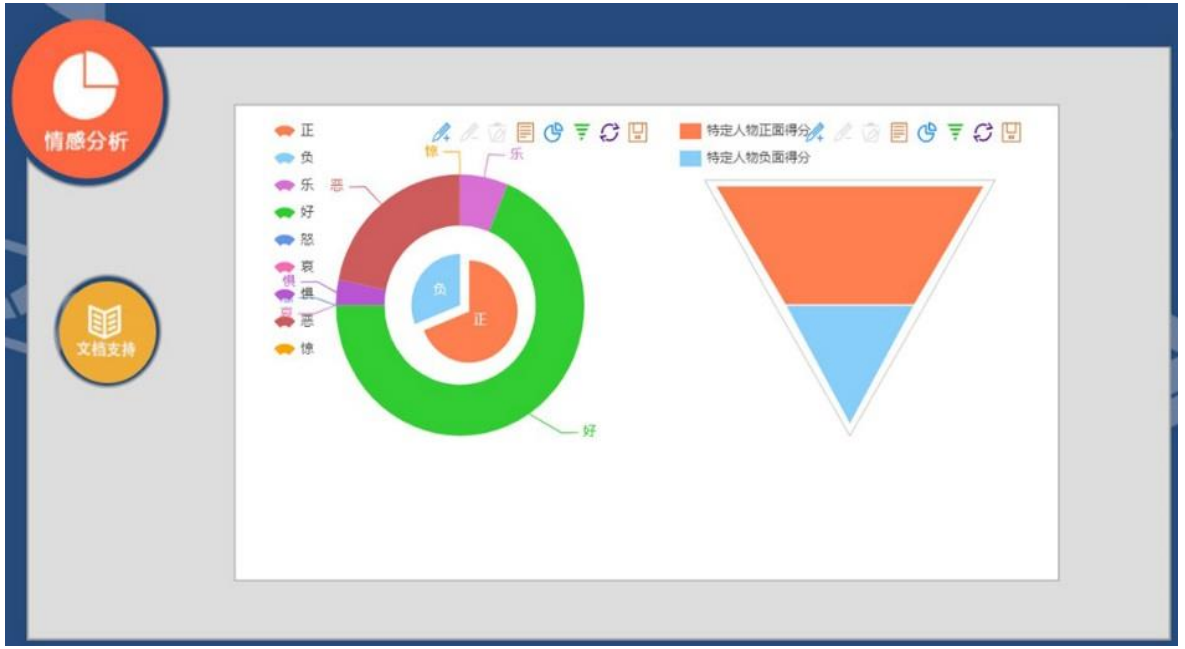
点击“词频统计”按钮，可以看到本新闻中出现最多的名词是“南海”，出现最多的动词是“推进”，出现最多的形容词是“稳定”，我们可以推断本篇新闻重点是南海稳定政策。



六、情感分析

操作演示6

情感分析是基于深度神经网络对情感词进行扩展计算，本新闻稿中“好”情绪占68.75%，“恶”情绪占21.88%，“乐”情绪点6.25%，“惧”情绪占3.13%，总体来说“正”面情绪占68.97%，“负”面情绪占31.03%。



七、关键词分析

操作演示7

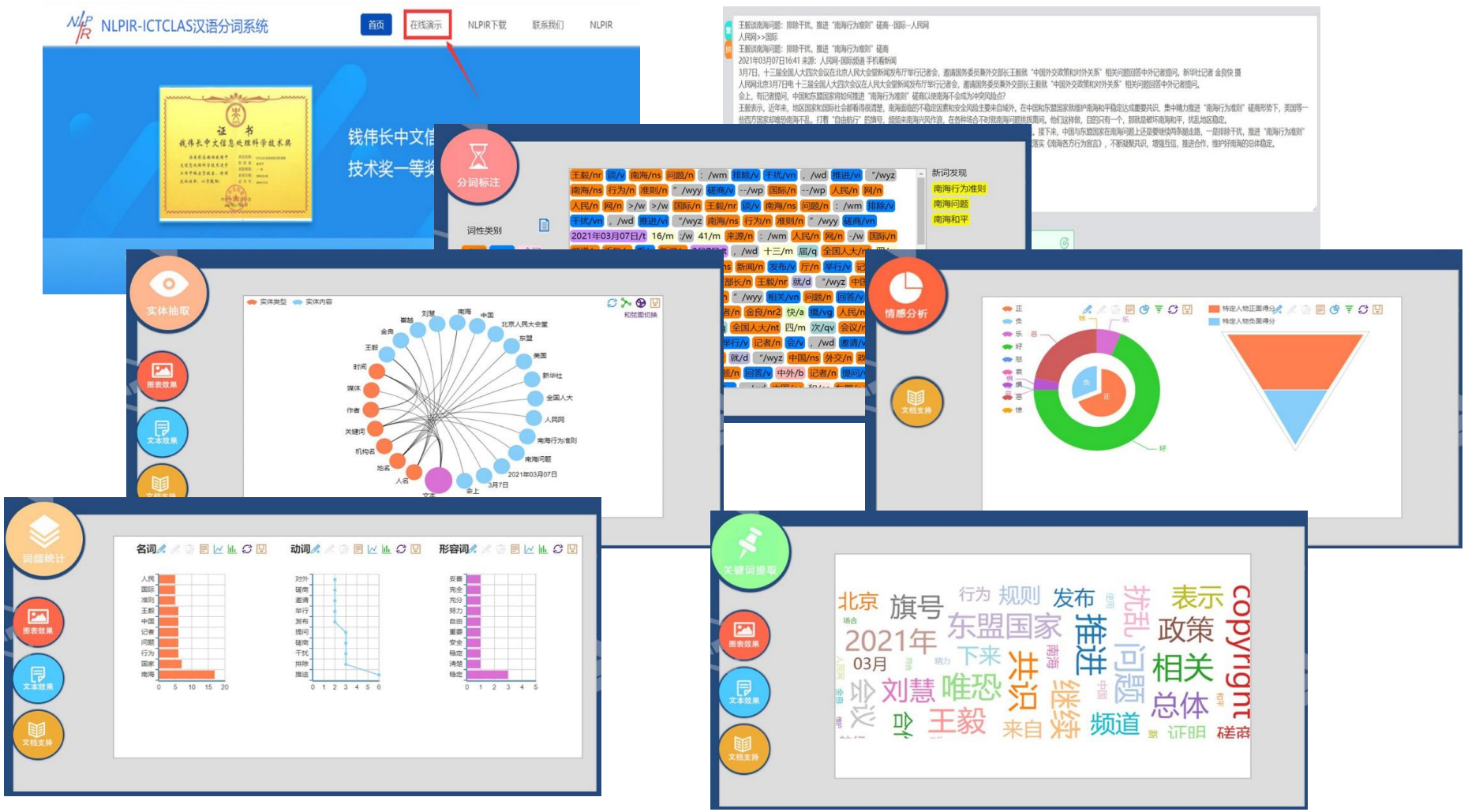
关键词分析是根据词频出现的次数展示的，词显示的越大，说明其出现的次数越多，词越小说明其出现的频次越低。

04

任务实现

学生在观看老师演示后开始动手实际操作，教师巡视指导，发现个别错误与一般错误。

任务实现



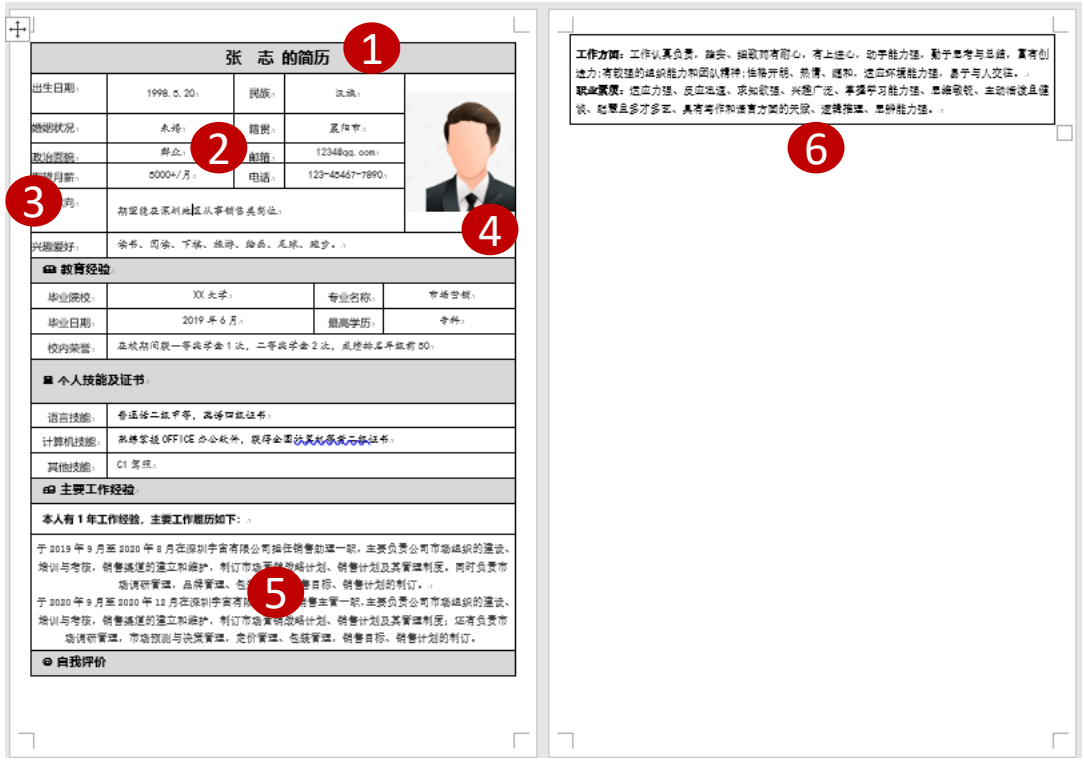
05

纠错重做

分析常见问题出错原因、讨论解决办法，及时纠正错误。

文档排版中常见问题

1. 标题格式错误
2. 行高不一致
3. 单元格对齐方式未居中
4. 照片有多余空白未裁剪
5. 单元格内容对齐方式不符合常规要求
6. 简历内容跨两页



06

总结评价

对操作步骤进行复述，特别提醒容易出错的步骤和环节，总结整堂课技能要点、方法要点和社会因素要点。

总结评价

分词准确性对搜索引擎来说十分重要，但如果分词速度太慢，即使准确性再高，对于搜索引擎来说也是不可用的，因为搜索引擎需要处理数以亿计的网页，如果分词耗用的时间过长，会严重影响搜索引擎内容更新的速度。因此对于搜索引擎来说，分词的准确性和速度，二者都需要达到很高的要求。