基于多平台的大学生就业态度分析

第11组第一次汇报

平台部署与数据处理

组员：

目 录
CONTENTS

01

项目设计思路

基于多平台的大学生就业态度分析

- 平台搭建：Hadoop + Spark 华为云服务器分布式平台
- 数据获取与存储：爬虫 + MongoDB
- 数据分析思路：词频统计 + KMeans++文本聚类 + BERT情感分析
- 结果展示：Echars + WordCloud + MatPlotlib

02
项目架构

项目架构

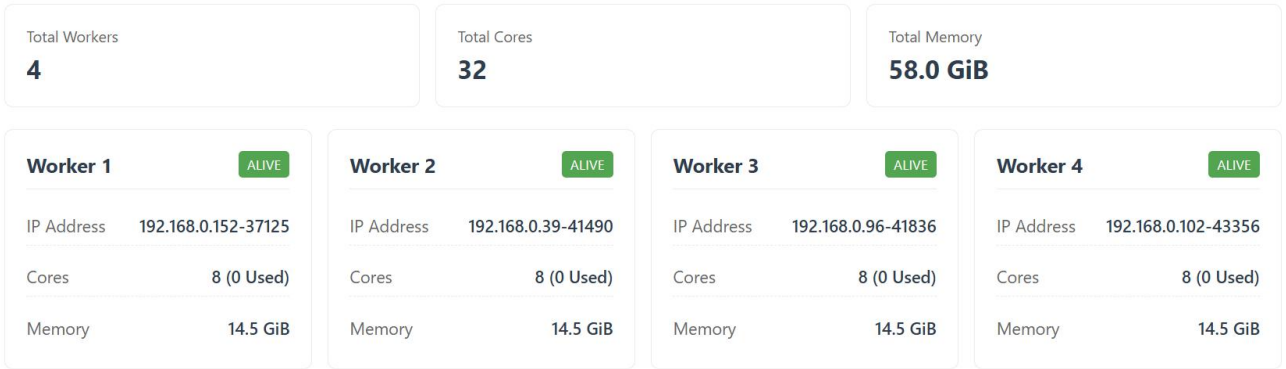
◆ 项目设备：

四台华为云服务器，均为x8cpu+16GB内存，通过内网高速通信

◆ 平台搭建：

Hadoop 2.10 + Spark 3.4 分布式平台

1个master节点，4个worker节点

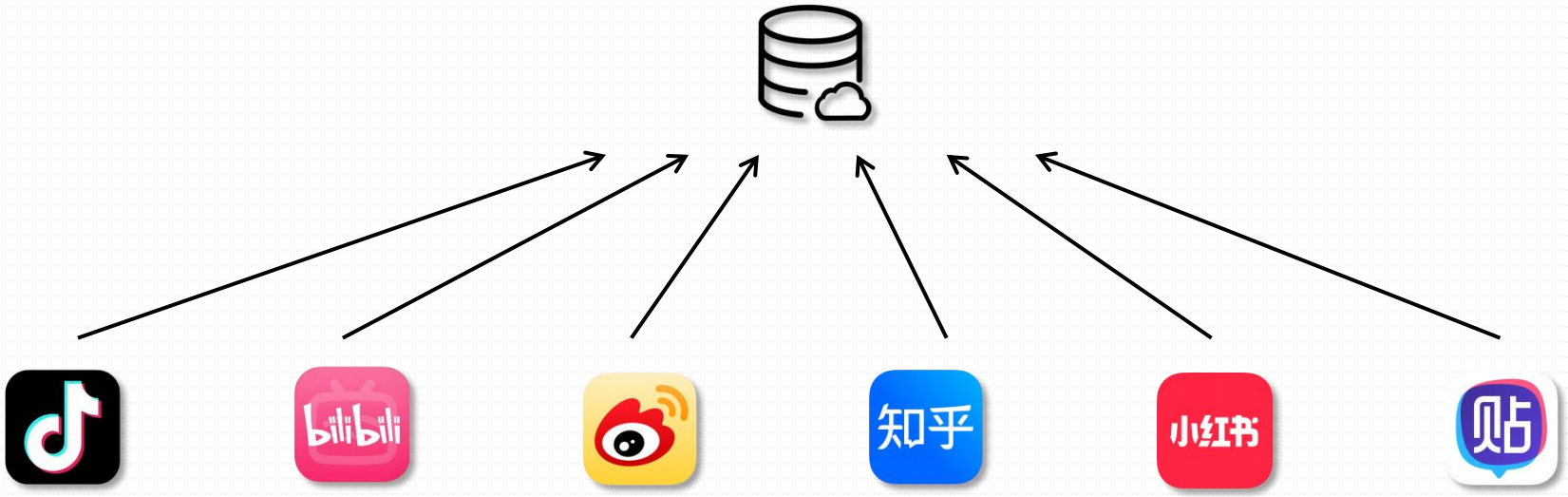


03

数据获取与存储

数据获取

◆ 多平台爬虫数据：约 29W 条



数据获取

◆ 关键词搜索与定位

通过模拟用户搜索行为，抓取与指定关键词相关的内容，包括帖子和评论

◆ HTML解析与数据提取

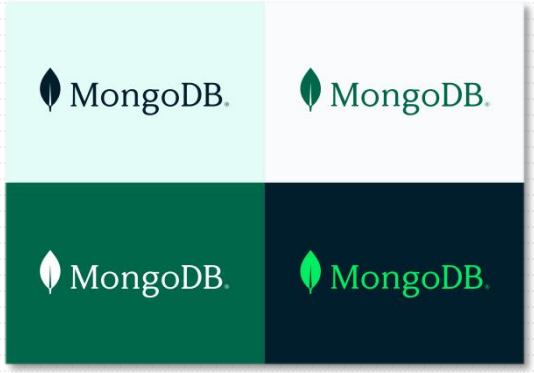
利用解析工具（如 BeautifulSoup 或 XPath），对抓取到的 HTML 内容进行结构化处理，从中提取关键信息，例如帖子标题、内容、评论内容、用户ip等

◆ 多平台适配与扩展

针对不同平台的数据结构和反爬机制，设计了定制化的解析策略，确保数据抓取的高效性与稳定性

数据存储

Nosql文档数据库 - MongoDB



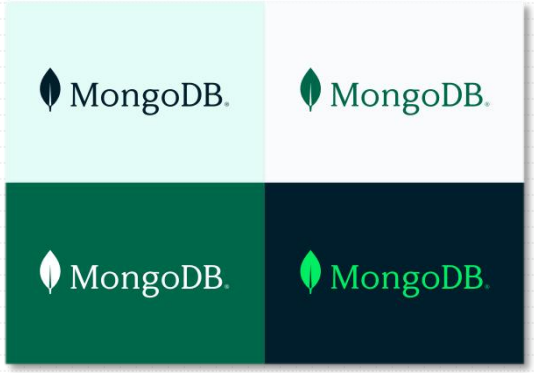
◆ 高效的JSON存储

MongoDB支持原生 JSON 格式存储，与爬取过程中生成的 JSON 数据实现无缝对接，减少了格式转换的复杂性和性能损耗

[{"aweme_id": "7431215260524186937", "aweme_type": "0", "title": "测评一巴掌，群面更是两巴掌。#秋招 #应届生求职 #应届生 #大学生就业 #面试", "desc": "测评一巴掌，群面更是两巴掌。#秋招 #应届生求职 #应届生 #大学生就业 #面试", "create_time": 1730214643, "user_id": "80820335543", "sec_uid": "MS4wLjABAAAAmItyY_hADzdtDKOiHN9ieGYLtoXXwexN2vFbFRiCxfU", "short_user_id": null, "user_unique_id": null, "user_signature": null, "nickname": "旺仔小丸子", "avatar": "https://p3-pc.douyinpic.com/aweme/100x100/aweme-avatar/tos-cn-avt-0015_5fc3b71c8da4e1e5487ed5933f0152f0.jpeg?from=327834062", "liked_count": "352173", "collected_count": "14560", "comment_count": "43184", "share_count": "349277", "ip_location": "", "last_modify_ts": 1731776329271, "aweme_url": "https://www.douyin.com/video/7431215260524186937", "source_keyword": "秋招"}, {"aweme_id": "7433095694619053349", "aweme_type": "0", "title": "秋招现状 # 秋招 #应届生 #985院校 #留学生", "desc": "秋招现状 #秋招 #应届生 #985院校 #留学生", "create_time": 1730652501, "user_id": "71149269092", "sec_uid": "MS4wLjABAAAAaZwUuFF0TLwRYw3FKG4BkhNnaDPSF_Kr8K2KhiZrw0c", "short_user_id": null, "user_unique_id": null, "user_signature": null, "nickname": "KKKKKKKK.", "avatar": "https://p3-pc.douyinpic.com/aweme/100x100/aweme-avatar/mosaic-legacy_38cc001b769d996e32ef.jpeg?from=327834062", "liked_count": "35764", "collected_count": "3149", "comment_count": "8418", "share_count": "65363", "ip_location": "", "last_modify_ts": 1731776329274, "aweme_url": "https://www.douyin.com/video/7433095694619053349", "source_keyword": "秋招"}, {"aweme_id": "7434174823221316899", "aweme_type": "0", "title": "#创作灵感 秋招结束了，大四工管0offer，发个长视频总结教训#秋招offer #秋招结束", "desc": "#创作灵感 秋招结束了，大四工管0offer，发个长视频总结教训#秋招offer #秋招结束", "create_time": 1730903721, "user_id": "91762828955", "sec_uid": "MS4wLjABAAAAm2WfbOvf0tTrx54g6KGpro3vnDODKFM4147Q4SbuBGQ", "short_user_id": null, "user_unique_id": null, "user_signature": null, "nickname": "红红一定拿offer！", "avatar": "https://p3-pc.douyinpic.com/aweme/100x100/aweme-avatar/tos-cn-avt-0015_fd2366a7461f0d159f6de3cc46deb43b.jpeg?from=327834062", "liked_count": "15756", "collected_count": "2171", "comment_count": "3178", "share_count": "4205", "ip_location": "", "last_modify_ts": 1731776329287, "aweme_url": "https://www.douyin.com/video/7434174823221316899", "source_keyword": "秋招"}, {"aweme_id": "7419581529296309558", "aweme_type": "0", "title": "秋季招聘会（秋招）还是太稳定了，稳定到要失业了 #精神状态belike #秋招 #研究生 #博士生", "desc": "秋季招聘会（秋招）还是太稳定了，稳定到要失业了 #精神状态belike #秋招 #研究生 #博士生", "create_time": 1727505960, "user_id": "95088163068", "sec_uid": "MS4wLjABAAAATx82Awawo5sJ-QHp4QB2goNlN6YUwlmBIwXyLguV-jM", "short_user_id": null, "user_unique_id": null, "user_signature": null, "nickname": "孤独的phd", "avatar": "https://p3-pc.douyinpic.com/aweme/100x100/aweme-avatar/tos-cn-avt-0015_e5e16fc5c00b659b311f44040793de4b.jpeg?from=327834062", "liked_count": "109256", "collected_count": "7804", "comment_count": "5303", "share_count": "70386", "ip_location": "", "last_modify_ts": 1731776329302, "aweme_url": "https://www.douyin.com/video/7419581529296309558", "source_keyword": "秋招"}, {"aweme_id": "7429972314017647930", "aweme_type": "0", "title": "秋招三个月0offer，我做对了什么 秋招三个月0offer，我做对了什么？\n一方面是想给未来秋招的同学一点参考\n另一方面就是反思反思吐吐槽\n总的来说还没有心仪 offer的同学别焦虑\n我们都有光明的未来 #应届生求职 #武汉大学 #面试 #秋招", "desc": "秋招三个月0offer，我做对了什么 秋招三个月0offer，我做对了什么？\n一方面是想给未来秋招的同学一点参考\n另一方面就是反思反思吐吐槽\n总的来说还没有心仪 offer的同学别焦虑\n我们都有光明的未来 #应届生求职 #武汉大学 #面试 #秋招", "create_time": 1729925258, "user_id": "107746917473", "sec_uid": "MS4wLjABAAAAyuIJtPkrAqbqA3VEtdJaILsKIKk1TMCUufmySWl1t4A", "short_user_id": null, "user_unique_id": null, "user_signature": null, "nickname": "asssuka", "avatar": "https://p3-pc.douyinpic.com/aweme/100x100/aweme-avatar/tos-cn-avt-0015_ed5a75aa83f16d756542ef80c296e7dc.jpeg?from=327834062", "liked_count": "33787", "collected_count": "3171", "comment_count": "4085", "share_count": "17806", "ip_location": "", "last_modify_ts": 1731776329318, "aweme_url": "https://www.douyin.com/video/7429972314017647930", "source_keyword": "秋招"}, {"aweme_id": "7293828083088379172", "aweme_type": "0", "title": "又是被秋招气到的一天#秋招 #大学生找工作现状 #应届生求职 #精神状态belike #地铁随拍", "desc": "又是被秋招气到的一天#秋招 #大学生找工作现状 #应届生求职 #精神状态belike #地铁随拍", "create_time": 1698226698, "user_id": "1764055443837102", "sec_uid": "MS4wLjABAAAAGrLxy-O63U2Vb4clwm1J0eXLrt0RgfIaAvQDk4rnaHr5fJkym6ogeWBlvjyYxJhd", "short_user_id": null, "user_unique_id": null, "user_signature": null, "nickname": "暴躁步步", "avatar": "https://p3-pc.douyinpic.com/aweme/100x100/

数据存储

Nosql文档数据库 - MongoDB



◆ 灵活的Schema设计

HTML解析与数据提取由于MongoDB无需预定义固定Schema，能够灵活存储结构化和非结构化数据，非常契合爬虫数据的多样性和非标准化特点

◆ 高效的扩展性与适配性

MongoDB的文档模型支持动态扩展字段，能够轻松应对不同平台爬取数据的异构性需求

04

数据预处理

数据去重　　MongoDB Query Language (MQL)

◆ 基于唯一标识去重

使用 comment_id 作为唯一标识字段，查询并删除重复记录，确保数据的唯一性和准确性。

```
db.comments.distinct("comment_id")

pipeline = [
    {
        "$group": {
            "_id": "$comment_id",  # 根据 comment_id 分组
            "doc": {"$first": "$$ROOT"}  # 保留第一个完整文档
        }
    },
    {
        "$replaceRoot": {"newRoot": "$doc"}  # 将完整文档提升为根级别
    }
]

result = db.comments.aggregate(pipeline)
```

去除无关数据

◆ 去除无关数据

针对初步收集到的内容及评论，剔除与大学生就业主题无关的数据，包括无意义的文字段落和评论

◆ 分词与预设关键词匹配

采用 Jieba 分词 技术对每条数据进行分词处理，细化文本内容

依据预设的大学生就业主题关键词词汇表，判别内容与主题的相关性

◆ 主题相关性判定方法

数据文本中出现关键词的频率或密度超过某一阈值即可判定为主题相关

数据处理

7437427530832858368	那些单位只需要到985、211学校去招聘就行了
7437427530832858368	拿到offer了，工资还可以但是其他的条件还是不满意，还在等，感觉找不到了
7433040098901445925	安徽六安或者合肥谁帮我内推懂得感恩
7437427530832858368	@拾贰號籃球街
7437427530832858368	无非就是想感动自己 想提升一下 然后在研究生继续发VLOG 找个存在感
7437427530832858368	别人招聘嘛，你说禁止限制985，211，选调嘛，你又……呵呵，好人坏人都让你做了
7437458805018873125	我找工作，都有公司特别要求不要22届的毕业生
7437458805018873125	孩子还那么多，大家都在偷偷生吗
7380301369338563890	都是一些我们实习工作的小经历和感受~说的不一定全对 毕竟实践出真知哈哈 之后
7437458805018873125	看来我要在明年的毕业季之前找工作了
7380301369338563890	24应届还没找工作好焦虑
7437458805018873125	这个平翘舌 听的我舌头要打结了
7380301369338563890	校招的工作 千万不要轻易放弃 因为那可能是你能去的最好的单位
7437458805018873125	别提了，今天失业了单位倒塌了
7380301369338563890	有没有大学生毕业生的群呀
7437458805018873125	别提了，现在已啃老
7380301369338563890	刷烂了
7437458805018873125	早就塌了
7380301369338563890	大专应届生，感觉自己什么也做不了，进过厂知道很累，虽然工资高，但是不想进厂
7437458805018873125	大学生应聘新生儿
7380301369338563890	应届生 学历大专专业是幼师 在考虑要不要干主播了
7437458805018873125	@beautiful
7380301369338563890	周一至周五看boss周末双休
7437458805018873125	@ohh_ @CC
7380301369338563890	快点快点，点开图片，然后缩小
7437458805018873125	还把退休调那么大，最后都是失业的
7380301369338563890	一定要先就业再择业！就业了才能真正和社会接轨给择业打下基础！不走出第一步怎
7380301369338563890	23届应届考的央国企，现在在烟草一年20
7380301369338563890	专升本二本刚毕业，到处都是很低要么被骗，卡经验，甚至800都不知道怎么办，天
7380301369338563890	25年毕业 我现在焦虑的不行
7414699192301538594	人间清醒啊，大学生太多了，现在工作确实是看能力不看学历

7437427530832858368	那些单位只需要到985、211学校去招聘就行了
7437427530832858368	拿到offer了，工资还可以但是其他的条件还是不满意，还在等，感觉找不到了
7433040098901445925	安徽六安或者合肥谁帮我内推懂得感恩
7437427530832858368	无非就是想感动自己 想提升一下 然后在研究生继续发VLOG 找个存在感
7437427530832858368	别人招聘嘛，你说禁止限制985，211，选调嘛，你又……呵呵，好人坏人都让你做了
7437458805018873125	我找工作，都有公司特别要求不要22届的毕业生
7380301369338563890	都是一些我们实习工作的小经历和感受~说的不一定全对 毕竟实践出真知哈哈 之后
7437458805018873125	看来我要在明年的毕业季之前找工作了
7380301369338563890	24应届还没找工作好焦虑
7380301369338563890	校招的工作 千万不要轻易放弃 因为那可能是你能去的最好的单位
7437458805018873125	别提了，今天失业了单位倒塌了
7380301369338563890	有没有大学生毕业生的群呀
7437458805018873125	别提了，现在已啃老
7380301369338563890	大专应届生，感觉自己什么也做不了，进过厂知道很累，虽然工资高，但是不想进厂
7437458805018873125	大学生应聘新生儿
7380301369338563890	应届生 学历大专专业是幼师 在考虑要不要干主播了
7380301369338563890	周一至周五看boss周末双休
7437458805018873125	还把退休调那么大，最后都是失业的
7380301369338563890	一定要先就业再择业！就业了才能真正和社会接轨给择业打下基础！不走出第一步怎
7380301369338563890	23届应届考的央国企，现在在烟草一年20
7380301369338563890	专升本二本刚毕业，到处都是很低要么被骗，卡经验，甚至800都不知道怎么办，天
7380301369338563890	25年毕业 我现在焦虑的不行
7414699192301538594	人间清醒啊，大学生太多了，现在工作确实是看能力不看学历
7380301369338563890	这年头，找工作真是太难了，好愁啊
7380301369338563890	毕业生有没有想创业的?
7380301369338563890	说实话，没学历，没背景，没家庭，还不如考一个军队文职稳定一点
7380301369338563890	去面试告诉我底薪2000，没好意思打车决定坐公交。走两步看到一个蛋糕店没忍住，
7380301369338563890	之前不知道为什么我公司对大学生敌意这么大，直到读了之后才知道
7380301369338563890	早九晚七单休 节假日正常这个工资水平可以嘛

经过去重和去除无关数据后，最终的数据量为：20W

_id	comment_id	ip_location	aweme_id	content	user_id	sub_comment_count
6738ba83a5a4c9166d63333b	7414425093205771017	云南	7414357233418521890	朋友专升本前工资1w+出差审计，我劝她升本，现在	56579552831	12996
6738ba83a5a4c9166d63333c	7395504050240291622	陕西	7395449615778204937	刷到的宝子们都能顺顺利利的上完大学，顺顺利利的	1450962281831447	12
6738ba83a5a4c9166d63333d	7414410853741953831	黑龙江	7414357233418521890	不得不拿出这张图了	92563099019	1898
6738ba83a5a4c9166d63333e	7398146813041083173	广东	7395449615778204937	早上boss直骗　中午前程堪忧　下午58同坑	67751682353	1057
6738ba83a5a4c9166d63333f	7414398417244390154	福建	7414357233418521890	985都这样了 我这种垃圾怎么办啊	536991365608051	5468
6738ba83a5a4c9166d633340	7395505931058643739	陕西	7395449615778204937	现在是干啥都不容易，我6800学了烧烤，三轮车360	3171407154643310	598
6738ba83a5a4c9166d633341	7414415808648889145	上海	7414357233418521890	985 理科 男生 有四大实习经验！！！ 我天啊还是人太	917438702817035	197
6738ba83a5a4c9166d633342	7395504378654507814	陕西	7395449615778204937	1、zg公共招聘网2、各省的就业网3、新职业4、毕业	62199418201	75
6738ba83a5a4c9166d633343	7414520789246280498	北京	7414357233418521890	给你们算一笔账	4340527466286823	1042
6738ba83a5a4c9166d633344	7395603775639749426	广东	7395449615778204937	投了，简历初筛过了，笔试过了，初面过了，终面被	103520292485	49
6738ba83a5a4c9166d633345	7414551482429981474	北京	7414357233418521890	也许你们真的在玩梗，但我是真的大专，外企1w上4	63825021265	1490
6738ba83a5a4c9166d633346	7414174857183396642	河南	7395449615778204937	我用的是鱼泡网，专科进的央企	1249462117742635	24
6738ba83a5a4c9166d633347	7414427648027411260	安徽	7414357233418521890	你好认识一下，我211本985硕➕一个美硕，刚找的	73796488786	4294
6738ba83a5a4c9166d633348	7401207793128751909	广东	7395449615778204937	不用看了 专科普本轮不到你	1494929582865063	77
6738ba83a5a4c9166d633349	7414404369217323810	上海	7414357233418521890	上海这边3000甚至可以招到复旦应届生	68488636085	1200
6738ba83a5a4c9166d63334a	7397325933781287717	四川	7395449615778204937	大专不用看了（别喷，我就是）	4099401625637031	86
6738ba83a5a4c9166d63334b	7414554513275421474	河南	7414357233418521890	上学中:做家教十几天2个w毕业上班:一月2k	3795125788215885	76
6738ba83a5a4c9166d63334c	7397327092642349878	广东	7395449615778204937	就业平台怎么能少了58同城呢，简称人生第一课	98985718807	5
6738ba83a5a4c9166d63334d	7414377827665527602	山东	7414357233418521890	你太敢要了1.2w，现在就业行情985工科硕士也有可	4424037026510451	298
6738ba83a5a4c9166d63334e	7398121711315632953	海南	7395449615778204937	建议出国找，国内不需要那么多劳动力	105831642739	22
6738ba83a5a4c9166d63334f	7414526695305347878	江苏	7414357233418521890	月薪3500的时代已经过去了，现在来到了月薪3000	100007531728	69
6738ba83a5a4c9166d633350	7413769395774440219	安徽	7395449615778204937	不需要找有技术我现在月入2w	98231179038	19
6738ba83a5a4c9166d633351	7414428774995165978	陕西	7414357233418521890	笑死，我是土木工程的女生，一下是我和hr的对话hr:	3835568415059837	299

谢谢！

组员：