数据科学在竞技世界的应用
——多模态数据分析及应用

竞技世界 / 巴川

巴川

- 竞技世界首席数据科学家；
- 北航软院兼职硕导；
- 西安交大研究生院授课专家；
- CCF TF数据科学SIG主席；
- 中国教育创新校企联盟专家委员会副主任；
- 中国国际“互联网+” 创新创业大赛专家评委；CCF科技创业秀等多个双创大赛评委；
- 曾就职于中国搜索、搜狐畅游等互联网公司从事数据挖掘、人工智能、知识图谱，风控体系、推荐系统、数据可视化相关工作；
- 多个技术峰会演讲嘉宾及出品人。
- 多所高校兼职教师及创新创业导师。

目录
contents

Part 01

多模态数据介绍

浅谈多模态数据——跨越感知界限的数据探索

➳ 多模态数据的定义

指包含多种感知模式（如图像、文本、声音等）的数据集合

在多模态数据中，不同感知模式之间存在交互和关联，提供了丰富的信息来源

例如：

一张照片可以同时包含图像和文本的信息，通过分析图像中的视觉特征以及图像上的文字描述

一段视频可以同时包含图像、声音和文本等多种感知模式

➳ 多模态数据的研究意义

1. 改善任务效果与性能

整合和融合不同感知模式的信息，可以提供更多的数据维度来支持机器学习、计算机视觉、自然语言处理等任务，并提高其性能

2. 拓宽数据研究的边界

多模态数据的研究可以帮助克服单一模态数据的局限性，突破传统的边界，引入更复杂、多样的数据特征和信息

3. 提供更全面且准确的信息

通过同时利用文本、图像、声音等多种模态，可以获取更丰富的上下文信息，从而改善对数据的理解和处理，挖掘出更精准、有价值的信息

Part 02

文本数据的分析与应用

文本数据分析的背景——基于电竞业务的文本分析

➳ 电竞业务说明

电竞赛事在各大直播短视频媒体平台上进行多渠道分发，KOL赛事转播，需要及时了解用户在共振场地（直播间、短视频、社群）中的问题反馈，需对文本内容进行分析

➳ 文本数据分析

对文本数据进行处理、挖掘和分析的过程

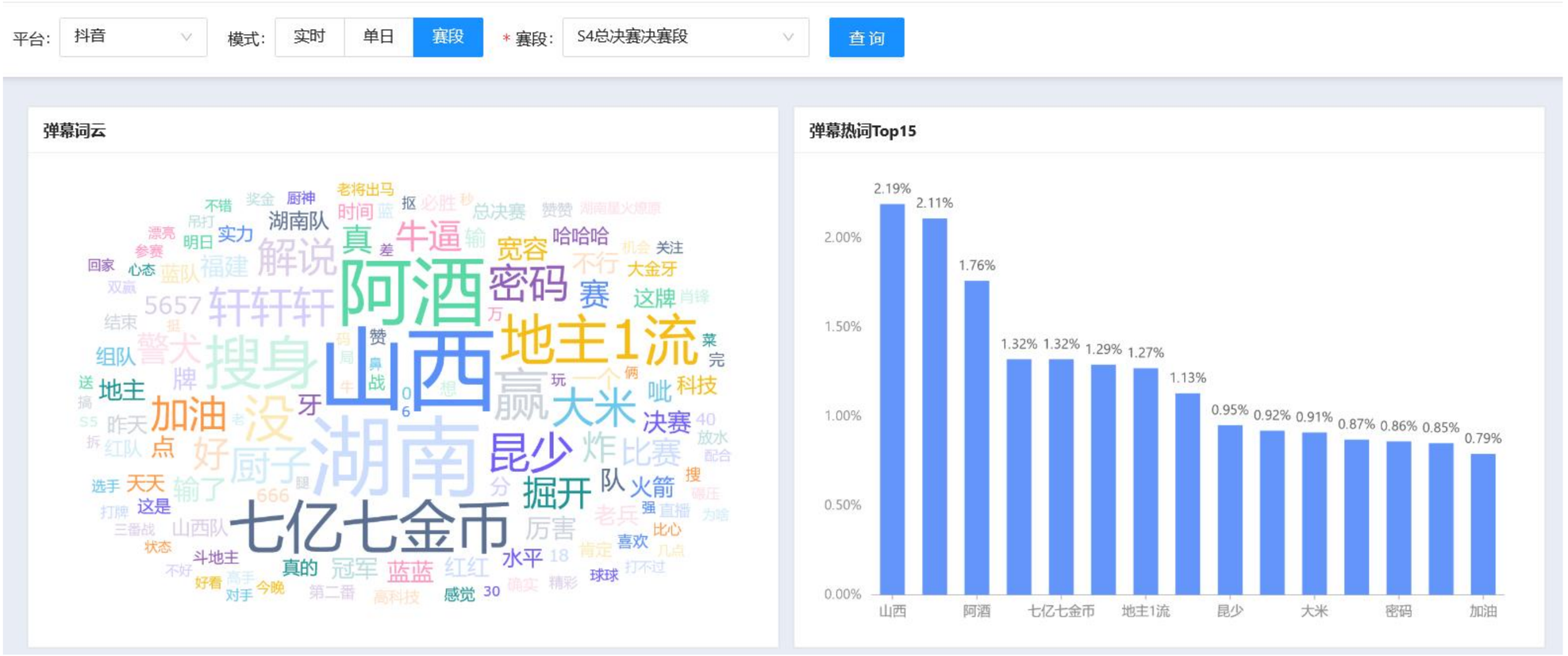
旨在从海量文本中提取有用信息、发在潜在模式和规律

通过文本数据分析，可以从文本中获取有关主题、情感和实体等方面的信息

文本数据分析的应用——关键词提取

应用案例：共振项目中弹幕关键词的词云展示

专有词加入到分词列表 → 分词 → 过滤停用词及特殊字符等 → 提取关键词 → 词频统计 → 绘制词云图



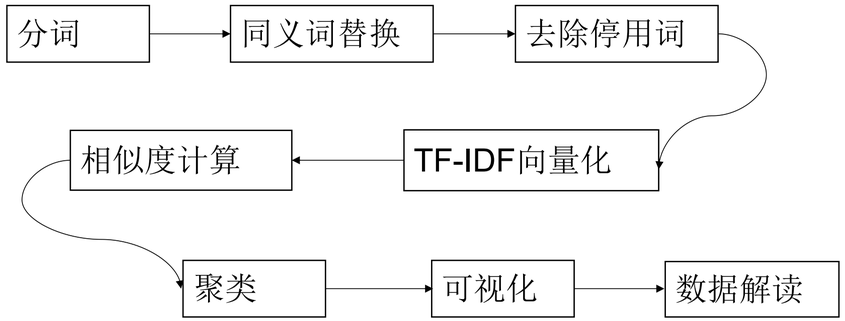
文本数据分析的应用——主题检测

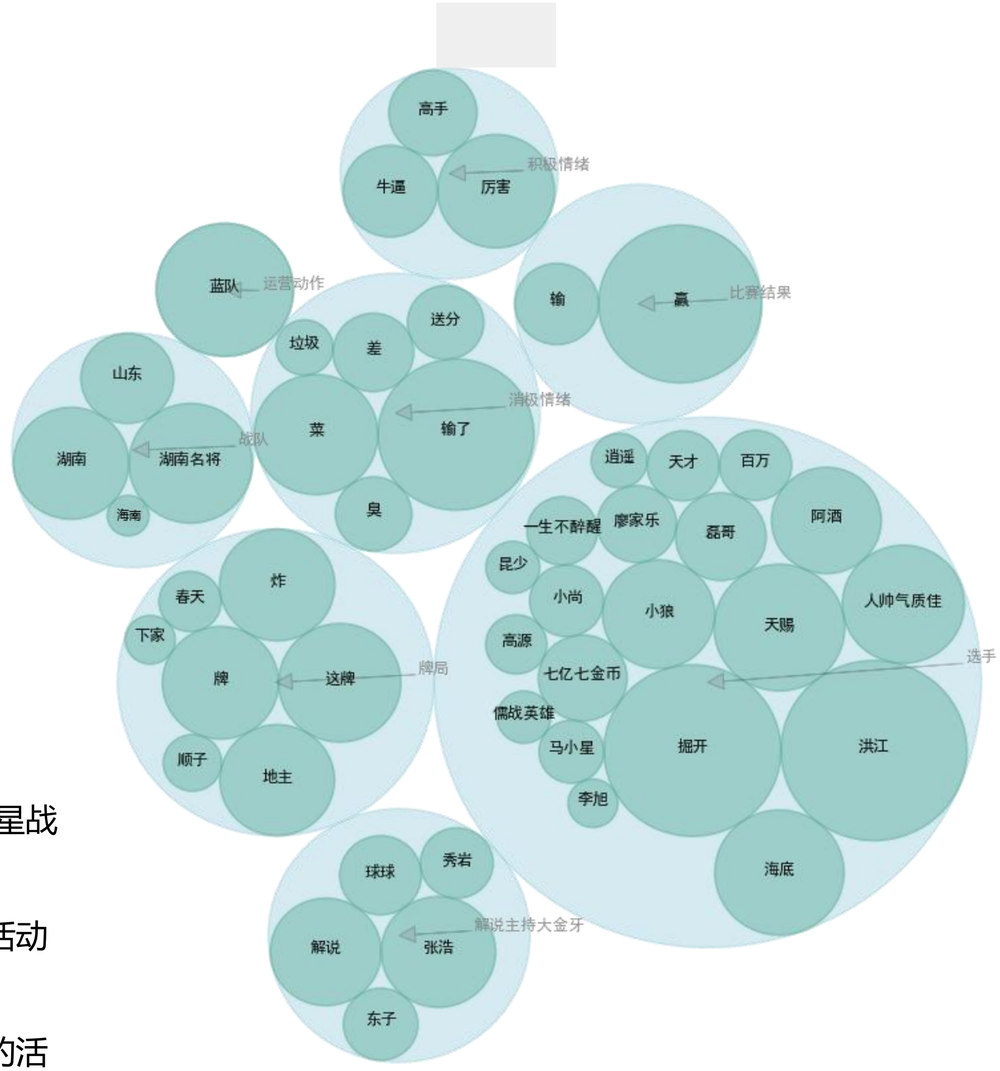
应用案例：抖音直播间内活跃用户的话题检测

冠军杯用户之间的距离存在特性：数据较为分散

DBSCAN作为基于密度的算法，可以找到样本点的全部密集区域

行为指标
首次评论日期
评论天数
评论次数
非活动弹幕评论次数
运营活动参与次数
评论高频词
非直播场次是否有过发言





➢ 结果说明：

1、除明星战队以外，核心用户对黑马战队（山西清徐学栋等）的关注度也较高，因此，除了为明星战队及选手投入主要营运资源以外，次级资源的投入可重点向黑马战队（如海南四海）及选手倾斜；

2、核心用户更多会在“春天”、“顺子”牌局出现时参与讨论，可针对此类牌局探索话题引导和活动设计（如：加入竞猜题目等），从而提高用户讨论度和活动参与率；

3、核心用户偏爱参与门槛低的运营活动，并且此类活动能够有效激发用户的评论欲望，所以后续的活动设计需将参与门槛作为重点考虑因素。

文本数据分析的应用——信息摘要生成

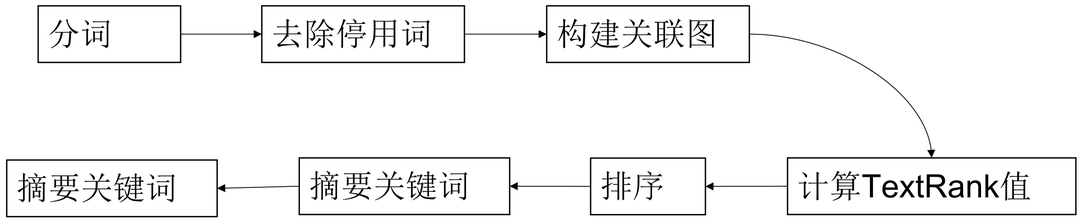
应用案例：热门短视频下用户评论的信息提取

➤ **数据说明**

本视频是对S5公开赛4强争夺战出结果后，在JJ斗地主官号发的视频内容

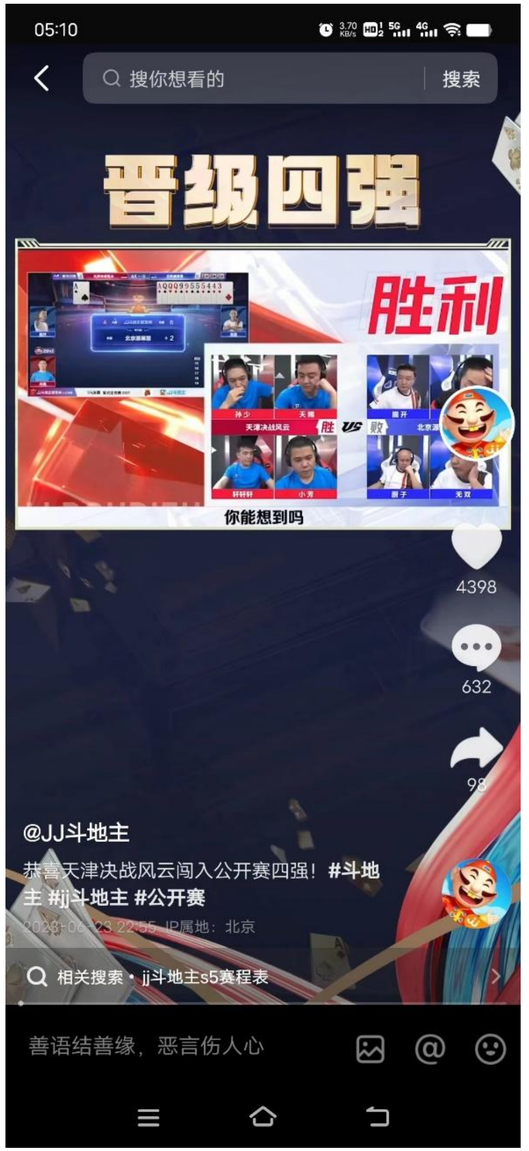
天津决战风云VS北京源莱慧，实力战队的较量

➤ **基于TextRand算法**




```
result = extract_summary(summary)
print(result)
```

```
天赐 和 轩 专克 掘开 小芳 跟轩 轩轩 搭档 配合 真好...
```



文本数据分析的应用——情感倾向分析

应用案例：用户评论的语义情感评分

➤ **数据处理：**

利用同义词、反义词、程度副词权重等方式进行文本增强

➤ **特征提取：**

基于TF-IDF、词频、n-gram的文本特征提取方法，调整建模的输入

➤ **算法模型：**

利用集成学习的思想，将多个算法的输出进行综合，提高分类准确率

直播弹幕文本：SVM+朴素贝叶斯

短视频评论文本：LSTM

关键词详情

导出

时间	平台	直播间	对象类型	账号	内容	关键词	情感评分
2023-08-16 21:26:17	douyin	直播间	官号	JJ斗地主	解说一说话我都烦死	解说	0.050
2023-08-16 20:40:44	douyin	直播间	官号	JJ斗地主	这俩解说毫无职业操守	解说	0.060
2023-08-16 19:51:03	douyin	直播间	官号	JJ斗地主	这俩解说真是搞笑	解说	0.060
2023-08-16 21:27:22	douyin	直播间	官号	JJ斗地主	脏洁解说就跟抬杠一样	解说	0.070
2023-08-16 20:04:27	douyin	直播间	官号	JJ斗地主	解说的闭上你那个坑	解说	0.080
2023-08-16 20:04:56	douyin	直播间	官号	JJ斗地主	解说太差了	解说	0.110
2023-08-16 21:07:37	douyin	直播间	官号	JJ斗地主	解说太差了	解说	0.110
2023-08-16 20:05:09	douyin	直播间	官号	JJ斗地主	解说不要胡说	解说	0.120
2023-08-16 21:34:45	douyin	直播间	官号	JJ斗地主	解说真的水平差	解说	0.130
2023-08-16 21:15:23	douyin	直播间	官号	JJ斗地主	解说都不公平	解说	0.140

< 1 2 3 4 5 ··· 21 > 10 条/页

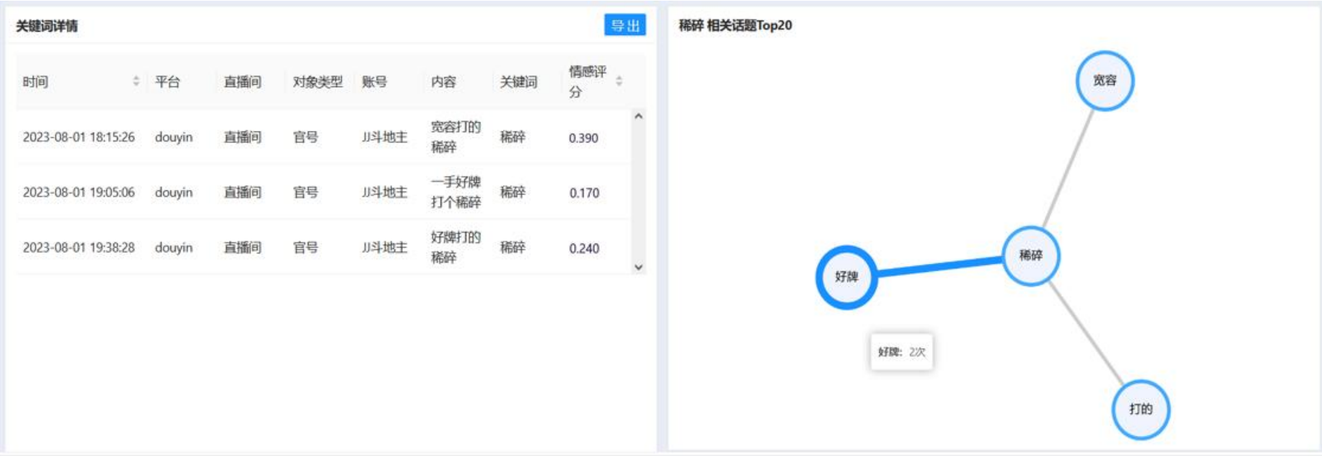
关键词详情

导出

时间	平台	直播间	对象类型	账号	内容	关键词	情感评分
2023-08-16 21:19:44	douyin	直播间	官号	JJ斗地主	酒傻子这牌不接是真菜	真菜	0.010
2023-08-16 21:25:54	douyin	直播间	官号	JJ斗地主	宁少你是真菜	真菜	0.160
2023-08-16 20:30:00	douyin	直播间	官号	JJ斗地主	宁少真菜	真菜	0.160
2023-08-16 13:06:51	douyin	直播间	官号	JJ斗地主	真菜自己先不打3个8	真菜	0.160
2023-08-16 20:31:20	douyin	直播间	官号	JJ斗地主	宁少真菜	真菜	0.160
2023-08-16 20:34:02	douyin	直播间	官号	JJ斗地主	宁少是真菜啊	真菜	0.160
2023-08-16 19:07:45	douyin	直播间	官号	JJ斗地主	掘开真菜	真菜	0.220
2023-08-16 19:41:25	douyin	直播间	官号	JJ斗地主	真菜的不行	真菜	0.240
2023-08-16 19:40:36	douyin	直播间	官号	JJ斗地主	真菜的不行	真菜	0.240
2023-08-16 21:09:55	douyin	直播间	官号	JJ斗地主	这个宁少是真菜的抠脚	真菜	0.250

< 1 2 3 4 5 ··· 13 > 10 条/页

文本数据分析的应用——关系抽取



排名	热度值		热度人物	关联事件	代表性评论	说明、结论	分类	属性		应对方案	来源	时间
1	0.116	574	马哲	打枪	哈哈。打枪就好好打枪去	CF（枪击类）游戏全国冠军选手-马哲跨界JJ斗地主，观众质疑马哲选手的实力。	赛事	非负向	用户自发	**可选择马哲选手的高光牌局（若有），消除质疑**	跨界来袭，6点不见不散#jj斗地主 #S5	2023/6/21
				跨界电竞	跨界电竞大佬们来袭，JJ官方越来越棒	观众对跨界选手的加入颇有兴趣，对官方持肯定态度。	赛事	非负向	自发+引导	**后续可考虑进一步推进和加强赛事/参赛选手的多元化，促发观众兴趣**	跨界来袭，6点不见不散#jj斗地主 #S5公开赛 #斗地主	2023/6/21
2	0.086	427	掘开	掘开 想吃鱼		**网友自制表情包。**画面：掘开入场 配文：告诉黑双，我想吃鱼 来源：《狂飙》。	赛事	非负向	用户自发	**赛场“花边”选取参考**	今晚的胜者才有资格争夺四强#斗地主 #jj斗地主 #S5公开赛	2023/6/22
				掘开 想太多	瞧掘开那把牌打得，什么玩意啊，能防炸的牌，给打成挨炸了，想的太多。	掘开出牌考虑时间过长，多次失误，实力发挥不稳定。粉丝心中对此产生较大**落差感。**	赛事	负向	用户自发	**官方不做评价，不予理会**	恭喜北京源莱慧顺利口 队强！#斗地主 #jj斗地主	2023/6/23
				掘开 笑	倔大豆腐你还舔脸笑呢	比赛结束，北京源莱慧止步八强，掘开同方位丢分较多，赛场失利，引起**粉丝不满。**	赛事	负向	用户自发	**赛后余韵避免踩雷**	恭喜天津决战风云闯入公开赛四强！#斗地主 #jj斗地主	2023/6/24
3	0.051	253	天赐	天赐 MVP	天赐最近状态太好了，连拿三个MVP，期待天赐拿到FMVP	天赐选手在公开赛中的表现亮眼，有实力有气场，**粉丝高度认可天赐，并期望夺冠。**	赛事	非负向	用户自发	**高光牌局选取参考/明星选手打造参考**	恭喜天津决战风云闯入公开赛四强！#斗地主 #jj斗地主	2023/6/24
3	0.025	124	小狼	小狼思维牛逼	我仔细模拟了半天，想了想，确实这思维模式高。断四门，他这么打是拆双王防一炸。而且防外面有仨2有小炸的打法。真的稳。	小狼精准拆王防炸2，出牌思路，打法果断，粉丝共鸣度高。	牌局	非负向	完全引导	**高光牌局选取思路参考（多种牌路取舍优劣判别的牌局类型）**	加入辽宁查理军团后又会有什么惊喜呢#斗地主 #JJ斗地主 #S5公开赛	2023/6/21
				小狼 秀	这副牌，小狼 ，真的是秀！	肖锋&老将出马的农民配合高光牌局，小狼作为地主，尽管输牌，但牌路，思考，引起粉丝热议，并受到称赞与认可。	牌局	非负向	用户自发	**高光牌局选取思路参考。**	看看牌做选择都不一定能选对……#斗地主 #JJ斗地主 #精彩	2023/6/25
4	0.011	52	轩轩轩	轩轩轩 开挂	轩轩轩打得最好	A&B组四强争夺战，轩轩轩地主胜率90%。	赛事	非负向	用户自发	**牌局选取参考**	恭喜天津决战风云闯入公开赛四强！#斗	2023/6/23
5	0.008	38	阿亮	阿亮 黑	阿亮这是从哪里回来的！晒这么黑[捂脸]	**阿亮回归**，粉丝调侃阿亮伊拉克挖石油、山西挖煤，煤老板。粉丝更关注**阿亮肤色变黑**的问题。	赛事	非负向	用户自发	**赛场“花边”素材选取参考**	别错过了，C组开始啦！#jj斗地主	2023/6/25
6	0.007	32	逍遥	逍遥 强	昨天晚上逍遥打的真好，这才是职业高手，逍遥气场强大，打的快节奏好，观感性强，看起来很爽！	S5公开赛小组赛B组：北京源莱慧VS辽宁查理军团。逍遥个人得分330，拿下当晚MVP，**粉丝认可逍遥的实力与赛场表现。**	赛事	非负向	自发+引导	**牌局选取参考**	恭喜北京源莱慧顺利口 队强！#斗地主 #jj斗地主	2023/6/23
7	0.005	27	小芳	小芳有水平	小芳有水平，提高的很快啊	高光牌局，小芳提7拆顺提高点位，打破常规，粉丝认可。	牌局	非负向	自发+引导	-	斗地主的魅力正是在这#斗地主 #JJ斗地	2023/6/23
8	0.004	20	磊哥	老兵揍磊哥	原来老兵和磊哥也在一个队待过...去年这哥俩矛盾挺大的	古早年间，选手间的个人矛盾。	八卦	负向	用户自发	**官方不予理会，不做评价。**	这个王炸的不得不服#斗地主 #JJ斗地主	2023/6/25
9	0.001	7	阿水	官方 速度	官方这速度都是被阿水逼迫的	阿水赛果视频更新速度一直快于官方，官方遭网友调侃	赛事	非负向	用户自发	**官方与阿水目的性不同，不予理会即可。**	恭喜山西清徐学栋拿下本场胜利！，#斗	2023/6/25
10	0.001	6	廖家乐	廖大师 威武霸气	廖大师威武霸气！	C组Day1，廖家乐选手拿下当晚的MVP，表现亮眼，粉丝高度认可。	赛事	非负向	用户自发	**高光牌局选取参考。**	恭喜山西清徐学栋拿下本场胜利！，#斗地主 #jj斗地主 #S	2023/6/25

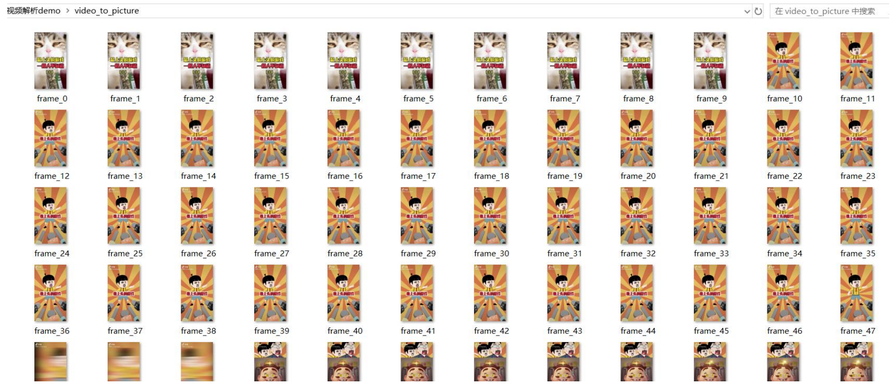
文本数据分析的应用——图片文本识别和视频文本识别

➤ 视频分帧（cv2）

```
video_path = "test.mp4"  # 将路径替换为你的视频文件路径
output_path = "video_to_picture"  # 将路径替换为输出图片的文件夹路径

video_to_frames(video_path, output_path)
```

视频已分割为863张图片



➤ 音频转换（moviepy）

➤ 图片中文本信息提取

安装好Tesseract OCR引擎，并配置好环境变量

（1）对图片进行灰度化处理

（2）应用OCR进行文本和数字信息识别


```
# 加载图片
image = cv2.imread('image.jpg')
# 灰度化处理
gray_image = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
# OCR识别
text = pytesseract.image_to_string(gray_image)
# 打印识别结果
print(text)
```

都有10分钟呢 最上头的游戏 抖音 抖音号：youxigangjing

Part 03

多模态数据分析及应用

直播视频—>短视频

AI智能剪辑

拆视频

将一场比赛视频拆分成一幅幅斗地主牌局

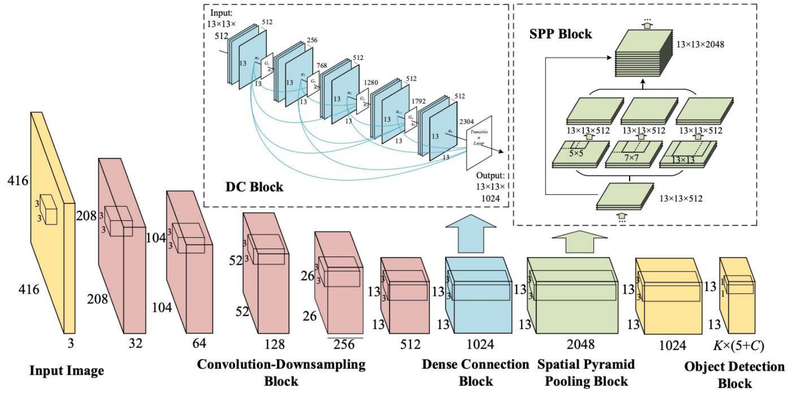
挑视频

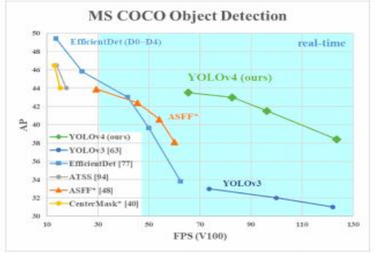
挑选精彩牌局视频

剪视频

剪辑并包装成短视频

拆视频：目标检测识别行牌过程——YOLO





Comparison of the proposed YOLOv4 and other state-of-the-art object detectors. YOLOv4 runs twice faster than EfficientDet with comparable performance. Improves YOLOv3's AP and FPS by 10% and 12%, respectively.

YOLOv4

Optimal Speed and Accuracy of Object Detection

识别错误与解决方案——OCR

Tesseract识别错误

- 位置变动（画面改版）
- /识别为1
- 6识别为8

文本检测精准定位

- 使用文本检测模型，准确定位文字位置
- 不受改版位置变动影响

Tesseract自定义训练

- 准备专有素材库
- 重新训练tesseract模型

CRNN等其他算法

- 基本解决前期出现的识别错误情况
- 采用CPU多核心加速，识别速度提高20倍

挑视频：多模态数据分析挑选精彩视频

牌局内容分析

牌局画像分析法

- 多炸牌局
- 剧情反转
- 农民配合
- ...

观众评论挖掘

文本挖掘手段

- 评论频率
- 语义分析
- 情绪识别
- ...

选手情绪识别

面部表情识别

- 高兴、愤怒、恐惧、悲伤、厌恶和惊奇
- 情绪波动
- ...

解说语音识别

情绪识别与文本分析

- 主持解说情绪波动
- 解说文本判断
- ...

专家综合打分

专业视角评判打分

- 1-5分制

Part 04

能力复用与思考

思考

一、舆情分析

二、自动摘要

三、端内外联动效果分析

四、识别带节奏的用户群体

五、舆情引导

六、复用时的工程效率

七、大模型应用……

思考


预警
报送
风险
判定
热点
发现
传播效
果分析
行业
分析
品牌口
碑洞察
舆情
预警
预警策
略制定
用户
洞察
舆情
感知
舆情
监测
舆情
分析
竞品
分析
舆情
跟踪
监测
大盘
危机
公关
大模型
AI机器人
评论
分析
事件
分析

THANKS !

THANKS

软件正在重新定义世界

Software Is Redefining The World