新疆棉事件-数据清洗分析

关键词搜索：

1. 新疆棉花
2. 强迫劳动 新疆
3. 新疆人权
4. 抵制新疆棉花
5. 制裁新疆棉花
6. 新疆纺织业
7. 新疆棉花争端
8. 国际贸易 新疆棉花
9. 人权组织 新疆
10. 中国与西方 新疆棉花
11. 维吾尔集中营

中文网站爬虫网站：

新华社：http://www.xinhuanet.com
人民网：http://www.people.com.cn
央视网：http://www.cctv.com
中国日报：http://www.chinadaily.com.cn
环球网：http://www.huanqiu.com
光明网：http://www.gmw.cn
中青网：http://www.cyol.com
中国新闻网：http://www.chinanews.com
新浪微博：http://www.weibo.com
知乎：http://www.zhihu.com
百度贴吧：http://tieba.baidu.com
豆瓣：http://www.douban.com
今日头条：http://www.toutiao.com
腾讯新闻：http://news.qq.com
网易新闻：http://news.163.com
搜狐新闻：http://news.sohu.com
百度：http://www.baidu.com
搜狗搜索：https://www.sogou.com/
360 搜索：https://www.so.com/
网易搜索：https://www.163.com/
雅虎搜索：https://www.yahoo.com/

总体要求

1、时间设定

按正常时间抓取（即 2020 年至现在）重点放在 2021 年 3 月-4 月。

目前已有部分数据，在原有数据上进行增加

2、需做过程处理

数据抓取——数据清洗——数据分析**（各阶段需要提交交付代码）**

3、原姓数据抓渠道事件，分词精精，可追溯。

4、具体分析模块如下：

（1） 数据收集模块

语言识别：自动识别和标记每条数据的语言。

数据量统计：各语言的数据量及其占比。

数据呈现：

数据收集概况：各语言数据量和来源的分布（条形图、饼图）。

数据质量报告：数据完整性、准确性评估（表格和柱状图）。

(2)情感分析模块的数据呈现

2.1 情感分类图

1. 情感分类饼图

描述：饼图展示不同来源（官媒与网民论坛）中情感分类（积极、消极、中立）的比例。

高级呈现：

多层饼图：使用多层饼图对比不同来源中的情感分类，内层表示官媒，外层表示网民论坛。

动态交互：使用 plotly 或 bokeh 库生成动态交互饼图，用户可以点击不同区域查看详细统计信息。

百分比显示：在每个饼图区域内显示具体的百分比值。

2. 情感分类条形图

描述：条形图展示各来源中的情感分类数量和比例。

高级呈现：

分组条形图：不同来源的情感分类使用分组条形图进行对比，条形图颜色区分情感类型（积极、消极、中立）。

堆积条形图：使用堆积条形图展示情感分类的绝对数量和比例，适合展示数据变化趋势。

动态过滤：使用 plotly 库添加交互式过滤器，使用户能够按时间段或具体关键词过滤数据。

2.2 情感得分趋势图

1. 情感得分折线图

描述：折线图展示情感得分随时间变化的趋势。

高级呈现：

多线折线图：为不同来源（官媒与网民论坛）绘制多条折线，显示各自的情感得分趋势。

趋势平滑：使用平滑曲线（如样条插值）来减少噪声，提高数据趋势的可读性。

时间轴标记：在时间轴上添加事件发生的关键日期和注释，以便对趋势变化进行详细分析。

2. 情感得分时间对比图

描述：对比不同来源的情感得分随时间的变化。

高级呈现：

双轴折线图：使用双轴折线图展示不同来源的情感得分趋势，便于对比分析。
事件标记：在图中标记事件的关键时间点和重要变化，以便识别情感得分的变化与事件的关联。

2.3 情感词云

1. 多语种情感词云

描述：词云展示每种语言中情感词汇的频率。
高级呈现：
动态词云：使用 WordCloud 库生成可交互的动态词云图，用户可以点击词汇查看相关情感分布。
情感色彩编码：根据情感得分为词汇添加不同的颜色（例如：正面词汇为绿色，负面词汇为红色），使词云图更具信息量。
多语言词云：为每种语言生成单独的词云图，并将其整合到一个多面板的可视化界面中。

2. 情感强度词云

描述：词云展示具有不同情感强度的词汇。
高级呈现：
强度映射：使用词汇的情感强度映射到字体大小和颜色，以突出显示高强度的情感词汇。
分组词云：按照情感强度将词汇分组，生成多个词云图，每个词云图展示一个情感强度范围。

2.4 高级数据展示

1. 情感分布热力图

描述：热力图展示情感分类在不同区域或时间段的分布情况。
高级呈现：
时间-情感热力图：显示情感分类在不同时间段的热力分布，便于识别情感变化趋势。
区域-情感热力图：显示不同地理区域的情感分布情况，帮助识别地域差异。

2. 情感分析报告

描述：综合展示情感分析结果的详细报告。
高级呈现：
动态报告生成：使用 Jupyter Notebook 或 Dash 生成交互式报告，用户可以选择不同的分析维度和时间范围查看结果。
自动摘要：根据分析结果生成自动化摘要，提供主要发现和数据驱动的洞见。
可视化整合：在报告中嵌入所有图表和数据可视化元素，确保信息的全面呈现和解释。

3、情感趋势分析

情感得分时间序列：
折线图：使用折线图展示情感得分在不同时间窗口的变化趋势。分为官媒和网民论坛的情感得分曲线，使用不同颜色和标记区分。
平滑曲线：应用平滑算法（如移动平均、Loess 回归）去除噪声，揭示潜在趋势。
情感波动分析：
波动率图：使用标准差或变异系数计算情感得分的波动率，绘制波动率图，以识别情感波动的强度和频率。
异常检测：应用异常检测算法（如基于统计学的 Z-score 方法或机器学习方法）识别情感得分的异常波动，并进行标注和解释。
趋势预测与建模

时间序列预测：
预测模型：使用时间序列预测模型（如 ARIMA、SARIMA、Prophet、LSTM 神经网络）对未来情感趋势进行预测。
模型评估：评估预测模型的性能，使用评估指标（如 MSE、RMSE、MAPE）进行模型优化。
趋势分析报告：
预测结果：提供未来情感趋势的预测结果，包括预测值、置信区间和趋势变化说明。
趋势图：在折线图中叠加预测曲线和实际数据，以便对比和验证预测准确性。
事件影响分析
事件节点分析：
关键事件标记：在时间序列图上标记关键事件的时间点，分析事件发生前后情感得分的变化。
事件影响分析：使用事件研究方法分析特定事件对情感得分的短期和长期影响，包括事件触发的情感波动。
多维度对比分析：
情感对比图：绘制多维度对比图，展示官媒与网民论坛在相同时间段内的情感对比，包括情感得分、情感分类比例等。
交互式面板：使用 Dash 或 Streamlit 等工具生成交互式面板，用户可以选择不同的时间段和数据维度进行对比分析。
高级数据可视化
情感热力图：
热力图生成：使用热力图展示不同时间段和不同来源的情感强度分布，便于识别热点时期和区域。
动态热力图：生成动态更新的热力图，以实时反映舆情的变化。
情感趋势动画：
动画展示：使用 matplotlib.animation 或 plotly 生成情感趋势的动画视频，展示情感得分随时间变化的动态过程。
事件演示：在动画中标记事件关键点，直观展示事件对情感趋势的影响。

3. 主题分析模块（区分官媒与网民论坛）

（1） 模块概述

主题分析模块旨在深入挖掘和分析新疆棉花事件中的主要主题和话题。该模块将对不同来源（官媒与网民论坛）中的文本数据进行深入的主题建模、关键词提取、话题演变分析，并提供多维度的可视化呈现。目标是揭示主要讨论议题、跟踪话题演变以及分析不同来源中的主题差异。

3.1 关键词提取与分析

TF-IDF（Term Frequency-Inverse Document Frequency）：
计算 TF-IDF：计算每个词汇在文本中的 TF-IDF 值，以评估关键词的重要性。
关键词排名：提取每个主题中的高 TF-IDF 值关键词，进行排名和分析。
关键词关联分析：
共现分析：计算关键词之间的共现频率，构建关键词共现网络。
网络可视化：使用 NetworkX 或 Gephi 等工具绘制关键词共现网络图，分析关键词的关联性。
词云生成：
主题词云：为每个主题生成词云图，直观展示主题中的重要关键词及其频率。
动态词云：使用 WordCloud 库生成动态词云，用户可以点击词云查看相关的文本和上下

文信息。

3.2 话题演变分析

时间序列分析：

主题时间序列图：绘制每个主题随时间变化的趋势图，分析不同话题在时间上的演变。

事件标记：在时间序列图中标记重要事件，分析事件对话题演变的影响。

热点话题分析：

热点检测：使用算法（如动态时间规整、趋势分析）检测话题热点及其变化。

热度图：绘制话题热度图，展示不同时间段和来源中的话题热度变化。

3.3 多维度对比分析

来源对比：

主题分布对比：对比官媒与网民论坛中的主题分布，识别不同来源中的主要讨论议题。

关键词对比：分析官媒与网民论坛中相同主题的关键词差异，揭示讨论焦点和观点差异。

情感与主题关联：

情感主题交互图：构建情感与主题的交互图，分析不同情感状态下的主题分布。

情感强度分析：结合情感分析结果，分析不同主题的情感强度变化。

3.4 高级数据可视化

主题趋势热力图：

热力图生成：使用热力图展示不同时间段内主题的热度和变化。

多维度热力图：结合来源（官媒与网民论坛）和语言进行多维度热力图展示。

交互式话题分析面板：

交互式界面：使用 Dash、Streamlit 等工具构建交互式分析面板，允许用户选择不同的主题、时间范围和数据源进行深入分析。

动态过滤：提供动态过滤选项，用户可以根据主题、情感、时间等条件筛选数据进行分析。

综合主题报告：

报告生成：自动生成包含主题分析结果的详细报告，包括主题分布、关键词提取、话题演变等内容。

报告可视化：在报告中嵌入图表、词云、热力图等可视化元素，以直观呈现分析结果。

4、多语种舆情分析软件详细模块及分析要点

4.1 数据收集模块

分析模块：

数据源配置：设置多语种数据源，包括社交媒体平台、新闻网站、论坛等。支持中、英、俄、法、阿拉伯语的数据抓取。

数据抓取与整合：使用爬虫技术或 API 接口抓取不同语言的数据，并将数据整合到统一的数据仓库中。

数据清洗与预处理：处理原始数据，去除噪音数据（如广告、垃圾信息），标准化不同语言的数据格式。

分析要点：

语言识别：自动识别和标记每条数据的语言。

数据量统计：各语言的数据量及其占比。

数据呈现：

数据收集概况：各语言数据量和来源的分布（条形图、饼图）。

数据质量报告：数据完整性、准确性评估（表格和柱状图）。

4.2 情感分析模块

分析模块：

情感模型训练：针对每种语言训练情感分析模型，包括情感分类（积极、消极、中立）和情感强度评估。

情感得分计算：使用训练好的模型对每条文本进行情感得分计算。

分析要点：

情感分布：各语言中不同情感类别的比例。

情感趋势：不同语言中情感得分随时间的变化。

情感词汇对比：各语言中情感词汇的分布和频率。

数据呈现：

情感分类图：各语言情感类别的比例（饼图）。

情感趋势图：各语言情感得分随时间的变化（折线图）。

情感词云：每种语言中的积极、消极关键词（词云图）。

4.3 主题分析模块

分析模块：

主题模型构建：使用 LDA（Latent Dirichlet Allocation）等主题模型对多语种数据进行主题提取。

主题标签生成：自动生成和标记各主题的标签。

分析要点：

主题分布：各语言中讨论的主要主题及其比例。

主题变化趋势：主题随时间的变化情况。

跨语言主题比较：不同语言中讨论主题的对比分析。

数据呈现：

主题分布图：各语言主题的比例（饼图或条形图）。

主题趋势图：各语言中主题随时间的变化（折线图）。

主题词云：各语言中主要主题的关键词（词云图）。

4.4 地理分布分析模块

分析模块：

地理信息提取：从多语言文本中提取和标记地理信息。

地理热力图生成：根据讨论量和情感分布生成地理热力图。

分析要点：

地理分布：各语言中讨论的地理分布情况。

地域情感分析：各地区的情感分布及其变化。

数据呈现：

地理热力图：各地区的讨论密度和情感分布（热力图）。

地域情感对比图：各地区情感的对比（地图标记图）。

4.5 用户分析模块

分析模块：

用户类型分类：按照用户类型（普通用户、专家、媒体等）进行分类。

用户贡献量统计：各用户类型的讨论量和互动量统计。

分析要点：

用户类型分布：各用户类型在不同语言中的分布。

用户互动分析：不同用户类型的互动量和情感贡献。

数据呈现:
用户类型分布图: 各语言中用户类型的比例（饼图或条形图）。
用户贡献量图: 各用户类型的讨论量和互动量（条形图）。

4.6 内容分析模块

分析模块:
关键词提取: 提取各语言中的主要关键词和短语。
关键词频率统计: 计算关键词在不同语言中的出现频率。
分析要点:
关键词分布: 各语言中主要关键词的分布情况。
关键词趋势: 关键词随时间的变化情况。
数据呈现:
关键词词云: 各语言中主要关键词的词云图。
关键词趋势图: 关键词在时间上的出现频率（折线图）。

4.7 综合分析模块

分析模块:
多语种对比分析: 比较不同语言中的情感、主题和用户贡献。
综合报告生成: 汇总所有分析结果，生成综合报告。
分析要点:
语言间差异: 不同语言中舆情分析结果的差异。
总体舆情概况: 综合各语言的讨论情况、情感分布和主题分析结果。
数据呈现:
综合对比图: 各语言分析结果的对比（条形图、雷达图）。
综合报告: 含所有分析结果的详细报告（PDF、HTML）。

以下为参考

上下文情感分析代码（基础功能仅供参照）：

```
import jieba

def is_positive_context(text, negative_word, window_size=5):
    """
    判断消极词汇是否出现在积极的上下文中。

    :param text: 输入文本
    :param negative_word: 消极词汇
    :param window_size: 用于判断上下文的窗口大小
    :return: 如果上下文是积极的，返回 True；否则返回 False
    """
    words = list(jieba.cut(text))

    if negative_word not in words:
        return False  # 如果消极词不在文本中，直接返回 False

    # 获取消极词汇的位置
    negative_word_idx = words.index(negative_word)

    # 定义积极词汇集
    positive_words = ['支持', '优秀', '认可', '进步', '成功', '积极']

    # 提取消极词汇的前后文
    start_idx = max(0, negative_word_idx - window_size)
    end_idx = min(len(words), negative_word_idx + window_size + 1)
    context = words[start_idx:end_idx]

    # 检查上下文中是否包含积极词汇
    for word in context:
        if word in positive_words:
            return True

    return False
```

应用情感转折检测（基础功能仅供参照）：

```
def has_sentiment_shift(text, negative_word):
    """
    检测消极词汇前是否存在情感转折词，如"但"、"虽然"等。
```

```
    :param text: 输入文本
    :param negative_word: 消极词汇
    :return: 如果有情感转折词，返回 True；否则返回 False
    """
    words = list(jieba.cut(text))

    # 定义情感转折词集
    shift_words = ['但', '虽然', '然而', '不过', '尽管']

    if negative_word in words:
        neg_idx = words.index(negative_word)
        # 检查消极词汇前是否有情感转折词
        if any(word in shift_words for word in words[:neg_idx]):
            return True

    return False
```

在整体情感分析中忽略特定消极词汇代码（基础功能仅供参照）：

```
def count_strings_in_content(row):
    content = str(row['正文'])
    words = list(jieba.cut(content))
    positive_strings = ['同胞', '抹黑', '支持国货', '坚持反对', '抵制', '棉不改色', '优质长绒棉', '舒适', '真相', '谎言', '世界顶级', '解约']
    negative_strings = ['强迫劳动', '拖欠工资', '工时长', '雇佣童工', '被迫', '压榨']

    positive_counts = sum(words.count(positive_string) for positive_string in positive_strings)
    negative_counts = 0

    for negative_string in negative_strings:
        if not is_positive_context(content, negative_string) and not has_sentiment_shift(content, negative_string):
            negative_counts += words.count(negative_string)

    return {"positive_counts": positive_counts, "negative_counts": negative_counts}
```

以上为基础版上下文分析代码，需要加入文本逻辑分析。如果缺乏文本分析，是无法进行舆情分析的，导致分析就不对

品牌声量的算法：

简单的品牌声量的算法，不是词频，而是转发数、曝光数、点赞数、踩数等的总和，而且分积极、消极

情感关键词的截距

问题：1. 目前的数据，我看品牌声量，统计是很困难的，本身也说明，原始数据有问题（有的网站爬不到转发数、曝光数、点赞数等信息，请问这要如何处理？）