目录

文档关键词的提取和文本相似度的计算

一、数据集的获取

数据集介绍

数据集使用搜狐新闻数据(SogouCS)Ver.2012，来自搜狐新闻 2012 年 6 月—7 月期间国内，国际，体育，社会，娱乐等 18 个频道的新闻数据，提供 URL 和正文信息。

数据集格式

<doc>
<url>页面 URL</url>
<docno>页面 ID</docno>
<contenttitle>页面标题</contenttitle>
<content>页面内容</content>
</doc>

二、数据的清洗

因为语料数据没有采用 utf-8 编码，这里我们使用 gb18030 编码方式，gb18030 拥有更广的中文编码，能够更好地读取语料数据。

我们数据集中的新闻数据进行遍历，将<content>页面内容</content>筛选出来，同时去除两边的<content>标识符，共筛选得到 1,367,845 个新闻语料。

接着我们对内容长度过少的语料数据进行清洗，规则是文本长度≤100。

接着再对新闻文本中的格式字符和 Unicode 字符编码进行处理，如\n、\u3000 和\ue40c 等特殊字符。

数据集语料数目	清洗后数据集语料数目
1,367,845	1,093,388

三、分词

分词我们选择 jieba 分词，停用词库我们选择了 stars 人数较多的中文常用停用词表（https://github.com/goto456/stopwords），再根据我们的新闻与了数据对停用词表进行适当的增删，以达到良好的分词效果。

节选的一片分词效果如下。

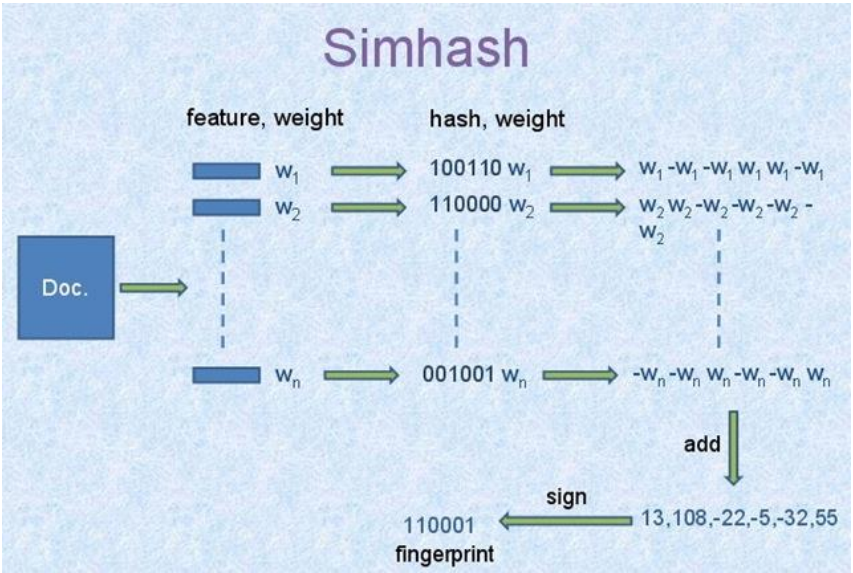
深圳市 法制办 网站 公布 深圳经济特区 社会 救助 条例 以下 简称 条例 再次 公开 征求 社会 意见 条例 增加 流浪 乞讨 人员 生活 救助 灾民 生活 救助 两章 内容 规定 救助站 应 乞讨 人员 提供 必需 食物 住处 医疗 救治 灾民 每人 每天 领取 元 基本 生活费 投靠 亲友 解决 住宿 每人每天 领取 元 补助 职业 行乞 进行 救助 教育 劝返 异地 分流 安置 条例 规定 流浪 乞讨 人员 实行 分类 救助 管理 无力解决 食宿 亲友 投靠 享受 最低 生活 保障 正在 流浪 乞讨 度日 生活 着落 自愿 求助 流浪 乞讨 人员 列为 求助 流浪 乞讨 人员 救助 管理 部门 给予 救助 护送 购票 返乡 二 十八岁 以下 未成年人 流浪 乞讨 行为 实行 保护性 救助 列为 未成年 流浪 乞讨 人员 此类 人员 及时 予以 进站 保护 三 乞讨 职业 流浪 乞讨 人员 列为 职业 流浪 乞讨 人员 救助 管理 部门 进行 救助 教育 劝返 异地 分流 安置 四 生命危险 流浪 乞讨 危重病 精神病人 列为 患病 流浪 乞讨 人员 先 救治 救助 原则 实行 人道主义 救治 民政 城管 公安 送往 定点医院 救治 五 违法犯罪 流浪 乞讨 人员 列为 有害 流浪 乞讨 人员 公安部门 依法 严厉 处置 打击 一时 处理 不了 无法 返乡 人员 进行 身份 核实 严加 管理 逐一 甄别 滞留 站 送 异地 安置 家属 抚慰金 每名 遇难者 低于 元 条例 规定 灾害 救助 采取 提供 救灾 生活 物资 基本 生活费 补助 临时 安置 住所 住宿 补助 医疗 救助 危机 心理咨询 服务 就业 服务 生产资料 救助 方式 自然灾害 救助 保障 灾民 基本 生活 需要 原则 困难 程度 每人每天 维持 基本 生活 费用 二十元 天 标准 二 住宿 补助 邻里 借住 投亲靠友 方式 解决 住宿 五十元 天 标准 补助 三 衣被 救助 保证 灾民 有衣 穿 天 冷 重点 救助 御寒衣 四 伤病 救助 伤病 家庭 困难 程度 适当 救助 五 遇难 人员 家属 抚慰金 每名 遇难者 低于 五千元 灾民 生活 特 困难 可予 临时 救助金 条例 提及 过渡性 安置 期 灾民 家庭 生活 特别 困难 人民政府 困难 程度 给予 一次性 临时 救助 补助金 本市 户籍 居民 持有 本市 居住证 深圳 连续 工作 生活 满 一年 申请 前 连续 本市 缴纳 养老保险费 超过 一年 本市 非 户籍 居民 遭受 疾病 意外事故 诉讼 失踪 死亡 突发 情况 致使 基本 生活 暂时 出现 较大 困难 申请 临时 救助 符合 临时 救助 条件 居民 遭遇 上述 突发 情况 申请 能力 无人 代理 申请 社会 救助 机构 了解 情况 后应 协助 申请 临时 救助 死亡 无遗 属 遗产 居民 应由 殡仪馆 办理 丧葬 事宜 财政 支付 丧葬费

四、相似度 simhash 算法与相似度的计算

相似度 simhash 算法

simhash 是由 Charikar 在 2002 年提出来的，论文名为《Similarity estimation techniques from rounding algorithms》。Google 基于 simhash 在海量网页中进行相似度计算并去重。通常对比两个文档是否相同时，会计算对应的 hash 值，常见的算法包括 md5 和 sha256。实际使用中，对于检测文档是否被篡改时，使用 hash 值具有不错的表现。但是当文档内容因为修改少许文字，插入广告甚至只是修改了标点符合和错别字，都会导致 hash 值改变，可是文档的核心内容并未发生改变。如何使用数学的方法表征这种文档相似性呢？simhash 的设计初衷就是使用一种所谓局部 hash 的方法，可以既可以敏感的识别文档的少许修改又可以识别出文档的大多数内容相同。

simhash 的一种典型实现就是将一个文档最后转换成一个 64 位的字节的特征字或者说 simhash 值，然后判断重复只需要判断他们的特征字的距离是不是小于 3，就可以判断两个文档是否相似。这个距离使用海明距离，即两个 simhash 值取异或后二进制中 1 的个数。大家可以结合自身业务特点修改 simhash 值的位数以及判断文档相似性的海明距离的值。



如图所示，计算 6 位 simhash 值典型的实现算法为：

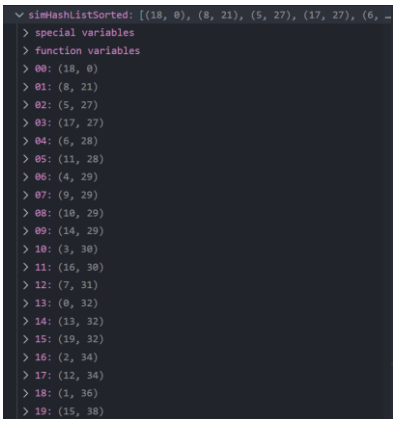
将 Doc 分词和计算权重，抽取出 n 个(关键词，权重)对，即图中的(feature, weight)。

计算关键词的 hash，生成图中的(hash,weight)，并将 hash 和 weight 相乘，这一过程是对 hash 值加权。

将 hash 和 weight 相乘的值相加，比如图中的[13, 108, -22, -5, -32, 55]，并最终转换成 simhash 值 110001，转换的规则为正数为 1 负数为 0。

相似度的计算

针对 1,093,388 条新闻语料，大范围的计算相似度需要耗费大量的时间，这显然是不可取的。我们采取的方法是，利用 random 库从 1,093,388 条新闻语料随机抽取 200 条新闻语料，这 200 条语料分别与 1,093,388 条新闻语料做相似度计算，记录下两两之间的新闻索引和计算得出的 score。



score 越低则代表着两条新闻越相似。我们将相似度最高的前 20000 条新闻文本以 txt 的形式保存下来，把他们的索引和相似度有 csv 格式存储。

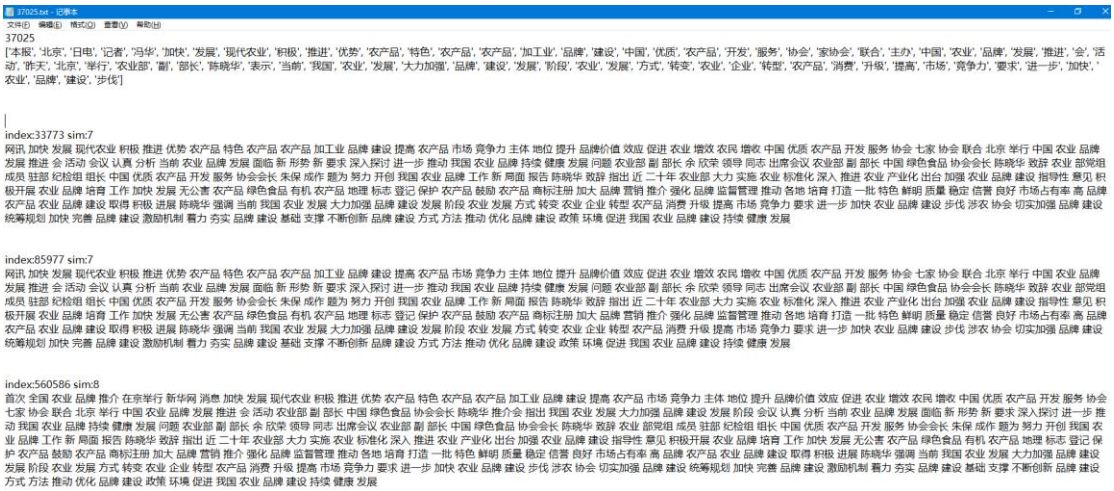
五、计算结果

相似度计算的结果由 3 部分组成，相似度 score 和新闻索引组成的 csv、新闻文本 txt、新闻和新闻的词云。

相似度和索引 csv

	index	sim
0	37025	0
1	33773	7
2	85977	7
3	560586	8
4	1012555	8
5	611862	10
6	262697	11
7	680565	11
8	143239	12
9	745701	12
10	1175	13
11	155568	13
12	281028	13
13	372854	13
14	419894	13
15	804060	13
16	983921	13
17	6183	14

新闻文本 txt



37025.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

37025

['本报', '北京', '日电', '记者', '冯华', '加快', '发展', '现代农业', '积极', '推进', '优势', '农产品', '特色', '农产品', '农产品', '加工业', '品牌', '建设', '中国', '优质', '农产品', '开发', '服务', '协会', '家协会', '联合', '主办', '中国', '农业', '品牌', '发展', '推进', '会', '活动', '昨天', '北京', '举行', '农业部', '副', '部长', '陈晓华', '表示', '当前', '我国', '农业', '发展', '大力加强', '品牌', '建设', '发展', '阶段', '农业', '发展', '方式', '转变', '农业', '企业', '转型', '农产品', '消费', '升级', '提高', '市场', '竞争力', '要求', '进一步', '加快', '农业', '品牌', '建设', '步伐', '涉农', '协会', '切实加强', '品牌', '建设', '统筹规划', '加快', '完善', '品牌', '建设', '激励机制', '看力', '夯实', '品牌', '建设', '基础', '支撑', '不断创新', '品牌', '建设', '方式', '方法', '推动', '优化', '品牌', '建设', '政策', '环境', '促进', '我国', '农业', '品牌', '建设', '持续', '健康', '发展']

index:33773 sim:7

网讯 加快 发展 现代农业 积极 推进 优势 农产品 特色 农产品 农产品 加工业 品牌 建设 提高 农产品 市场 竞争力 主体 地位 提升 品牌价值 效应 促进 农业 增效 农民 增收 中国 优质 农产品 开发 服务 协会 七家 协会 联合 北京 举行 中国 农业 品牌 发展 推进 会 活动 会议 认真 分析 当前 农业 品牌 发展 面临 新 形势 新 要求 深入探讨 进一步 推动 我国 农业 品牌 持续 健康 发展 问题 农业部 副 部长 余 欣荣 领导 同志 出席会议 农业部 副 部长 中国 绿色食品 协会会长 陈晓华 致辞 农业 部党组 成员 驻部 纪检组 组长 中国 优质 农产品 开发 服务 协会会长 朱保 成作 题为 努力 开创 我国 农业 品牌 工作 新 局面 报告 陈晓华 致辞 指出 近 二十年 农业部 大力 实施 农业 标准化 深入 推进 农业 产业化 出台 加强 农业 品牌 建设 指导性 意见 积极开展 农业 品牌 培育 工作 加快 发展 无公害 农产品 绿色食品 有机 农产品 地理 标志 登记 保护 农产品 鼓励 农产品 商标注册 加大 品牌 营销 推介 强化 品牌 监督管理 推动 各地 培育 打造 一批 特色 鲜明 质量 稳定 信誉 良好 市场占有率 高 品牌 农产品 农业 品牌 建设 取得 积极 进展 陈晓华 强调 当前 我国 农业 发展 大力加强 品牌 建设 发展 阶段 农业 发展 方式 转变 农业 企业 转型 农产品 消费 升级 提高 市场 竞争力 要求 进一步 加快 农业 品牌 建设 步伐 涉农 协会 切实加强 品牌 建设 统筹规划 加快 完善 品牌 建设 激励机制 看力 夯实 品牌 建设 基础 支撑 不断创新 品牌 建设 方式 方法 推动 优化 品牌 建设 政策 环境 促进 我国 农业 品牌 建设 持续 健康 发展

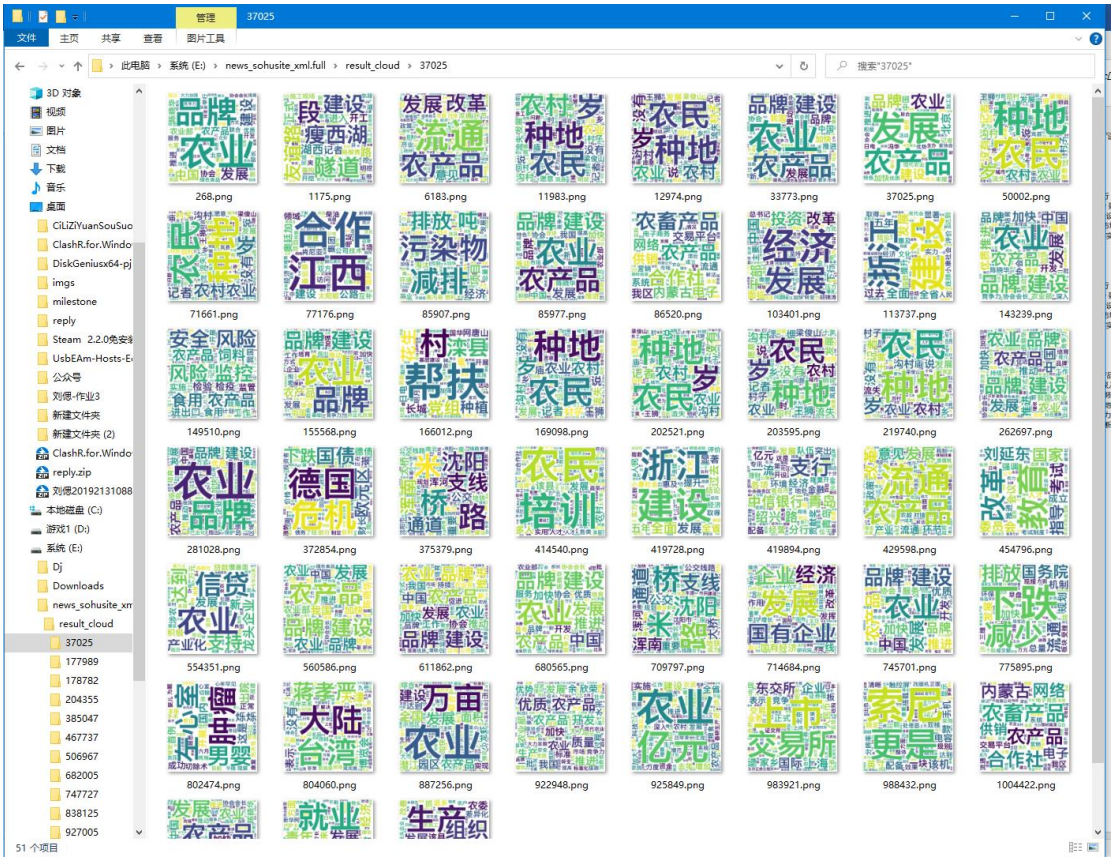
index:85977 sim:7

网讯 加快 发展 现代农业 积极 推进 优势 农产品 特色 农产品 农产品 加工业 品牌 建设 提高 农产品 市场 竞争力 主体 地位 提升 品牌价值 效应 促进 农业 增效 农民 增收 中国 优质 农产品 开发 服务 协会 七家 协会 联合 北京 举行 中国 农业 品牌 发展 推进 会 活动 会议 认真 分析 当前 农业 品牌 发展 面临 新 形势 新 要求 深入探讨 进一步 推动 我国 农业 品牌 持续 健康 发展 问题 农业部 副 部长 余 欣荣 领导 同志 出席会议 农业部 副 部长 中国 绿色食品 协会会长 陈晓华 致辞 农业 部党组 成员 驻部 纪检组 组长 中国 优质 农产品 开发 服务 协会会长 朱保 成作 题为 努力 开创 我国 农业 品牌 工作 新 局面 报告 陈晓华 致辞 指出 近 二十年 农业部 大力 实施 农业 标准化 深入 推进 农业 产业化 出台 加强 农业 品牌 建设 指导性 意见 积极开展 农业 品牌 培育 工作 加快 发展 无公害 农产品 绿色食品 有机 农产品 地理 标志 登记 保护 农产品 鼓励 农产品 商标注册 加大 品牌 营销 推介 强化 品牌 监督管理 推动 各地 培育 打造 一批 特色 鲜明 质量 稳定 信誉 良好 市场占有率 高 品牌 农产品 农业 品牌 建设 取得 积极 进展 陈晓华 强调 当前 我国 农业 发展 大力加强 品牌 建设 发展 阶段 农业 发展 方式 转变 农业 企业 转型 农产品 消费 升级 提高 市场 竞争力 要求 进一步 加快 农业 品牌 建设 步伐 涉农 协会 切实加强 品牌 建设 统筹规划 加快 完善 品牌 建设 激励机制 看力 夯实 品牌 建设 基础 支撑 不断创新 品牌 建设 方式 方法 推动 优化 品牌 建设 政策 环境 促进 我国 农业 品牌 建设 持续 健康 发展

index:560586 sim:8

首次 全国 农业 品牌 推介 在京举行 新华网 消息 加快 发展 现代农业 积极 推进 优势 农产品 特色 农产品 农产品 加工业 品牌 建设 提高 农产品 市场 竞争力 主体 地位 提升 品牌价值 效应 促进 农业 增效 农民 增收 中国 优质 农产品 开发 服务 协会 七家 协会 联合 北京 举行 中国 农业 品牌 发展 推进 会 活动 农业部 副 部长 中国 绿色食品 协会会长 陈晓华 推介会 指出 我国 农业 发展 大力加强 品牌 建设 发展 阶段 会议 认真 分析 当前 农业 品牌 发展 面临 新 形势 新 要求 深入探讨 进一步 推动 我国 农业 品牌 持续 健康 发展 问题 农业部 副 部长 余 欣荣 领导 同志 出席会议 农业部 副 部长 中国 绿色食品 协会会长 陈晓华 致辞 农业 部党组 成员 驻部 纪检组 组长 中国 优质 农产品 开发 服务 协会会长 朱保 成作 题为 努力 开创 我国 农业 品牌 工作 新 局面 报告 陈晓华 致辞 指出 近 二十年 农业部 大力 实施 农业 标准化 深入 推进 农业 产业化 出台 加强 农业 品牌 建设 指导性 意见 积极开展 农业 品牌 培育 工作 加快 发展 无公害 农产品 绿色食品 有机 农产品 地理 标志 登记 保护 农产品 鼓励 农产品 商标注册 加大 品牌 营销 推介 强化 品牌 监督管理 推动 各地 培育 打造 一批 特色 鲜明 质量 稳定 信誉 良好 市场占有率 高 品牌 农产品 农业 品牌 建设 取得 积极 进展 陈晓华 强调 当前 我国 农业 发展 大力加强 品牌 建设 发展 阶段 农业 发展 方式 转变 农业 企业 转型 农产品 消费 升级 提高 市场 竞争力 要求 进一步 加快 农业 品牌 建设 步伐 涉农 协会 切实加强 品牌 建设 统筹规划 加快 完善 品牌 建设 激励机制 看力 夯实 品牌 建设 基础 支撑 不断创新 品牌 建设 方式 方法 推动 优化 品牌 建设 政策 环境 促进 我国 农业 品牌 建设 持续 健康 发展

词云


管理
37025
文件
主页
共享
查看
图片工具
此电脑 › 系统 (E:) › news_sohusite_xml.full › result_cloud › 37025
搜索"37025"
3D 对象
视频
图片
文档
下载
音乐
桌面
CiLiZiYuanSouSuo
ClashR.for.Windo
DiskGeniusx64-pj
imgs
milestone
reply
Steam 2.2.0免安
UsbEAm-Hosts-E
公众号
刘傅-作业3
新建文件夹
新建文件夹 (2)
ClashR.for.Windo
reply.zip
刘傅20192131088
本地磁盘 (C:)
游戏1 (D:)
系统 (E:)
Dj
Downloads
news_sohusite_xm
result_cloud
37025
177989
178782
204355
385047
467737
506967
682005
747727
838125
927005
51 个项目
268.png
1175.png
6183.png
11983.png
12974.png
33773.png
37025.png
50002.png
71661.png
77176.png
85907.png
85977.png
86520.png
103401.png
113737.png
143239.png
149510.png
155568.png
166012.png
169098.png
202521.png
203595.png
219740.png
262697.png
281028.png
372854.png
375379.png
414540.png
419728.png
419894.png
429598.png
454796.png
554351.png
560586.png
611862.png
680565.png
709797.png
714684.png
745701.png
775895.png
802474.png
804060.png
887256.png
922948.png
925849.png
983921.png
988432.png
1004422.png