目录

[文档关键词的提取和文本相似度的计算 1](#_Toc74758574)

[一、数据集的获取 1](#_Toc74758575)

[数据集介绍 1](#_Toc74758576)

[数据集格式 1](#_Toc74758577)

[二、数据的清洗 2](#_Toc74758578)

[三、分词 2](#_Toc74758579)

[四、相似度simhash算法与相似度的计算 3](#_Toc74758580)

[相似度simhash算法 3](#_Toc74758581)

[相似度的计算 4](#_Toc74758582)

[五、计算结果 4](#_Toc74758583)

[相似度和索引csv 5](#_Toc74758584)

[新闻文本txt 5](#_Toc74758585)

[词云 6](#_Toc74758586)

# 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。

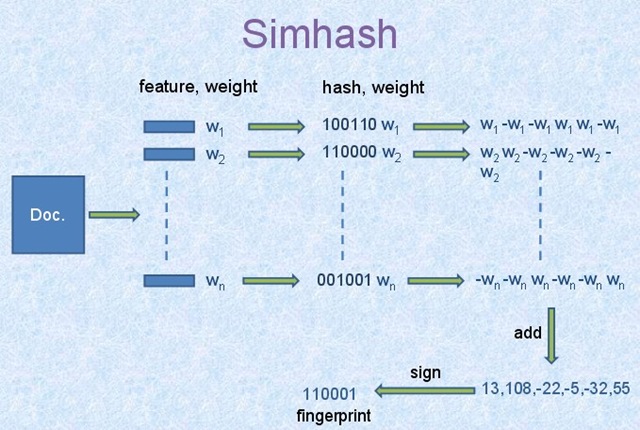
深圳市 法制办 网站 公布 深圳经济特区 社会 救助 条例 以下 简称 条例 再次 公开 征求 社会 意见 条例 增加 流浪 乞讨 人员 生活 救助 灾民 生活 救助 两章 内容 规定 救助站 应 乞讨 人员 提供 必需 食物 住处 医疗 救治 灾民 每人每天 领取 元 基本 生活费 投靠 亲友 解决 住宿 每人每天 领取 元 补助 职业 行乞 进行 救助 教育 劝返 异地 分流 安置 条例 规定 流浪 乞讨 人员 实行 分类 救助 管理 无力解决 食宿 亲友 投靠 享受 最低 生活 保障 正在 流浪 乞讨 度日 生活 着落 自愿 求助 流浪 乞讨 人员 列为 求助 流浪 乞讨 人员 救助 管理 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。

文本

描述已自动生成

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



### 新闻文本txt

图片包含 图形用户界面

描述已自动生成

### 词云

表格

描述已自动生成