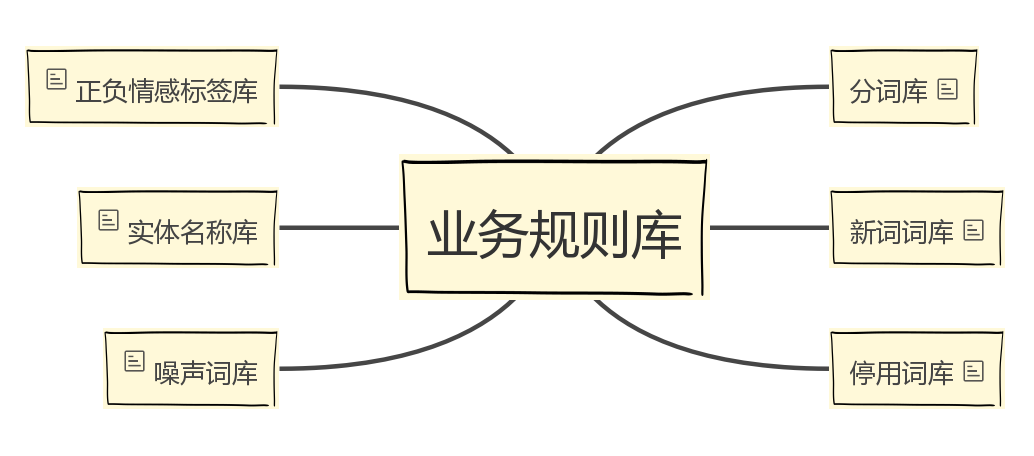
文本分析模型设计文档

1. **业务规则库设计**
   1. **业务规则库概述**

业务规则库涉及分词库、新词词库、停用词库、噪声词库、实体名称库以及正负面情感标签库。



* 1. **业务规则库说明**

|  |  |  |
| --- | --- | --- |
| **规则库** | **规则库主要属性** | **规则库说明** |
| 分词库 | 词语、词频、词性 | 用于文本分词。 |
| 新词词库 | 词语、词频（可选）、词性（可选） | 用于及时补充新颖词汇，提高文本分词的准确率和效果。需要频繁更新。 |
| 停用词库 | 停用词语、停用符号 | 用于在处理文本之前或之后会自动过滤掉某些字或词。 |
| 噪声词库 | 噪声词语 | 用于在处理文本之前或之后，根据用户的需求，过滤掉某些字或词。 |
| 实体名称库 | 股票名称、股票所属公司全称、公司简称、公私募公司名称 | 用于关联分析。 |
| 正负情感标签库 | 情感标签，情感分数 | 用于情感分析。 |

词性格式规范如下：

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **一类** | **二类** | **符号** |
| 1 | 名词 | 名词 | n |
| 2 | 人名 | nr |
|  | 汉语姓氏 | nr1 |
|  | 汉语名字 | nr2 |
|  | 日语人名 | nrj |
|  | 音译人名 | nrf |
|  | 地名 | ns |
|  | 音译地名 | nsf |
|  | 机构团体名 | nt |
|  | 其它专名 | nz |
|  | 名词性惯用语 | nl |
|  | 名词性语素 | ng |
|  | 处所词 | 处所词 | s |
|  | 方位词 | 方位词 | f |
|  | 时间词 | 时间词 | t |
|  | 时间词性语素 | tg |
|  | 动词 | 动词 | v |
|  | 副动词 | vd |
|  | 名动词 | vn |
|  | 动词“是” | vshi |
|  | 动词“有” | vyou |
|  | 趋向动词 | vf |
|  | 形式动词 | vx |
|  | 不及物动词（内动词） | vi |
|  | 动词性惯用语 | vl |
|  | 动词性语素 | vg |
|  | 形容词 | 形容词 | a |
|  | 副形词 | ad |
|  | 名形词 | an |
|  | 形容词性语素 | ag |
|  | 形容词性惯用语 | al |
|  | 区别词 | 区别词 | b |
|  | 区别词性惯用语 | bl |
|  | z 状态词 | 状态词 | z |
|  | 代词 | 代词 | r |
|  | 人称代词 | rr |
|  | 指示代词 | rz |
|  | 时间指示代词 | rzt |
|  | 处所指示代词 | rzs |
|  | 谓词性指示代词 | rzv |
|  | 疑问代词 | ry |
|  | 时间疑问代词 | ryt |
|  | 处所疑问代词 | rys |
|  | 谓词性疑问代词 | ryv |
|  | 代词性语素 | rg |
|  | 量词 | 量词 | q |
|  | 时量词 | qt |
|  | 动量词 | qv |
|  | d 副词 | 副词 | d |
|  | 介词 | 介词 | p |
|  | 介词“把” | pba |
|  | 介词“被” | pbei |
|  | 连词 | 连词 | c |
|  | 并列连词 | cc |
|  | 助词 | 助词 | u |
|  | 着 | uzhe |
|  | 了 喽 | ule |
|  | 过 | uguo |
|  | 的 底 | ude1 |
|  | 地 | ude2 |
|  | 得 | ude3 |
|  | 所 | usuo |
|  | 等 等等 云云 | udeng |
|  | 一样 一般 似的 般 | uyy |
|  | 的话 | udh |
|  | 来讲 来说 而言 说来 | uls |
|  | 之 | uzhi |
|  | 连 （“连小学生都会”） | ulian |
|  | 叹词 | 叹词 | e |
|  | 语气词 | 语气词(delete yg) | y |
|  | 拟声词 | 拟声词 | o |
|  | 前缀 | 前缀 | h |
|  | 后缀 | 后缀 | k |
|  | 字符串 | 字符串 | x |
|  | 非语素字 | xx |
|  | 网址URL | xu |
|  | 标点符号 | 标点符号 | w |
|  | 左括号，全角：（ 〔 ［ ｛ 《 【 〖 〈 半角：( [ { < | wkz |
|  | 右括号，全角：） 〕 ］ ｝ 》 】 〗 〉 半角： ) ] { > | wky |
|  | 左引号，全角：“ ‘ 『 | wyz |
|  | 右引号，全角：” ’ 』 | wyy |
|  | 句号，全角：。 | wj |
|  | 问号，全角：？ 半角：? | ww |
|  | 叹号，全角：！ 半角：! | wt |
|  | 逗号，全角：， 半角：, | wd |
|  | 分号，全角：； 半角： ; | wf |
|  | 顿号，全角：、 | wn |
|  | 冒号，全角：： 半角： : | wm |
|  | 省略号，全角：…… … | ws |
|  | 破折号，全角：—— －－ ——－ 半角：--- ---- | wp |
|  | 百分号千分号，全角：％ ‰ 半角：% | wb |
|  | 单位符号，全角：￥ ＄ ￡ ° ℃ 半角：$ | wh |
|  |  |  |  |

1. **文本分词模型设计**
2. **热词提取模型设计**
3. **词云展示模型设计**
4. **关联分析模型设计**