http://www.chinascope.com

新闻分析系统

用户手册

2019

ChinaScope 數庫

目录

数库新闻分析系统

NEWS

数库新闻分析系统说明

概要介绍

数库新闻分析系统是一个集新闻采集、分析、标签结果输出和数据存储为一体的综合性系统。

资讯覆盖面

（1） 数库利用自身在爬虫技术上的积累，对于中国大陆主流财经新闻媒体进行覆盖。目前已经覆盖 37 个核心财经媒体，100+个新闻媒体，1000+的新闻版面。
（2） 数库能够每 5 分钟对于站点进行一次抓取，保证新闻获取的及时性。
（3） 数库能够对外提供标准接口，方便客户接入自己的中文新闻源，进行新闻分析。

资讯分析维度

（1） 数库利用自身在 NLP 技术上的积累，使用多套自有专利技术对采集的新闻进行实时分析，并快速匹配 Smart-Tag。
（2） 数库 Smart-Tag 已经建立完整的标签体系，具体包括公司，行业，产品，概念，事件，人物，文章情感等七个维度，对于新闻给出立体化和科学化的分类。
（3） 数库 Smart-Tag 除了对文章情感给出情感标签，还能对文章中出现的公司主体和人物主体作出情感判断。
（4） 数库 Smart-Tag 能够提供标签热度，方便客户了解该标签的市场热度情况。
（5） 数库 Smart-Tag 能够提供版本值，版本值和生成算法的版本号一一对应，方便维护。

分析结果统计

数库资讯分析结果的统计数据显示，准缺率和召回率处于业界领先水平。

标签统计	公司	行业	产品	概念	事件	人物	情感
数量	275,250	115	2,868	241	1,821	437,876	0/1/2
准确率	89%	89.7%	90.4%	94.9%	95%	95%	84%
召回率	100%	98.67%	99.19%	93.07%	95%	87.69%	NA

特色优势

（1） 领先的中文自然语言处理技术。
（2） 实时捕捉中国市场动向。
（3） Smart-Tag 实现将新闻的分析结果-标签与公司信息披露形成的结构化数据有机关联。例如：新闻中的产品可以对应到主营该产品的相关上市公司。

推送及操作说明

数库可提供三种推送方式：FTP、AWS S3、客户端
新闻分析（标准版）文件格式 为 txt。内容格式为 JSON。

新闻字典表

标签基础表-概念字典表

表名：[concept_dictionary]-概念字典表

该表记录了概念的名称及编码，概念一般是由新闻中的热点事件而来。

字段	说明	描述
operation	记录标识	记录标识（A:新增、U:更新：D 删除）
code	概念编码	数库自定义的概念标准代码
name	概念名称	概念的名称
name_en	概念名称_英文	概念对应的英文名称
concept_id	概念记录 id	概念的记录 id，唯一标识

案例：可以看到“工业富联”是与“工业 4.0”概念匹配的一家 A 股上市公司

字段	说明	案例
operation	记录标识	A
code	概念编码	CP0058
name	概念名称	工业 4.0
name_en	概念名称_英文	Industry 4.0 Concept
concept_id	概念记录 id	58

标签基础表-概念成分股表

表名：[concept _stock] -概念成份股表

该表记录了每个概念包含的股票，目前股票为 A 股市场的股票。

字段	说明	描述
operation	记录标识	记录标识（A:新增、U:更新：D 删除）
code	证券代码	数库自定义的股票代码 唯一标识：code+ concept_code
name	证券简称	概念成分股的股票简称
name_en	证券简称_英文	概念成分股的股票英文简称
concept_code	概念编码	数库自定义的概念标准代码 关联 [concept_dictionary] “code”

案例：可以看到“工业富联”是与“工业 4.0”概念匹配的一家 A 股上市公司

字段	说明	案例
operation	记录标识	A
code	证券代码	601138_SH_EQ
name	证券简称	工业富联
name_en	证券简称_英文	Foxconn Industrial
concept_code	概念编码	CP0058

标签基础表-事件字典表

数库的事件指中国企业经营和资本市场上的活动行为。数库主要根据中国证监会发布的《上市公司信息披露管理办法》以及交易所给出的指引细则，将企业披露公告划分 11 种一级分类，109 种二级分类，以及 964 种三级分类，更加准确的定义企业活动行为，例如：一级分类为重组，二级分类为重组方法，三级分类为重组提议；其次根据各个行业关心的核心指标变动提供了诸如人民币贬值之类的 857 中非公告事件类型。

表名：[event_dictionary]-事件字典表

该表记录了事件的名称及层级关系。

字段	说明	描述
operation	记录标识	记录标识（A:新增、U:更新：D 删除）
code	事件编码	数库自定义的事件编码，唯一标识
name	事件名称	事件名称
name_en	事件名称_英文	事件英文名称
parent_code	事件父级	事件的直属父层级
level	事件层级	事件的层级

案例：“A 股上市”对应的事件分类为“IPO 事项”

字段	说明	案例
operation	记录标识	A
code	事件编码	CA001001
name	事件名称	A 股上市
name_en	事件名称_英文	A-Share Listing
parent_code	事件父级	CA001
level	事件层级	4

标签基础表-地区字典表

表名：[region_dictionary]-地区字典表

该表记录了中国地区的名称及层级关系。

字段	说明	描述
operation	记录标识	记录标识（A:新增、U:更新：D 删除）
code	地区编码	数库自定义的事件编码，唯一标识
name	地区名称	地区名称
name_en	地区名称_英文	地区英文名称
level	层级	地区层级
parent_code	地区直属父级	地区的直属父层级
ancestors	地区所有父层级	地区的所有父层级

新闻分析

数库新闻分析结果会按照统一的文件格式存储和传输。用户能够方便的解析文件，获得详细的信息。

文件格式：

属性名称	子属性	数据类型	是否必须	格式	备注
NewsResult		Object			
	newsInfo	NewsInfo	Y		新闻基本信息
	newsTags	Array	Y		分析出的新闻标签
	emotionInfos	Array	Y		分析出的舆情结果
NewsInfo		Object			
	newsId	String	Y		新闻 Id
	newsTitle	String	Y		当 newsTitle_cn 有值时，显示新闻英文标题 当 newsTitle_cn 为空或 null 时，显示新闻中文标题
	newsTitle_cn	String	Y		新闻中文标题
	newsTs	Date	Y	YYYY-MM-DDThh:mm:ssZ	新闻时间戳
	newsFetchTs	Date	Y	YYYY-MM-DDThh:mm:ssZ	新闻抓取时间戳
	newsOriginalTs	Date	N	YYYY-MM-DDThh:mm:ssZ	新闻披露时间
	newsUrl	String	Y		新闻 Url

	newsSource	String	N		新闻源
	newsSummary	String	N		新闻摘要
	whitelistFlag	Integer	N		是否加入白名单，1- 是，0-否
	newsExtId	String	N		新闻内部 Id
emotionInfos	EmotionInfos	Array			
EmotionInfos		Object			
	emotionEntity	String	Y		舆情的实体类别
	entityRefId	String	N		舆情的实体 Id
	entityName_cn	String	Y		舆情的中文实体名称
	entityName	String	Y		舆情的英文实体名称
	entityCode	String	N		舆情的实体编码
	emotionIndicator	Integer	Y		舆情指标: 0 中性 1 正面 2 负面
	emotionWeight	Double	Y		舆情指标权重
	emotionDetail	String	Y		三类舆情指标对应的权重
	emotionAlgoVersion	String	N		算法版本
newsTags	NewsTag	Array	Y		新闻分析结果
NewsTag		Object	Y		
	itemType	Enum	Y		标签类型: Company - 公司 People - 人物 Industry - 行业 Product - 产品 Event - 事件 Concept - 概念 Region-地区 Event_entity-事件-实体关系

	itemName_cn	String	Y		标签对应类型的中文名称
	itemName	String	Y		标签对应类型的英文名称
	itemHotIndex	Long	N		标签热度，是否存在取决于算法
	itemAlgoVersion	String	N		算法版本
	itemId	String	N		标签对应类型的内部 id
	ItemExtId	String	N		标签对应类型的外部 id
	ItemDesc	String	N		标签的描述
	ItemRelevance	Double	N	0-1 之间	标签的相关度
	itemRelevanceType	int	N	0/1/2	标签重要程度: 1:所有的公司 0:比较相关的公司 2 ： 最相关的公司
	itemExtType	String			事件标签的相关主体类别： Company：公司 Person：人物

案例：例如我们可以查询到 2018 年 11 月 9 日东方财富关于工业富联的一篇报道。对于该报道新闻分析的输出文件提供了：

（1） 新闻标题，时间，URL，摘要等基本信息；
（2） 数库分析的各个维度的标签结果，包含了：“富士康工业互联网股份有限公司”，“互联网软件与服务”，“云计算服务”，“工业 4.0”，“大数据”，“李军旗”，“李杰”。
（3） 新闻，公司，人物三个舆情实体的情感标签和权重值。
（4） 通过标签的“itemId” 代码能够关联字典表获取详细信息。


```
{
    "newsInfo": {
    "newsId": "217536",
    "newsTitle": "China Telecom Acquisition and Cataloguing Subsidiaries Pave the Way for Restructuring and Transforming",
    "newsTitle_cn": "国电信收编子公司 为重组转型铺路",
    "newsTs": "2018-11-09T14:25:17Z",
    "newsFetchTs": "2018-11-09T1015:05:17Z",
    "newsUrl": " http://finance.eastmoney.com/a/20181109981690445.html ",
    "newsSource": "东方财富网",
```

```
    "newsSummary": "11 月 9 日，工业富联(601138)在深圳召开 2018 年第三次临时股东大会。采用现场投票的方式，表决通过了选举李杰、吴惠锋为公司董事的议案。至此，由公司董事长李军旗、副董事长李杰、吴惠锋、内部董事兼总经理郑弘孟及独立董事薛健、孙中亮组成的工业富联新一届董事会正式亮相。在股东大会上，针对公司当前面临的机遇和挑战，董事会成员做出详细回应。工业富联此次新增选的两位董事中，李杰是知名工业大数据专家，现任美国辛辛那提大学特聘讲座教授，美国国家科学基金会智能维护系统产学合作中心创始主任，着有《工业大数据》、《从大数据到智能制造》、《CPS》以及《云上的工业智能》等著作，曾在 2016 年被美国制造学会选为美国 30 位最有远见的智能制造人物。",
    "newsSimHash": "5be59e7d83a7940e3a2a8c5d ",
    "newsRefLocation": null
  },
  "emotionInfos": [
    {
      "emotionEntity":"News",
      "entityRefId":217536,
      "emotionIndicator":1,
      "emotionWeight": 0.92,
      "emotionDetail ": {0=0.0723, 1=0.9201, 2=0.0076}
    },
    {
      "emotionEntity":"Company",
      "entityRefId": "CSF0000124539",
      "entityName_cn":"富士康工业互联网股份有限公司",
      "entityName":"Foxconn Industrial Internet Co., Ltd.",
      "entityCode":"601138_SH_EQ" ,
      "emotionIndicator":1,
      "emotionWeight": 0.70,
      "emotionDetail ": {0=0.0605, 1=0.70, 2=0.2395}
    },
    {
      "emotionEntity":"People",
      "entityRefId": "P13151560" ,
      "entityName_cn":"李军旗",
      "entityName":"Li Junqi",
      "emotionIndicator":1,
      "emotionWeight": 0.85,
      "emotionDetail ": {0=0.0505, 1=0.8503, 2=0.0992}
    },
    {
      "emotionEntity":"People",
      "entityRefId": "P13155891" ,
      "entityName_cn":"李杰",
      "entityName":"Li Jie",
      "emotionIndicator":1,
      "emotionWeight": 0.76,
      "emotionDetail ": {0=0.0501, 1=0.7654, 2=0.1845}
    },
    {
```

```
        "itemType":"Concept",
        "itemName_cn": "大数据",
        "itemName":"Big Data Concept" ,
        "itemId":"31"
    },
    {
        "itemType":" Concept",
        "itemName_cn": "工业 4.0",
        "itemName":"Industry 4.0 Concept" ,
        "itemId":"122"
    },
    {
        "itemType":"Company",
        "itemName_cn": "富士康工业互联网股份有限公司",
        "itemName": "Foxconn Industrial Internet Co., Ltd.",
        "itemId":"CSF0000005764"
    },
    {
        "itemType":"Industry",
        "itemName_cn": "互联网软件与服务",
        "itemName ": "Internet Software and Services",
        "itemId":"CSF_45101010"
    },
    {
        "itemType":"Product",
        "itemName_cn": "云计算服务",
        "itemName":"Cloud Computing Services" ,
        "itemId":"IT001004"
    }
        ]
}
```


新闻分析(标签区分)

新闻公司标签表

表名：[news_company_label]-新闻公司标签表

该表记录的是公司标签结果。

属性名称	子属性	数据类型	是否必须	格式	备注
	stockCode	String	Y		股票代码
	companyId	Array	Y		公司 id
	chineseName	String	Y		公司名称
	englishName	String	Y		公司名称_英文
	newsId	String	Y		新闻 Id
	newsTs	Date	Y	YYYY-MM-DDThh:mm:ssZ	新闻时间戳
	Relevance	Double	N	0-1 之间	标签的相关度
	itemRelevanceType	int	N	0/1/2	标签重要程度: 1:所有的公司 0:比较相关的公司 2:最相关的公司
	emotionIndicator	Integer	Y		舆情指标: 0 中性 1 正面 2 负面
	emotionWeight	Double	Y		舆情指标权重
	emotionDetail	String	Y		三类舆情指标对应的权重

新闻事件标签表

表名：[news_event_label]-新闻事件标签表

该表记录的是事件标签结果。

属性名称	子属性	数据类型	是否必须	格式	备注
	eventCode	String	Y		事件代码

	chineseName	String	Y		事件名称
	englishName	String	Y		事件名称_英文
	newsId	String	Y		新闻 Id
	newsTs	Date	Y	YYYY-MM-DDThh:mm:ssZ	新闻时间戳
	emotionIndicator	Integer	Y		新闻舆情指标: 0 中性 1 正面 2 负面
	emotionWeight	Double	Y		新闻舆情指标权重
	emotionDetail	String	Y		三类舆情指标对应的权重

新闻行业标签表

表名：[news_industry_label]-新闻行业标签表

该表记录的是行业标签结果

属性名称	子属性	数据类型	是否必须	格式	备注
	industryCode	String	Y		行业代码
	chineseName	String	Y		行业名称
	englishName	String	Y		行业名称_英文
	newsId	String	Y		新闻 Id
	newsTs	Date	Y	YYYY-MM-DDThh:mm:ssZ	新闻时间戳
	Relevance	Double	N	0-1 之间	标签的相关度
	emotionIndicator	Integer	Y		新闻舆情指标: 0 中性 1 正面 2 负面
	emotionWeight	Double	Y		新闻舆情指标权重
	emotionDetail	String	Y		三类舆情指标对应的权重

新闻产品标签表

表名：[news_product_label]-新闻产品标签表

该表记录的是产品标签结果

属性名称	子属性	数据类型	是否必须	格式	备注
	productCode	String	Y		产品代码
	chineseName	String	Y		产品名称
	englishName	String	Y		产品名称_英文
	newsId	String	Y		新闻 Id
	newsTs	Date	Y	YYYY-MM-DDThh:mm:ssZ	新闻时间戳
	Relevance	Double	N	0-1 之间	标签的相关度
	emotionIndicator	Integer	Y		舆情指标: 0 中性 1 正面 2 负面
	emotionWeight	Double	Y		新闻舆情指标权重
	emotionDetail	String	Y		三类舆情指标对应的权重

新闻概念标签表

表名：[news_concept_label]-新闻概念标签表

该表记录的是概念标签结果

属性名称	子属性	数据类型	是否必须	格式	备注
	conceptCode	String	Y		概念代码
	chineseName	String	Y		概念名称
	englishName	String	Y		概念名称_英文
	newsId	String	Y		新闻 Id
	newsTs	Date	Y	YYYY-MM-DDThh:mm:ssZ	新闻时间戳
	Relevance	Double	N	0-1 之间	标签的相关度
	emotionIndicator	Integer	Y		舆情指标: 0 中性 1 正面 2 负面
	emotionWeight	Double	Y		新闻舆情指标权重
	emotionDetail	String	Y		三类舆情指标对应的权重

新闻主体表

表名：[news_info]-新闻主体表

该表记录的是新闻标题、新闻摘要、来源及舆情信息的表。

属性名称	子属性	数据类型	是否必须	格式	备注
	newsId	String	Y		新闻 Id
	newsTitle	String	Y		当 newsTitle_cn 有值时，显示新闻英文标题 当 newsTitle_cn 为空或 null 时，显示新闻中文标题
	newsTitle_cn	String	Y		新闻中文标题
	newsTs	Date	Y	YYYY-MM-DDThh:mm:ssZ	新闻时间戳
	newsOriginalTs	Date	N	YYYY-MM-DDThh:mm:ssZ	新闻披露时间
	newsUrl	String	Y		新闻 Url
	newsSource	String	N		新闻源
	newsSummary	String	N		新闻摘要
	whitelistFlag	Integer	N		是否加入白名单，1- 是，0-否
	newsExtId	String	N		新闻内部 Id
	emotionIndicator	Integer	Y		舆情指标: 0 中性 1 正面 2 负面
	emotionWeight	Double	Y		新闻舆情指标权重

	emotionDetail	String	Y		三类舆情指标对应的权重

新闻事件-实体表

表名：[news_event_entity]-新闻事件-实体表

该表记录的是新闻中关于事件和人物、公司之间的关系。

属性名称	子属性	数据类型	是否必须	格式	备注
	eventCode	String	Y		事件代码
	chineseName	String	Y		事件名称
	englishName	String	Y		事件名称_英文
	newsId	String	Y		新闻 Id
	newsTs	Date	Y	YYYY-MM-DDThh:mm:ssZ	新闻时间戳
	itemExtType	String			事件标签的相关主体类别： Company：公司 Person：人物
	ItemExtId	String	N		对于事件标签来说，是itemExtType对应的实体id
	emotionIndicator	Integer	Y		舆情指标： 0 中性 1 正面 2 负面
	emotionWeight	Double	Y		新闻舆情指标权重
	emotionDetail	String	Y		三类舆情指标对应的权重

新闻人物标签表

表名：[news_people_label]-新闻人物标签表

该表记录的是新闻中人物标签结果。

属性名称	子属性	数据类型	是否必须	格式	备注
	personCode	String	Y		人物代码
	chineseName	String	Y		人物名称
	englishName	String	Y		人物名称_英文
	newsId	String	Y		新闻 Id
	newsTs	Date	Y	YYYY-MM-DDThh:mm:ssZ	新闻时间戳
	emotionIndicator	Integer	Y		舆情指标: 0 中性 1 正面 2 负面
	emotionWeight	Double	Y		人物舆情指标权重
	emotionDetail	String	Y		三类舆情指标对应的权重

新闻地区标签表

表名：[news_region_label]-新闻地区标签表

该表记录的是新闻中地区标签结果。

属性名称	子属性	数据类型	是否必须	格式	备注
	RegionCode	String	Y		地区代码
	chineseName	String	Y		地区名称
	englishName	String	Y		地区名称_英文
	newsId	String	Y		新闻 Id
	newsTs	Date	Y	YYYY-MM-DDThh:mm:ssZ	新闻时间戳
	emotionIndicator	Integer	Y		舆情指标: 0 中性 1 正面 2 负面
	emotionWeight	Double	Y		新闻舆情指标权重
	emotionDetail	String	Y		三类舆情指标对应的权重

JSON 格式新闻 PYTHON 解析

字段信息介绍

数库新闻分析结果会按照统一的文件格式存储和传输。为使用户能够方便的解析文件，将数据传输的格式和 python 类进行了映射，关系如下：

Class NewsResult()类：对应的是一条完整数据，一条完整的数据下面有三部分，分别是新闻基本信息，分析出的舆情结果,分析出的新闻标签，分别抽象为 python 的三个类：

class NewsInfo():新闻基本信息

class EmotionInfo():分析出的舆情结果

class NewsTag():分析出的新闻标签

这三个类对应的具体字段参考智能新闻分析用户手册_V3.4.pdf 中的文件格式模块对应的内容

代码业务逻辑介绍

首先是读取文件，并遍历每一行数据：

1.with open(file_path, 'r', encoding='utf-8') as fp:
    for num,i in enumerate(fp):

2.读取文件之后，先将每一行的数据转换成 python 对象

```
rawNews = json.loads(i.strip(),strict=False)
```

3.然后将数据的三部分内容和 python 的三个类 class NewsInfo(), class EmotionInfo() 和 class NewsTag()进行关系映射：

```
newsInfo = NewsInfo(**(rawNews.get("newsInfo")))
tags = []
if rawNews.get('newsTags', []):
    for tag in rawNews.get('newsTags'):
        newstag = NewsTag(**(tag))
        tags.append(newstag)
emotions = []
if rawNews.get("emotionInfos"):
    for emotion in rawNews.get("emotionInfos"):
        newemotion = EmotionInfo(**(emotion))
        emotions.append(newemotion)
```

4.最后得出的 Class NewsResult()就是单条数据所对应的 python 类

```
newsResult = NewsResult(newsInfo, tags, emotions)
```

使用说明

PYTHON 环境安装

1. 下载 python(https://www.python.org/ python 官网)，下载与自己电脑对应的 python 版本即可（python 必须是 python3.4+以上版本）

   Windows 上安装 python 参考：

   https://blog.csdn.net/mingzhuo_126/article/details/81239156

   mac 上安装 python 参考：

   https://blog.csdn.net/xuanlv_haoshao/article/details/82316766

   linux 上安装 python 参考：

   https://www.cnblogs.com/xinjinfu/p/7887255.html

代码的使用方法

1. 代码中的 file_path 是需要解析的文件路径，该路径既可以是绝对路径，也可以是相对路径，运行不同的文件时，只需要修改该路径即可。

```
if __name__ == '__main__':
    file_path = 'test.json'
    main(file_path)
```

2. 添加自己的业务逻辑：添加自己的业务逻辑的时候，只需要对 newsResult 进行操作即可，newsResult 是根据单条新闻格式映射的 python 类，通过获取 newsResult 的属性即可获取新闻的相关字段：如 newsResult.emotionInfos[0].emotionEntity
   newsResult.newsInfo ,等

```
newsResult = NewsResult(newsInfo, tags, emotions)
## TODO，  you can put your logic here
print(newsResult)
```

3.代码的运行：

Windows:

打开 cmd 命令窗口然后切换到 NewsFileReader.py 所在的文件，执行：

python NewsFileReader.py 即可运行文件

mac, linux:

打开终端，切换到 NewsFileReader.py 所在的文件，执行：

python NewsFileReader.py 即可运行文件

联系我们

电话：（86）021-36359360-511

邮箱：business@chinascope.com

ChinaScope 数庫