

大数据解决方案

# TEXTOM MANUAL

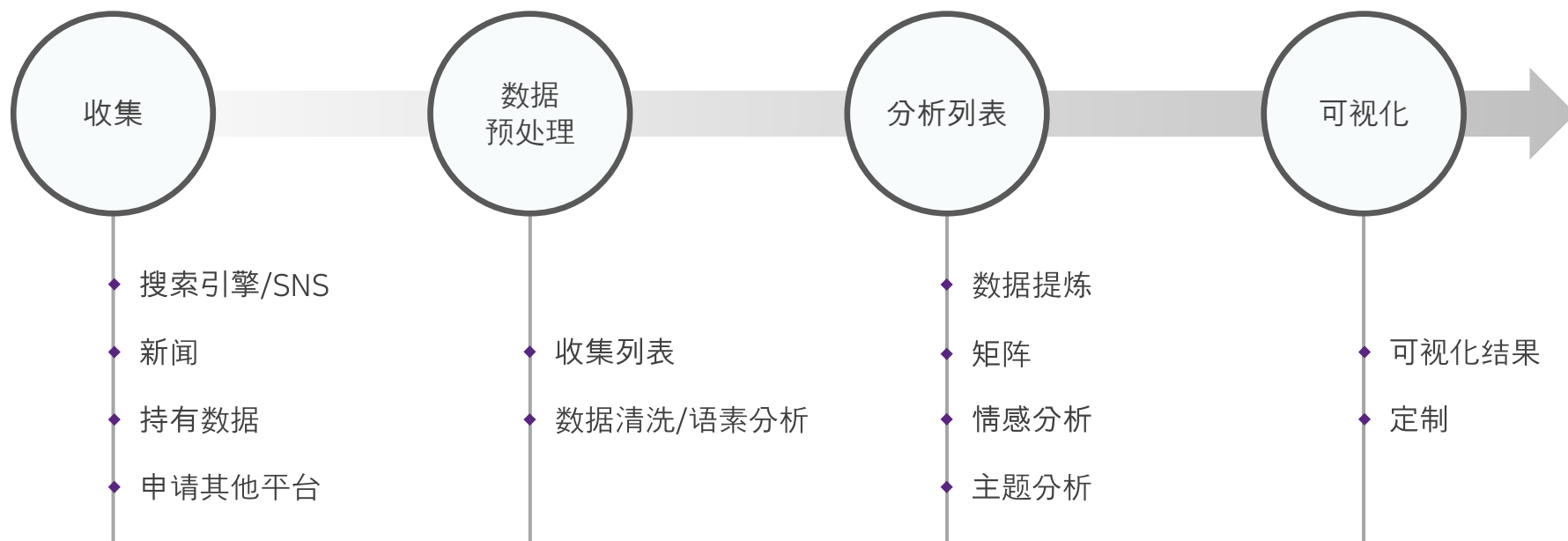
从大数据收集到数据可视化!

v5.0

钛思通  
TEXTOM

# 一目了然的TEXTOM

庞大又复杂的文本资料，  
通过钛思通进行高效地分析  
并在研究、营销、舆论分析等多个领域中使用！



经过优化，TEXTOM在chrome浏览器的环境下运行时可以有更快更好的体验。

# Contents

1. [注册 / 登录](#)
2. [收集数据](#)
3. [数据预处理](#)
4. [数据分析列表](#)
5. [数据可视化](#)

按下每个主题的话，移动每段的第一页。

# 注册 / 登录

钛思通  
TEXTOM

About

Manual

Communication

Partners

TEXTOM Edu

利用TEXTOM教育版的无限数据容量进行快速又简单的数据分析教育课程。

钛思通

快速分析大数据，实现可视化  
获取多样的见解并灵活运用。

TEXTOM

可以对韩国多种平台上的数据进行收集、分析以及可视化。

## 大数据一贯的解决方案, 钛思通

大数据的一贯解决方案 提供数据收集,  
数据清洗, 数据矩阵, 可视化



Collecting



Storage



Cleaning



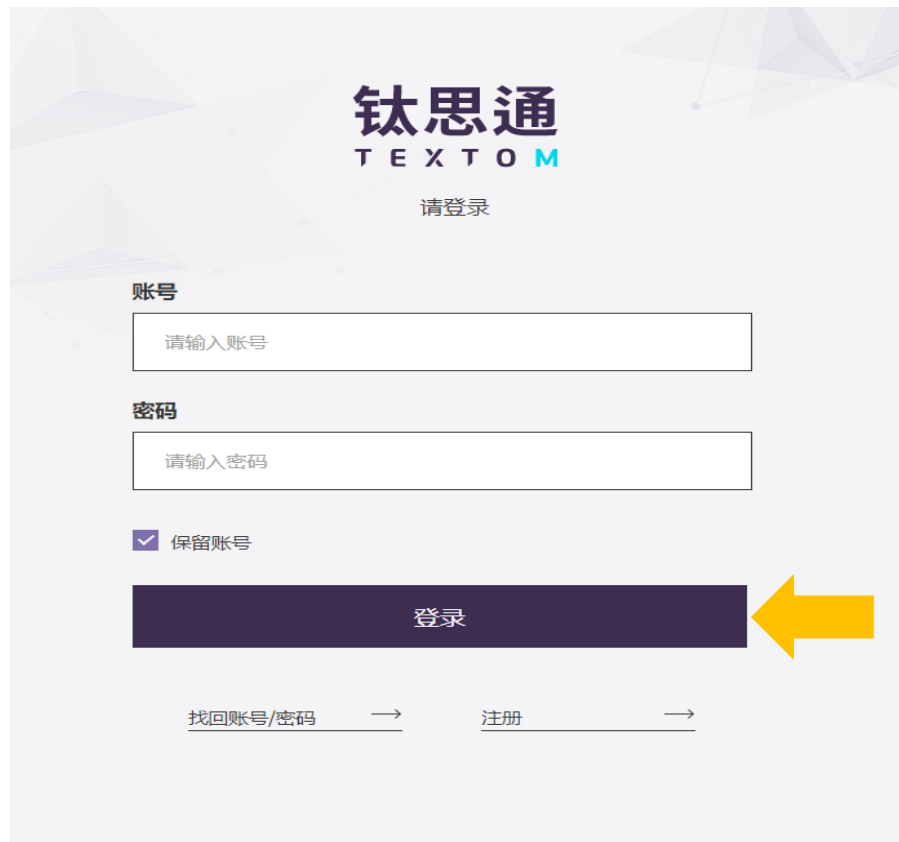
Matrix



Visualization

点击**钛思通**后移动至注册&登录界面。

# 注册 / 登录



钛思通  
TEXTOM

请登录

账号

请输入账号

密码

请输入密码

☒ 保留账号

登录

[找回密码/密码](#) → [注册](#) →



钛思通  
TEXTOM

더아이엠씨 您好!

登录

[钛思通管理](#) → [MY PAGE](#) →

[LOGOUT](#) →

注册成功后，登录后即可使用解决方案。

以工作日为准，注册许可需要1~3日。

# 登录的第一页

1. 收集数据

钛思通  
TEXTOM

≡ 收集

容量充值 尊敬的 더아이엠씨

收集

数据预处理  
数据列表  
提炼/语义分析

分析列表  
数据提炼  
矩阵  
凝聚子群分析  
情感分析  
主题分析

可视化  
可视化结果  
定制

搜索引擎/SNS 持有数据 请求频道

关键词

添加关键词

算法

初始化

时间跨度

一周

三个月

一年

初始化

平台

收集信息

Baidu

全部

网络文件

新闻

所有学术信息

cnki

全部

文献

期刊

博硕士

报纸

图书

Google

全部

网页

学术

人民网

全部

新浪微博

全部

WeChat

全部 (公众号)

万方数据

全部

在收集功能界面可收集要分析的数据或也可上传持有数据。

6



- ① **搜索引擎/SNS** 可收集百度、中国知网、谷歌、人民网、微博、微信万方数据的数据。
- ② **持有数据** 可上传内容为文字的pdf, txt, xlsx形式的文件。
- ③ **申请其他平台** 除了在钛思通提供的平台之外，需要收集其他平台的数据的话可另外付费使用服务。

※ 数据收集阶段不会扣除流量。

三

收集

容量充值 尊敬的 더아이엠씨

9

搜索引擎/SNS

持有数据

请求频道

关键词

1

添加关键词

算法

初始化

添加关键词可以一次性生成多个收集列表(收集条件相同)

时间跨度

2

~

一周

三个月

一年

初始化

百度新闻, 人民网, 万方数据平台不能设置时间跨度。

平台

3

收集信息

Baidu

全部

网络文件

新闻

所有学术信息

cnki

全部

文献

期刊

博硕士

报纸

图书

Google

全部

人民网

全部

新浪微博

全部

全部 (公众号)

万方数据

全部

### ① 关键词

输入关键词开始收集数据。选择所需的频道进行关键词收集，确认所收集数据信息是否正确

※添加关键词-在同样的条件（时间，收集单位，平台）下，同时收集多个关键词的情况。

### ② 时间跨度

设定数据生成的时间跨度。

#### <各频道数据采集量>

| 频道名  | 收集量       |
|------|-----------|
| 百度   | 最大 2,000件 |
| 谷歌   | 最大 1,000件 |
| 微博   | 最大 2,100件 |
| 微信   | 最大 2,000件 |
| 知网   | 最大 200件   |
| 万方数据 | 最大 320件   |
| 人民日报 | 最大 1,100件 |

### ③ 平台

请选择收集频道或区域各频道收集的数据内容存在差异。各频道收集的数据内容可通过“各频道收集信息”确认。

※ 设定结束后，请点击 开始收集 → 按钮。在点击的同时数据收集开始进行。

※：※ 收集所需的时间平均为30分钟左右,根据数据收集量和用户数可能会存在差异。



# 上传持有数据



❶ 将持有数据上传至TEXTOM上后，可进行数据清洗、分析、以及可视化。

※ 可适用的文件类型: txt, pdf, xls, xlsx

※ txt格式的文件需要以UTF-8格式上传 [UTF-8 编码设定方法](#) ▶

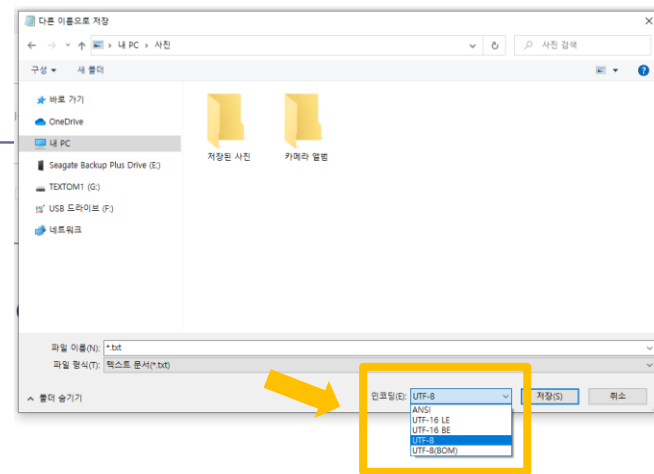
※ 不能识别pdf文件中包含的图片

❷, ❸ 文件拓展名为xls,xlsx的时候，请选择需要分析的行列。

※ 可进行多重选择，如若没有特定要选择项目则将会自动选择行列，直接选择时，复数的行列需要用逗号隔开 例) G,H

※ 设定结束后，请点击「开始收集」按钮。在点击的同时数据收集开始进行。

※ 上传所需的时间平均为10分钟左右,但根据数据容量不同可能会存在差异。上传完毕的保留数据可在“收集完成”页面查看。



钛思通 TEXTOM

收集完成

容量充值 尊敬的 더아이엠씨

关键词搜索 搜索结果 927 / 927

删除 提炼/语义分析 →

数据预处理

收集列表

提炼/语义分析

分析列表

数据提炼

矩阵

情感分析

主题分析

可视化

可视化结果

定制

搜索引擎/SNS 持有数据 请求频道

| <input type="checkbox"/> | 关键词 | 频道         | 持续时间                    | 收集日期       | 文件大小     | 收集状态 |
|--------------------------|-----|------------|-------------------------|------------|----------|------|
| <input type="checkbox"/> | 首尔  | Wechat     | 2011-01-21 ~ 2019-09-05 | 2019-09-05 | 19.39 MB | 收集完成 |
| <input type="checkbox"/> | 济州岛 | Wechat     | 2011-01-21 ~ 2019-09-05 | 2019-09-05 | 12.16 MB | 收集完成 |
| <input type="checkbox"/> | 首尔  | Sina weibo | 2009-08-01 ~ 2019-09-05 | 2019-09-05 | 6.27 MB  | 收集完成 |
| <input type="checkbox"/> | 济州岛 | Sina weibo | 2009-08-01 ~ 2019-09-05 | 2019-09-05 | 2.48 MB  | 收集完成 |
| <input type="checkbox"/> | 首尔  | Baidu(Web) | 2000-01-01 ~ 2019-09-05 | 2019-09-05 | 267 KB   | 收集完成 |
| <input type="checkbox"/> | 济州岛 | Baidu(Web) | 2000-01-01 ~ 2019-09-05 | 2019-09-05 | 274 KB   | 收集完成 |
| <input type="checkbox"/> | 韩国  | Sina weibo | 2019-06-04 ~ 2019-09-04 | 2019-09-04 | 1.31 MB  | 收集完成 |
| <input type="checkbox"/> | 吴亦凡 | Sina weibo | 2019-08-28 ~ 2019-09-04 | 2019-09-04 | 537 KB   | 收集完成 |

预览数据

点击用量即可预览部分收集的数据原文

吴亦凡

2019-08-28 ~ 2019-09-04

| 频道        | 分项   | 收集量(个) | 文件大小   |
|-----------|------|--------|--------|
| sinaweibo | 新浪微博 | 1,000  | 537 KB |

收集量可视化

提炼/语义分析 →

❶可以确认正在生成的收集列表若收集完成，则会在列表中消失，并在“收集完成”中出现

❷可查看已收集完成的数据列表

❸可确认各平台和项目的收集量和容量。

❹点击各平台所收集的文件大小可提前查看收集数据。

※ 收集完成阶段可预览结果，在‘文本挖掘’阶段可下载原文数据。

❺再次确认所收集的数据和容量,选择想要进行分析的数据 点击 ‘数据清洗/语素分析’按钮进行数据分析。

提炼/语义分析

容量充值 尊敬的 더아이엠씨

分离提炼 1

☒ 全部(标题+原文)

☐ 标题

☐ 原文

标题与原文可拆分收集

关键词过滤 2

使用 不使用

☒ 清除

☐ 抽提

关键词添加

去除或提取包含特定关键词的文件,反映在提炼/词素分析结果中.

去除重复内容 3

使用 不使用

语义分析

分析语言 4

中文

英語

用户词典 5

使用 不使用

将特定单词改为在用户词典中登记的单词并进行提炼.  
例) 把 "词语1" 变更为 "词语2" 进行提炼.

选择的收集列表

☐ 全部选择

选择除外

☐ ▶ 吴亦凡

2019-08-28 ~ 2019-09-04

537 KB

6 列表综合生成

把选择的数据列表合并为一个 分析列表生成  
(收集列表不统一)

≡ 收集列表

分析列表生成 →

## ❶ 分离提炼

标题与原文分离或整合分析。

## ❷ 关键词过滤

去除或提取包含特定关键词的文本。

## ❸ 去重

以URL或以内容为准去除相同内容的文件。

## ❹ 分析语言

设置分析语言。

## ❺ 使用者词典

之前进行过的相同的词语提炼或者类似的数据内容时，使用用户词典会更加便利。请提前变更“用户词典设定”中要清洗的单词。

## ❻ 生成综合列表

把选择的收集列表合并成一个分析清单。  
(收集列表不合并)

※ 在“选择的收集列表”中没有的数据,可通过点击右下角的“收集列表”,从收集列表中重新选择。

★★ 分析列表生成的同时根据选择的收集清单的数据容量会从原有容量中直接扣除

请再次确认选择的数据设置内容是否正确

11

钛思通  
TEXTOM

数据提炼

容量充值 尊敬的 더아이에씨

收集

数据预处理

数据预分析

分析列表

数据提炼

矩阵

情感分析

主题分析

可视化

关键词搜索 搜索结果 615 / 615

删除

搜索引擎/SNS 持有数据 请求频道

|                          | 关键词    | 收集日期       | 文件大小   |
|--------------------------|--------|------------|--------|
| <input type="checkbox"/> | 可穿戴机器人 | 2020-11-17 | 125 KB |
| <input type="checkbox"/> | 钧瓷瓶    | 2020-11-16 | 416 KB |
| <input type="checkbox"/> | 无性别风格2 | 2020-11-16 | 778 KB |
| <input type="checkbox"/> | 无性别2   | 2020-11-16 | 488 KB |
| <input type="checkbox"/> | 复古波点时尚 | 2020-11-13 | 535 KB |
| <input type="checkbox"/> | 复古波点时尚 | 2020-11-13 | 170 KB |
| <input type="checkbox"/> | 留白时尚   | 2020-11-12 | 839 KB |
| <input type="checkbox"/> | 中国风    | 2020-11-12 | 643 KB |
| <input type="checkbox"/> |        |            | 658 KB |
| <input type="checkbox"/> |        |            | 64 KB  |

数据提炼 矩阵 情感分析 主题分析

如果完成了词素分析,请通过"直接编辑"或"上传"功能对数据进行编辑(筛选).  
如果您希望在网页进行快速编辑(筛选),请使用"直接编辑"功能;如果您希望将数据下载下来并进行手动操作,请使用"上传"功能

原文数据 1

预览 下载(Excel) 下载(txt)

提炼数据 2

预览 下载(Excel) 下载(txt)

数据编辑 3 编辑的数据已应用

直接编辑 在网页上编辑数据

文件上传

对下载的分词结果(注意,这里下载的不是原文数据)进行进一步筛选,然后上传编辑(筛选)后的单词列表  
- 将EXCEL表格数据转换成txt格式的记事本文件(编码形式选择"UTF-8")并对数据进行编辑(筛选)后上传  
显示内容,便可以使用上传数据功能

分析结果

单词的预览数据

|              |                      |
|--------------|----------------------|
| 以'完全一致'形式来变更 | 葡萄 苹果树 青苹果 苹果果实 苹果箱子 |
| 以'部分一致'形式来变更 | 葡萄 葡萄树 青葡萄 葡萄果实 葡萄箱子 |

①原文数据收集到的数据原文可进行预览或者以xlsx形式的文件直接下载。

②数据清洗显示数据清洗/语素分析数据结果。清洗的数据可以预览或者已xlsx文件进行下载。

※ 根据'数据清洗/语素分析'的设置为基础实现数据的预处理, 为了使数据的分析更加准确化, 可在'数据编辑'中对数据进行编辑。

③数据编辑去除或修改固有名次, 复合名词, 同义词, 非用语的工作。

※ '直接编辑'下编辑内容

示例)

|                            |              |                      |
|----------------------------|--------------|----------------------|
| 苹果 苹果树 青苹果 苹果果实 苹果箱子       | 以'完全一致'形式来变更 | 葡萄 苹果树 青苹果 苹果果实 苹果箱子 |
| - 变更的词语: 苹果<br>- 修正的词语: 葡萄 | 以'部分一致'形式来变更 | 葡萄 葡萄树 青葡萄 葡萄果实 葡萄箱子 |

\* 编辑完成后请点击'上传'进行数据的更新。

数据提炼 矩阵 情感分析 主题分析

数据编辑 编辑的数据已应用

直接编辑

直接编辑 在网页上编辑数据

文件上传

对下载的分词结果(注意,这里下载的不是原文数据)进行进一步筛选,然后上传编辑(筛选)后的单词列表

- 将EXCEL表格数据转换成txt格式的记事本文件(编码形式选择"UTF-8")并对数据进行编辑(筛选)后上传
- 当您看到"编辑(筛选)好的数据已适用"的提示内容,便可以使用上传数据功能

分析结果

单词频度数 4

预览

下载(Excel)

下载(txt)

N-gram 5

预览

下载(Excel)

下载(txt)

TF-IDF 6

预览

下载(Excel)

下载(txt)

连接中心性 7

预览

下载(Excel)

下载(txt)

个体名称识别 8

预览

编辑清洗后的数据, 已多种方式确认结果值。

## 4 单词频度数

确认提取的单词和频度数。

## 5 N-gram

随机性地表现n个单词的连锁性,可以确认实际表现的文章记录。在文本中对两个词进行分析。(bigram,同时出现的单词频率数)

## 6 TF-IDF

TF(词语频度、Term Frequency)和IDF(Inverse Document Frequency,文件频率的倒数),是一个单词在特定的文章中的重要性的统计数据。在特定文件内,单词频率越高且文件中包含单词的文章越少,TF-IDF数据会越高,因为可以起到过滤单词的效果所以TF-IDF数值经常被运用在文章分析中。

## 7 连接中心性

说明特定单词与其他单词有多少联系。连接到NODE的链接越多,相应NODE的连接中心性越高。

## 8 个体识别

根据14个个体的范围可以确认单词的分类和频度数

(人,学问,对象物,机关,地区,文明,日期,时间,数字,事件/思考,动物,植物,金属,术语)

※ 个体名称识别并非提炼数据,而是通过自身形态分析来反映原文数据。

### 3. 数据分析列表

**❷** 可以查找并预览要变更的数据，也可提前确认要修改的数据的修改明细。

### 3. 数据分析列表

钛思通  
TEXTOM

数据预处理

收集

数据预处理

收集列表

提炼/语义分析

分析列表

数据提炼

矩阵

情感分析

主题分析

可视化

可视化结果

定制

关键词搜索

删除

搜索引擎/SNS

街头时尚

疫情

疫情

智能

中国

中国

中国男装设计师

中国男装设计师品牌

中国

涂鸦时尚

提炼数据 (文本挖掘)

关键词

收集日期

文件大小

街头时尚

2020-11-25

757 KB

预览

直接编辑

完全一致

部分一致

用户词典: 未指定

应用

变更词语

修改后词

替换

用户词典可在菜单中修改 进入用户词典 >

修改明细

数据提炼

矩阵

情感分析

主题分析

如果完成了词素分析,请通过"直接编辑"或"上传"功能对数据进行编辑(筛选).

如果您希望在网页进行快速编辑(筛选),请使用"直接编辑"功能;如果您希望将数据下载下来并进行手动操作,请使用"上传"功能

原文数据

预览

下载(Excel)

下载(txt)

提炼数据

预览

下载(Excel)

下载(txt)

数据编辑

编辑的数据已应用

直接编辑

在网页上编辑数据

文件上传

对下载的分词结果(注意,这里下载的不是原文数据)进行进一步筛选,然后上传编辑(筛选)后的单词列表

- 将EXCEL表格数据转换成txt格式的记事本文件(编码形式选择"UTF-8")并对数据进行编辑(筛选)后上传

- 当您看到"编辑(筛选)好的数据已适用"的提示内容,便可以使用上传数据功能

单词频度数

预览

下载(Excel)

下载(txt)

③ 完全一致/部分一致

完全一致是只改变完全一致的字符。部分一致是将部分一致的字符全部变更。



钛思通 TEXTOM

数据提炼

提炼数据 (文本挖掘)

关键词 收集日期 文件大小

街头时尚 2020-11-25 757 KB

7 上传更改内容

3 完全一致 部分一致

6 用户词典: 未指定 适用

4 变更词语 + 修改后词 替换

2 用户词典可在菜单中修改 进入用户词典

5 修改明细

数据提炼 矩阵 情感分析 主题分析

如果完成了词素分析,请通过"直接编辑"或"上传"功能对数据进行编辑(筛选).  
如果您希望在网页进行快速编辑(筛选),请使用"直接编辑"功能;如果您希望将数据下载下来并进行手动操作,请使用"上传"功能

原文数据

预览 下载(Excel) 下载(txt)

提炼数据

预览 下载(Excel) 下载(txt)

直接编辑 编辑的数据已应用

1 直接编辑

直接编辑 在网页上编辑数据

④ 在改变单词的左侧格输入想要变更的单词，右侧格输入修改后的结果单词。  
完全一致是只改变完全一致的字符。部分一致是将部分一致的字符全部变更。

※ 如果要多个要变更的词改为同一个修改单词时，点击加号（+）按钮即可添加想变更的单词。  
※ 参考单词频度和N-gram分析结果，可有效找出需要修改的关键词、需要去除的停用词、需要结合的单词。

⑤ 点击修改明细的倒转图标，即可取消该词的变更。

※ 如果按下“应用已变更内容”按钮后则无法返回之前的操作。



钛思通

TEXTOM

收集

数据预处理

收集列表

提炼/语义分析

分析列表

数据提炼

矩阵

情感分析

主题分析

可视化

定制化

数据提炼

关键词搜索

删除

搜索引擎/SNS

街头时尚

疫情

疫情

智能

中国

中国

中国男装设计师

中国男装设计师品牌

提炼数据 (文本挖掘)

| 关键词  | 收集日期       | 文件大小   |
|------|------------|--------|
| 街头时尚 | 2020-11-25 | 757 KB |

7 上传更改内容

3 预览

直接编辑

完全一致

部分一致

6 用户词典: 未指定

适用

4 变更词语

修改后词

替换

2 用户词典可在菜单中修改 进入用户词典

5 修改明细

数据提炼

矩阵

情感分析

主题分析

如果完成了词素分析,请通过"直接编辑"或"上传"功能对数据进行编辑(筛选).

如果您希望在网页进行快速编辑(筛选),请使用"直接编辑"功能;如果您希望将数据下载下来并进行手动操作,请使用"上传"功能

原文数据

预览

下载(Excel)

下载(txt)

提炼数据

预览

下载(Excel)

下载(txt)

数据编辑

编辑的数据已应用

1 直接编辑

直接编辑 在网页上编辑数据

文件上传

对下载的分词结果(注意,这里下载的不是原文数据)进行进一步筛选,然后上传编辑(筛选)后的单词列表

将EXCEL表格数据转换成txt格式的记事本文件(编码形式选择"UTF-8")并对数据进行编辑(筛选)后上传

单词频度数

预览

下载(Excel)

下载(txt)

⑥使用用户词典时，设置用户词典后可使用的功能，点击使用后可修改数据。

⑦如将修改内容应用到数据中，请点击“应用修改内容”按钮。

※ 应用变更内容后则不能返回修改前，建议提前下载修改前内容及清洗数据。

钛思通  
TEXTOM

数据提炼

容量充值 尊敬的 더아이엠씨

收集

数据预处理

收集列表

提炼/语义分析

分析列表

数据提炼

矩阵

情感分析

主题分析

可视化

关键词搜索 搜索结果 630 / 630

删除

搜索引擎/SNS 持有数据 请求频道

|                          | 关键词     | 收集日期       | 文件大小   |
|--------------------------|---------|------------|--------|
| <input type="checkbox"/> | 街头时尚    | 2020-11-25 | 757 KB |
| <input type="checkbox"/> | 疫情艺术    | 2020-11-24 | 375 KB |
| <input type="checkbox"/> | 疫情艺术    | 2020-11-24 | 375 KB |
| <input type="checkbox"/> | 智能服饰    | 2020-11-24 | 814 KB |
| <input type="checkbox"/> | 中国男装时尚  | 2020-11-24 | 455 KB |
| <input type="checkbox"/> | 中国传统图案  | 2020-11-23 | 427 KB |
| <input type="checkbox"/> | 中国男装设计师 | 2020-11-20 | 753 KB |

数据提炼 矩阵 情感分析 主题分析

如果完成了词素分析,请通过"直接编辑"或"上传"功能对数据进行编辑(筛选).  
如果您希望在网页进行快速编辑(筛选),请使用"直接编辑"功能;如果您希望将数据下载下来并进行手动操作,请使用"上传"功能

原文数据

预览 下载(Excel) 下载(txt)

提炼数据

预览 下载(Excel) 下载(txt)

数据编辑 2 编辑的数据已应用

直接编辑

直接编辑 在网页上编辑数据

1

应用

① 上传档案 - 是先下载清洗数据文件后再编辑词汇，之后将编辑好的档案重新上传的数据编辑方式。

- ※ 要下载编辑的不是原文数据文件是清洗数据文件。
- ※ excel文件保存时另存为 txt 档案保存，txt档案要设定为UTF-8编码档案进行保存。

② 可以确认编辑好的数据应用进行情况。

- ※ 数据编辑好时‘编辑的数据已应用’提示消息会出现，之后可再次进行编辑。
- ※ 但，在显示‘正在适用编辑的内容’的提示消息的状态下不能进行编辑。

**1 分析结果 (文本挖掘)**

| 关键词 | 收集日期       | 文件大小   |
|-----|------------|--------|
| 雪花秀 | 2020-11-27 | 293 KB |

单词频数 N-gram TF-IDF 连接中心性 Topic Modeling

以词频为准，显示全文中前200个单词。通过下载可以查看全部单词

| 单词  | 词频   | 百分比(%) |
|-----|------|--------|
| 雪花  | 1154 | 14.018 |
| 秀   | 879  | 10.678 |
| 品牌  | 225  | 2.733  |
| 人   | 90   | 1.093  |
| 韩国  | 86   | 1.045  |
| 价格  | 70   | 0.85   |
| 正品  | 64   | 0.777  |
| 化妆品 | 60   | 0.729  |
| 百度  | 59   | 0.717  |
| 雅诗  | 59   | 0.717  |
| 茉莉  | 55   | 0.668  |
| 好   | 53   | 0.644  |
| 面膜  | 50   | 0.607  |
| 太平洋 | 50   | 0.607  |
| 产品  | 50   | 0.607  |

**2 分析结果 (文本挖掘)**

| 关键词 | 收集日期       | 文件大小   |
|-----|------------|--------|
| 雪花秀 | 2020-11-27 | 293 KB |

单词频数 N-gram TF-IDF 连接中心性 Topic Modeling

以词频为准，显示全文中前200个单词。通过下载可以查看全部单词

| 单词1 | 单词2 | 词频  |
|-----|-----|-----|
| 雪花  | 秀   | 872 |
| 秀   | 雪花  | 79  |
| 品牌  | 雪花  | 64  |
| 百度  | 快照  | 43  |
| 秀   | 品牌  | 38  |
| 正品  | 雪花  | 38  |
| 茉莉  | 太平洋 | 34  |
| 价格  | 正品  | 30  |
| 图片  | 价格  | 30  |
| 雪花  | 雪花  | 24  |
| 秀   | 韩国  | 21  |
| 网   | 雪花  | 21  |
| 韩国  | 雪花  | 20  |
| 公益  | 限定版 | 18  |
| 精华  | 雪花  | 18  |

**分析结果**

单词频数 **1**

预览 下载(Excel) 下载(txt)

N-gram **2**

预览 下载(Excel) 下载(txt)

TF-IDF **3**

预览 下载(Excel) 下载(txt)

连接中心性 **4**

预览 下载(Excel) 下载(txt)

个体名称识别 **5**

预览

**1 单词频数** 表示所提取的单词和数据中相应单词的频率。

※ 词频数值高表示提炼数据中相应的单词出现频率高

※ 百分比是指该单词占总单词量的百分比。

**2 N-gram** n个单词关联出现。

※ 单词1和单词2的频率高意味着两个词同时出现的频率高。

数据提炼

矩阵

情感分析

主题分析

可视化

可视化结果

定制

☐yunja

☐街头时

☐疫情艺

☐疫情艺

☐智能服

☐中国男

☐中国传

☐中国男

☐中国男

关键词

收集日期

文件大小

雪花秀

2020-11-27

293 KB

67 KB

757 KB

375 KB

375 KB

814 KB

455 KB

427 KB

753 KB

85 KB

分析结果 (文本挖掘)

单词频度数

N-gram

TF-IDF

连接中心性

Topic Modeling

以词频为准。显示全文中前200个单词。通过下载可以查看全部单词

下载

| 单词  | TF-IDF        |
|-----|---------------|
| 品牌  | 247.18776495  |
| 人   | 197.75021196  |
| 正品  | 191.726865507 |
| 韩国  | 178.831972584 |
| 价格  | 173.943465485 |
| 面膜  | 156.774710796 |
| 精华  | 151.33201209  |
| 化妆品 | 149.094398987 |
| 茉莉  | 141.07221466  |
| 图片  | 140.048032135 |
| 梅花  | 130.323861521 |
| 太平洋 | 128.247467873 |
| 限量  | 124.570309251 |
| 好   | 122.037009929 |
| 百度  | 114.808698794 |

分析结果

单词频度数 1

预览

下载(Excel)

下载(txt)

N-gram 2

预览

下载(Excel)

下载(txt)

TF-IDF 3

预览

下载(Excel)

下载(txt)

连接中心性 4

预览

下载(Excel)

下载(txt)

个体名称识别 5

预览

❸ TF-IDF TF-IDF：TF（单词频度）和IDF（频度的倒数）的乘值，表示单词在特定文件中的重要性。

※ TF:文件内特定单词的频率数/DF:各种文件内特定单词的频率数/IDF:DF的倒数

※  $TF - IDF = TF \times 1/DF$

※ 在特定范围内求所有词的频率数和包含词的条目（信息）的频率数的倒数相乘后得到的值，以求条目（信息）的重要性的方法。

yunjac

| 关键词 | 收集日期       | 文件大小   |
|-----|------------|--------|
| 雪花秀 | 2020-11-27 | 293 KB |

分析结果 (文本挖掘)

关键词

收集日期

文件大小

雪花秀

2020-11-27

293 KB

单词频度数

N-gram

TF-IDF

连接中心性

Topic Modeling

以词频为准，显示全文中前200个单词。通过下载可以查看全部单词

| 单词  | 连接中心性 |
|-----|-------|
| 雪花  | 0.25  |
| 秀   | 0.18  |
| 品牌  | 0.10  |
| 人   | 0.05  |
| 产品  | 0.04  |
| 韩国  | 0.038 |
| 好   | 0.038 |
| 百度  | 0.034 |
| 精华  | 0.034 |
| 面膜  | 0.031 |
| 化妆品 | 0.030 |
| 佳人  | 0.028 |
| 限量  | 0.027 |
| 中国  | 0.024 |
| 大   | 0.024 |

分析结果 (文本挖掘)

关键词

收集日期

文件大小

雪花秀

2020-11-27

293 KB

以词频为准，显示全文中前200个单词。通过下载可以查看全部单词

下载

| 集团 | 单词 | Topic Modeling |
|----|----|----------------|
| 1  | 茉莉 | 55             |
| 1  | 图片 | 46             |
| 1  | 快照 | 43             |
| 1  | 梅花 | 40             |
| 1  | 人参 | 35             |
| 1  | 护肤 | 26             |
| 1  | 面霜 | 24             |
| 1  | 粉底 | 23             |
| 1  | 公益 | 22             |
| 1  | 肌肤 | 18             |
| 1  | 商城 | 17             |
| 1  | 上市 | 15             |
| 1  | 专题 | 15             |
| 1  | 最大 | 15             |
| 1  | 智慧 | 13             |

### 分析结果

#### 单词频度数 ①

[预览](#)
[下载\(Excel\)](#)
[下载\(txt\)](#)

#### N-gram ②

[预览](#)
[下载\(Excel\)](#)
[下载\(txt\)](#)

#### TF-IDF ③

[预览](#)
[下载\(Excel\)](#)
[下载\(txt\)](#)

#### 连接中心性 ④

[预览](#)
[下载\(Excel\)](#)
[下载\(txt\)](#)

#### 个体名称识别 ⑤

[预览](#)

④ 度中心性 A单词和B单词之间连接的程度

⑤ 命名实体识别 每个单词分布在14个实体集团中，指示该频率。

※ 14个集团：人物、修饰词、学问、成语、数字、生物、食品、医疗/健康、机关、物品、职业、专用名词、场所、时间

※ 词频, N-gram, TF-IDF, 度中心性, 命名实体识别 菜单根据每次提取的数据排序和内容会发生改变。

数据提炼

矩阵

情感分析

主题分析

选择矩阵单词

直接选择/可以通过选择性上传来决定生成矩阵的单词

1-mode

2-mode

选择单词已适用

1

直接选择

2

适用

示例文件下载

将适用txt文件编辑完成的内容转换为Excel文件上传

矩阵单词

选择单词

预览

下载

矩阵结果

矩阵(词频)

预览

下载

边缘列表

预览

下载

欧几里得系数

预览

下载

余弦系数

预览

下载

杰卡德相似系数

预览

下载

相关系数

预览

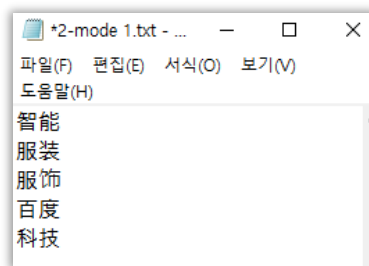
下载

## ① 直接选择

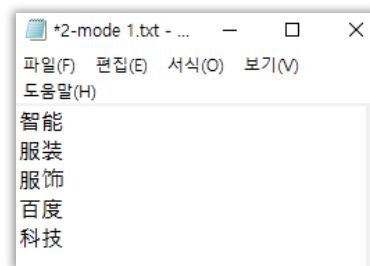
选择要分析的单词,点击“适用”。Matrix（矩阵）生成完成后,在"立即选择"上端会标出"已适用单词选择"的字样。

## ② 文件上传

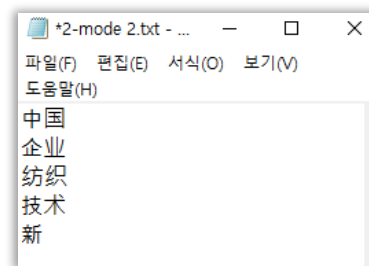
文本挖掘（Text Mining）可参考分析结果中的“单词频度数”，制作单词列表并上传UTF-8编码的txt文件。



1-mode 上传文件事例



2-mode上传文件事例（左-行 / 右-列）



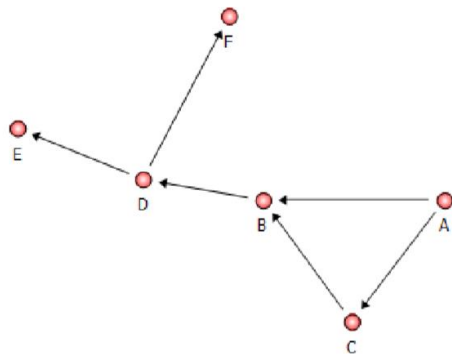
## ③ 分析结果

单词间通过同时出现的相似度系数, 提供4种不同计算方式的结果系数表。

欧几里得系数 余弦系数 Jaccard 系数 相关系数

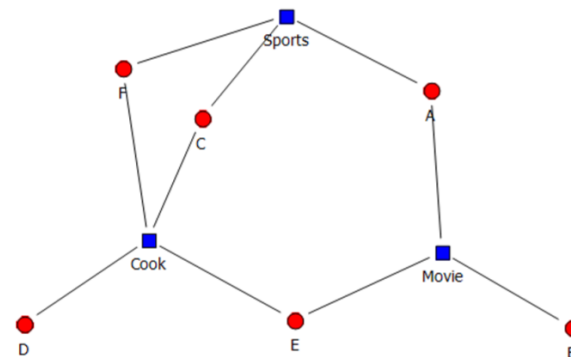
※ 共现：在特定范围内同时出现相同的单词（节点）时，这一范围内的所有单词（节点）之间在意义上存在相互关联的关系。

## ① 1-mode 分析方法



|   | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | 1 | 1 | 0 | 0 | 0 |
| B | 0 | 0 | 0 | 2 | 0 | 0 |
| C | 0 | 3 | 0 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 0 | 1 | 2 |
| E | 0 | 0 | 0 | 0 | 0 | 0 |
| F | 0 | 0 | 0 | 0 | 0 | 0 |

## ② 2-mode分析方法



|   | Sports | Movie | Cook |
|---|--------|-------|------|
| A | 1      | 1     | 0    |
| B | 0      | 1     | 0    |
| C | 1      | 0     | 1    |
| D | 0      | 0     | 1    |
| E | 0      | 1     | 1    |
| F | 1      | 0     | 1    |

※ 1-mode和 2-mode有什么差别？

- 1-mode: 显示所选关键词间的关系，行与列输入的单词目录相同。
- 2-mode: 显示特定关键词和其他关键词间的关系。行与列输入的单词目录不相同。

## 生成用于网络分析的矩阵数据

分析结果 (1-mode) - 欧几里得系数

|   | 大              | 强              | 人              | 好              | 渠              | 全              | 宽              |      |
|---|----------------|----------------|----------------|----------------|----------------|----------------|----------------|------|
| 大 | 0              | 0.967313977477 | 0.980726416388 | 0.96621313108  | 0.908712907082 | 0.919935923097 | 0.938686066052 | 0.95 |
| 强 | 0.967313977477 | 0              | 0.91608186417  | 0.958114609171 | 0              | 0.886772296586 | 0.764297739604 | 0.9  |
| 人 | 0.980726416388 | 0.91608186417  | 0              | 0.979519201987 | 0              | 0.944615122438 | 0.980083474477 | 0.95 |
| 好 | 0.96621313108  | 0.958114609171 | 0.979519201987 | 0              | 0.953067674554 | 0.929465438414 | 0.946083613398 | 0.94 |
| 渠 | 0.908712907082 | 0              | 0              | 0.953067674554 | 0              | 0              | 0              |      |
| 全 | 0.919935923097 | 0.886772296586 | 0.944615122438 | 0.929465438414 | 0              | 0              | 0.711324865405 | 0.85 |
| 宽 | 0.938686066052 | 0.764297739604 | 0.980083474477 | 0.946083613398 | 0              | 0.711324865405 | 0              | 0.5  |
| 高 | 0.930662475472 | 0.93584997009  | 0.957398567716 | 0.943295202288 | 0              | 0.896857875374 | 0.9105572809   |      |
| 指 | 0.985295707558 | 0.5            | 0.98551485053  | 0.984716339357 | 0              | 0.984838039128 | 0.985330555656 | 0.9  |
| 新 | 0.986635744614 | 0.97591836735  | 0.98519776884  | 0.984342473791 | 0.958333333333 | 0.98349304498  | 0.985537591372 | 0.98 |
| 业 | 0.966347320593 | 0.895742792971 | 0.971647601144 | 0.96201314118  | 0              | 0.954685674557 | 0.961985703937 | 0.95 |

分析结果 (1-mode) - 余弦函数

|   | 大                | 强                 | 人                | 好                | 渠                | 全                | 宽       |  |
|---|------------------|-------------------|------------------|------------------|------------------|------------------|---------|--|
| 大 | 0                | 0.0302315720325   | 0.0199232865067  | 0.0642161510172  | 0.00933927325624 | 0.087075063123   | 0.0588  |  |
| 强 | 0.0302315720325  | 0                 | 0.00527943931557 | 0.0218499127207  | 0                | 0.0346108192167  | 0.00909 |  |
| 人 | 0.0199232865067  | 0.00527943931557  | 0                | 0.0172469597772  | 0                | 0.0304124119754  | 0.0225  |  |
| 好 | 0.0642161510172  | 0.0218499127207   | 0.0172469597772  | 0                | 0.0151016352897  | 0.0480002252816  | 0.0309  |  |
| 渠 | 0.00933927325624 | 0                 | 0                | 0.0151016352897  | 0                | 0                |         |  |
| 全 | 0.087075063123   | 0.0346108192167   | 0.0304124119754  | 0.0480002252816  | 0                | 0                | 0.0476  |  |
| 宽 | 0.058899908383   | 0.00909747623385  | 0.0225283261237  | 0.0309942193659  | 0                | 0.047642145074   |         |  |
| 高 | 0.0497772625551  | 0.0238840383178   | 0.0193724540315  | 0.0480152442259  | 0                | 0.0767300491173  | 0.0364  |  |
| 指 | 0.0044703961313  | 0.000957461212933 | 0.00252395397021 | 0.00236064161871 | 0                | 0.00551549191451 | 0.00421 |  |
| 新 | 0.0428392180554  | 0.0142304109535   | 0.015109259641   | 0.0382040916547  | 0.00976916128906 | 0.0352947168499  | 0.0269  |  |
| 业 | 0.0210217456637  | 0.00760375737381  | 0.00715863896587 | 0.0176759602463  | 0                | 0.0312869535665  | 0.0119  |  |

分析结果 (1-mode) - 杰卡德相似系数

|   | 大               | 强               | 人               | 好               | 渠               | 全               | 宽              |  |
|---|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|----------------|--|
| 大 | 0               | 0.0714285714286 | 0.0521390374332 | 0.0985048372911 | 0.0185979971388 | 0.0945454545455 | 0.083636363636 |  |
| 强 | 0.0714285714286 | 0               | 0.0216802168022 | 0.0523513753327 | 0               | 0.0592592592593 | 0.020370370370 |  |
| 人 | 0.0521390374332 | 0.0216802168022 | 0               | 0.0462074978204 | 0               | 0.0571428571429 | 0.055357142857 |  |
| 好 | 0.0985048372911 | 0.0523513753327 | 0.0462074978204 | 0               | 0.0300546448087 | 0.0474183350896 | 0.040042149631 |  |
| 渠 | 0.0185979971388 | 0               | 0               | 0.0300546448087 | 0               | 0               | 0              |  |
| 全 | 0.0945454545455 | 0.0592592592593 | 0.0571428571429 | 0.0474183350896 | 0               | 0               | 0.055248618784 |  |
| 宽 | 0.0836363636364 | 0.0203703703704 | 0.0553571428571 | 0.0400421496312 | 0               | 0.0552486187845 | 0              |  |
| 高 | 0.0645756457565 | 0.0488721804511 | 0.0434782608696 | 0.0563230605739 | 0               | 0.0819209039548 | 0.050847457627 |  |
| 指 | 0.0215439856373 | 0.0073126142596 | 0.021164021164  | 0.010460251046  | 0               | 0.0216802168022 | 0.021680216802 |  |
| 新 | 0.0800915331808 | 0.0414269275029 | 0.0494880546075 | 0.0912901723335 | 0.0234055002926 | 0.0397435897436 | 0.039743589743 |  |
| 业 | 0.0473484848485 | 0.027027027027  | 0.0278810408922 | 0.0355987055016 | 0               | 0.0588235294118 | 0.029411764705 |  |

分析结果 (1-mode) - 相关系数

|   | 大               | 强                | 人                | 好               | 渠                | 全               | 宽           |  |
|---|-----------------|------------------|------------------|-----------------|------------------|-----------------|-------------|--|
| 大 | 0               | 0.0840868213676  | 0.177150674046   | 0.411111939277  | 0.0764854641348  | 0.188519324722  | 0.27243715  |  |
| 强 | 0.0840868213676 | 0                | 0.00568954967495 | 0.0737148579461 | 0.0264985324753  | 0.0442832747044 | 0.004672732 |  |
| 人 | 0.177150674046  | 0.00568954967495 | 0                | 0.1207505666657 | 0.0245415840325  | 0.136527859474  | 0.17404385  |  |
| 好 | 0.411111939277  | 0.0737148579461  | 0.1207505666657  | 0               | 0.146982595449   | 0.119614581467  | 0.14375223  |  |
| 渠 | 0.0764854641348 | 0.0264985324753  | 0.0245415840325  | 0.146982595449  | 0                | 0.0389861491035 | 0.02950427  |  |
| 全 | 0.188519324722  | 0.0442832747044  | 0.136527859474   | 0.119614581467  | 0.0389861491035  | 0               | 0.09563078  |  |
| 宽 | 0.272437152389  | 0.00467273225545 | 0.174043853417   | 0.143752235793  | 0.0295042750347  | 0.095630786289  | 0           |  |
| 高 | 0.148910993438  | 0.0640604030233  | 0.23212582869    | 0.245298970962  | 0.0313801755296  | 0.23779508141   | 0.09029212  |  |
| 指 | 0.0708746190404 | 0.00574250383134 | 0.254701542891   | 0.0738481202671 | 0.00875542471698 | 0.173346971276  | 0.06479693  |  |
| 新 | 0.465003038222  | 0.085766102794   | 0.288866987407   | 0.504850677249  | 0.0542278263767  | 0.285952017253  | 0.25349130  |  |
| 业 | 0.144853082308  | 0.00728694392921 | 0.573683887451   | 0.113904204458  | 0.0168592846249  | 0.240789757352  | 0.08348785  |  |

矩阵数据适用于多种网络分析程序，利于进行数据补充分析



## 选择矩阵单词

直接选择/可以通过选择性上传来决定生成矩阵的单词

1-mode

2-mode

1

正在适用编辑的内容

直接选择



✓ 适用

示例文件下载

将适用txt文件编辑完成的内容转换为Excel文件上传

## 矩阵单词

选择单词

2

预览

下载

## 矩阵结果

矩阵(词频)

预览

下载

边缘列表

预览

下载

欧几里得系数

预览

下载

余弦系数

预览

下载

数据预处理

收集列表

提炼/语义

≡ 分析列表

数据提炼

矩阵

情感分析

主题分析

≡ 可视化

可视化结果

定制

收集日期

文件大小

2020-11-27

293 KB

2020-11-27

67 KB

2020-11-25

757 KB

2020-11-24

375 KB

2020-11-24

375 KB

2020-11-24

814 KB

2020-11-24

455 KB

2020-11-23

427 KB

2020-11-20

753 KB

2020-11-19

85 KB

1

2

3

4

5

6

7

8

9

10

下一个块 →

>>

① 当出现‘选择单词已适用’提示时开始分析下方的矩阵单词，并确认分析结果。

② 选择单词和频度值可下载也可预览。下载只支持excel格式。

## 矩阵结果

### 1 矩阵(词频)

预览

下载

### 2 边缘列表

预览

下载

### 欧几里得系数

预览

下载

### 余弦系数

预览

下载

### 杰卡德相似系数

预览

下载

### 相关系数

预览

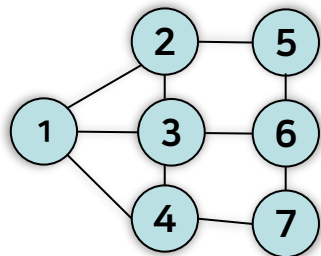
下载

## 1 词频的矩阵

| 分析结果 (1-mode) - 词频 |      |      |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|--------------------|------|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|                    | 雪花   | 秀    | 品牌  | 人   | 韩国  | 价格  | 正品  | 化妆品 | 百度  | 精华  | 茉莉  | 好   | 面膜  | 太平洋 | 产品  | 图片  | 限量  | 快照  | 护肤品 |
| 雪花                 | 0    | 3987 | 871 | 343 | 411 | 412 | 333 | 188 | 272 | 249 | 169 | 177 | 183 | 209 | 202 | 244 | 157 | 214 | 208 |
| 秀                  | 3987 | 0    | 723 | 256 | 345 | 343 | 263 | 170 | 238 | 215 | 149 | 138 | 156 | 164 | 178 | 199 | 90  | 180 | 174 |
| 品牌                 | 871  | 723  | 0   | 45  | 86  | 55  | 60  | 109 | 29  | 34  | 63  | 10  | 14  | 44  | 36  | 51  | 13  | 29  | 30  |
| 人                  | 343  | 256  | 45  | 0   | 4   | 4   | 0   | 0   | 3   | 0   | 0   | 10  | 37  | 0   | 3   | 0   | 4   | 3   | 10  |
| 韩国                 | 411  | 345  | 86  | 4   | 0   | 54  | 20  | 125 | 36  | 43  | 90  | 14  | 1   | 60  | 40  | 15  | 0   | 26  | 32  |
| 价格                 | 412  | 343  | 55  | 4   | 54  | 0   | 180 | 1   | 11  | 24  | 2   | 4   | 0   | 5   | 18  | 135 | 0   | 11  | 2   |
| 正品                 | 333  | 263  | 60  | 0   | 20  | 180 | 0   | 0   | 1   | 20  | 0   | 3   | 0   | 0   | 1   | 180 | 0   | 1   | 0   |
| 化妆品                | 188  | 170  | 109 | 0   | 125 | 1   | 0   | 0   | 19  | 50  | 96  | 0   | 4   | 57  | 14  | 0   | 0   | 16  | 16  |
| 百度                 | 272  | 238  | 29  | 3   | 36  | 11  | 1   | 19  | 0   | 24  | 6   | 9   | 11  | 8   | 22  | 1   | 0   | 45  | 41  |
| 精华                 | 249  | 215  | 34  | 0   | 43  | 24  | 20  | 50  | 24  | 0   | 32  | 2   | 11  | 36  | 20  | 15  | 0   | 2   | 2   |
| 茉莉                 | 169  | 149  | 63  | 0   | 90  | 2   | 0   | 96  | 6   | 32  | 0   | 3   | 0   | 100 | 2   | 0   | 0   | 0   | 0   |
| 好                  | 177  | 138  | 10  | 10  | 14  | 4   | 3   | 0   | 9   | 2   | 3   | 0   | 0   | 5   | 1   | 0   | 0   | 0   | 0   |
| 面膜                 | 183  | 156  | 14  | 37  | 1   | 0   | 0   | 4   | 11  | 11  | 0   | 0   | 0   | 26  | 4   | 0   | 0   | 0   | 0   |
| 太平洋                | 209  | 164  | 44  | 0   | 60  | 5   | 0   | 57  | 8   | 36  | 100 | 5   | 26  | 0   | 2   | 0   | 0   | 0   | 0   |

## 2 边缘列表的矩阵

| 分析结果 (1-mode) - 边缘列表 |       |        |
|----------------------|-------|--------|
| word1                | word2 | weight |
| 雪花                   | 秀     | 3987   |
| 雪花                   | 品牌    | 871    |
| 雪花                   | 人     | 343    |
| 雪花                   | 韩国    | 411    |
| 雪花                   | 价格    | 412    |
| 雪花                   | 正品    | 333    |
| 雪花                   | 化妆品   | 188    |
| 雪花                   | 百度    | 272    |
| 雪花                   | 精华    | 249    |
| 雪花                   | 茉莉    | 169    |
| 雪花                   | 好     | 177    |
| 雪花                   | 面膜    | 183    |
| 雪花                   | 太平洋   | 209    |
| 雪花                   | 产品    | 202    |
| 雪花                   | 图片    | 244    |
| 雪花                   | 限量    | 157    |
| 雪花                   | 快照    | 214    |
| 雪花                   | 护肤品   | 208    |
| 雪花                   | 梅花    | 158    |



<Mini Social graph>

## 1 词频

在全文件内单词出现的频率，从高到低排序表示。

## 2 边缘列表(Edge List)

单词和单词，节点和节点配对表示的目录。

钛思通  
TEXTOM

≡

收集

数据预处理

收集列表

提炼/语义分析

分析列表

数据提炼

矩阵

凝聚子群分析

情感分析

主题分析

可视化

可视化结果

定制

关键词搜索

搜索结果 1991 / 1991

10个

凝聚子群分析

| 单词   | 词频  |
|------|-----|
| 领导   | 440 |
| 大学生  | 199 |
| 辞职信  | 153 |
| 职场   | 132 |
| 辞职信  | 126 |
| 胡萝卜  | 81  |
| 社会   | 75  |
| 工资   | 74  |
| 努力   | 67  |
| 忍气吞声 | 67  |
| 前哨   | 66  |
| 走人   | 66  |
| 任性   | 65  |
| 明白   | 64  |
| 天真   | 64  |
| 思想   | 64  |
| 学生   | 64  |
| 钱    | 63  |
| 冲动   | 63  |

| 收集日期       | 文件大小      |
|------------|-----------|
| 2022-10-20 | 849.02 KB |
| 2022-10-16 | 1.03 MB   |
| 2022-10-16 | 288 KB    |
| 2022-10-16 | 336 KB    |
| 2022-10-12 | 1.29 MB   |
| 2022-10-04 | 232 KB    |
| 2022-09-04 | 1.06 MB   |
| 2022-09-25 | 284 KB    |
| 2022-08-24 | 70.15 KB  |
| 2022-09-04 | 1.04 MB   |

更了解一下凝聚子群分析 >

1

要显示文件内同时出现（共现）相同的单词之间在意义上存在相互关联的关系，因此使用单词之间相关关系的模式来进行群集的分析方式。

凝聚子群数

2个 4个 8个 16个

训练数据已适用 应用

2

分析单词

选择单词(1-mode)

预览 下载

选择列单词(2-mode)

预览 下载

选择行单词(2-mode)

预览 下载

3

分析结果

结果(1-mode)

预览 下载

密度值(1-mode)

预览 下载

结果(2-mode)

预览 下载

密度值(2-mode)

预览 下载

- 凝聚子群个数** 凝聚子群是为了衡量每个关键字的相似度，重复执行相关分析，通过设置和重复块的数量，同时根据相关系数的值对它们进行分组，从而找到最佳的块数。
- 分析词汇** 是抽出在矩阵阶段的预览词汇。
- 分析结果** 在相关分析导出相关系数的密度价值于 -1 和 1 之间。一般来说，相关分析的价值结果是0、-1和1。

钛思通 TEXTOM

情感分析

关键词搜索 搜索结果 629 / 629

数据预处理

收集列表

提炼/语义分析

分析列表

数据提炼

矩阵

情感分析

主题分析

可视化

可视化结果

定制

关键词

收集日期

文件大小

疫情艺术

疫情艺术

智能服饰

中国男装时尚

中国传统图案

中国男装设计师

中国男装设计师品牌

中国

涂鸦时尚

Y2K时尚

2020-11-24

2020-11-24

2020-11-24

2020-11-24

2020-11-23

2020-11-20

2020-11-19

2020-11-19

2020-11-19

2020-11-19

375 KB

375 KB

814 KB

455 KB

427 KB

753 KB

85 KB

40 KB

1.19 MB

下载excel格式可能会因特殊字符导致数据遗漏

原文数据 1

预览 下载(Excel) 下载(bat)

训练数据 2

下载示例文件

情感分析需上传训练数据

请参考示例文件 将最少100个 最多1000个数据按内容的 积极/中性/消极分类并上传 (的比率越接近,分析结果越准确)

分析结果

全部 (积极/中性/消极)

预览 下载

积极

预览 下载 追加分析

中性

预览 下载 追加分析

消极

预览 下载 追加分析

预览 可确认部分数据,如需全部数据请利用下载功能

追加分析 可进行语义分析和制作网络矩阵图等额外的分析

情感分析可视化的结果可以在“可视化结果”页面确认

以机械学习技巧为基础对肯定,否定,中立的文章进行归类。

**①原文数据** 可以对收集到的原文数据进行预览或下载 (xlsx文件)

**②训练数据** 以上传的部分训练数据为依据,对全部原文进行自动分类。

- A列: 情感分析对象的文本信息

※.请删除原文数据中日期, URL等的情感分析中不必要的列

- B列: 积极, 中立, 消极中选择并标记

※ 支持对最少100个, 最多1000个数据 (行) 进行标记, 且保持积极消极中性的比例尽量保持平衡。

※ 只支持xlsx文件格式上传

### ③ 分析结果

可确认分类后的原文数据结果

- 补充分析: 标记极性词语后的数据将会转移到数据挖掘页面

|    | A   | B |
|----|---|---|
| 1  | 杭州老字号 经典百货的“变与不变”“浙江第一店”老底子的时候,杭州人把解放路沿线积极      |   |
| 2  | 广州公开课 招商实战大全 精准的商业定位是招商成功的基础。那么如何找到市场差?消极       |   |
| 3  | 12月23-24日 广州《购物中心、文商旅街区、社区商业招商实战研修班》报名热线:400 消极 |   |
| 4  | 去韩国买买买、网红店拍拍拍,看这篇攻略就够了!一转眼已经11月了,年底也快到 消极       |   |
| 5  | 热烈祝贺大盛被评为2017年“浙江省商贸流通业诚信示范企业”2017年“浙江省商贸流通业 消极 |   |
| 6  | 韩国不止“泡菜”:这10家美食餐厅必去 据说来首尔的女孩儿有3种:追韩剧的、不追韩 中立    |   |
| 7  | 购物中心招商失败的11大症状,你知道吗?“招商难”的根本原因并不在于承租商资源 中立      |   |
| 8  | 2017年12月23-24日 广州《购物中心、文商旅街区、社区商业招商实战研修班》报名热 消极 |   |
| 9  | 你知道中百大厦,却不知道中百大厦有多少店?(山东商业171216期) 山东潍坊百货集 中立   |   |
| 10 | 招聘!宿迁最任性的商场大商百货招人啦,事少,钱多~大商百货微信号: sqdsbh 隶属 消极  |   |

(训练数据文件示例)

以Word-level Semantic Clustering文件中单词共现关系为基准进行向量化，将相邻单词汇编成同一个群集。

- ❶ **群集数 (K值)**：输入想要设定的群集的个数。
- ❷ **各群集的单词数**：输入要包含至群集的单词个数
- ❸ **应用**：设置群集数 (K值) 和群集别单词数后点击应用按钮
- ❹ **分析结果**：应用完毕后，关联性高的单词之间被分类为群集，可以查看详细结果。

分析结果 (文本挖掘)

| 关键词  | 收集日期       | 文件大小   |
|------|------------|--------|
| 街头时尚 | 2020-11-25 | 757 KB |

以词频为准，显示全文中前200个单词。通过下载可以查看全部单词

下载

| 集团 | 单词   | Topic Modeling |
|----|------|----------------|
| 1  | 街头时尚 | 0.042          |
| 1  | 时尚   | 0.037          |
| 1  | 设计   | 0.016          |
| 1  | 文化   | 0.016          |
| 1  | 外套   | 0.015          |
| 1  | 日本   | 0.015          |
| 1  | 牛仔   | 0.015          |
| 1  | 潮流   | 0.012          |
| 1  | 风格   | 0.011          |
| 1  | 品牌   | 0.009          |
| 1  | 服装   | 0.009          |
| 1  | 经典   | 0.007          |
| 1  | 互联网  | 0.007          |
| 1  | 百度   | 0.007          |
| 1  | 中国   | 0.007          |

分析结果 (文本挖掘)

| 关键词  | 收集日期       | 文件大小   |
|------|------------|--------|
| 街头时尚 | 2020-11-27 | 757 KB |

以词频为准，显示全文中前200个单词。通过下载可以查看全部单词

下载

| 集团 | 单词   | Topic Modeling |
|----|------|----------------|
| 1  | 范    | 15             |
| 1  | 大比拼  | 7              |
| 1  | 多元化  | 6              |
| 1  | 页片   | 5              |
| 2  | 时尚   | 1219           |
| 2  | 街头时尚 | 1171           |
| 2  | 风格   | 418            |
| 2  | 潮流   | 301            |
| 2  | 品牌   | 204            |
| 2  | 日本   | 203            |
| 2  | 文化   | 188            |
| 2  | 设计   | 157            |
| 2  | 新浪网  | 128            |
| 2  | 服饰   | 122            |
| 2  | 视频   | 119            |

**LDA主题模型分析**  
主题模型分析作为从大量文本中自动查找主题而开发的算法，将拥有类似含义的单词进行聚类而推算出主题内容。

1 > 主题数  个

2 > 主题内单词数  个

3 > 随机值  不使用

选择是否使用样本数据中随机生成主题功能。如果选择使用，可能会降低主题模型的再现性。

4 学习数据已适用。

**分析结果**

> 词级语义聚类分析

预览 5 下载(Excel) 下载(txt)

LDA主题模型分析

预览 5 下载(Excel) 下载(txt)

> LDA Topic Origin Text

预览 5 下载(Excel) 下载(txt)

LDA Topic Modeling 是自动从大量文件群中找出主题（话题）的算法，将具有类似意义的单词聚类。

- ❶ topic数：输入想要设定的组合个数。
- ❷ topic 单词数：输入包含至各组合内的单词个数
- ❸ 随机值：进行随机分配时，分析结果的再现性会降低，因此，即使用同样的数据进行分析，结果导出的值也会有所不同。想要确保分析结果的再现性时，请选择不使用。
- ❹ 应用：选择好Topic数和Topic单词数、随机值后点击应用按钮。
- ❺ 分析结果：应用完毕后，即可确认关键词被分类为哪一组。

※ Topic Modeling 数值是以百分比形式表示相应话题中加入单词的概率。

## 4. 可视化



- ❶ 选择颜色
- ❷ 选择单词 以词频或关键词的降序为基准排列。关键词可以设定为用户认为重要的词。
- ❸ 变更词云背景 上传没有背景的 jpg, png 图片, 将生成图片模样的词云。

※ 选择尽量有意义的、频度高的单词比较好。

※ 选好的关键词当中，有数字或英文的话，按一下数字和英文的按钮，相关单词会强调，另外其他关键词表示地变成灰色。



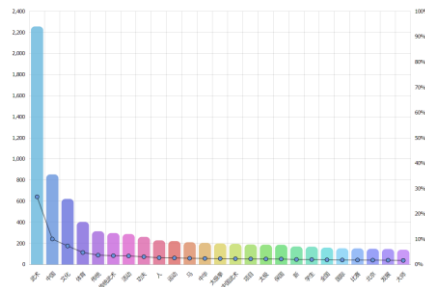
# 查看可视化结果

4. 可视化

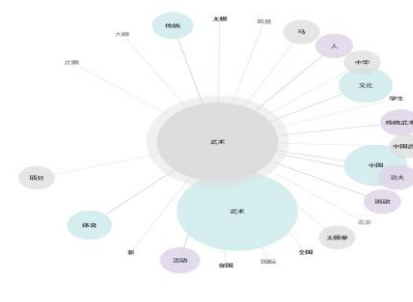
► 一个能将词频数差异视觉化的图表



[ 词云 ]



[ 条形图 ]



[ 自我网络分析图 ]



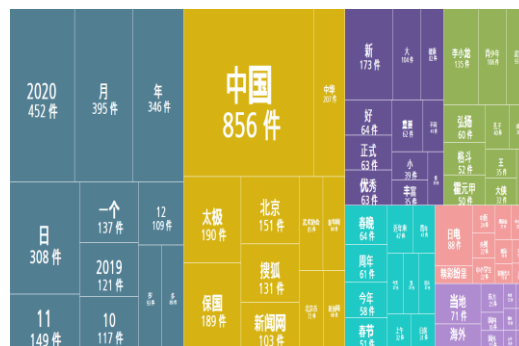
[ 2-way 词树 ]

► N-gram 结果可用视觉化表示



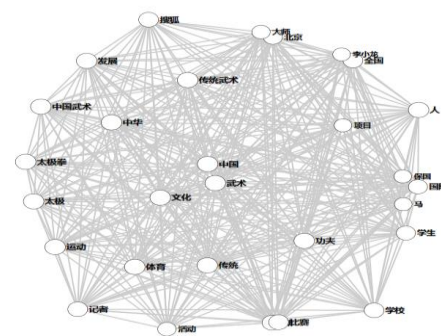
[ N-gram网络 ]

► 个体名识别结果的视觉化图表



[ 树状图 ]

► 关键词前后关系可以用视觉化图表来说明



[ 矩阵网络图 ]



# 查看可视化结果

4. 可视化

钛思通  
TEXTOM

视觉化结果

容量充值 尊敬的 더아이엠씨

收集

词云 条形图 自我中心网络 网络 树形图 LDA主题分布 聚类分析 矩阵网络图 文本情感分析

Selected Topic: 0 Previous Topic Next Topic Clear Topic

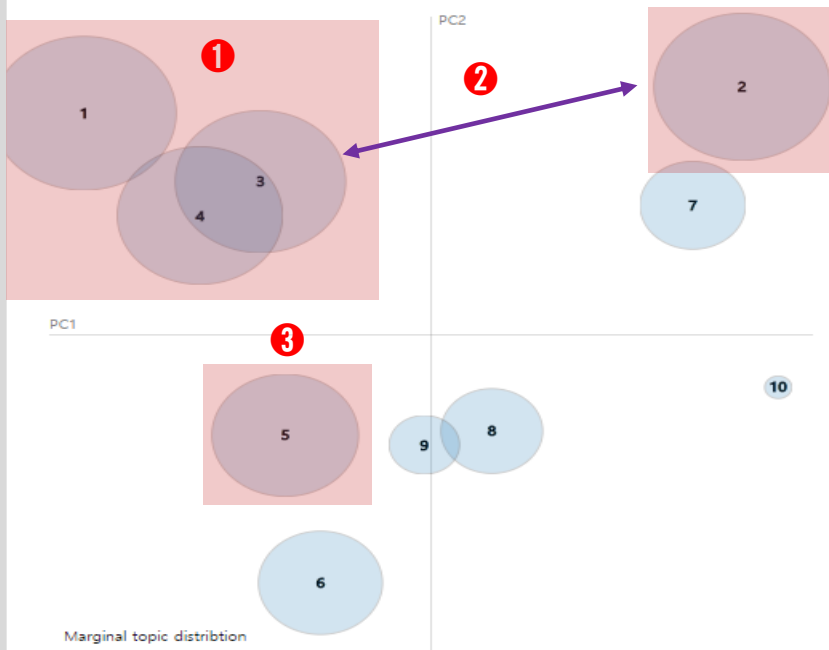
4

Slide to adjust relevance metric:(2)

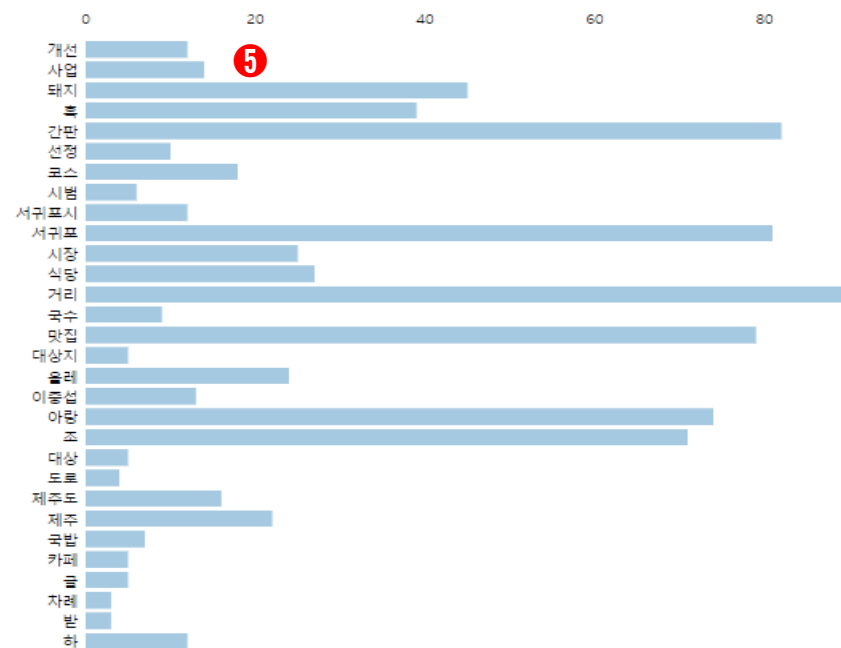
$\lambda = 1$

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Salient Terms<sup>1</sup>



Overall term frequency  
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) \* [sum\_t p(t | w) \* log(p(t | w)/p(t)) for topics t; see Chuang et. al (2012)  
2. relevance(term w | topic t) =  $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$ ; see Sievert & Shirley (2014)

LDA提供LDA Topic Modeling分析结果的可视化。  
(先进行LDA Topic Modeling分析, 才可以看到可视化结果。)

## LDA Topic Modeling 可视化说明

- ❶ **选择群组** 群组分布图中, 点击群组或者在'Selected Topic'中直接输入群组编号进行选择就可查看各群组组成部分的前30个单词。
- ❷ **群组间的距离** 群组之间距离越远,区分效度 (discriminant validity) 越高,主题分得越明显。如果群组之间距离较近或重叠,则区分效度较低, 说明两个群组的主题相近。
- ❸ **群组的大小** 表示群组的圆圈的大小是表示群组内单词频度的高低, 最大的圆圈一般是主话题。
- ❹  **$\lambda$ 数值设定** 通过调整 $\lambda$ 数值可以设定特定话题当中词语的构成条件
- ❺ **可查看构成Topic的单词** 蓝色条状图表示所有单词的频率,红色条状图表示在相应Topic中的频率。

※ 数值接近1, 说明以TF (词语频度数) 值为侧重点; 接近0, 说明以IDF (频度反函数) 值为侧重点进行分析。

※  $\lambda$ 数值越低 构成话题的单词的差异越明显, 但通常频度较低。

※ 因此,为了提高对群组的区分效度,对低频单词也需要精确地提炼。

# 查看可视化结果

4. 可视化

钛思通  
TEXTOM



视觉化结果

容量充值

尊敬的 더아이엠씨



收集

词云

条形图

自我中心网络

网络

树形图

LDA主题分布

聚类分析

矩阵网络图

文本情感分析

数据预处理

收集列表

提炼/语义分析

分析列表

数据提炼

矩阵

情感分析

主题分析

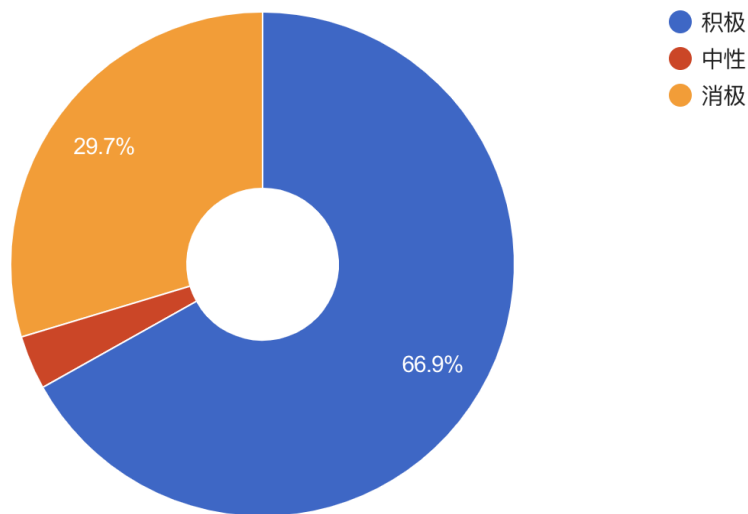
可视化

可视化结果

| 关键词  | 收集日期       | 收集单位   |
|------|------------|--------|
| 三峡大坝 | 2020-07-23 | 550 KB |

下载

词频 情感分析(文本)



文本情感分析是上传全原文数据的基于学习数据分三类极性，由饼图表示的可视化结果。

※ 移动鼠标在饼图上，可看到文件数量和百分比价。

※ 未上传学习数据的话，无法显示可视化结果。

钛思通  
TEXTOM

佣金

容量充值 尊敬的 더아이엠씨

收集

数据处理

数据列表

提炼/语义分析

分析列表

数据提炼

矩阵

凝聚子群分析

情感分析

主题分析

可视化

可视化结果

定制

佣金

词云

条形图

自我中心网络

饼状图

条形图

N-gram 网络

1-way 词树

树形图

2-way 词树

定制分析的可视化结果不需要消耗额外容量。

用你现有的数据尝试完成更多快捷又丰富的数据可视化吧。

词云

按照下面形式，上传包括词频内容的表格文件，将会生成前100个单词的词云

| Sample |     |      |   |   |   |   |
|--------|-----|------|---|---|---|---|
|        | A   | B    | C | D | E | F |
| 1      | 世界杯 | 3299 |   |   |   |   |
| 2      | 天   | 905  |   |   |   |   |
| 3      | 百度  | 705  |   |   |   |   |
| 4      | 俄罗斯 | 675  |   |   |   |   |
| 5      | 快照  | 576  |   |   |   |   |
| 6      | 视频  | 536  |   |   |   |   |
| 7      | 旦   | 376  |   |   |   |   |
| 8      | 客户端 | 267  |   |   |   |   |
| 9      | 多   | 260  |   |   |   |   |
| 10     | 小时  | 245  |   |   |   |   |

词云

A列：关键词 (100个)

B列：词频

查看结果

※ 上传的Excel文件所需的容量从剩余数据中扣除。

※ 可视化结果不会保存,请务必在关闭窗口前下载可视化图片。

36