文本挖掘小探索：避孕药都有哪些内容主题

冯大福 | 2017-08-31 10:29

挖掘 避孕 内容 主题

【数据猿导读】舆情监测一直是众多品牌关注的地方，尤其品牌想知道在品牌推广，品牌策略，品牌广告中出现的问题，从而能进行策略上的改进，但是现在很多人都是读帖子，笔者在4年前做舆情分析时候就是读帖子

舆情监测一直是众多品牌关注的地方，尤其品牌想知道在品牌推广，品牌策略，品牌广告中出现的问题，从而能进行策略上的改进，但是现在很多人都是读帖子，笔者在4年前做舆情分析时候就是读帖子，至今没有太多改善，关注舆情监测中的主题挖掘部分，主题挖掘可以使数据分析师，减轻工作量，去掉读帖子等一系列等的复杂工作，大致了解主题规律。

本文是笔者早前发在某网站上的，由于笔者最近太忙，将本文修改下呈现给大家：

本文分析逻辑：

数据处理

1.数据源：

从各大网站论坛，微博等爬虫关于某避孕药的内容

关键字段名称包含

content Author: 发帖作者(第D列)

Content Forward: 转发的内容(第F列)

Content_Main: 发帖内容(第G列)

Title：发帖内容(第H列)

精品栏目

[2017/07/27] 大数据24小时 More>
[2017/07/24-28] 大数据周周看 More>
[2017/07/24-28] 大数据投融资 More>
[2017/07/24-28] 大咖周语录 More>
[2017/07/24-28] 大数据周聘汇 More>
[2017/07/24-28] 每周一本书 More>
[2016/08-10] 大数据活动公告 More>

专家推荐 More >

涂子沛 傅志华 马亮
苏萌 崔晓波 韩涵
车品觉 刘雷鸣 董飞
郭炜 张涵诚 陈运文

人物专访 More >

活动推荐 More >

ACS 2017中国汽车CIO峰会10 2017-10-25
2017金融科技价值—数据驱动 2017-10-19

其他字段和本文不想关，不阐述

2.加载数据包(r语言)和需要在中文分词中插入的中文词语：

Rwordseg：(4年前用的分词包，不知道现在更新与否)，分词包就是让R语言识别中文，按照单词来视为一个值

插入单词：因为Rwordseq中文词性包含不了其他奇怪词汇，例如： 妈富隆、优思明、短期避孕药、治疗多囊等。插入单词作为模型的变量值

```
insertWords("毓婷")
insertWords("短期避孕药")
insertWords("多囊")
insertWords("治疗多囊")
insertWords("长期避孕药")
insertWords("调经")
insertWords("紧急避孕药")
insertWords("口服避孕药")
insertWords("Yasmin")
insertWords("Yas")
insertWords("Yaz")
insertWords("短效避孕药")
```

3.读入文本分析处理

去掉数字、特殊字符、标准符号

数据探索：大概了解下数据现状

1.根据变量值(单词)统计各个单词出现的次数

避孕药	避孕	口服避孕药	服用	月经	女性	可以
5320	3571	3154	3066	2565	2542	2064
优思明	使用	作用	没有	激素	孕激素	紧急避孕药
2055	1782	1435	1434	1424	1423	1373
副作用	如果	可能	药物	周期	影响	雌激素
1359	1339	1260	1201	1198	1151	1136
卵巢	子宫	安全	一个	排卵	怀孕	增加
1121	1046	1033	1024	985	943	935
妇女	效果	开始	这个	方法	问题	目前
919	907	898	892	849	839	810
身体	发胖	出血	PostsEnd	医生	复方	健康
804	799	787	778	778	761	758
但是	自己	治疗	体重	导致	时间	因为
751	744	743	738	736	732	731
降低	发生	引起	出现	剂量	很多	我们
727	715	715	699	685	676	671
长期	所以	服药	妈富隆	什么	生育	那么
644	641	627	612	608	595	592
主要	风险	减少	建议	一种	一些	妊娠
591	584	579	559	555	542	541

2.根据单词量画词云图

3.重新转化用于聚类的数据格式

根据以上数据探索的词频，词作为colname，词频表示数值，每一行是帖子内容作为id标示

例如：

即每个帖子出现了某词的词频的次数，帖子1中出现避孕药2次，优思明4次，囊中1次

R语言tm包来作处理

即：分词之后生成一个列表变量，用列表变量构建语料库。

由于tm包中的停用词()都是英文(可以输入stopwords()查看)，所以大家可以去网上查找中文的停用词，用removeWords函数去除语料库中的停用词：

生成语料库之后，生成词项-文档矩阵(Term Document Matrix，TDM)，顾名思义，TDM是一个矩阵，矩阵的列对应语料库中所有的文档，矩阵的行对应所有文档中抽取的词项，该矩阵中，一个[i,j]位置的元素代表词项i在文档j中出现的次数。

4.注意：

默认的加权方式是TF，即词频，这里采用Tf-Idf，该方法用于评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度：

在一份给定的文件里，词频 (term frequency, TF) 指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化，以防止它偏向长的文件。

逆向文件频率 (inverse document frequency, IDF) 是一个词语普遍重要性的度量。某一特定词语的IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的TF-IDF。因此，TF-IDF倾向于保留文档中较为特别的词语，过滤常用词。

同时，需要用removeSparseTerms()函数进行降维

数据挖掘

1.查看频率&基本统计

其实就是在数据挖掘查看数据基本统计，目的看下数据逻辑符合不符合社会认知

2.LDA

LDA(Latent Dirichlet Allocation)是一种文档主题生成模型，也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”

具体的算法核心在这里略，因为写太多可能读者看不懂。

VEM = LDA(sample.dtm2, k=10, control = list(seed= SEED)),

根据上面所形成的raw data形成主题词：

解读：

主题分类10个主题，在这个文本中，但是还需要优化

第一个主题是女性服用避孕药后作用会不会增加流产风险

第二个主题是女性避孕和激素的关系(需要优化)

第三个主题医生推荐优思明会不会影响月经(需要优化)

第四个主题口服避孕药会不会影响月经

第五个主题治疗痤疮，激素，多囊

第六个主题和第四个主题一样(需要优化)

第七个主题同上

第八个主题，杂文帖子

第九个主题，会不会是吃完避孕药后发胖

第十个主题，优思明女性服用避孕

确切来说，这10个主题还需要优化，文本经过人工看完应该提炼的的是优思明使用目的，大部分集中在避孕，安全，发胖，治疗痤疮等，少部分会集中副作用等。

结束语

由于4年前做脚本，因此好多需要优化，之后会将优化的和大家分享

优化内容包含

需要在文本库中添加月经不调，治疗痤疮，青春痘等词语

文本还需要继续处理改进去掉postend

主题数目需要加大

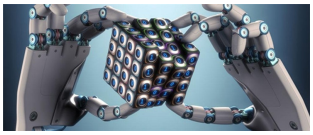
以及主题内容维度需要增加可以让他成为一句话

来源：36大数据

相关文章

刷新

七夕后大数据：CBNData发布线上避孕套十大品牌



如何更好地阅读很多数学相关内容的机器学习论文？

【大咖周语录】深耕数据挖掘，实现互联互通，在银行风控管理中 ...

我要评论

我想要评论.....

提交评论

ok