# 文本分析初步

李锋 中共中央党校(国家行政学院)



# 文本数据的分布

合 首页

三 浏览

□ 深度

📃 📃 回复反馈

月 月 用户中心

☑ 快速留言

当前位置:人民网 >> 地方领导留言板



地方领导留言板 Message Board for Local Leaders 54位 省委书记、省长 公开回应地方领导留言板网友留言





□Ⅰ最新动态

### 居民区屠宰场:其杀猪声不断 小孩不敢住

#### 教师投诉黑作坊浓烟污染学校 河南新野县接报查处

- ・内地在藏教师反映昌都市机票紧张问题获官方回复
- •甘肃平凉网友咨询农村危房改造补助政策获官方回复

### 大学生反映学校周边彩钢房存火灾隐患 官方整顿

- 河南太康县网友反映农村公路桥年久失修问题获官方回复
- ·四川省委书记王东明、省长魏宏回复9条网友留言

#### 桥面漏洞、两侧无护栏 河南网友反映进村危桥

- ·安徽省委书记张宝顺集中回复36条网友留言
- · 陕西省委书记赵正永答复学校乱收费等问题

▼ > 请选择 ▼ > 请选择

快速浏览

北京 | 天津 | 河北 | 山西 | 内蒙古 | 辽宁 | 吉林 | 黑龙江 | 上海 |

|浙江||安徽||福建||江西||山东||河南||湖北||湖南||广东||

广西 海南 重庆 四川 贵州 云南

宁夏 | 新疆 | 香港 | 澳门 | 台湾 |

「点击地区进入留言板留言]

昨日留言 175条 今年留言 39402条 历史总留言 607029条 昨日回复 370条 今年回复 27289条

历史总回复 334533条



### 人民日报图文数据库(1946-2018)

欢迎您:中共中央党校 | 关闭退出 | 个人中心 | 搜索中心

2018年 10月 22日 星期一 今日24版 文章114篇

请输入关键词

检索

检索选项 ● 标题 ● 正文 ● 标题+正文

#### 第1版 [要闻] 文章8篇

- 1 心无旁骛创新创造 踏踏实实办好企业
- 2 在新时代党的真挚关怀下不断奋进
- 3 回信
- 4 唱响新时代奋斗者之歌(社论)
- 5 敢为砥柱立中流
- 6 丰收时节稻谷香
- 7 当好政治生态的"护林员"
- 8 十九届中央第二轮巡视已全部进驻





22

ARTHHER

---



STREET BY AS AN ART STREET, IN-SERVICE STREET OF ROSE STREET STREET, IN-SERVICE STREET, INSTRUMENT STREET, IN-SERVICE STREE STATE CARACTER STATE OF THE STA

十九届中央第二轮 巡视已全部进驻

PERSONAL PROPERTY AND PROPERTY

NAME OF TAXABLE PARTY.

\*\*\*\*\*\*\*\*\*\* minul, equera

\*\*\*\*.................. AND THE RESERVE OF THE PARTY OF

#### 在新时代党的真挚关怀下不断奋进 一 习近平总书记关心工人阶级和工会工作纪实

を担ぐ込み来、2・2回です。 でする。有点、内容とへのでける。 ではる。有点、内容とへのでける。 では3・2を10である。ながある。 と、を含む・・・・ことである。有例で を、を含む・・・ことである。 をはませ、上が自然をはなる。 を対しているはましたもの。 は、ためをなるとはなる。 ALERGE CO. WERLING CARROLL PROBLEMS OF ABOUT AND THE AND

大力弘扬安进精神。 保証时代包存者 TOPINETIES PONTE TORREST TORRE \*\*\*\*\*\*\*\*\*\*\*\*\*\*

·其大田中國市,後日二首衛行政上 OWNERS AND PROPERTY OF THE PARTY OF THE PART 1289404 #1794

### 唱响新时代奋斗者之歌

— 热型祝贺中国工会第十七次全国代表大会开幕



\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\* ACCOUNTED FARMANCE FARMANCE FACTOR #2. F#2. ##\*\* LA. ##4. 

当

\*\*\*\*\*\*\*\*\*\*\*

#### 凝聚起亿万职工的磅礴力量 中侧工会第十六次全国代表大会以来工作标述

好 # 政 成力不减、泵弧节症。 治 正报刊创始终高格 MEREZHAUSE, NT 态 工作。 "在现象由企业可,需要出出, 设定于有有效的原则。"他们不会 在现在社会是不同じ来说。一个 年1月,以是有管理工作证明成功。一 年1月,以是有管理工作证明成功。 新,每个证明电流形下中提供的 新可以出版的成功和。 整的

\*\*\*\*\*\*\*\*\*

879



### 中国裁判文书网

China Judgements Online

★ 首页 刑事案件 民事案件 行政案件 赔偿案件 执行案件 民族语言文书

高级检索。

输入案由、关键词、法院、当事人、律师

搜索

?

# 开放 动态 透明 便民



真的很惊喜了哈,第一是速度,上午买了下午到,当日达效果不错,看来京东的物流确实很强大!照顾到买新机的人很着急!谢谢!速度666第二,打开很惊艳,手感超级好!试用了一个多小时,系统流畅,拍照效果好,666.第三,性价比很高!5.5英寸,4G运行内存,64存储内存,电池容量大!很不错,处理器也还可以哈!如果你考虑在1500左右的手机这款就是值得购买的!好吧!上几张图片供大家看吧!



### 头条校园 🗸 🀽



#高考作文#【全国卷(Ⅱ):6个古诗句选2个或者3个自行立意】

- 1、天行健, 君子以自强不息《周易》
- 2、露从今夜白,月是故乡明(杜甫)
- 3、何须浅碧深红色, 自是花中第一流(李清照)
- 4、受光于庭户见一堂,受光于天下照四方(魏源)
- 5、必须敢于正视,这才可望,敢想,敢说,敢做,敢当(鲁迅)
- 6、数风流人物,还看今朝(毛泽东)

你们怎么看? 😌 收起全文 🧥





# 案例:自媒体指责马蜂窝数 据造假

全球旅游消费指南

覆盖全球200多个国家和地区

100,000,000 位旅行者

920,000 家国际酒店

21,000,000 条真实点评

382,000,000 次攻略下载

中国旅游行业第一部"玩法"

'只占"2.91%却放在首页



#### 展望未来的夏天

不好吃啊。普遍太咸, 北方菜特点, 这几年一直没变!??, 饮食要少盐啊。炸酱面也太咸, 凉

太一般般。一般都是冷菜,冷点心,连羊肉串也是冷的。当然热菜也有。豆汁第一次吃,其臭味有点类似南方的千里飘香,我是可以多吃几瓢羹,家人。价格贵,不便宜。

本人胃口好, 吃得杂, 应该不是吃不习惯的原因。 点评往往只有简短几行字





#### 孤独星球-冰岛游记

这些年跑过很多地方,如果说 圣托里尼 是浪漫, 北海道 是梦幻, 西藏是苍茫,那么 冰岛 的关键词就是孤寂。 准备工作 签证:签证大概需要15个工作日。从 丹麦 入境,办的 丹麦 的 申根 签...

游记则以图片为主

○ 冰岛, by 🕅 然若 ⊙ 980/26



(桐乡永利广场后) 2016/12/14, 她表示每

210 关注

643 粉丝

### 餐饮点评占比

■ 其他用户 ■ 抄袭账号





中心新年大采购! 逛累了 ② 来坐会儿! ❸ ♥ ② 圖 ♦ 正好有张买2赠一券 ▮

克力和冰瑶柠檬茶! 标配 基本每次都是这样♥ 老公·艮喜欢摩卡可可碎冰星冰乐。 休息完继续逛哇哈哈

### 小班长10号今年2月初陪

### 老公和妈妈在上海喝星巴克



今天中午题老婆王附近找吃的,找来找去,最后选的这家,不是他们家下面标注音,我 环境一般般吧, 灯光稍暗, 不过虽然店小, 居然还有二 是直接柜台点餐,服务员还是比较热情的,会主动推荐,介绍,最后我 辣,感觉还是有点辣好吃一点,香一点,挺不错,可以一试。

### 而去年11月中旬她还

### 跟老婆在深圳吃面



Dirty Lickins

★ ● ● ★ 値得一去

男朋友和我在这里吃饭可能每个月有一个星期天,但老实说,我希望我能每天都有 他们的鸡招标。我拿到了甜辣椒的三片组合。我最喜欢的酱是情人的咬,但他们没有在 一段时间:(鸡的招标是伟大的, 我其实并不偏向鸡招标/条/手指正常, 但他们的招标是巨 大的优质鸡肉块轻轻地面包屑和酱汁。

唯一的缺点是它有点贵,有时他们不接电话。我们喜欢提前拿货。上一次,我等了好五 分钟,但没有人拿起。我等了几分钟,然后又试了一次,过了大约15分钟后再次尝试, 但是线路只是响了, 响了。

### 去年圣诞前她与男友在夏威夷



Theny restauran

達荐

误一并抄了下来

喜欢这个地方! 大气氛, 美味的食物, 农场新鲜的食材, 难以置信的服务!

食物很好&我喜欢他们使用来自当地农场的原料(他们使用来自Happy Boy Farms的胡萝卜……相信我,这些是您尝过的最好的胡萝卜)。我试过的两道菜让我难忘:1.布鲁塞尔豆芽松露(无板帕玛森芝士)小盘子。真的很喜欢这些。很好的味道& 蜂甘蓝是我的最爱之一

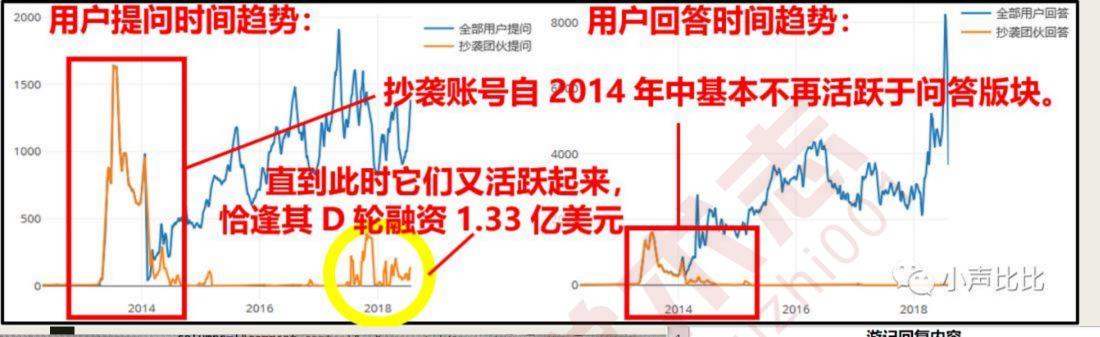
来自"简明<mark>英汉词典</mark>"我的主菜有箭鱼菜。这不是我尝过的最好的箭鱼菜,因为口味(刺山 相利相檔)的组合不如我想象的那么美味,但是我清理了我的盘子,所以它当然好!

> 抄袭二:因为Yelp上的 评论是英文的

### 抄袭三:

### 不慎把不该抓取的内容抓了进来





columns=["comment_content", "comment_id", "poi-typeid", "uid"],	1	游记回复内容	回复次数
filter_fn=get_content_len	2	有没有详细介绍?	3238
22]: review_text_count["wordcount"] = review_text_count.comment_content.str.len()   22]: print("游记总字数", youji_word_count.word_count.sum())   print("点评总字数", review_text_count["wordcount"].sum())   print("被删除字数", review_text_count[review_text_count.uid.isin(copy_dogs_uids)]["wordcount"].   游记总字数 4897982033   点评总字数 1744779058   被删除字数 1346405176   In []:	3	没更新完吧,等楼主更新哦~	3234
	4	好棒的游记,楼主多写点吧,写完记得通知我,哈哈	3188
	5	没懂呐?有没有详细介绍?	3184
	6	好美的照片,文字再多点就更好啦	3179
	7	不错不错,楼主快出来写点详细攻略哇~	3168
	8	只看图片不过瘾啊!	3165
	9	光看照片就觉得不错啊,有没有详细攻略?	3144
	10	没完呢吧? 感觉没写完哈	3136
	11	求详细攻略啊!	3132
	12	强烈关注楼主~请继续!	3130
	13	lz也多多分享一下攻略嘛,想去看看。	3129
	14	有点兴趣,要有详细的攻略就好啦。	3124
	15	想知道楼主的评价啊,怎么样?	3124
	16	关注! 等楼主都更完了再看! 楼主加油!	3114
	17	有没有什么需要注意的?	3112
	18	我有个小建议,楼主把攻略写详细点吧才会吸引更多读者呀。	3106

供表了成了 左插环次中代的最兴 nana

Bright, J. (2012). The Dynamics of Parliamentary Discourse in the UK: 1936-2011.

- 利用议会文本记录分析英国议会中议会争论的发展特点:
- 利用英国议会解析网站提供的议会资料构建了1936-2011 年间英国国会下院发布的由7.4亿单词所构成的数据库
- 利用自动编码技术对法律、国防、环境、卫生、就业、权利、教育、农业、经济等关键词在75年间的出现频率进行了描绘
- 发现这些关键词的出现频率具有一定稳定性,但也存在很大变化。例如,争论变得更加激烈;环境议题更为突出,而农业等议题则逐渐衰落;女性议员倾向于较长的发言时间;而贵族议员被打断的频率更高一些。

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases.

- 利用文本分析方法研究美国参议员与选民的政治沟通;
- 采用了贝叶斯分层主题模型;
- 搜集美国参议院2007年来发布的24000余份新闻通告,利用无监督学习法 (unsupervised learning methods)进行文本分析;
- 发现:每个参议员的关注重点与其他参议员的关注事件之间存在显著相关性;关注重点的地域分布具有一定的集聚性;议员对参议院拨款法案的关注程度与他们对德敏特—麦凯恩修正案的反对票呈现正相关关系。

### 反腐新政:19000多贪腐官员的文本分析

### 贪官来源



### 涉腐原因



### 新闻联播中的"国际"与"国内"地域关注

国际地域关注

国内地域关注





# 文本数据的普遍性

- 微信、微博
- 论坛、博客
- 商品评价
- 投诉文本
- 电子邮件
- 调查问卷开放问题

- 新闻媒体
- 法律判决文书
- 档案材料
- 议会发言
- 政府文件

# 文本分析的重要性

- □ 与结构化数据结合,提升决策和预测模型的准确性。
  - 结合互联网舆情与法院判决信息、股价信息等,评判企业发展前景;
  - 结合交易流水信息,提高消费者画像精确度;
  - 预测国会选举、投票率等

- □ 人工智能系统
  - 自然语言处理(NLP)、文本分析、语音分析是核心技术

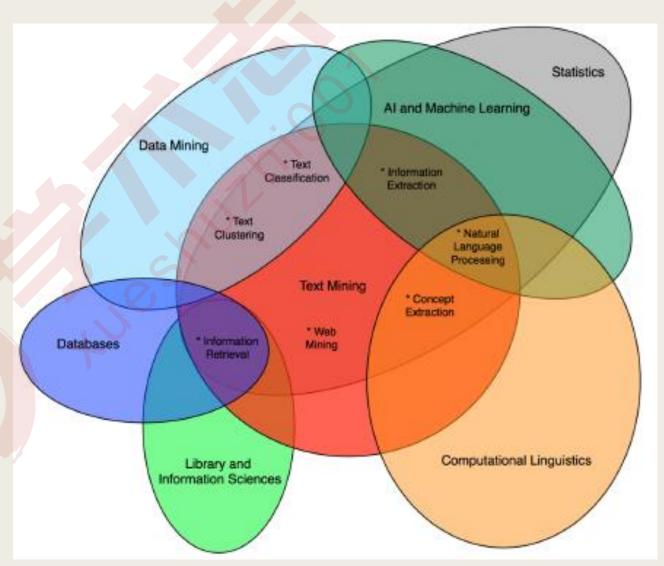
# 文本分析的挑战性

- □ 分析对象局限性
- □ 分析语句歧义大
  - 一词多义:需要结合上下文才能理解,例如"这款车油耗很高""电脑性价比很高"
  - 多词多义:"发货速度快"、"物流超级赞"、"物流迅速"
- □ 维度灾难
  - 向量空间模型极其高维,而且非常稀疏
- □ 语言表达随意性
  - 老司机 十动然拒



# 文本分析中的主要问题

- 分词和词性标注
- 信息抽取
- 文本分类和聚类
- 语义分析
- 主题分析
- 情感分析
- 机器翻译



# 文本分类的预处理

- 文本分词
- 去停用词
- 文本特征提取
- 文本向量化

# 文本分词

- 文本分词是预处理过程中必不可少的一个操作,因为后续的分类操作需要使用文本中的单词来表征文本。
- 文本分词包括两个主要步骤,第一个是词典的构造,第二个是分词算法的操作。 Eg:我的家乡是广东省广州市-->我/的/家乡/是/广东省/广州市
- 具体实现: python 中安装jieba

# 去除停用词

- 去停用词也是预处理过程中不可缺少的一部分。
- 因为并不是文本中每一个单词或字符都能够表征这个文本,比如说"这个"、"的"、 "一二三四"、"我你他"、"012……9"等等,那么这些词就应当从文本中清除掉。

# 特征提取

- 特征项选取,把词条中最能代表某类文本信息的词条挑选出来,作为文本的特征项。 实验结果表明简化特征项不但不会使分类结果准确率降低,而且还会使结果更加准 确。
- 在文本处理中,一些常用特征提取评估函数有文档频数(document frequency)、信息增益(information gain)、期望交叉熵(expected cross entropy)、互信息(mutual information)、统计(CHI)、文本证据权(the weight of evidence for text)等

# 向量空间模型

- □ 向量空间模型(VSM:Vector Space Model)由Salton等人于20世纪70年代 提出,并成功地应用于著名的SMART文本检索系统,目前广泛应用于自然语言处理中。
  - 具体而言,每一个词语代表一个维度,取值表示该词语的出现频次或者是否出现

# 每个词语如何在向量空间取值?

■ 词袋模型 (bag of word )

将文本表示成为词典维度大小的高维向量,用0和1来表示。通俗地讲,就是把一篇文档看成词袋,里面装着一个个的词。从而,将一篇文档转化成了一个个的词(或者称之为term),显然地,文档转化成一个个的词之后,词与词之间的顺序关系丢失了。

■ TF模型 (Term Frequency )

tf(term frequency),是指该词语(term)在某篇文档中出现的频率。Tf(t,d)代表了term在document中出现的词数,反映出该词语的重要程度。

■ TF-IDF模型

使用词语在整个文档中的逆文档频率来归一化TF权重,以此来消除常用词在文档中权重过大的现象,并由此突出不同文档中的差异

# 举例

■ 文档1:小明的父亲认真教育了小明。

■ 文档2:教育问题是关系国家发展的大问题,必须提高教育经费支出。

■ 分词处理:

小明\的\父亲\认真\教育\了\小明\。

教育\问题\是\关系\国家\发展\的\大\问题\,,\必须\提高\教育\经费支出\。

■ 停用词处理:

小明\父亲\认真\教育\小明\

教育\问题\关系\国家\发展\大\问题\必须\提高\教育\经费支出\

### 教育\问题\关系\国家\发展\大\问题\必须\提高\教育\经费支出\

词典	词袋模型	TF模型
教育	1	2
问题	1	2
关系	1	1
国家	1	1
发展	1	1
大	1	1
必须	1	1
提高	1	1
经费支出	1	1
小明	0	0
父亲	0	0
认真	0	0

## 多个词特征:N-gram

### □把多个词多做单一的特征值

- N表示考虑将多少词进行组合
- 1-gram (unigrams): 公司
- 2-gram (bigram):有限公司
- 3-gram (trigrams):科技有限公司

马尔科夫假设:一个词的出现仅仅依赖于它前面出现的有限的一个或者几个词。

- □ 将词序纳入考虑
- □ 维度指数级增长,空间稀疏程度高

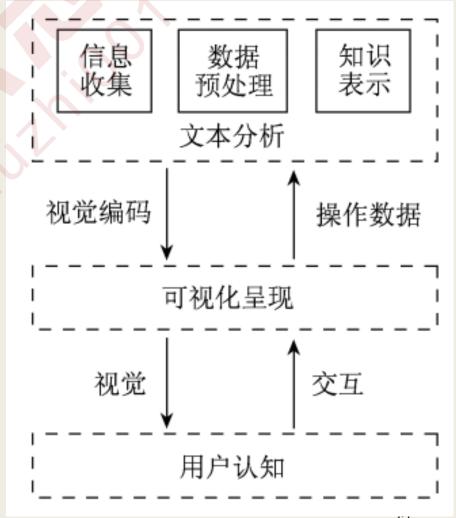
# 文本降维

- □在机器学习中需要考虑维度灾难和数据系数的问题
  - 假设词表中有20000个词,如果是bigram那么可能的N-gram就有4000000个,如果是trigram,那么可能的N-gram就有800000000个!
- □ 对文本的降维处理非常必要,并非越多越好
  - 只选取具体主题的词语
  - 停用词和常用词删除(了,呢)
  - 模型方法 (Latent Semantic Analysis、 probabilistic Latent Semantic Analysis )

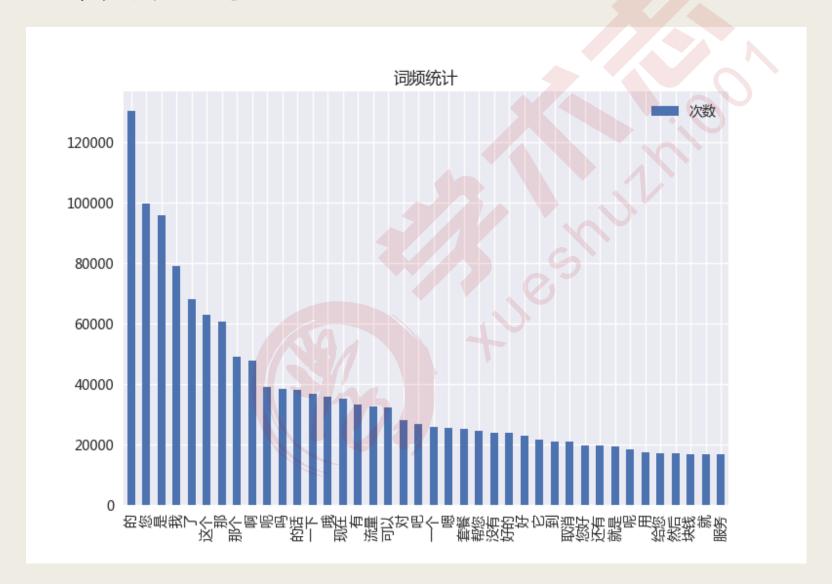


# 文本可视化

□ 文本可视化技术综合了文本分析、 数据挖掘、数据可视化、计算机 图形学、人机交互、认知科学等 学科的理论和方法,为人们理解 复杂的文本内容、结构和内在的 规律等信息的有效手段。



# 词频统计



# 案例:政府工作报告











新华网数据新闻部 出品

资料来源:十二届全国人大四次会议开幕会直播实录,新华网 新华网版权作品,未经许可请勿转载。 联系电话:010-88050730 中日期一性和 长江事名转形由突 谢谢!