数据库系统概论新技术篇

文本大数据分析及应用案例

窦志成

中国人民大学信息学院

2017年7月

文本大数据分析及应用案例

- 课程内容
  - 交互式文本大数据分析系统：时事探针
  - 自然语言处理与文本挖掘基础算法
  - 文本搜索、文本分析系统构建

互联网文本大数据

搜索引擎

1998

搜索引擎真的够用了吗？

2017

Google 谷歌

写一篇关于雾霾的报告

雾霾 百度一下

日本担忧北京雾霾飘到日本
中国日报 11小时前
启动空气重污染"红色预警",是北京自2013年10月引入预警机制以来的第一次。 日本山
和东北大学的研究团队对美国航天局(NASA)的人造卫星拍摄的图像进行分析后...
42条相同新闻 - 百度快照

北京雾霾预计中午消散 周末将再度来袭
新华网 15小时前
北京雾霾预计中午消散 周末将再度来袭---昨天,此轮污染过程进入最重的一天,北京市大
区的空气质量陷入六级严重污染水平。好在今天一早,冷空气快马加鞭进京,自...
37条相同新闻 - 百度快照

北京市治理雾霾 五年总投资7700亿
中国经济网 4小时前
中国经济网北京12月10日讯 (记者 祝惠春) 北京雾霾现红色预警。大气治理再次成为人
的热点。 12月10日,在国务院新闻办新闻发布会上,北京市常务副市长李士祥、...
11条相同新闻 - 百度快照

北京副市长回应"治理雾霾是否受到周边省份阻力"
人民网 9小时前
资料图。中新网记者金硕摄 视频:京津冀等六省区欲打破行政区划共同治理雾霾来源:辽
中新网12月10日电谈及"北京市政府是否在实施对抗雾霾措施时感受到来自...
48条相同新闻 - 百度快照

北京雾霾天什么最畅销?外媒:不是口罩是安全套

新浪新闻 11小时前
原标题:北京雾霾天什么最畅销?外媒:不是口罩是安全套点击图
下一页2015年12月7日,汽车在雾霾笼罩下的北京东三环国贸桥上
驶。新华社记者罗晓光摄参考消息网... 30条相同新闻 - 百度快照

人民网 people.cn 人民网 >> 时政 >> 滚动新闻

sina 新闻中心 综合

北京雾霾天什么最畅销？外媒：不是口罩

24

原标题：北京雾霾天什么最畅销？外媒：不是口罩是安全套

点击图片进入下一页

2015年12月7日，汽车在雾霾笼罩下的北京东三环国贸桥上行驶。新华社记者 罗晓光 摄


雾霾
百度一下
北京雾霾持续 天安门武警首次戴口罩执勤
搜狐 15小时前
12月9日,北京,在天安门广场执勤的武警战士首次在雾霾天佩戴口罩执勤。北京市应急办于12月
至12月10日将启动雾霾红色预警...
为您找至
590,000个
找到相
50,000篇
其实在
前景更
:陆金所 拍拍贷 万富宝
寸前
:出行少开私家车,尽量使用公共交通工具;发现身边
环保部门举报。同理,在"问题平台"频出的P2P"行业
的就... 2条相同新闻 - 百度快照
算算雾霾的经济账!看雾霾消耗了多少经济增长前景
圣济账。【经济
和印度造成的经
百度快照
现有"搜索引擎"不能满足用户对大规
模文本数据的深层次聚合分析的需求
到了5元一斤。(记者史林静、陈子夏、天桂峰、倪元锦、宋晓东、李明、丁...
54条相同新闻 - 百度快照

搜索引擎的缺陷

比利-博斯沃思，DataStax CEO

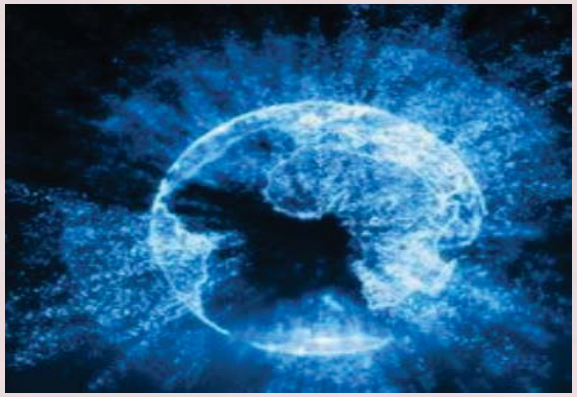
从现在开始的10年内，
当我们回顾大数据时代是如何发展时，
我们会震惊于
在以往做出决策时信息的匮乏

以及信息获取工具的匮乏

引自“互联网女皇”2015年全球互联网趋势报告

互联网文本分析引擎

- 互联网文本分析引擎：在搜索的基础上，提供文本的**深度分析**，帮助用户从文本中获取**高阶知识**
- 就像一个“超人”，帮助用户完成对大规模文本的阅读和理解，并对其中所包含的关键信息与知识进行抽取、挖掘并汇总，并最终通过交互式的分析过程让用户对挖掘到的高阶知识进行浏览和分析，进而为用户决策提供支持



窦志成, 文继荣. 大数据时代的互联网分析引擎[J]. 大数据, 2015, 1(3):36-47.

websensor.playbigdata.c

websensor.playbigdata.com/fss3

时事探针 互联网分析引擎 by 中国人民大学 窦志成，文继荣

douzc 退出 日志 话题

Search 查询 示例查询

参考资料

大数据时代的互联网分析引擎 互联网分析引擎PPT User Manual（用户手册） 演示视频（密码bigdata）

目前该系统已被300余家政府和企业使用

对比分析


正面政策


争议政策


负面政策


期待政策


热点社会话题


负面话题


正面话题


争议话题


2015年政策

2015网络流行语排

2015网络流行语


中国人民大学学院

银行

websensor.playbigdata.c

websensor.playbigdata.com/fss3

时事探针 互联网分析引擎 by 中国人民大学 窦志成，文继荣

douzc 退出 日志 话题

Search 查询 示例查询

wu'mai 工具箱(分号)

1. 雾霾 2. 五脉 3. 无脉 4. 唔买 5. 无

大数据时代的互联网分析引擎 互联网分析引擎PPT User Manual（用户手册） 演示视频（密码bigdata)

对比分析


正面政策


争议政策


负面政策

期待政策


热点社会话题


负面话题


正面话题


争议话题


2015年政策

2015网络流行语排

2015网络流行语


中国人民大学学院

银行

websensor.playbigdata.c

websensor.playbigdata.com/fss3

时事探针 互联网分析引擎 by 中国人民大学 窦志成，文继荣

douzc 退出 日志 话题

雾霾 查询 示例查询

参考资料

大数据时代的互联网分析引擎 互联网分析引擎PPT User Manual（用户手册） 演示视频（密码bigdata)

对比分析


正面政策


争议政策


负面政策

期待政策


热点社会话题


负面话题


正面话题


争议话题


2015年政策

2015网络流行语排

2015网络流行语


中国人民大学学院

银行

http://websensor.playbigdata.com/fss3/Search2.aspx

雾霾

websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

时事探针 互联网分析引擎 by 中国人民大学 窦志成，文继荣

douzc 退出 日志 话题

雾霾 查询 示例查询 维度大小

雾霾 Finished.

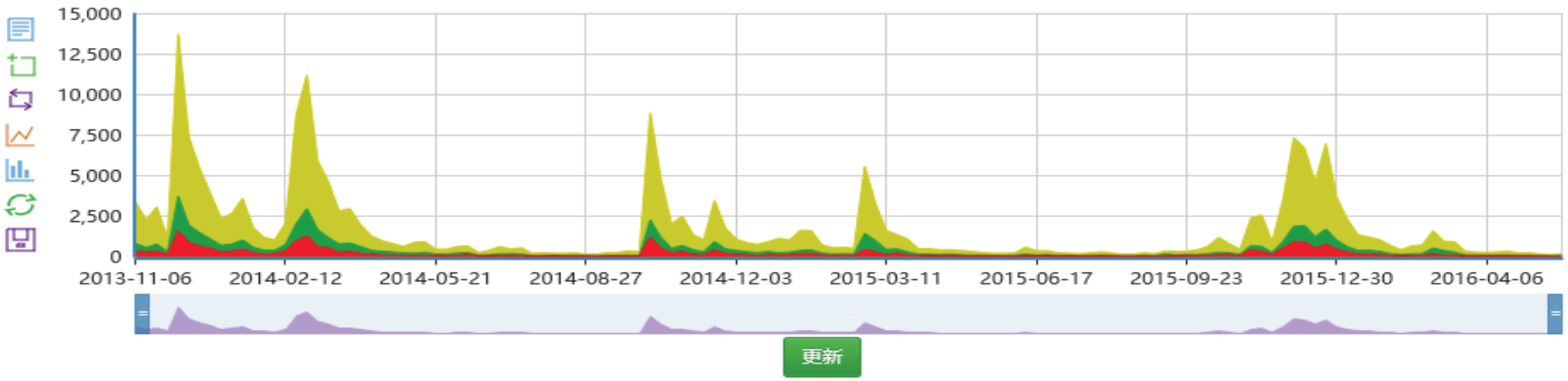
取消关注 巨屏版

深度分析 比较

雾霾

220929	36286	26113	-0.815	2568641
报道总数	正面报道	负面报道	倾向性评分	互联网搜索次数

雾霾 媒体报道趋势

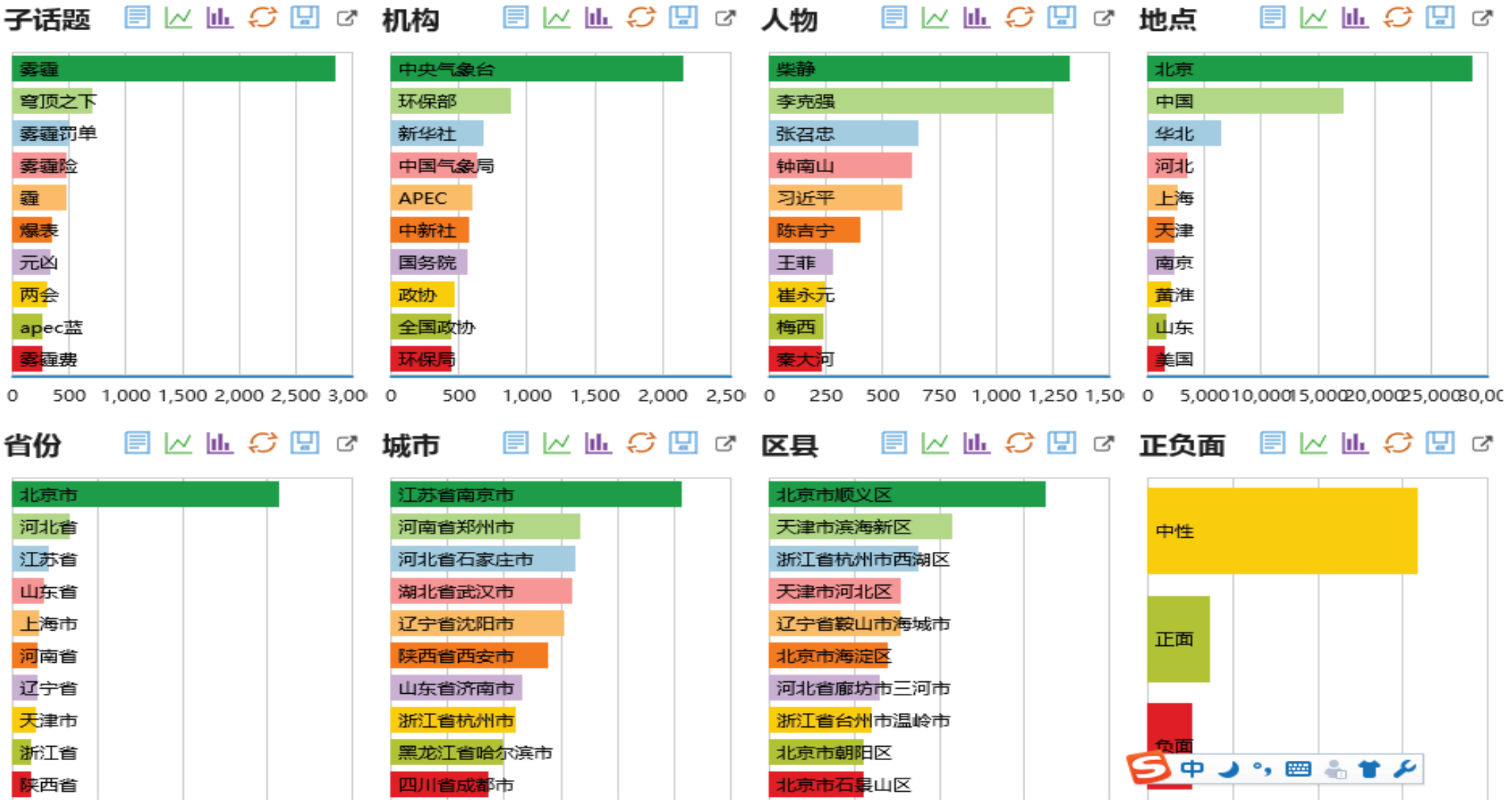


更新

雾霾

websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

Total 220929 records for 雾霾.



雾霾

websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

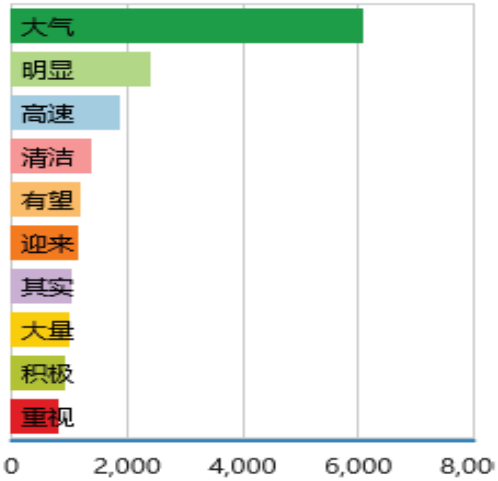
0 10,000 20,000 30,000 40,0(

0 500 1,000 1,500 2,000 2,500 3,00

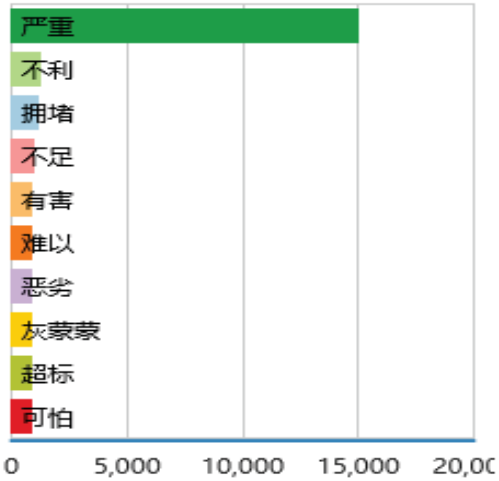
0 20 40 60 80

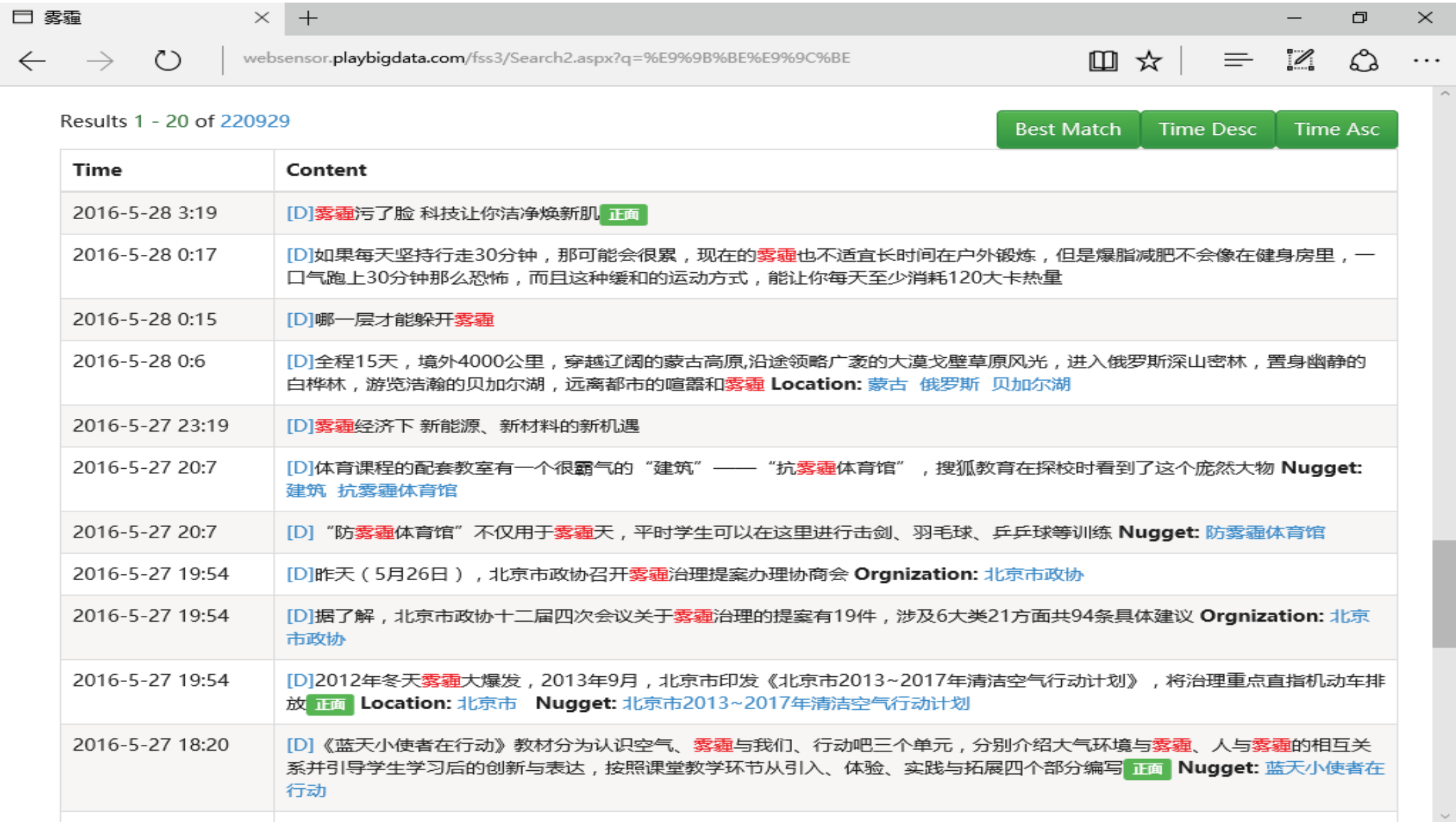
0 50,000 100,000 150,000 200,0

正面词



负面词




雾霾
websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE
Results 1 - 20 of 220929
Best Match
Time Desc
Time Asc
Time
Content
2016-5-28 3:19
[D]雾霾污了脸 科技让你洁净焕新肌 正面
2016-5-28 0:17
[D]如果每天坚持行走30分钟，那可能会很累，现在的雾霾也不适宜长时间在户外锻炼，但是爆脂减肥不会像在健身房里，一口气跑上30分钟那么恐怖，而且这种缓和的运动方式，能让你每天至少消耗120大卡热量
2016-5-28 0:15
[D]哪一层才能躲开雾霾
2016-5-28 0:6
[D]全程15天，境外4000公里，穿越辽阔的蒙古高原,沿途领略广袤的大漠戈壁草原风光，进入俄罗斯深山密林，置身幽静的白桦林，游览浩瀚的贝加尔湖，远离都市的喧嚣和雾霾 Location: 蒙古 俄罗斯 贝加尔湖
2016-5-27 23:19
[D]雾霾经济下 新能源、新材料的新机遇
2016-5-27 20:7
[D]体育课程的配套教室有一个很霸气的“建筑”——“抗雾霾体育馆”，搜狐教育在探校时看到了这个庞然大物 Nugget: 建筑 抗雾霾体育馆
2016-5-27 20:7
[D]“防雾霾体育馆”不仅用于雾霾天，平时学生可以在这里进行击剑、羽毛球、乒乓球等训练 Nugget: 防雾霾体育馆
2016-5-27 19:54
[D]昨天（5月26日），北京市政协召开雾霾治理提案办理协商会 Orgnization: 北京市政协
2016-5-27 19:54
[D]据了解，北京市政协十二届四次会议关于雾霾治理的提案有19件，涉及6大类21方面共94条具体建议 Orgnization: 北京市政协
2016-5-27 19:54
[D]2012年冬天雾霾大爆发，2013年9月，北京市印发《北京市2013~2017年清洁空气行动计划》，将治理重点直指机动车排放 正面 Location: 北京市 Nugget: 北京市2013~2017年清洁空气行动计划
2016-5-27 18:20
[D]《蓝天小使者在行动》教材分为认识空气、雾霾与我们、行动吧三个单元，分别介绍大气环境与雾霾、人与雾霾的相互关系并引导学生学习后的创新与表达，按照课堂教学环节从引入、体验、实践与拓展四个部分编写 正面 Nugget: 蓝天小使者在行动

	系并引导学生学习后的创新与表达，按照课堂教学环节从引入、体验、实践与拓展四个部分编写 正面 **Nugget:** 蓝天小使者在行动
2016-5-27 18:20	[D]在这过程中孩子们会观察生活，产生了不少很有启发的研究课题，如《烟花爆竹对环境的影响》、《鲜花保鲜剂对环境的影响》、《雾霾的形成与原因》等 **Nugget:** 烟花爆竹对环境的影响 鲜花保鲜剂对环境的影响 雾霾的形成与原因
2016-5-27 16:26	[D]发达国家历史上也经历了这个阶段，但是像伦敦、东京、美国的很多地方，五六十年代污染一样是很严重，一样是空气雾霾、光化学污染等很严重，但是进入六七十年代以后，发达国家通过严格的立法、执法，有效地改善了 正面 **Location:** 伦敦 东京 美国
2016-5-27 15:34	[D]这是在中国做生意 一条 几乎人人共知的规则，无论从小札面带笑容地跑过雾霾满天的天安门广场，还是此次苹果掷资10亿美金砸向滴滴后库克马不停蹄地赴京搞一场开发者活动示好，都免不了被猜测带有获得商业政策化 正面 **Location:** 中国 天安门广场 **Person:** 库克
2016-5-27 14:52	[D]2015年两会前，前央视记者柴静通过多个网络平台推出自费拍摄制作的雾霾调查纪录片《穹顶之下》，让空气质量和污染治理议题成为舆论焦点，也引起立法和监管部门、能源企业、环保组织的高度关注 **Person:** 柴静 **Nugget:** 穹顶之下
2016-5-27 14:6	[D]从 这个意义上说，雾霾成为灾害性天气，不是治理乏力的借口，而应该成为给治理者 “增压” 的阀门 **Nugget:** 增压
2016-5-27 14:6	[D]雾霾出现于何时还有争议，但早在商代的甲骨 上，就已经留下了对 “霾” 的记载，《诗经》中也有 “终风且霾” 的诗句 **Nugget:** 霾 诗经 终风且霾
2016-5-27 14:6	[D]从这个意义上，对雾霾性质的清晰界定也应该成为治理的工具，最终的目标，是让雾霾在我们写下的历史中 绝迹 正面
2016-5-27 14:6	[D]文章关键词： 雾霾 环境保护 争议
2016-5-27 14:6	[D]雾霾：是 “天灾” 也是 “人祸” **Nugget:** 天灾 人祸

« Previous 1 2 3 4 5 Next »

The static address for the result

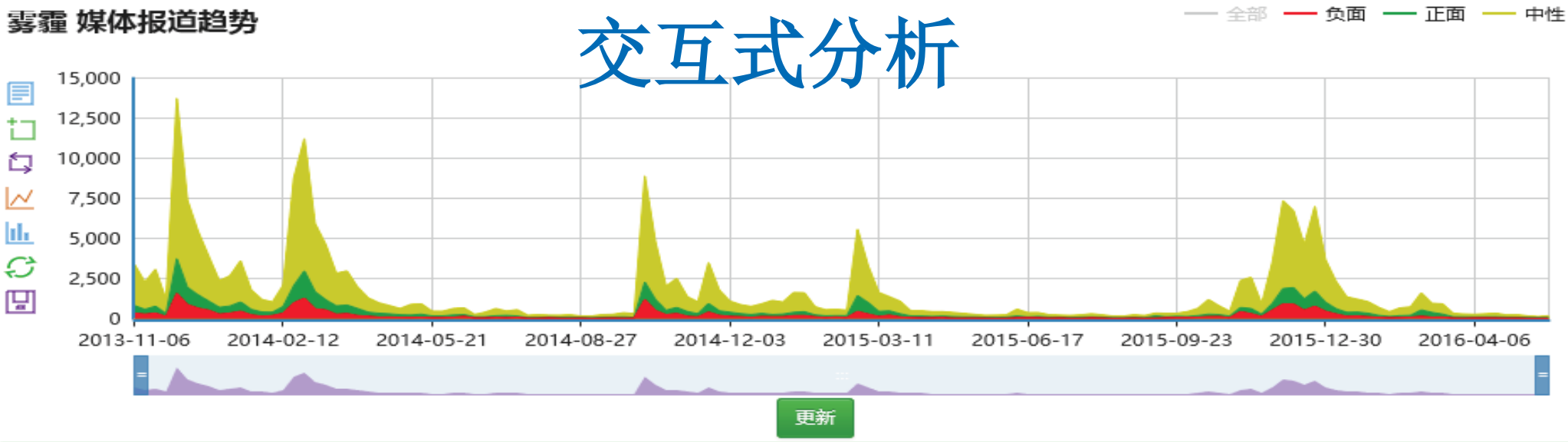
雾霾

websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

220929	36286	26113	-0.815	2568641
报道总数	正面报道	负面报道	倾向性评分	互联网搜索次数

雾霾 媒体报道趋势

交互式分析



Total 220929 records for 雾霾.

子话题	机构	人物	地点
雾霾	中央气象台	柴静	北京
穹顶之下	环保部	李克强	中国
雾霾罚单	新华社	张召忠	华北
雾霾险	中国气象局	钟南山	河北
霾	APEC	习近平	上海
爆表	中新社	陈吉宁	天津

雾霾

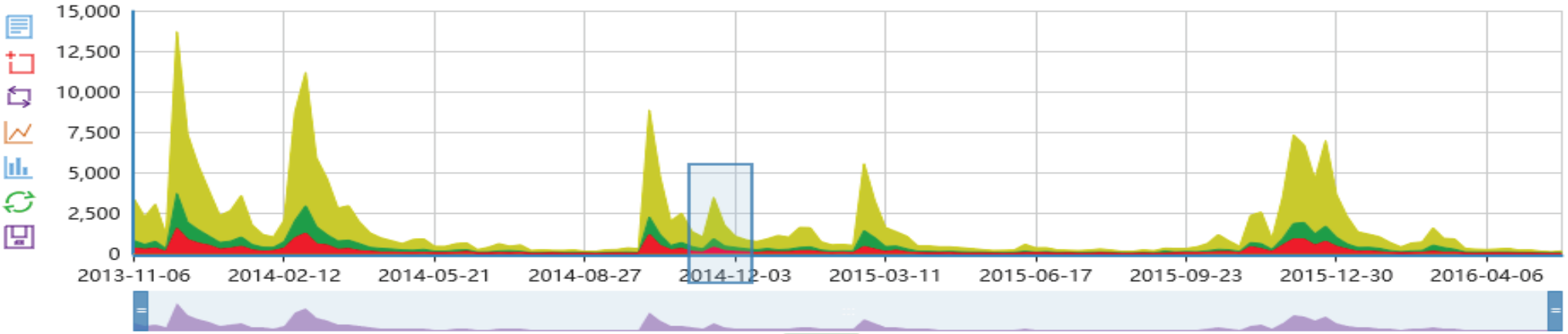
websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

雾霾

220929	36286	26113	-0.815	2568641
报道总数	正面报道	负面报道	倾向性评分	互联网搜索次数

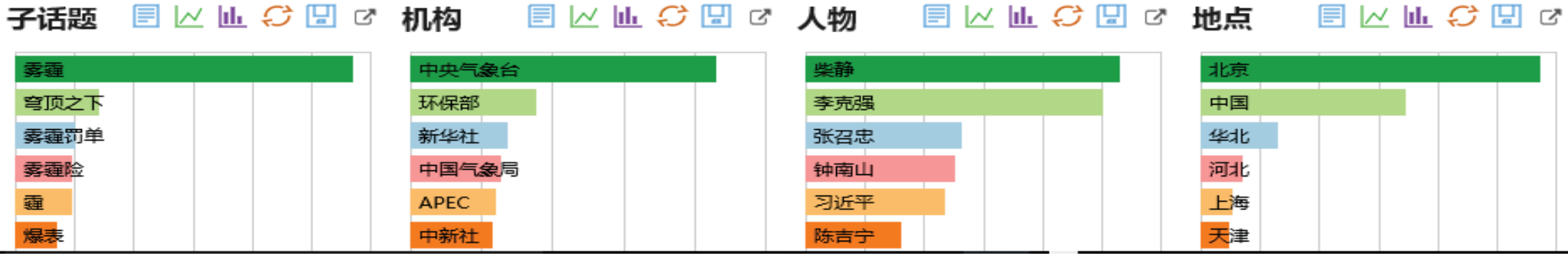
雾霾 媒体报道趋势

全部 负面 正面 中性



更新

Total 220929 records for 雾霾.



雾霾

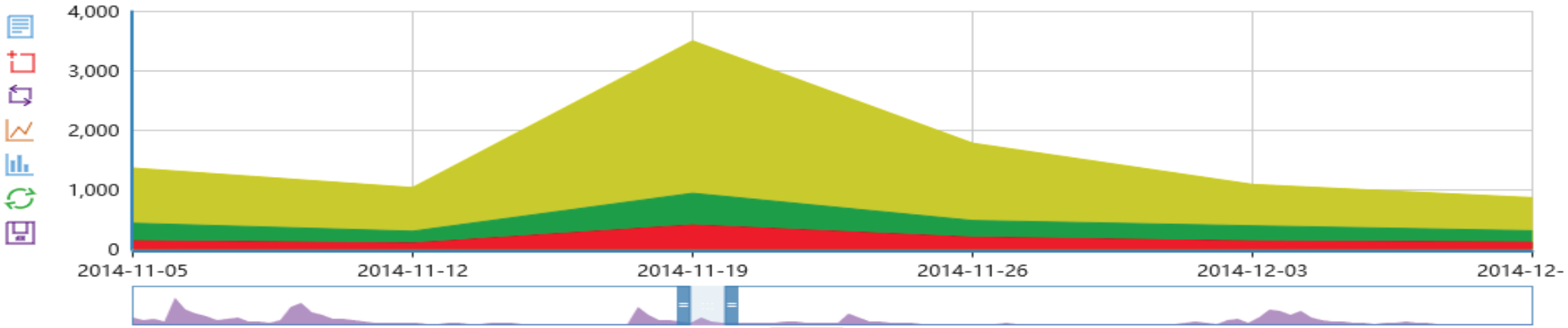
websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

雾霾

220929	36286	26113	-0.815	2568641
报道总数	正面报道	负面报道	倾向性评分	互联网搜索次数

雾霾 媒体报道趋势

全部 负面 正面 中性

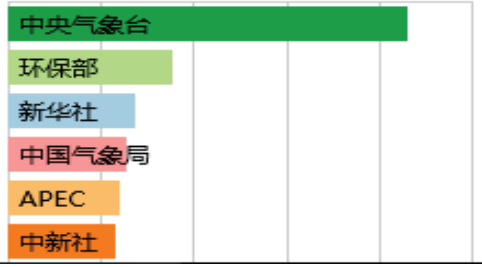


更新

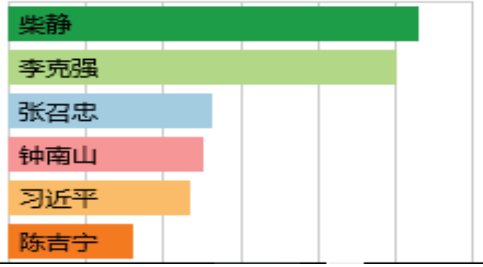
Total 220929 records for 雾霾.

子话题

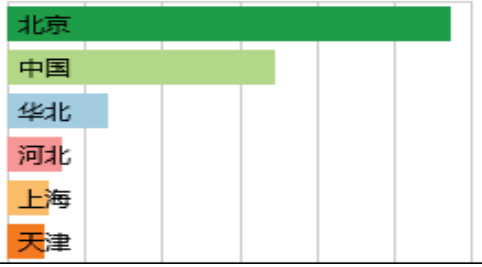
机构



人物



地点



雾霾

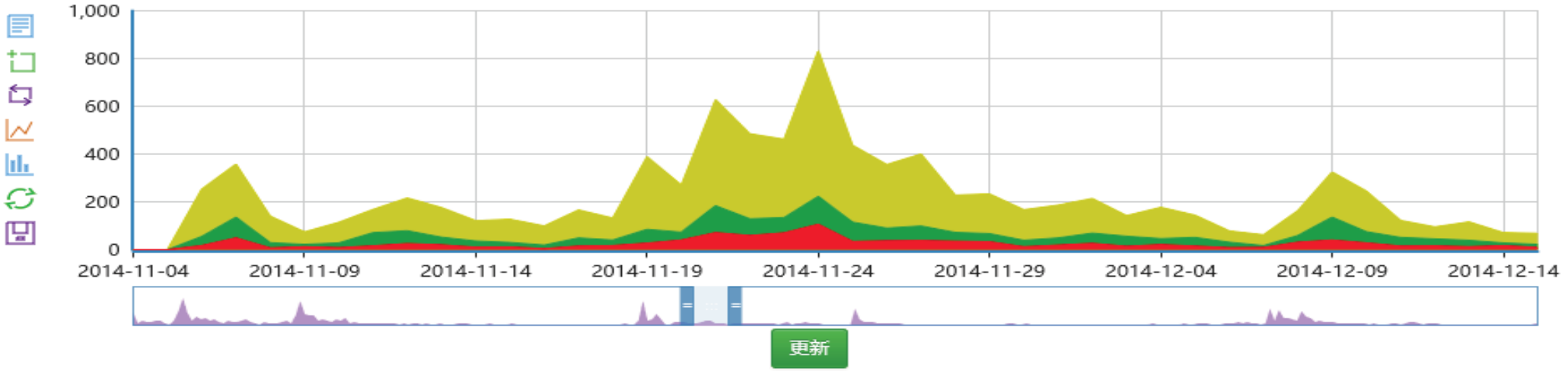
websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

雾霾

220929	1920	1157	-0.715	2568641
报道总数	正面报道	负面报道	倾向性评分	互联网搜索次数

雾霾 媒体报道趋势

全部 负面 正面 中性



更新

Total 10235 records for 雾霾.

子话题

深度分

- apec蓝
- 雾霾
- 大花猫
- 致郑州市长马懿的公开信
- 雾霾补贴
- 地铁雾霾超地上16倍

机构

- APEC
- 中央气象台
- ＡＰＥＣ蓝"
- 中新社
- APEC蓝
- 国务院

人物

- 马懿
- 习近平
- 贾跃亭
- 张小曳
- 李国发
- 马军

地点

- 北京
- 中国
- 华北
- 郑州
- 河北
- 郑州市

雾霾

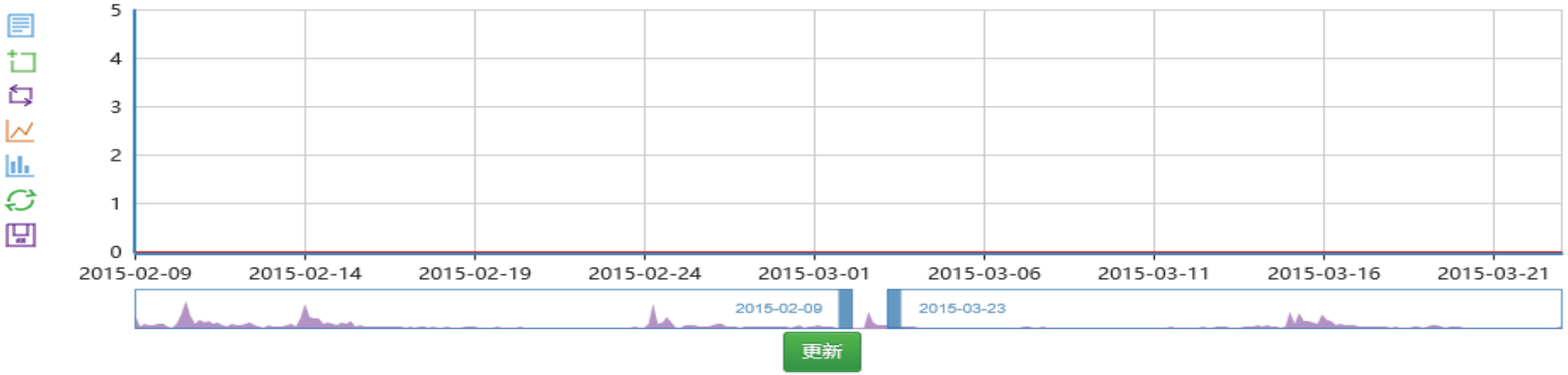
websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

雾霾

220929	1920	1157	-0.715	2568641
报道总数	正面报道	负面报道	倾向性评分	互联网搜索次数

雾霾 媒体报道趋势

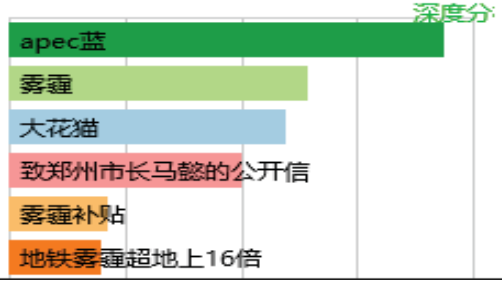
全部 负面 正面 中性



Total 10235 records for 雾霾.

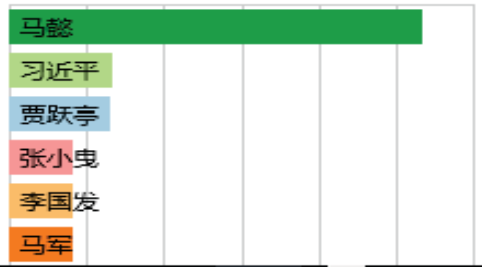
子话题

深度分

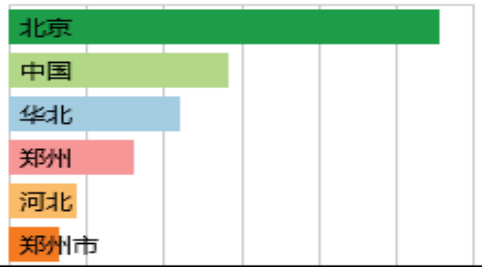


机构

人物



地点



雾霾

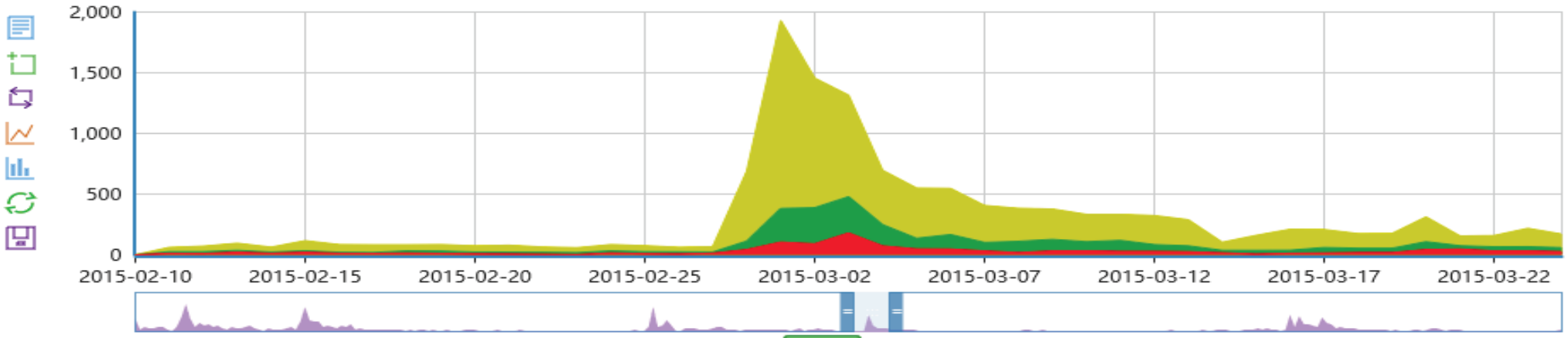
websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

雾霾

220929	2549	1252	-0.561	2568641
报道总数	正面报道	负面报道	倾向性评分	互联网搜索次数

雾霾 媒体报道趋势

全部 负面 正面 中性



更新

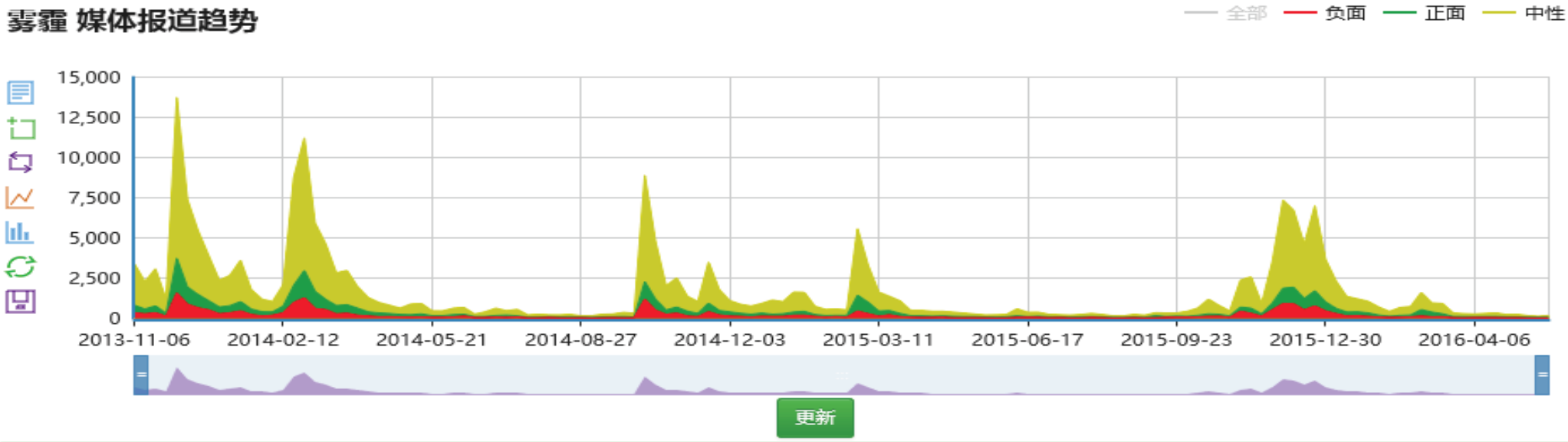
Total 13230 records for 雾霾.

子话题	机构	人物	地点
穹顶之下	全国政协	柴静	中国
雾霾	全国人大	陈吉宁	北京
553876	环保部	钟南山	华北
柴静雾霾调查：穹顶之下	政协	李克强	巴黎
两会	人大	陈凯歌	美国
造大风扇吹雾霾	中石化	韩红	英国

雾霾

websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

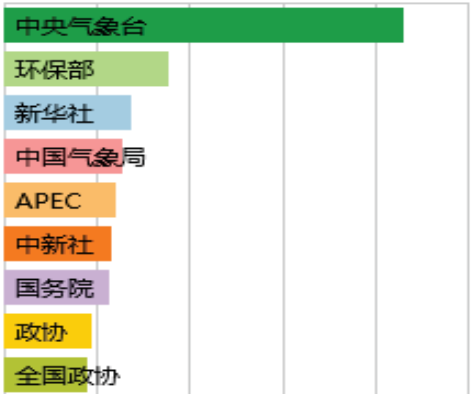
雾霾 媒体报道趋势



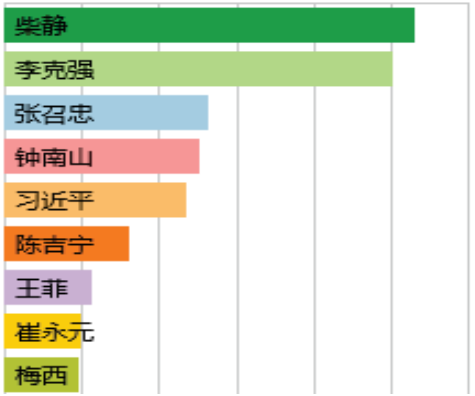
Total 220929 records for 雾霾.

子话题

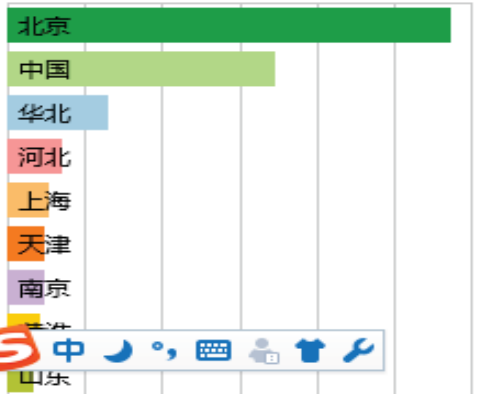
机构



人物



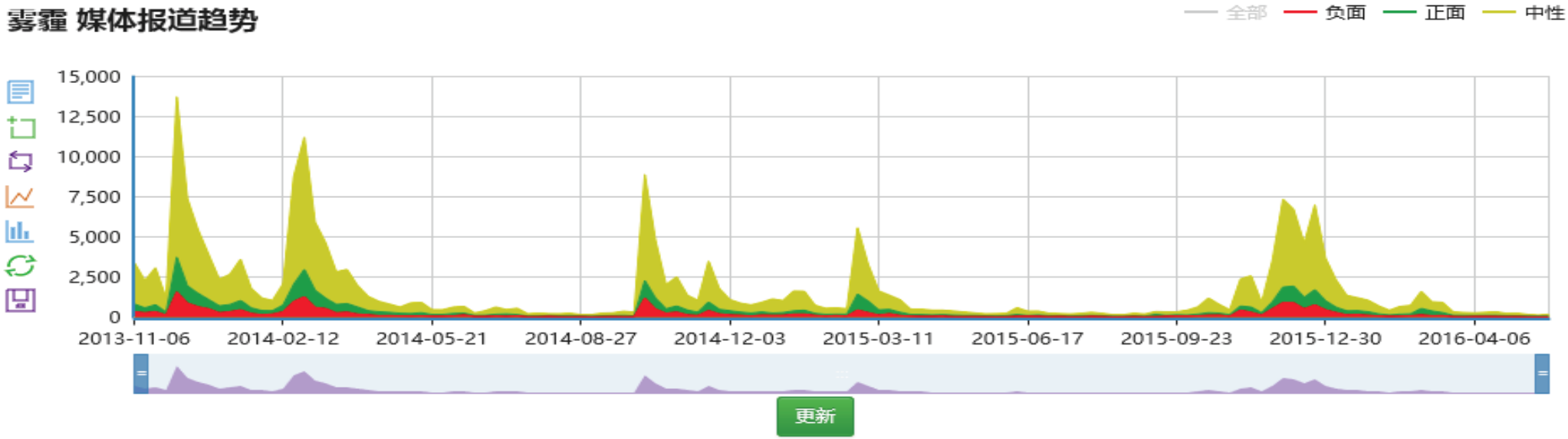
地点



雾霾

websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

雾霾 媒体报道趋势

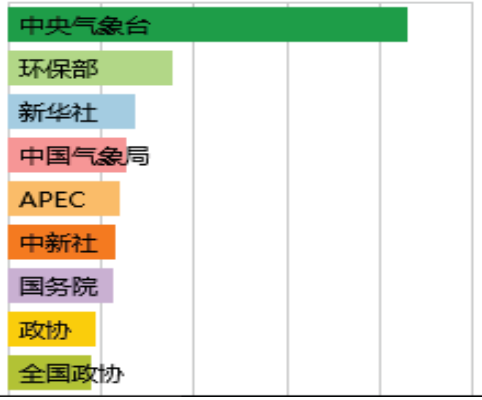


更新

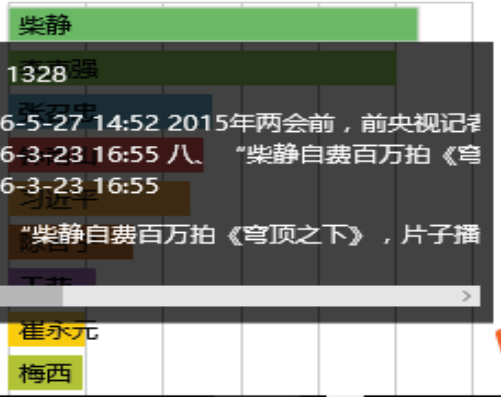
Total 220929 records for 雾霾.

子话题

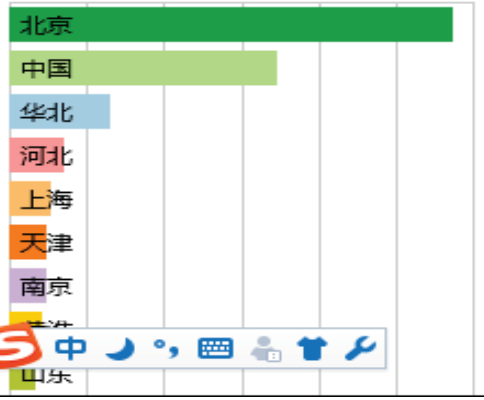
机构



人物



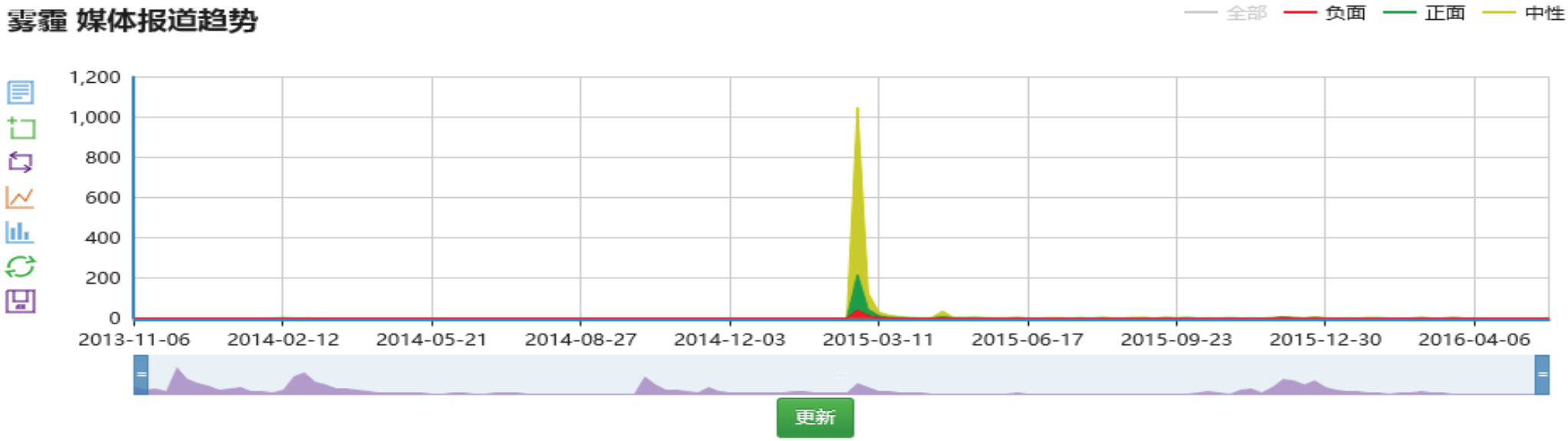
地点



雾霾

websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

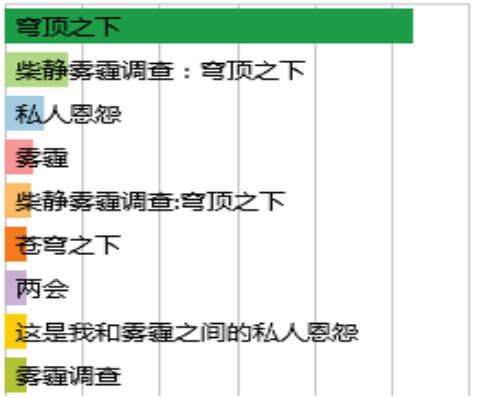
雾霾 媒体报道趋势



Total 1328 records for 雾霾.

Filters:[×柴静][×ALL]

子话题



机构

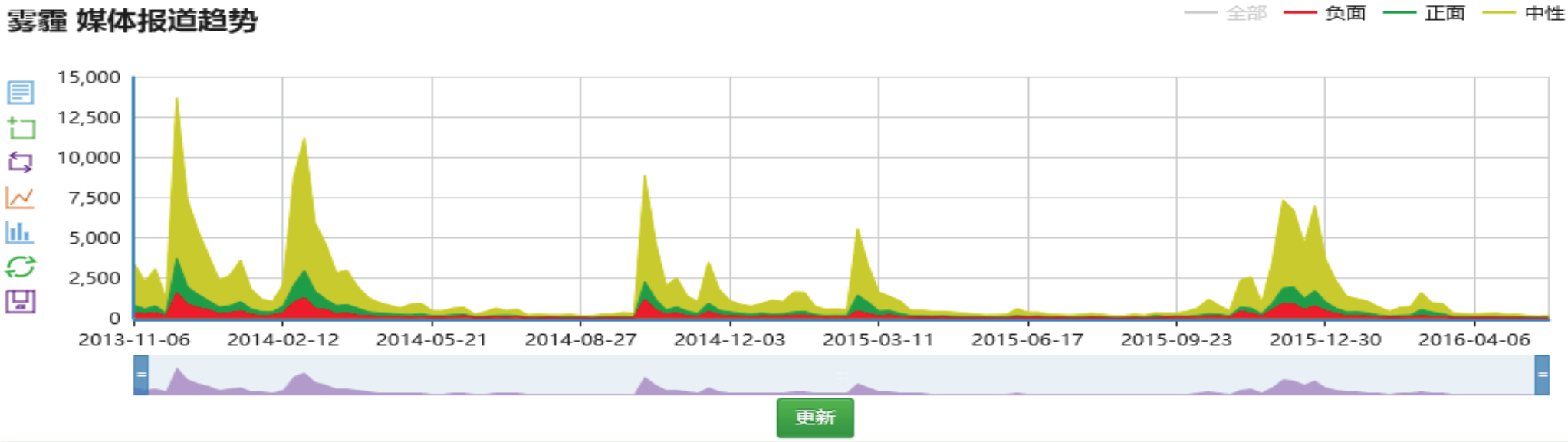
人物

地点

雾霾

websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

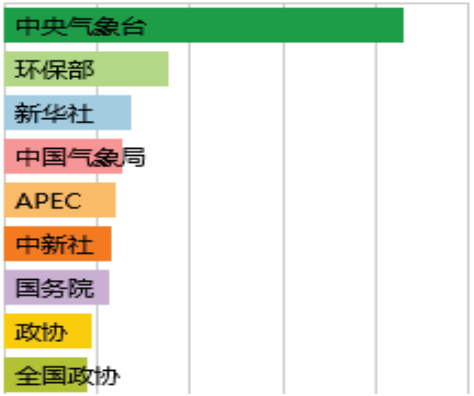
雾霾 媒体报道趋势



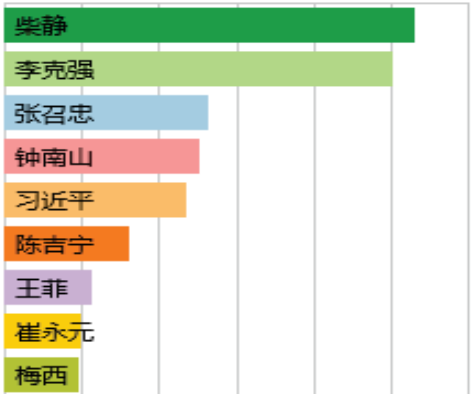
Total 220929 records for 雾霾.

子话题

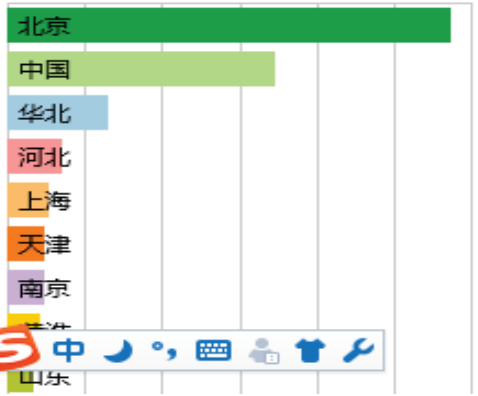
机构



人物



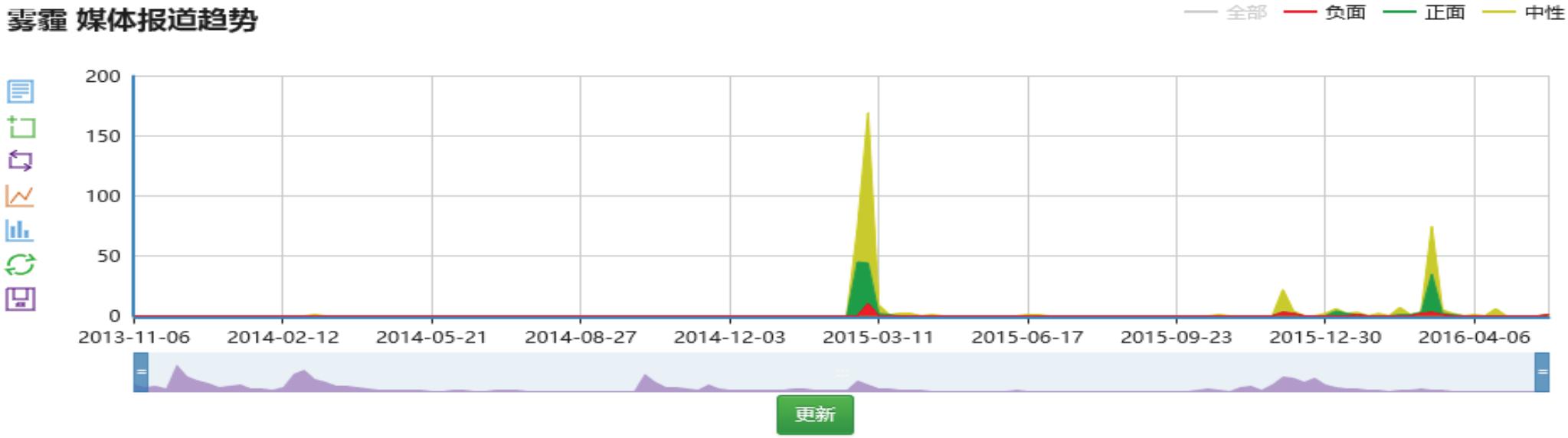
地点



雾霾

websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

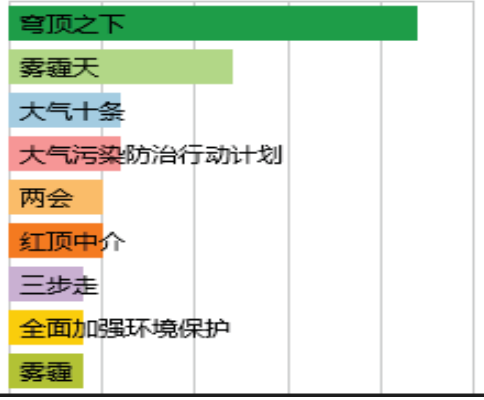
雾霾 媒体报道趋势



Total 404 records for 雾霾.

Filters:[×陈吉宁][×ALL]

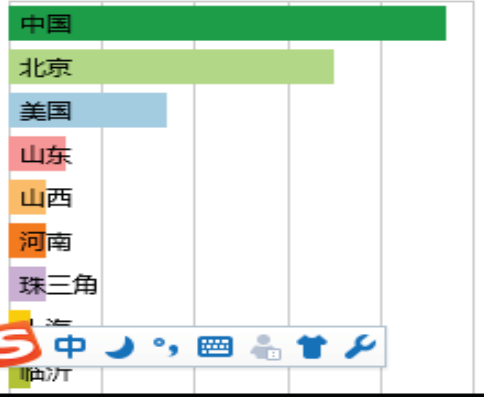
子话题



机构

人物

地点

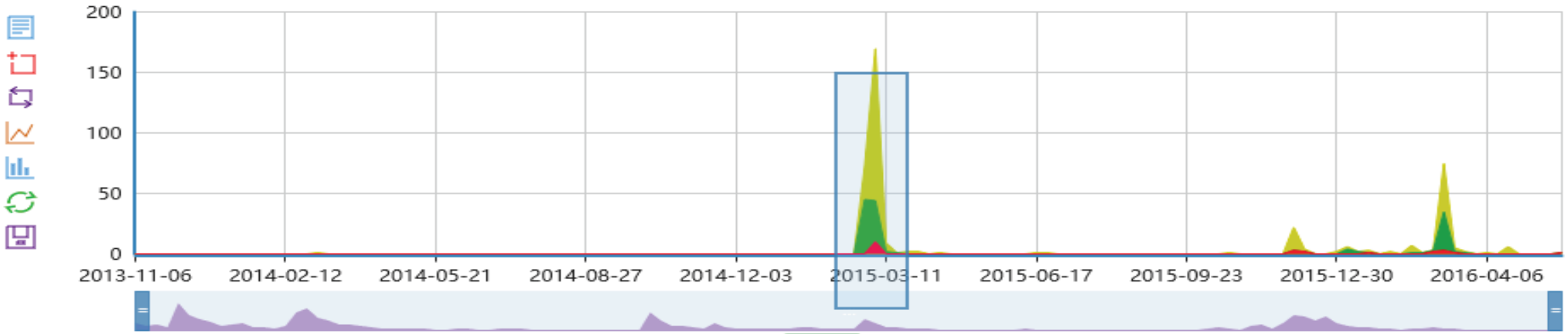


雾霾

websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

雾霾 媒体报道趋势

全部 负面 正面 中性

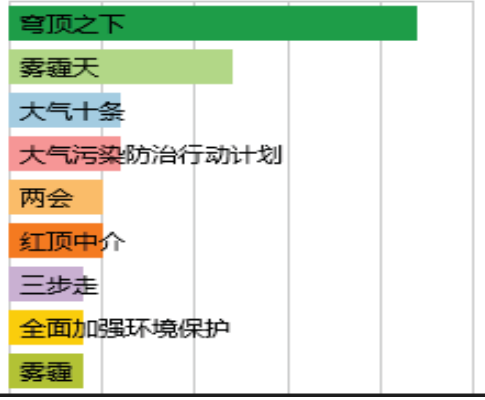


更新

Total 404 records for 雾霾.

Filters:[×陈吉宁][×ALL]

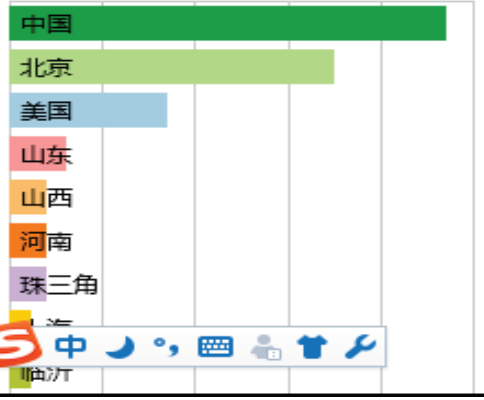
子话题



机构

人物

地点

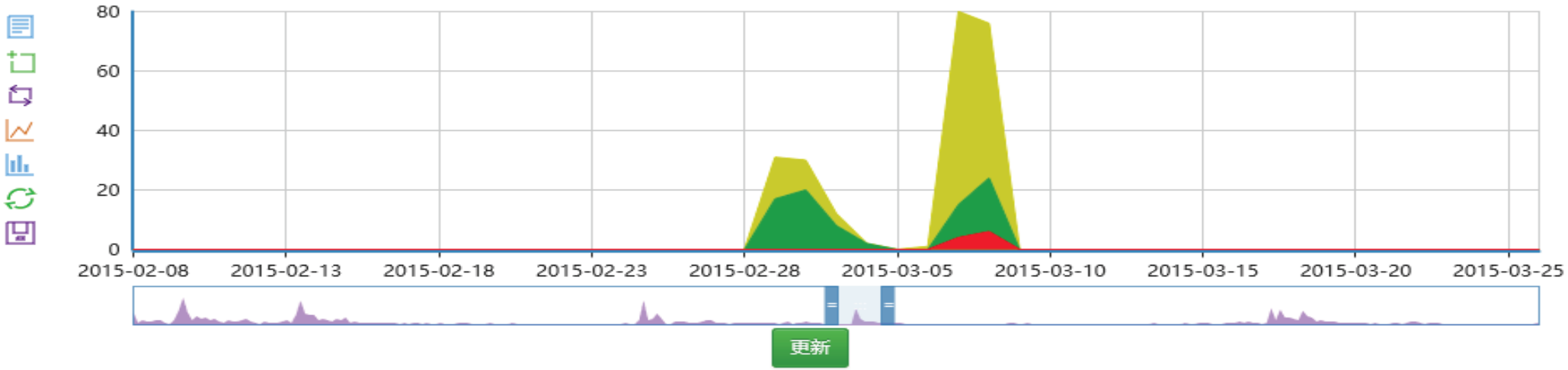


雾霾

websensor.playbigdata.com/fss3/Search2.aspx?q=%E9%9B%BE%E9%9C%BE

雾霾 媒体报道趋势

全部 负面 正面 中性



Total 232 records for 雾霾.

Filters:[×陈吉宁][×ALL]

子话题

- 穹顶之下
- 雾霾天
- 大气十条
- 全面加强环境保护
- 大气污染防治行动计划
- 雾霾
- 中国雾霾飘到了美国
- 中国雾霾飘到美国
- 对于驱散雾霾，还老百姓一片蓝天有没有

机构

- 环保部
- 新华社
- 环保 部 部
- 环保 部新 部
- 王申 摄 环保部
- 环境保护部
- 中国环保部
- 十二届全国人大
- 国务院

人物

- 陈吉宁
- 柴静
- 邹春霞
- 库兹涅茨
- 蔡敏
- 陈君
- 周凯
- 周琳
- 杜洋

地点

- 中国
- 北京
- 美国
- 山东
- 上海
- 临沂
- 北极