为价值发现提供线索
——爬虫与大数据在投研场景的应用

罗军 S0260511010004
安宁宁 S0260512020003
文巧钧 S0260517070001
陈原文 S0260517080003
邮箱：wenqiaojun@gf.com.cn
广发证券金融工程
2018年12月7日

资产管理机构开始应用大数据辅助投资决策

---RS Metrics通过高分辨率卫星影像，对零售店、办公楼等的停车场进行车流监控，预估企业运营状况； Cargo Metrics用卫星监控航运数据

---iSentium通过推特情绪指标进行择时

---2018年，JP 摩根强制要求新入职的分析师学习Python

大数据如何获取？什么叫Python，其在大数据技术中的角色？

大数据获取

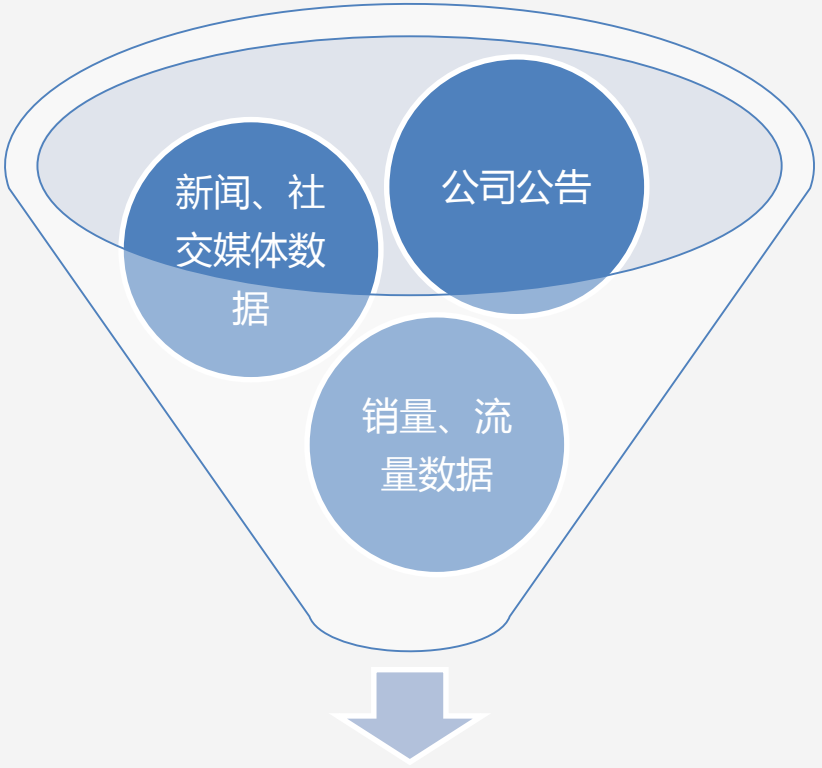
数据获取

- 公司公告、调研事件、研究报告
- 新闻、媒体舆情
- （消费类公司）销量、网站流量数据

传统投研 手工获取数据

智能投研 爬虫获取数据

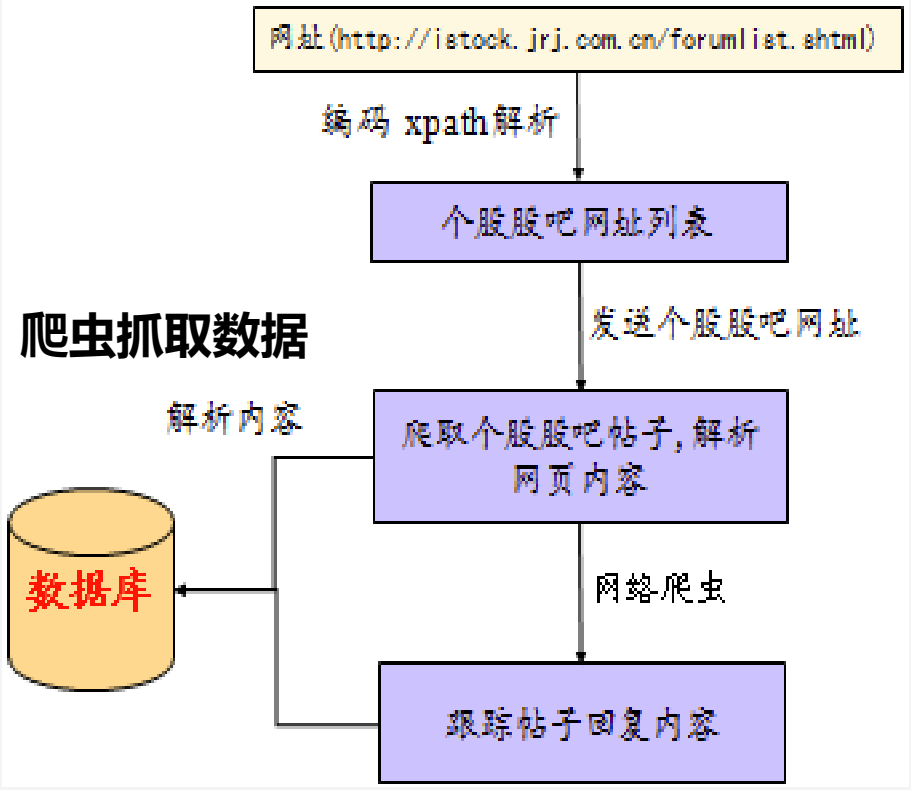
爬虫：自动获取网页内容的程序，模拟人的操作将网页信息采集下来



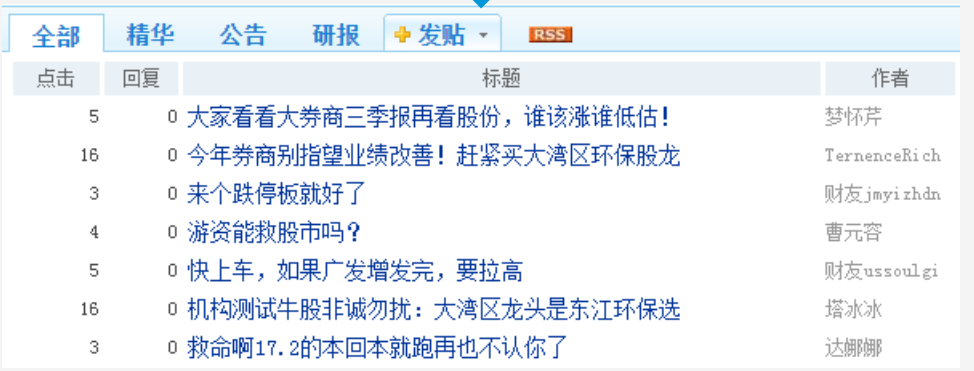
大数据获取

爬虫：自动获取网页内容的程序，模拟人的操作将网页信息采集下来

- 通过爬取股吧数据，获取市场舆情



输入“广发证券”股票代码，进入股吧

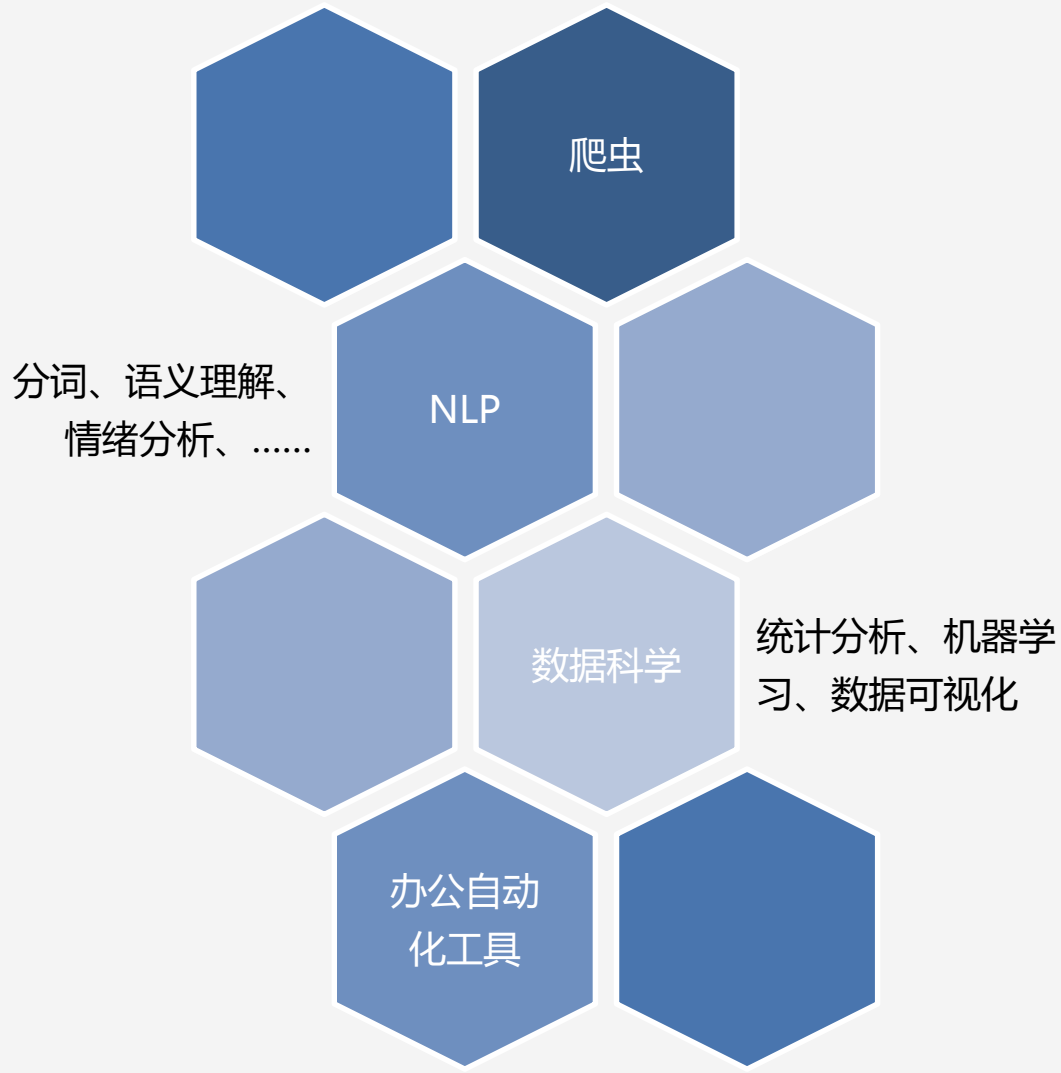


将股吧内容存储为txt

手工抓取数据

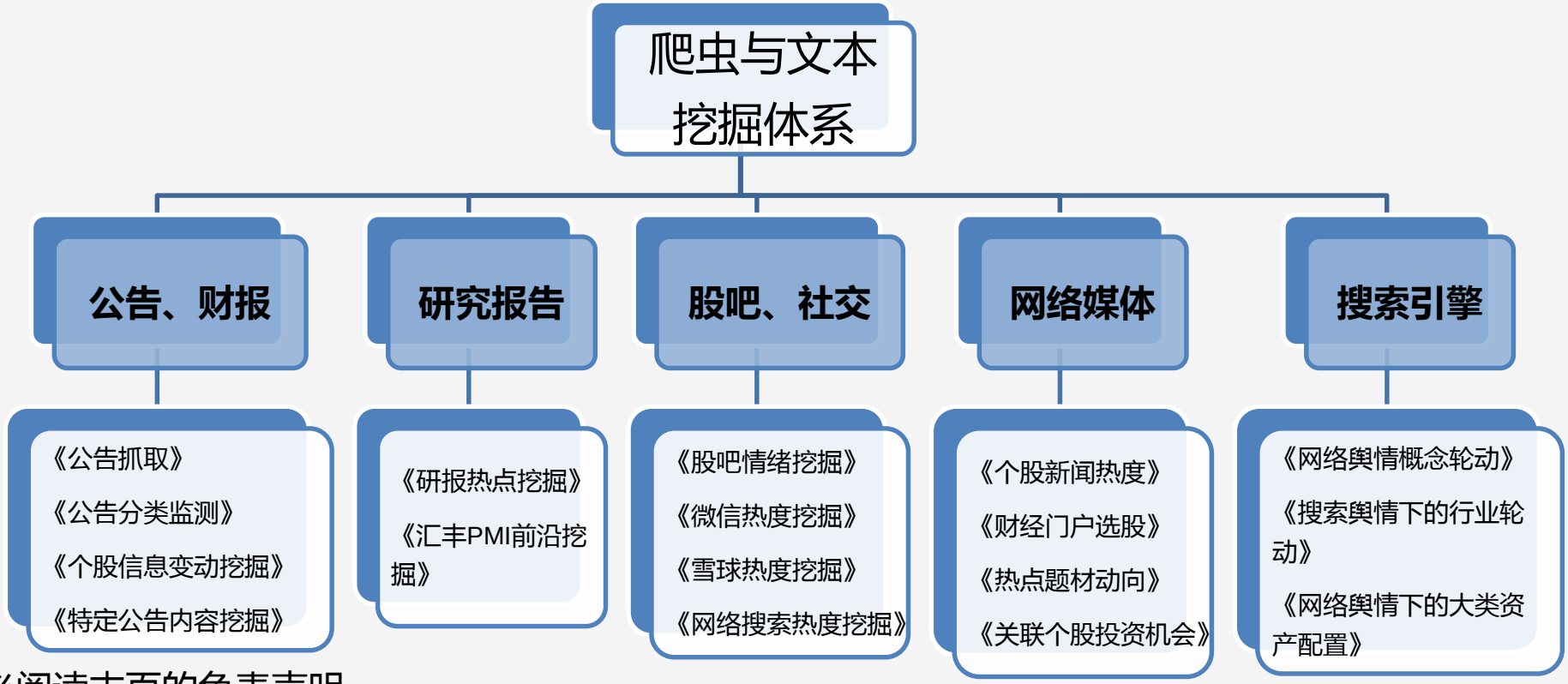
大数据获取

- Python是近年来最热门的编程语言
- Python用途广泛，包括：网络爬虫、自然语言处理（NLP）、数据科学、办公自动化、等等



广发金工大数据爬取与分析框架

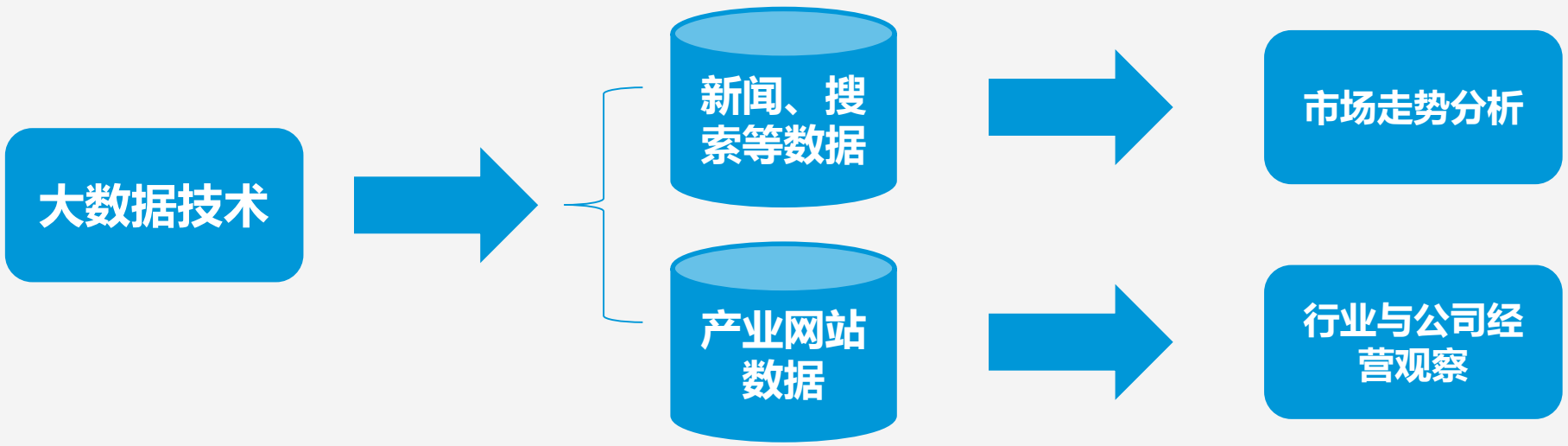
- 广发金工通过网络爬虫和文本挖掘技术，覆盖了公司公告、研究报告、社交网络、门户网站、搜索热度等方面的另类数据。



应用框架

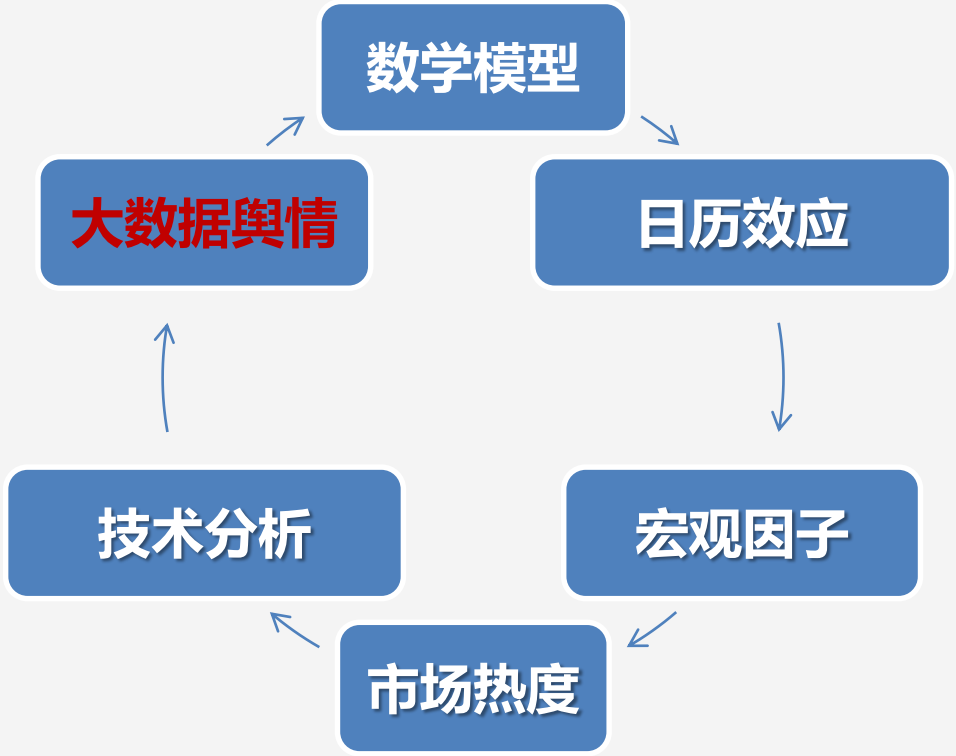
大数据应用

通过大数据技术，获取新闻、搜索引擎、产业等网站数据，有助于对市场走势、产业与公司经营趋势进行分析



应用场景1：大数据择时

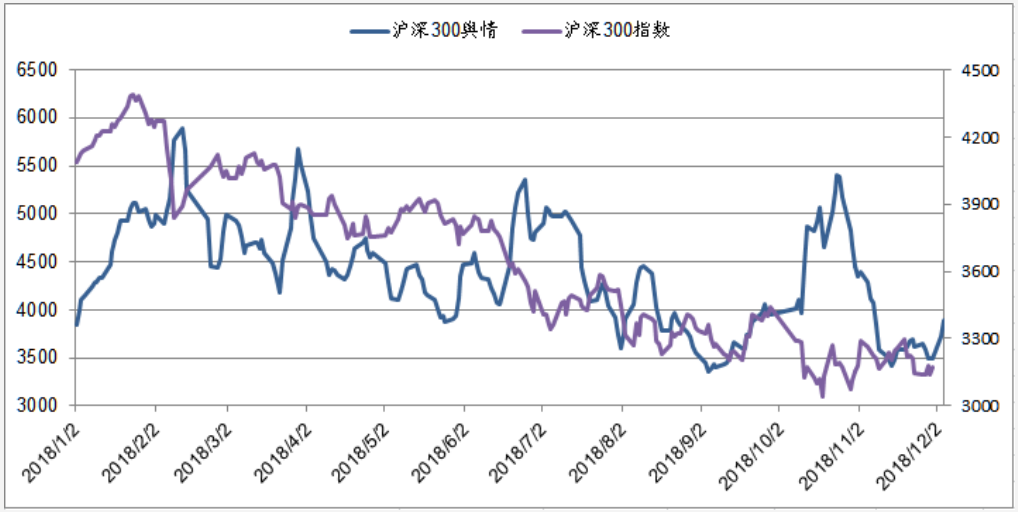
量化择时体系：6个维度



应用场景1：大数据择时

舆情数据和市场走势的关联

通过监控各大搜索引擎（百度、360），新闻网站（东财、雪球）上等互联网上关于指数、个股等的新闻、关注度等舆情数据，从网络舆情角度看市场的情绪高涨程度。



以百度指数为例：百度指数与沪深300指数走势十分相近，二者之间相关系数达到0.69

新闻量与指数涨跌呈正相关关系

- 指数成份股新闻量剧增时，指数更容易上涨
- 指数成份股新闻量剧减时，指数倾向于下跌

数据来源：Wind，百度指数，广发证券发展研究中心

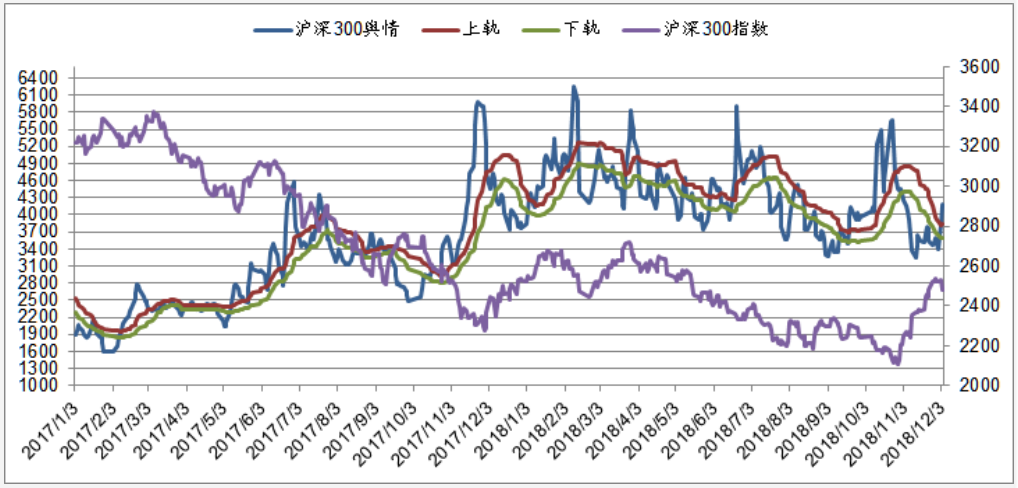
应用场景1：大数据择时

舆情数据和市场走势的关联

最新信号看涨！

新闻量与指数涨跌呈正相关关系

- 指数成份股新闻量剧增时，指数更容易上涨，突破上轨，发出做多信号
- 指数成份股新闻量剧减时，指数倾向于下跌，突破下轨，发出做空信号



分年度指标统计	累计收益率	最大回撤	胜率	年度盈亏
2011	71.82%	-6.41%	61.85%	盈
2012	47.78%	-10.31%	56.46%	盈
2013	18.11%	-10.73%	55.38%	盈
2014	55.79%	-4.72%	58.76%	盈
2015	115.53%	-18.11%	55.50%	盈
2016	15.65%	-12.47%	51.26%	盈
2017	0.45%	-8.63%	50.25%	盈
2018年至今	2.65%	-13.30%	52.54%	盈
整体	1100.77%	-18.11%	55.18%	盈

数据来源：Wind，百度指数，广发证券发展研究中心

应用场景2：中观观察之电商

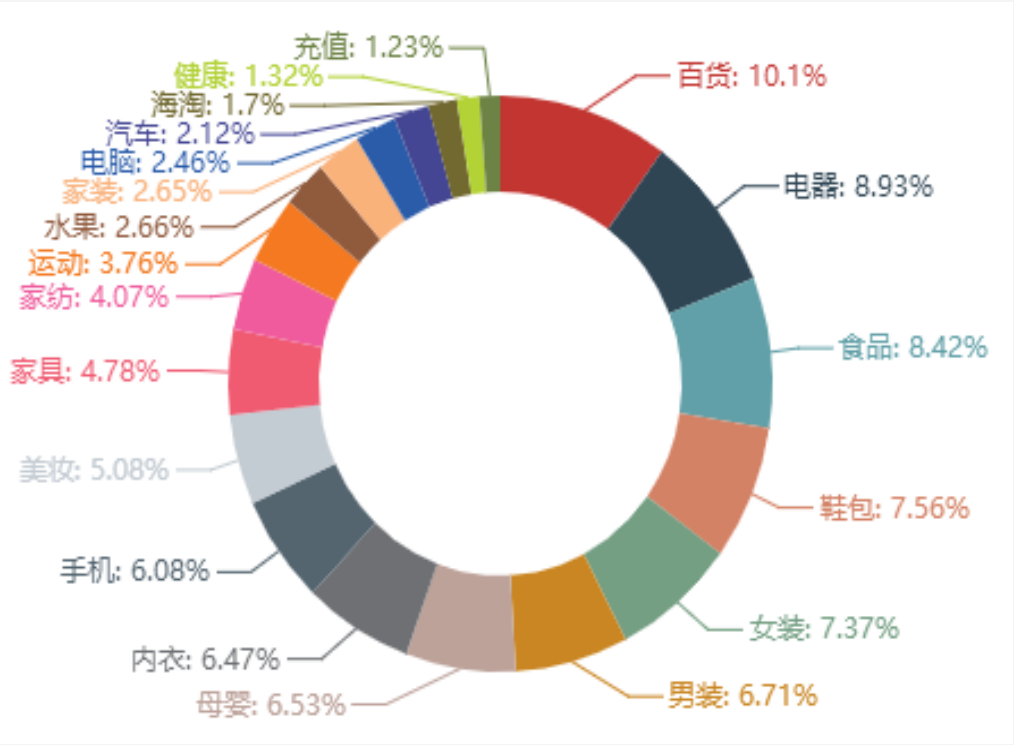
拼多多数据

采集app上30万个商品数据，统计不同类别商品的销售额。对销售额数据的持续跟踪，有助于我们把握不同行业的消费信息。

商品数据

销售额

食品类商品销售额



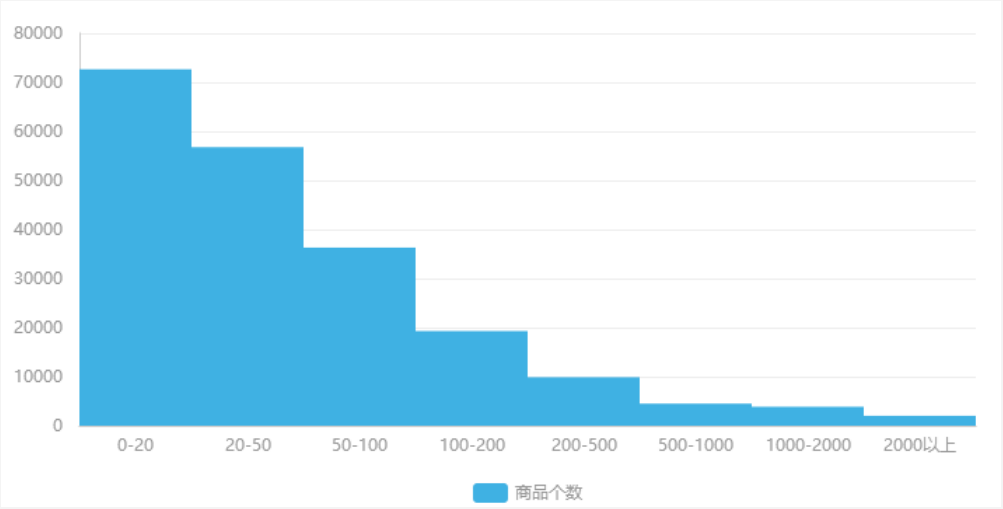
拼多多销售额统计

数据来源：拼多多，广发证券发展研究中心

应用场景2：中观观察之电商

从拼多多数据观察三四线城市消费习惯

- 在拼多多平台上销售的商品价格主要集中在0-20元。
- 销量最高的产品主要集中在百货商品栏目上，其中以纸巾的销量最高。



拼多多销量前十产品一览

商品名称	价格（元）	栏目
【800 万家庭的选择】香约原木抽纸餐巾纸卫生纸纸巾 3 层面巾纸批发	12.9	百货
【40 卷 24 卷 10 卷可选】富豪 5.5 斤 40 卷天然竹浆本色卫生纸卷纸	8.9	百货
【27 包 18 包 8 包】丝飘本色竹浆抽纸 300 张/包餐巾纸面纸	8.9	百货
【500 万人见证好品质】遇水开花晴雨伞三折防晒紫外线遮太阳男女	10.3	百货
【直冲 400W 销量】思宏红满疆枣夹核桃 500g 新疆特产枣想和你在一起	16.9	食品
【卖爆 7 千 8 百万包】330 张植护原木抽纸 24 包/6 包 110 抽卫生纸批发箱	9.9	百货
【200 万销量见证】点断式加厚彩色垃圾袋家用平口垃圾袋 50*45cm	5.1	百货
【30 包 300 张/包】心逸原木抽纸纸巾餐巾纸面巾纸纸抽卫生纸批发	29.9	百货
【35 卷 24 卷 12 卷可选】高品质 5.5 斤 35 卷 4 层加厚 臻木本色卫生纸	8.4	百货
【加赠手帕纸不加价】亲爽 300 张/包原木抽纸面巾纸纸巾餐巾纸批发	9.9	百货

数据来源：拼多多，广发证券发展研究中心

应用场景2：中观观察之电商

从拼多多数据观察三四线城市消费习惯

通过分词发现：

-----消费群体为妇女、学生等群体，以及二三十岁的已婚群体

-----促销方式：“折扣相关信息里价格最低、成本最低、性价比最高等关键词频率高

数据来源：拼多多，广发证券发展研究中心

应用场景2：中观观察之电商

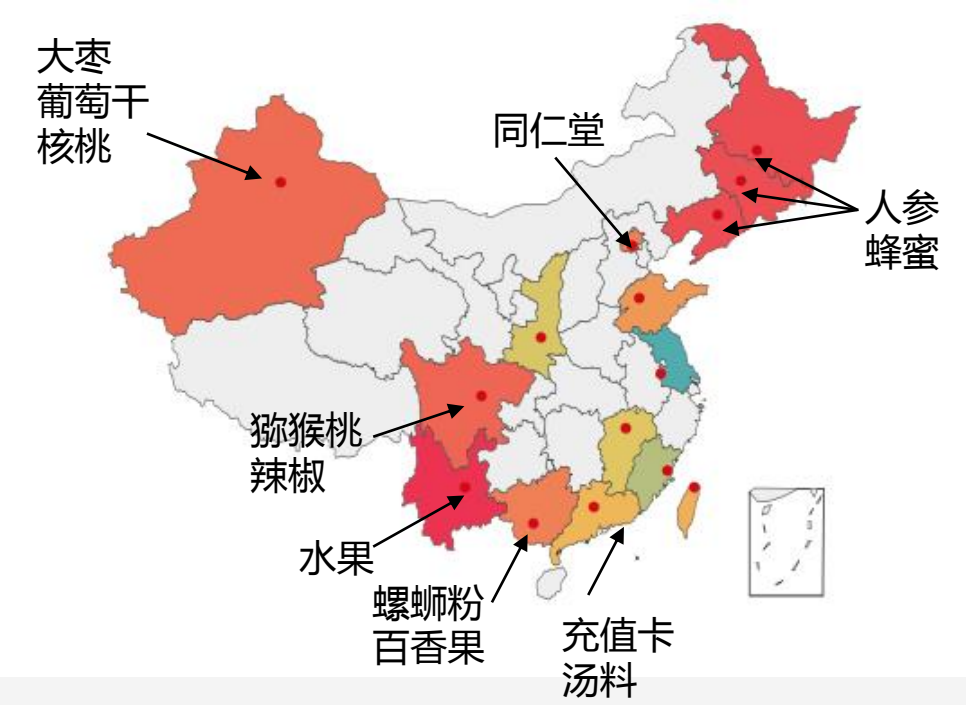
从拼多多数据观察三四线城市消费习惯

通过对商品名称分词发现：

-----商家主要来自东北、西北、西南等省份

-----这些省份的热销商品与我们传统印象中这些省份的生活方式及盛产物有很明显的对应关系

拼多多商品主要产地一览



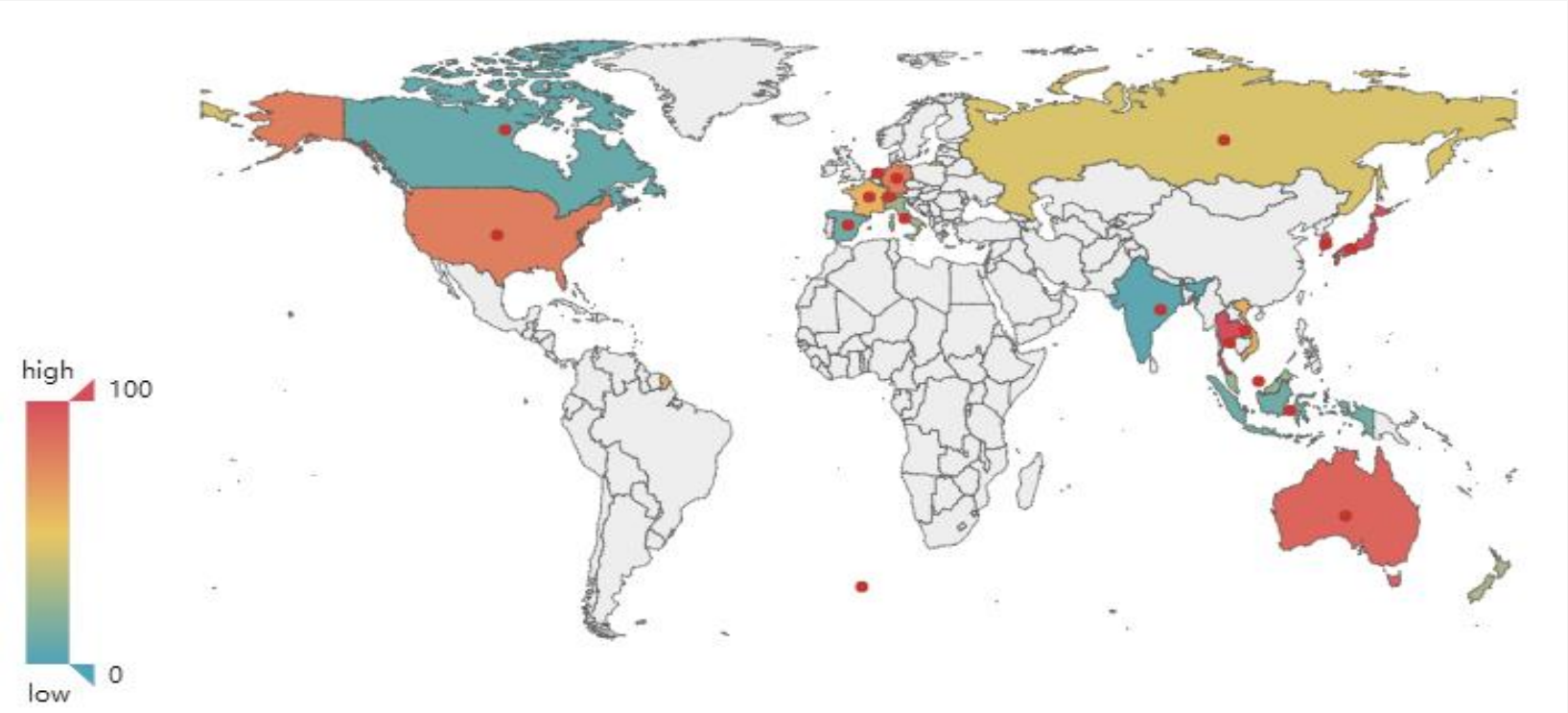
数据来源：拼多多，广发证券发展研究中心

应用场景2：中观观察之电商

拼多多进口相关数据：

--- 泰国、韩国、日本、澳洲、美国、德国、越南是主要的商品来源地

--- 日韩美妆、泰国的榴莲芒果、澳洲保健品、美国的男装

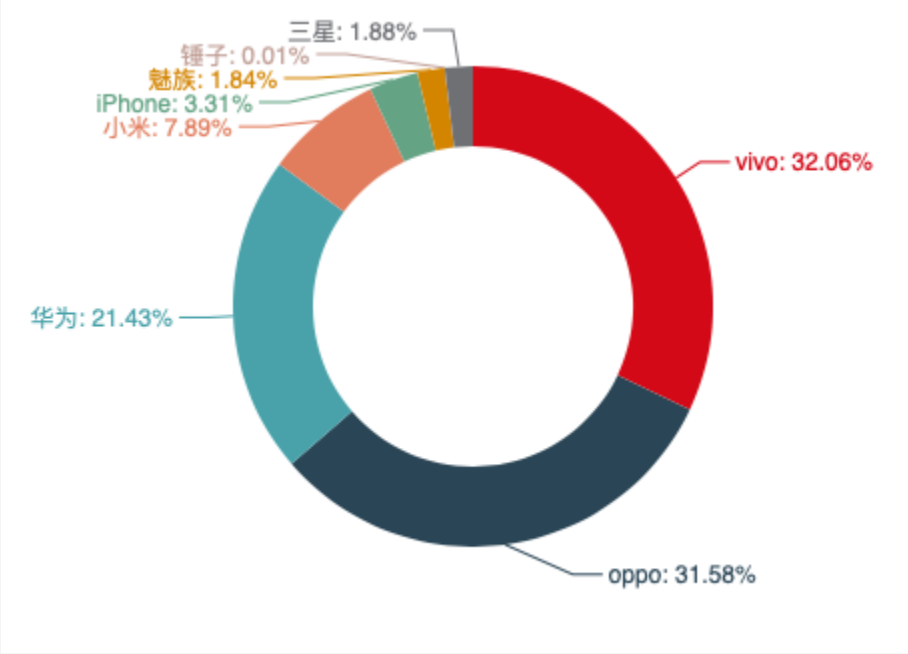


数据来源：拼多多，广发证券发展研究中心

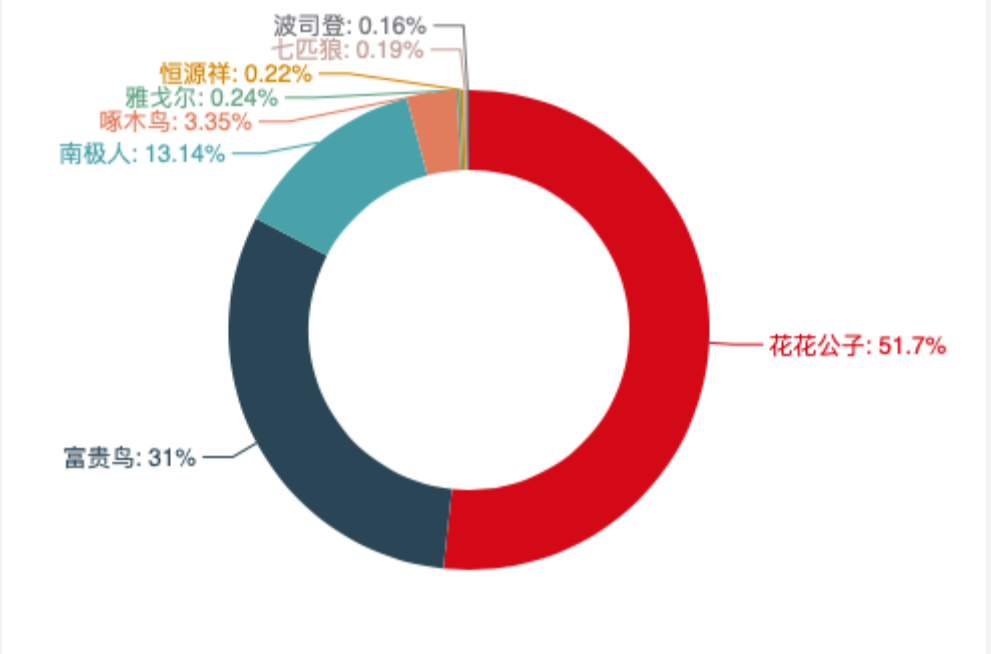
应用场景2：中观观察之电商

从拼多多数据观察不同品牌影响力

手机销量：vivo、oppo、华为
男装销量：花花公子、富贵鸟、南极人



拼多多手机品牌销量统计



拼多多男装品牌市场份额统计

数据来源：拼多多，广发证券发展研究中心

应用场景2：中观观察之房价

差异化的分析房价

地产行业研究员的房地产数据一般来自Wind，从统计局公布信息获取，主要问题：

- 数据滞后：次月月中公布
- 数据单一：只有笼统的城市房价

通过地产中介网站获取房价数据的主要优势：

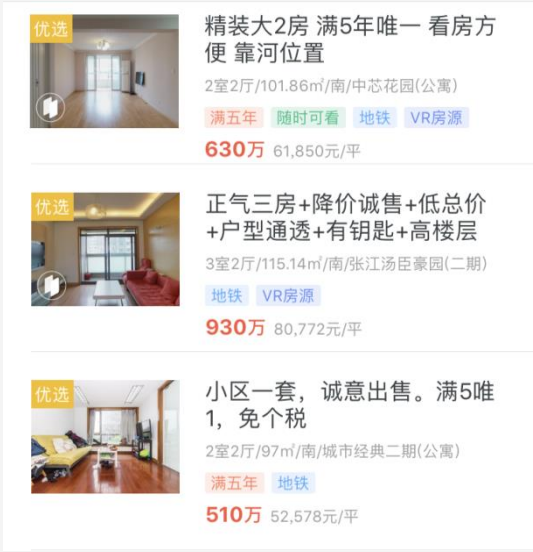
- 每天都可以获得最新房价
- 可以考察精细化数据，如学区房、城区与郊区房价，特定小区房价

通过约70万条链家北京区域房价数据进行不同问题的定制化研究：

- 北京不同区域房价走势怎么样？
- 怎么看北京名校学区房房价走势？

应用场景2：中观观察之房价

链家数据信息



二手房挂牌数据

解析价格数据

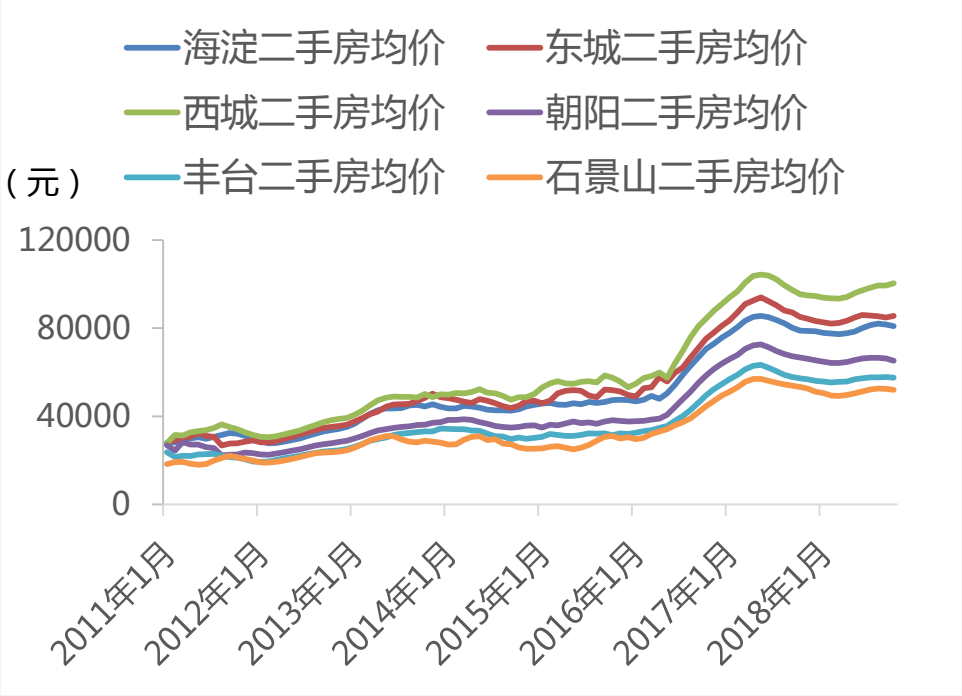
不同区域均价

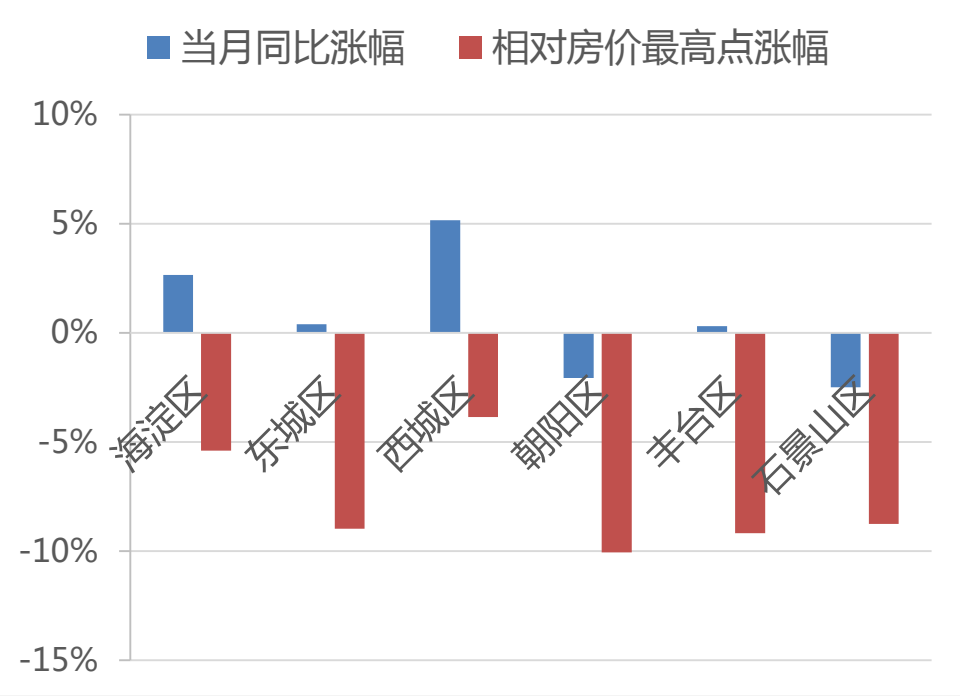
数据来源：链家网，广发证券发展研究中心

应用场景2：中观观察之房价

链家数据研究：北京不同区域二手房价走势怎么样？

西城区二手房成交价格是六个区域中最高的
过去一年中，西城区房价最强势，石景山房价表现最弱



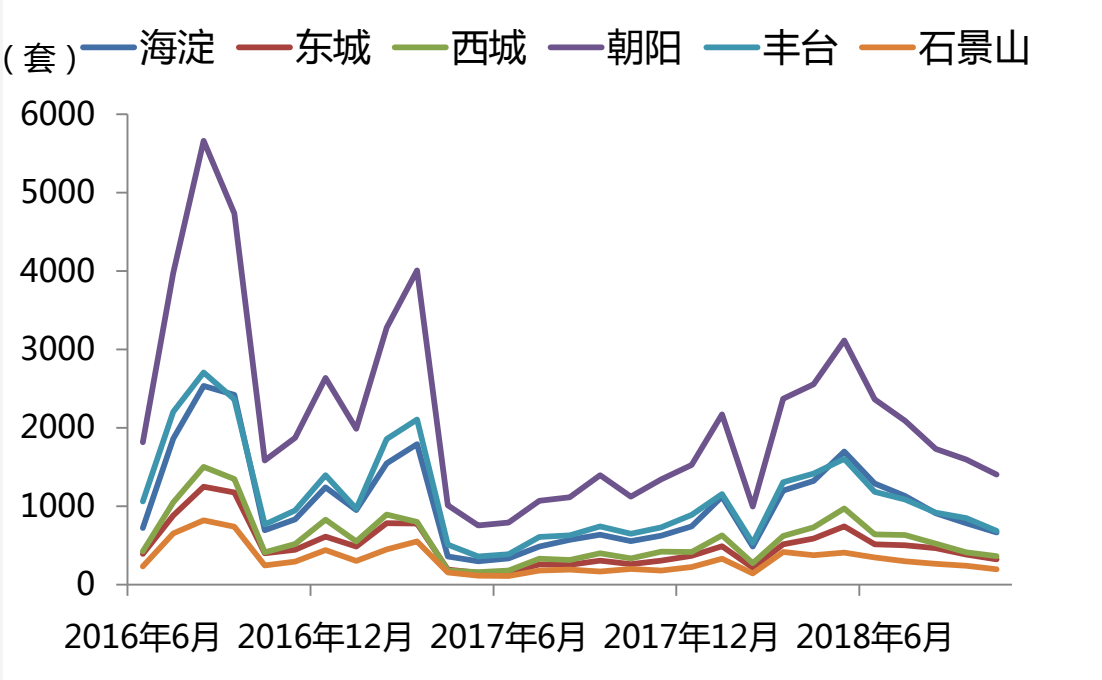


数据来源：链家网，广发证券发展研究中心

应用场景2：中观观察之房价

链家数据研究：北京不同区域二手房价走势怎么样？

近半年来，链家成交量持续下滑
当前成交量与2016年相比，下滑明显



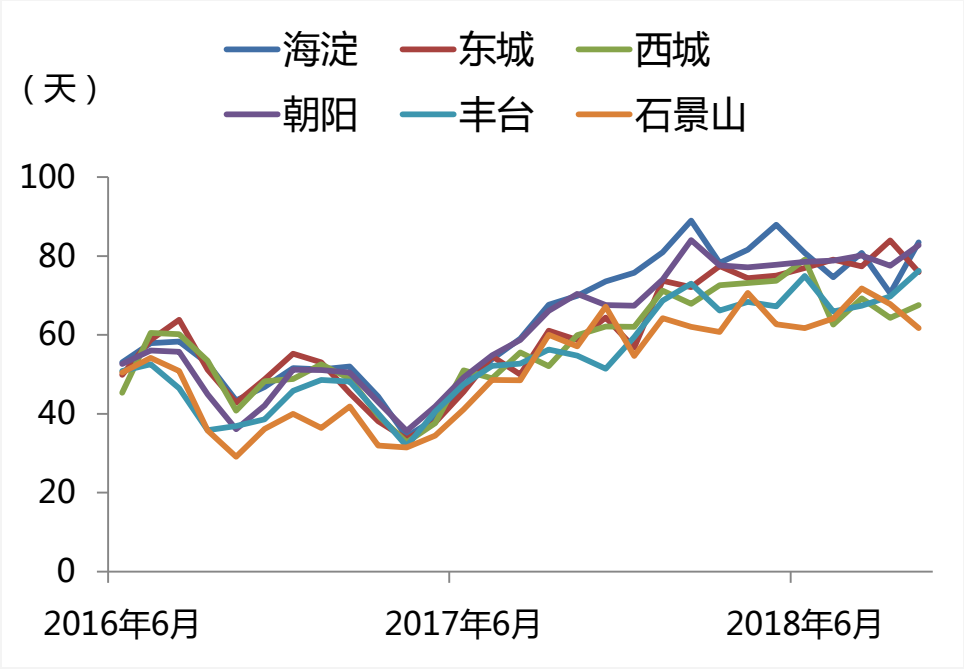
北京不同区域二手房成交量（套）

数据来源：链家网，广发证券发展研究中心

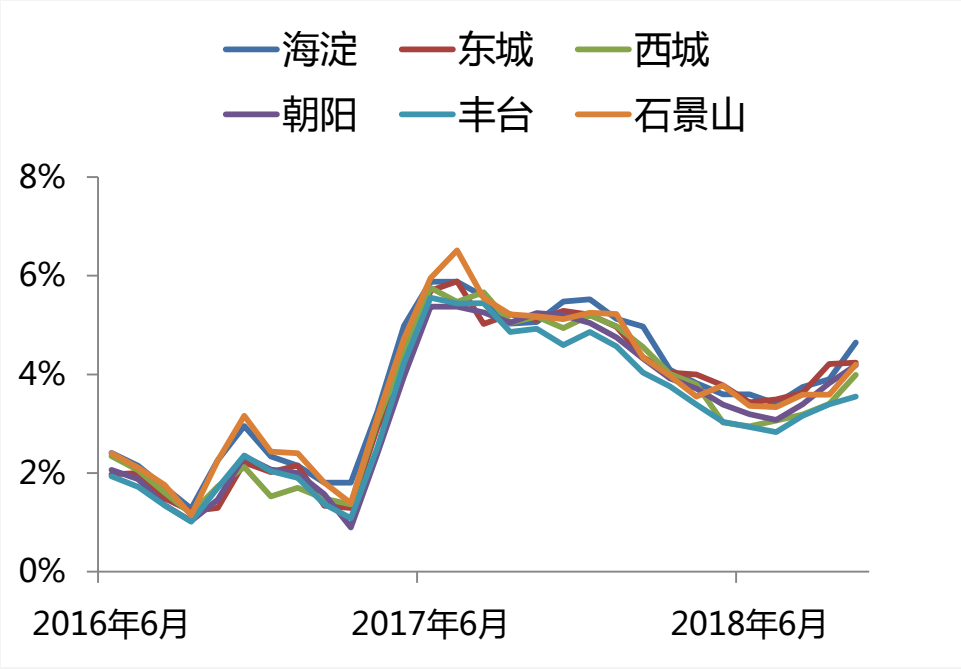
应用场景2：中观观察之房价

链家数据研究：北京不同区域二手房价走势怎么样？

成交周期增大：说明房子不容易卖出
交易价格相对挂牌价格降幅增大：说明目前房地产交易是买方市场，买房者处于主动地位



二手房平均成交周期



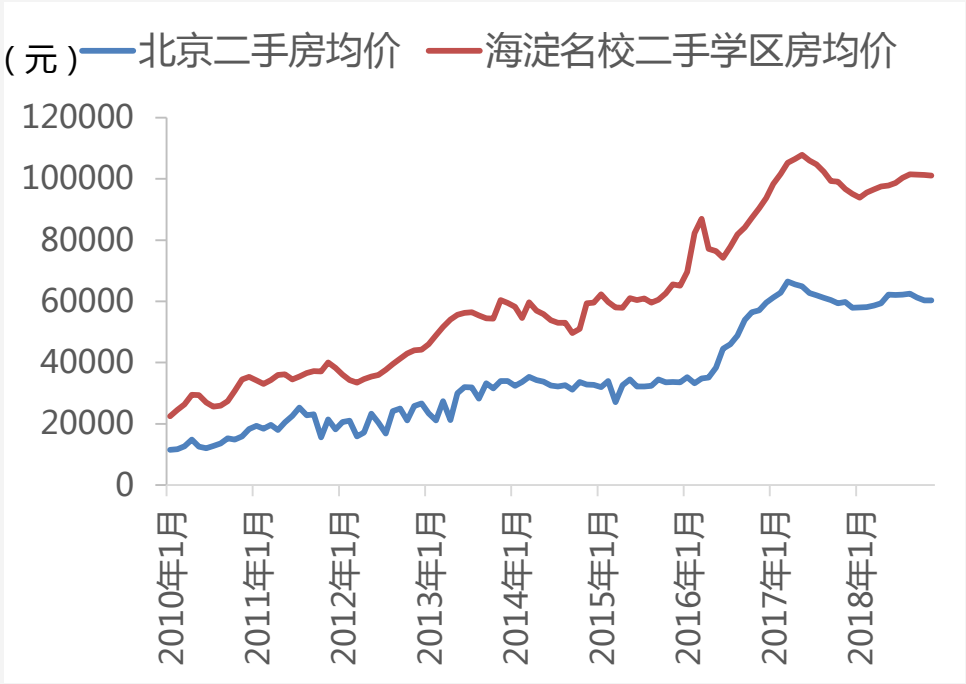
二手房交易价格相对挂牌价格平均降幅

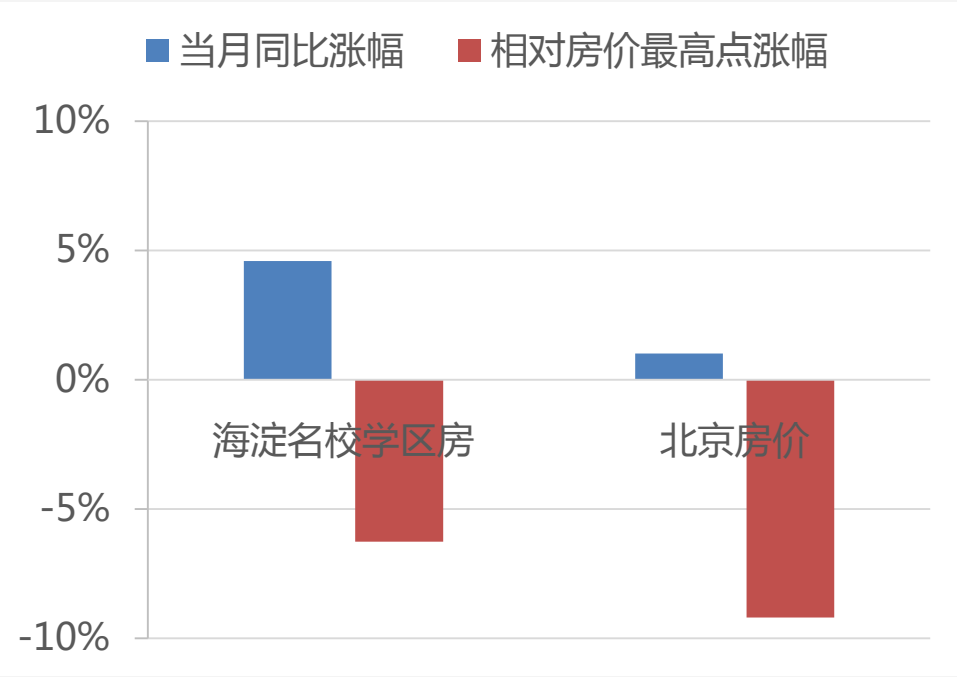
数据来源：链家网，广发证券发展研究中心

应用场景2：中观观察之房价

链家数据研究：怎么看北京名校学区房房价走势？

选取海淀区排名前十的小学的学区房，观察价格走势
海淀区二手房成交均价比北京二手房成交均价有4万元/平的溢价
名校学区房价格相对最高点有6%的跌幅，但相对比较保值



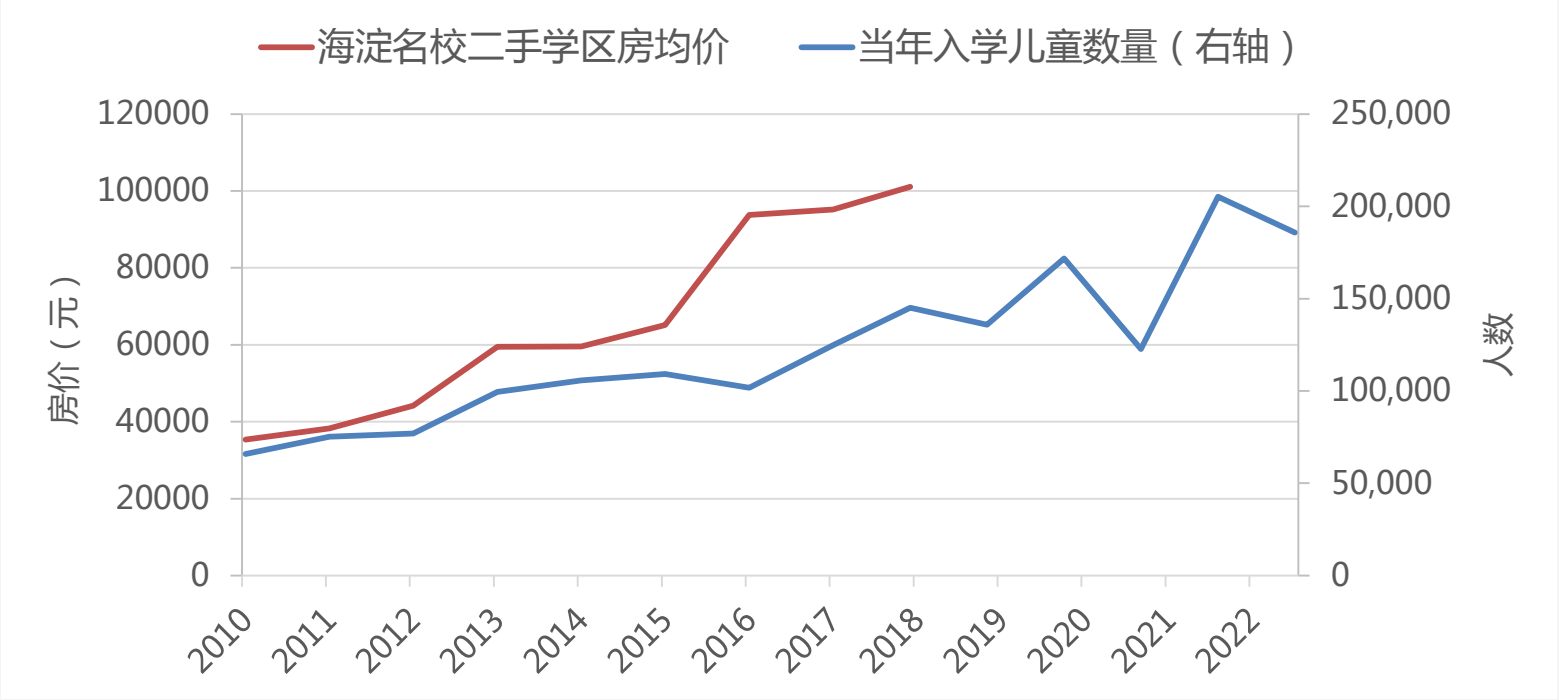


数据来源：链家网，广发证券发展研究中心

应用场景2：中观观察之房价

链家数据研究：怎么看北京名校学区房房价走势？

未来学区房的走势判断：通过学龄儿童数量来进行分析

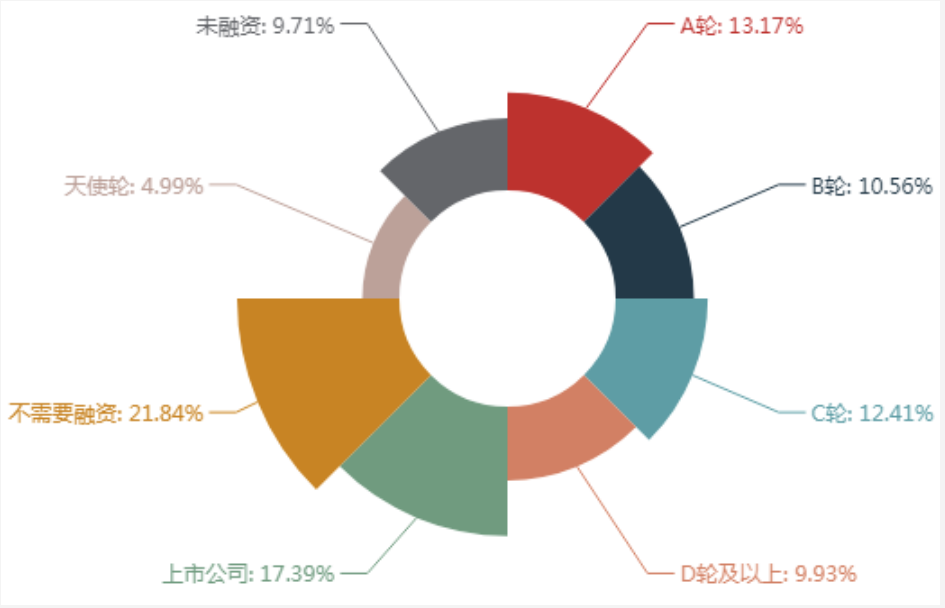


数据来源：链家网，广发证券发展研究中心

应用场景2：中观观察之招聘

拉勾网数据

- 获取AI相关岗位的2万条招聘信息，观察AI在企业的应用情况。
- 初创型企业（从天使轮到ABCD轮的公司）对AI相关岗位的需求较大，占到了全部招聘岗位的51%，而上市公司招聘的岗位占比为17.4%。
- **初创型企业对AI人才需求量大，这些企业可能是新的AI应用领域，或许未来诞生独角兽公司**



数据来源：拉勾网，广发证券发展研究中心

应用场景2：中观观察之招聘

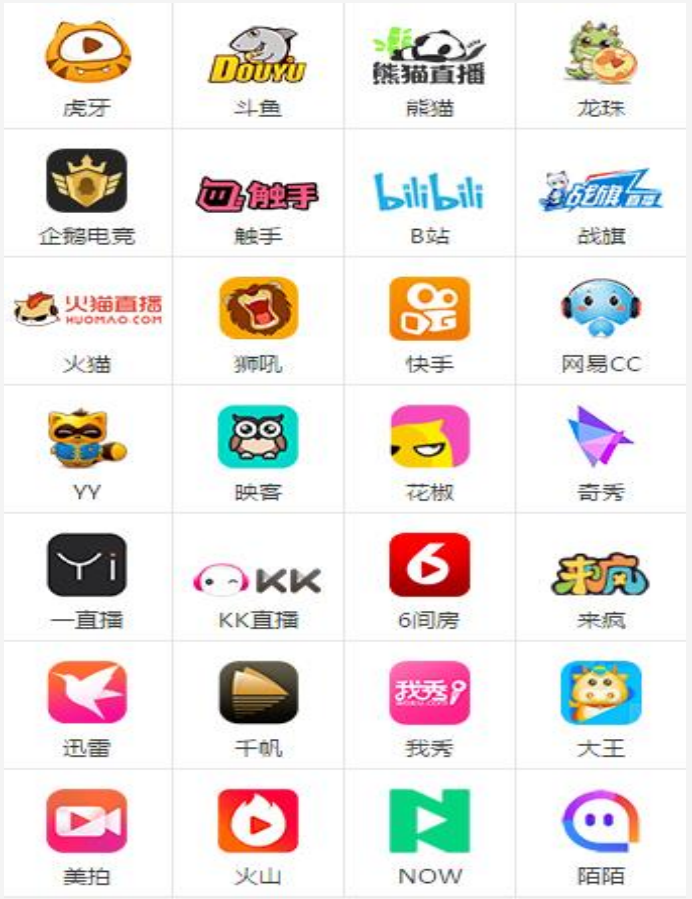
上市公司情况

- 通过对上市公司招聘AI岗位的统计，可以看到不同公司的AI招聘人数。
- 研发支出对上市公司保持和提高市场竞争力至关重要，AI人才招聘数量能够从一方面反映公司的研发水平。
- 京东、搜狗、腾讯、阿里巴巴等公司是从拉勾网招聘AI人才最多的上市公司。
- 大部分公司境外上市；中国平安、泛微网络是招聘AI人才最多的A股上市公司。

公司代码	公司名	招聘数量（个）
JD.O	京东	198
SOGO.N	搜狗	153
0700.HK	腾讯控股	134
BABA.N	阿里巴巴	108
TAL.N	好未来	101
NTES.O	网易	101
SINA.O	新浪	73
PDD.O	拼多多	71
MOMO.O	陌陌	70
0992.HK	联想集团	68
IQ.O	爱奇艺	67
601318.SH	中国平安	52
603039.SH	泛微网络	51
3888.HK	金山软件	39
YY.O	欢聚时代	34

数据来源：拉勾网，广发证券发展研究中心

应用场景3：微观观察



平台流量

公司业绩

娱乐平台	对应上市公司	股票代码
虎牙	虎牙直播	HUYA.N
B站	哔哩哔哩	BILI.O
YY	欢聚时代	YY.O
映客	映客	3700.HK
美拍	美图公司	1357.HK
陌陌	陌陌	MOMO.O
奇秀	爱奇艺	IQ.O
企鹅电竞	腾讯控股	0700.HK
NOW	腾讯控股	0700.HK
网易CC	网易	NTES.O
一直播	新浪	SINA.O
6间房	宋城演艺	300144.SZ
来疯	阿里巴巴	BABA.N
迅雷	迅雷	XNET.O

数据来源：小葫芦，广发证券发展研究中心

应用场景3：微观观察

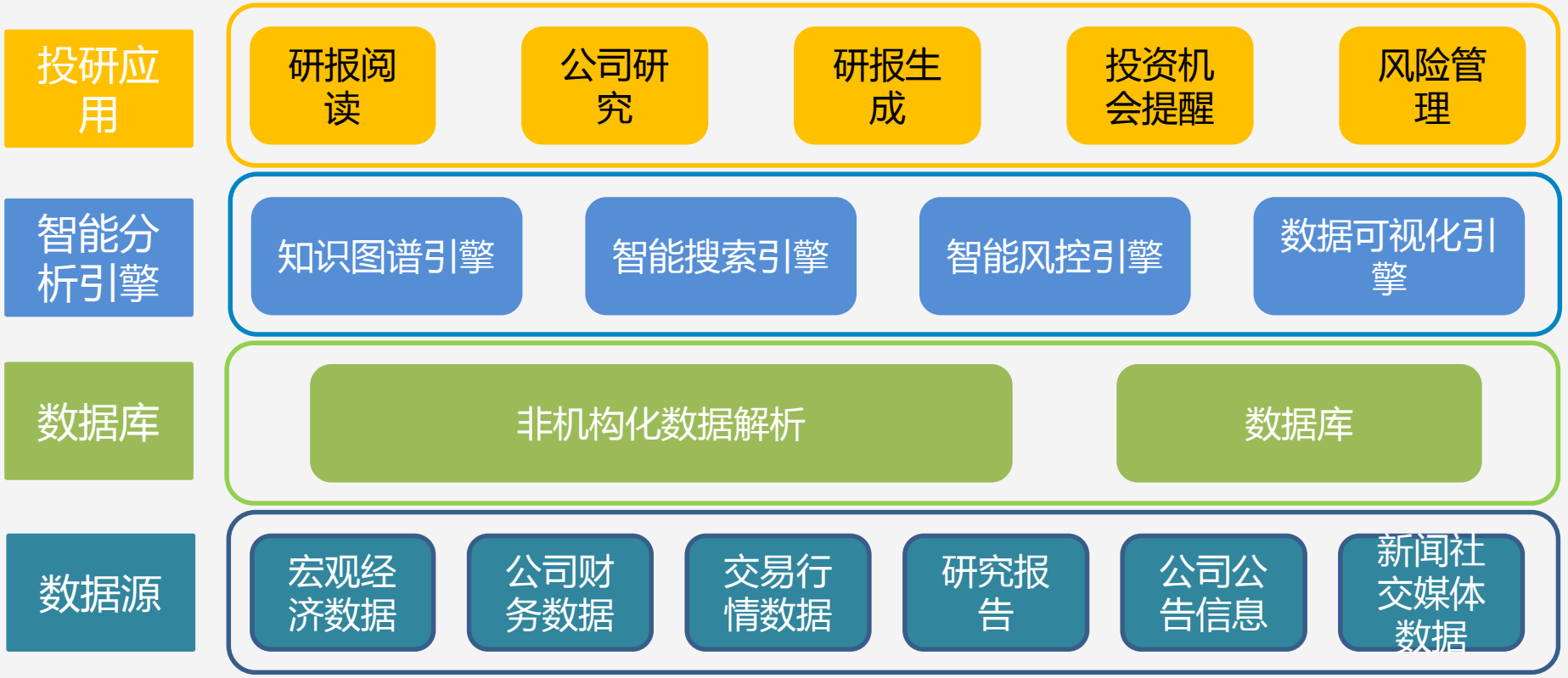
小葫芦网站流量监测

通过小葫芦网站监测不同娱乐平台网站的流量，包括开播数量、弹幕人数、弹幕条数和礼物收入等，用来估计上市公司的流量和营收。

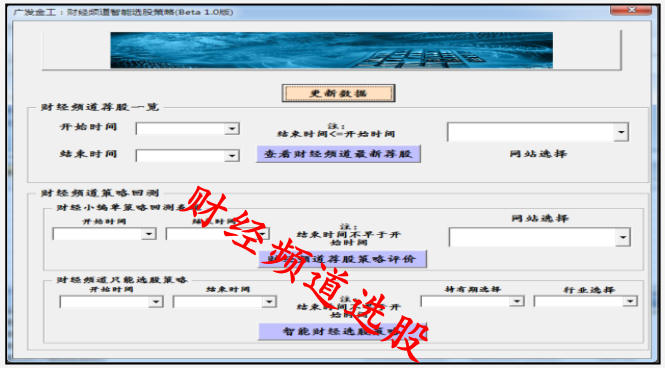
数据来源：小葫芦，广发证券发展研究中心

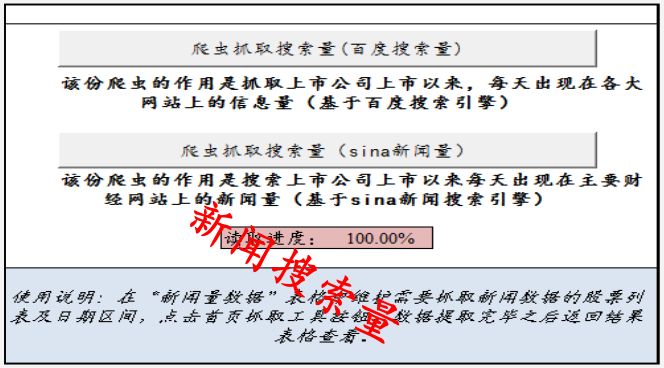
总结

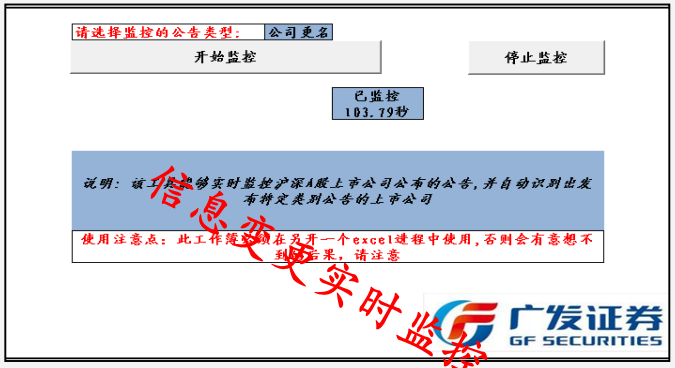
- 人工智能与大数据技术正在影响资产管理行业，数据、算法、算力可能成为未来资产管理公司的核心能力
- 投资研究领域，我们认为可行的路径是人机结合：机器会完成信息获取、数据处理、量化分析，为基金经理和研究员提供决策支持

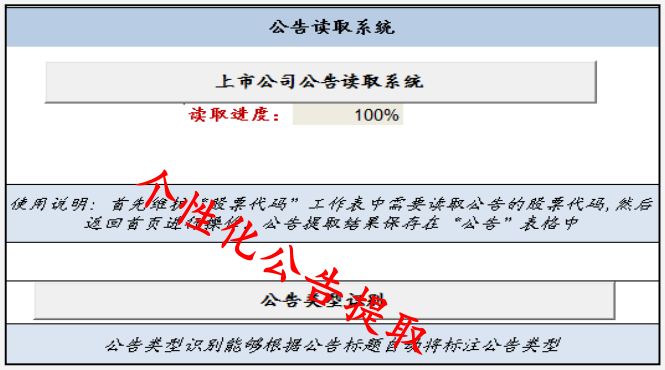


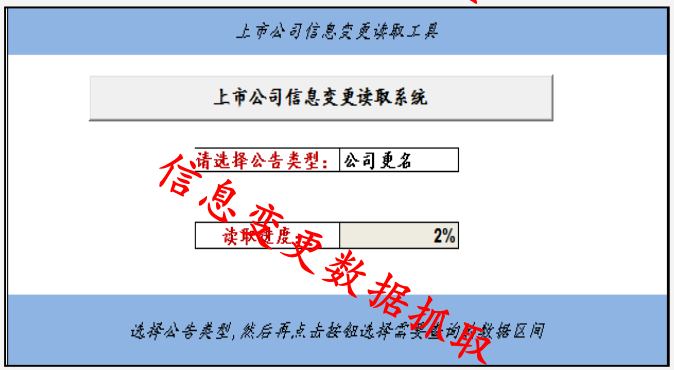
服务：研究、数据与爬虫小工具提供











本文旨在对所研究问题的主要关注点进行分析，因此对市场及相关交易做了一些合理假设，但这样会导致建立的模型以及基于模型所得出的结论并不能完全准确地刻画现实环境。而且由于分析时采用的相关数据都是过去的时间序列，因此可能会与未来真实的情况出现偏差。本文内容并不是适合所有的投资者，客户在制定投资策略时，必须结合自身的环境和投资理念。

免责声明

广发证券股份有限公司（以下简称“广发证券”）具备证券投资咨询业务资格。本报告只发送给广发证券重点客户，不对外公开发布，只有接收客户才可以使用，且对于接收客户而言具有相关保密义务。广发证券并不因相关人员通过其他途径收到或阅读本报告而视其为广发证券的客户。本报告的内容、观点或建议并未考虑个别客户的特定状况，不应被视为对特定客户关于特定证券或金融工具的投资建议。本报告发送给某客户是基于该客户被认为有能力独立评估投资风险、独立行使投资决策并独立承担相应风险。

本报告所载资料的来源及观点的出处皆被广发证券股份有限公司认为可靠，但广发证券不对其准确性或完整性做出任何保证。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价。广发证券不对因使用本报告的内容而引致的损失承担任何责任，除非法律法规有明确规定。客户不应以本报告取代其独立判断或仅根据本报告做出决策。

广发证券可发出其它与本报告所载信息不一致及有不同结论的报告。本报告反映研究人员的不同观点、见解及分析方法，并不代表广发证券或其附属机构的立场。报告所载资料、意见及推测仅反映研究人员于发出本报告当日的判断，可随时更改且不予通告。

本报告旨在发送给广发证券的特定客户及其它专业人士。未经广发证券事先书面许可，任何机构或个人不得以任何形式翻版、复制、刊登、转载和引用，否则由此造成的一切不良后果及法律责任由私自翻版、复制、刊登、转载和引用者承担。

Thanks！

谢谢

地址: 广州市天河北路183号大都会广场 P.C.510075 电话: 020-87555888 传真: 020-87553600 WWW.GF.COM.CN