segmentfault 思否 | HACK

大赛主题合作方 Google Developer Summit China 2022

CODE FOR BETTER

Hackathon

Code For Better_Hackthon

RiskTechist

风险侦察者

——基于文本数据（tensorflow）的企业风险评估系统

制作团队：哈尔滨工业大学
PraMet金融科技实验室
联系方式：15395099536
负责人微信：Sunliankun_HIT

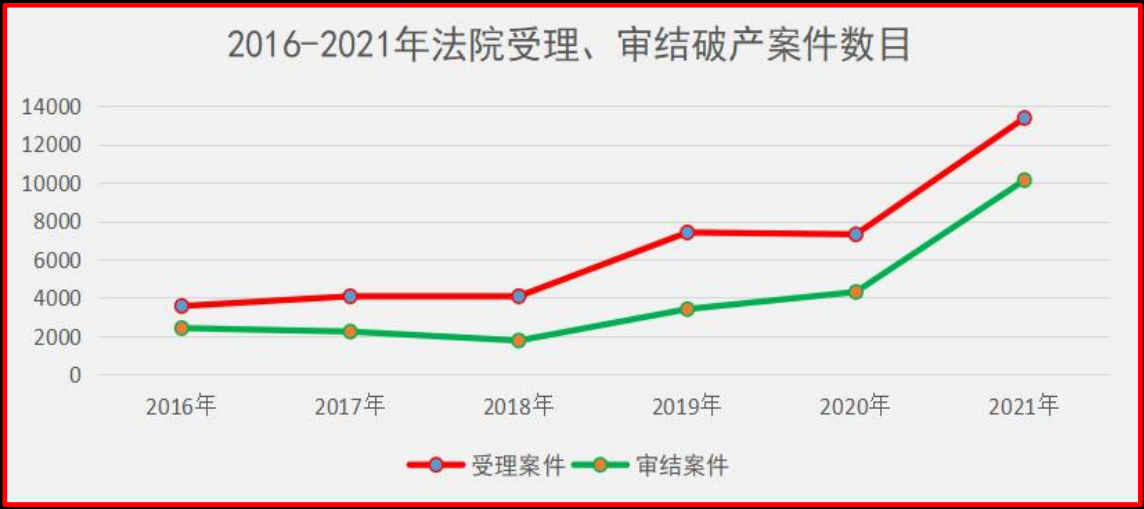
PART 01

需 求 分 析

1.1 问题价值：为什么要对公司执行风险评估程序？

企业风险暴雷，情况频发！

随之疫情常态化席卷，企业生存环境愈加恶劣，越来越多的**企业因资金链断裂，商誉暴雷等风险**，而面临破产清算...
从资本市场上来看，2022年退市公司已超过过去两年总和，**常态化清算退市**渐行渐近...



1月-从年入3.5亿到负债6亿的知名运动品牌德尔惠宣布停业
2月-无人货架gogo小超倒闭。
5月-浙江绍兴的中国500强企业盾安集团爆发出450亿元债务危机濒临破产
6月-康佳集团4.55亿收购破产倒闭的昔日电器霸主河南新飞100%股权。
7月-"大豆之王"、"中国民营企业500强"的山东晨曦被裁定破产。

企业风险暴雷，后果严重！

企业风险，包含市场风险、流动性风险、行业风险、政治风险、信用风险等
风险，在**恶劣的企业生存环境**下，不再是一次次小波折，而可能是企业的**直接颠覆和轰然垮台**...


中国证券监督管理委员会
CHINA SECURITIES REGULATORY COMMISSION
证券期货监督管理信息公开目录
索 引 号:40000895X/ 分类:行政处罚;行政处罚决定
发布机构:证监会 发文日期:2021年06月17日
名 称:中国证监会行政处罚决定书（永城煤电控股集团有限公司、强岱民等7名责任主体）
文 号:（2021）44号 主 题 词:
中国证监会行政处罚决定书（永城煤电控股集团有限公司、强岱民等7名责任主体）

*自去年以来，**AAA级永煤债"暴雷"**引发**债券"地震"**并引发连锁反应，债券**市场信心严重受损。***
8月3日下午，#永煤控股账上861亿现金全是假的#一度登上微博热搜话题榜，一石惊起千层浪，截至发稿，阅读量达到1169万，永煤控股再度被推上风口浪尖。

1.2 需求痛点：我们需要解决风险评估程序中的什么问题？

现有的风险量化评估，**只局限于**对反馈迟缓的**财务指标分析**

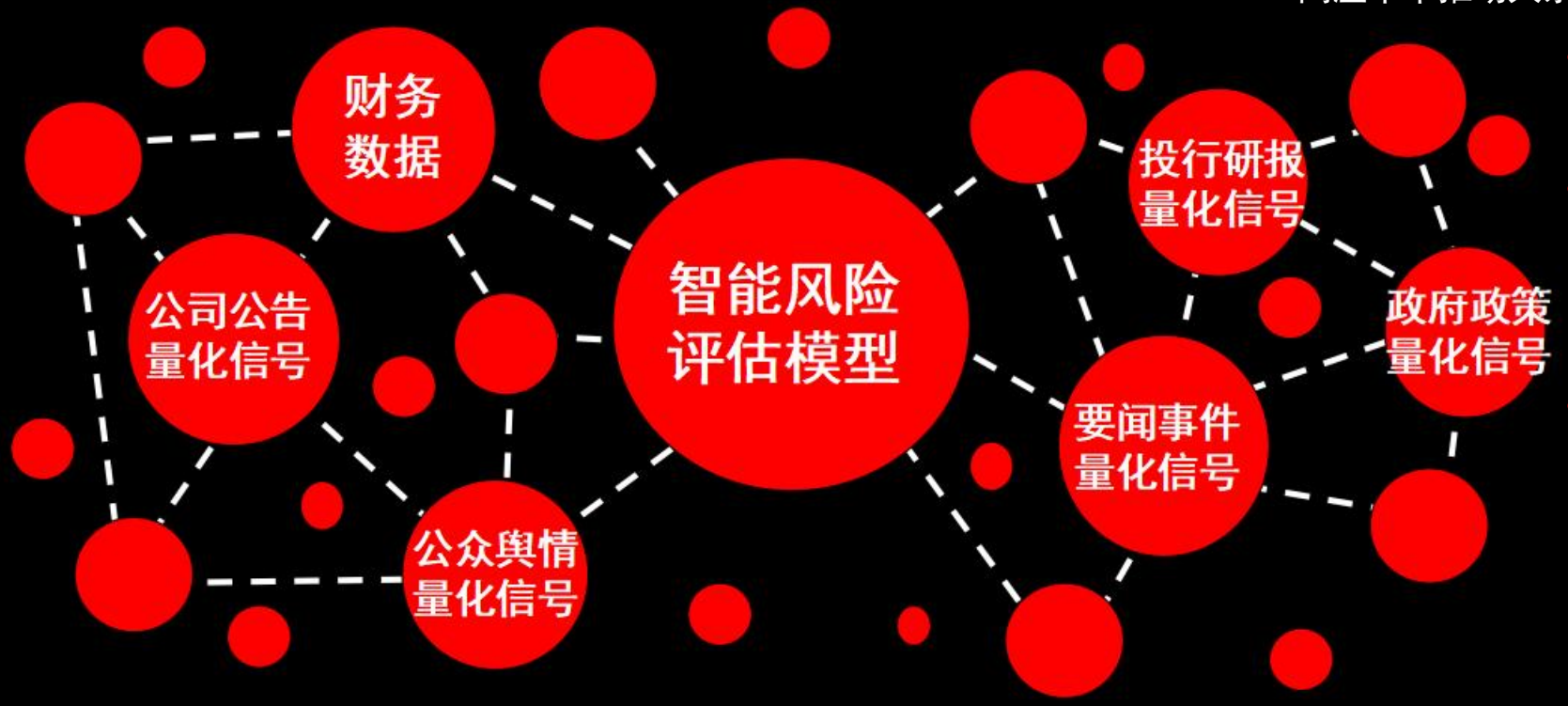
只局限于财务指标分析，会有什么后果？

- 财务指标多以季度或年为单位，反馈周期迟缓，**难以实时洞察**和监测可能的风险

现有的风险量化评估，**需要**宏中微观的**市场、要闻、舆情等**综合信息帮助判别

需要**更多的信息源**，来帮助风险等级的划分

- **公司公告量化信号**：《2022 年上半年甘肃定西持续干旱》→→天气情况：上半年定西气温偏高降水少→→行情变化：高温干旱推动大宗行情上行
- **政府政策量化信号**：《销售“四类药品”放开！这类情况下无需实名登记》→→不得以各种理由一刀切，行情上行
- **要闻事件量化指标**：《钧达股份：关于 2022 年股票期权激励计划首次授予完成的公告》→→企业发展战略和长期规划一定程度得到保障→→企业资金问题得到缓解
- **公众舆情量化信号**：“股海天天游：涨这么多了来说，不怀好意”“散户不愿来接盘，估计只到半山腰”→→公众情绪对于目前上行态势存疑，怀疑可能有刻意套路的风险



1.3 解决思路：利用“自然语言处理”+“文本数据”，助力智能风险评估新时代！

主要客户群体

主要帮助**金融机构**投资、放贷业务中，涉及到风险评估的环节

商业银行
花旗银行，宁波银行，摩根大通，摩根士丹利
...

证券公司
中金公司，中信建投，华泰证券，广发证券
...

财务公司
安永财务，毕马威财务
德勤财务，普华永道...

解决共同需求

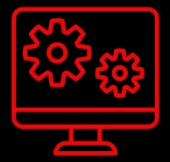
金融机构在对企业**投资、放贷**时，迫切的需要新型信息源，**更加准确、实时**的评估企业风险

准确、实时的**量化**企业风险

明确产品定位

一款企业风险的**量化数据分析**软件

产品**形式多样化**：软件、系统、网页、平台、插件...



产品不同形式
软件、系统、平台、插件

使用技术路线

使用爬虫爬取文本数据，然后利用基于**谷歌Tensorflow框架**的自然语言处理&深度学习模型训练

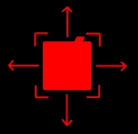
爬虫**爬取文本数据源**
新浪新闻，中国政府网，巨潮资讯...

利用**谷歌的Tensorflow框架**
搭建深度学习模型
LSTM, RNN, CNN, GNN...

训练**自然语言处理**模型
Word2Vec, Glove, Skip-gram...

取得期望效果

在风险评估环节
更加**准确**的测度风险
更加**省力**的评估风险
更加**实时**的量化风险



准确

省力

实时

PART 02

业务设计

2.1 寻找风险识别的“新水源”，企业文本数据

为什么文本数据值得深入挖掘？

企业风险的影响变动无非通过行业龙头的新闻资讯或者公告，以及政府发布的行业政策，大众传递的论坛舆情这几部分组成， 毕竟**文本（新闻、政策、评论）**，是**事件与信息的主要载体**。

有哪些种类的文本数据，值得深入挖掘？

与企业风险紧密关联的文本信息，主要由**企业新闻、政府政策、公司公告、公众评论**四部分组成，其在**文本长度、速度&权威、发布主体、情绪强度、案例示范**，五个维度对企业风险特征进行评估

属性\类别	新闻	政策	公告	评论
文本长度	可长可短	可长可短，关键信息短小	可长可短，关键信息短小	总体极为短小
速度&权威	捕风捉影，速度最快	官方权威性最强，但发布速度较慢	公司选择性发布非负面事件	速度次于新闻
发布主体	媒体	政府	公司	股民
情绪强度	有情绪倾向，但不明显	无情绪倾向，但有专业关键词	无情绪倾向，但有专业关键词	有明显的情绪倾向
案例示范	短线见此信号大胆买入，精细化工龙头中报业绩大爆发	减速器成机器人最强分支！龙头股9天6板创历史新高	莱茵生物-《关于控股股东，实际控制人部分股份质押地公告》	《抄牛牛：供不应求，3万吨碳酸锂，科力远，严重低估》

2.2 梳理不同文本类别的推理路径

不同种类的文本类别，在**风险推演路径**中，**有什么区别？**

不同类型的文本信息，也会各自体现出不同的风险推演路径，也就是我们人类面对不同 的文本来源，对于可能的风险将会处于不同的推理路径或者方式，接下来我们将对几种主要文本信息，进行**风险推演路径的分析和总结**，方便后续指标体系的建立：

新闻类

《2022 年上半年甘肃定西持续干旱》

→→天气情况：上半年定西气温偏高降水少

→→行情变 化：高温干旱推动大宗行情上行

→→原料解析：党参、当归关注度最高

政策类

《销售“四类药品”放开！这类情况下无需实名登记》

→→保证疫情风险区域群中正常用药 需求

→→不得以各种理由一刀切，行情上行

公告类

《钧达股份：关于 2022 年股票期权激励计划首次授予完成的公告》

→→企业发展战略和长期 规划一定程度得到保障

→→企业资金问题得到缓解

评论类

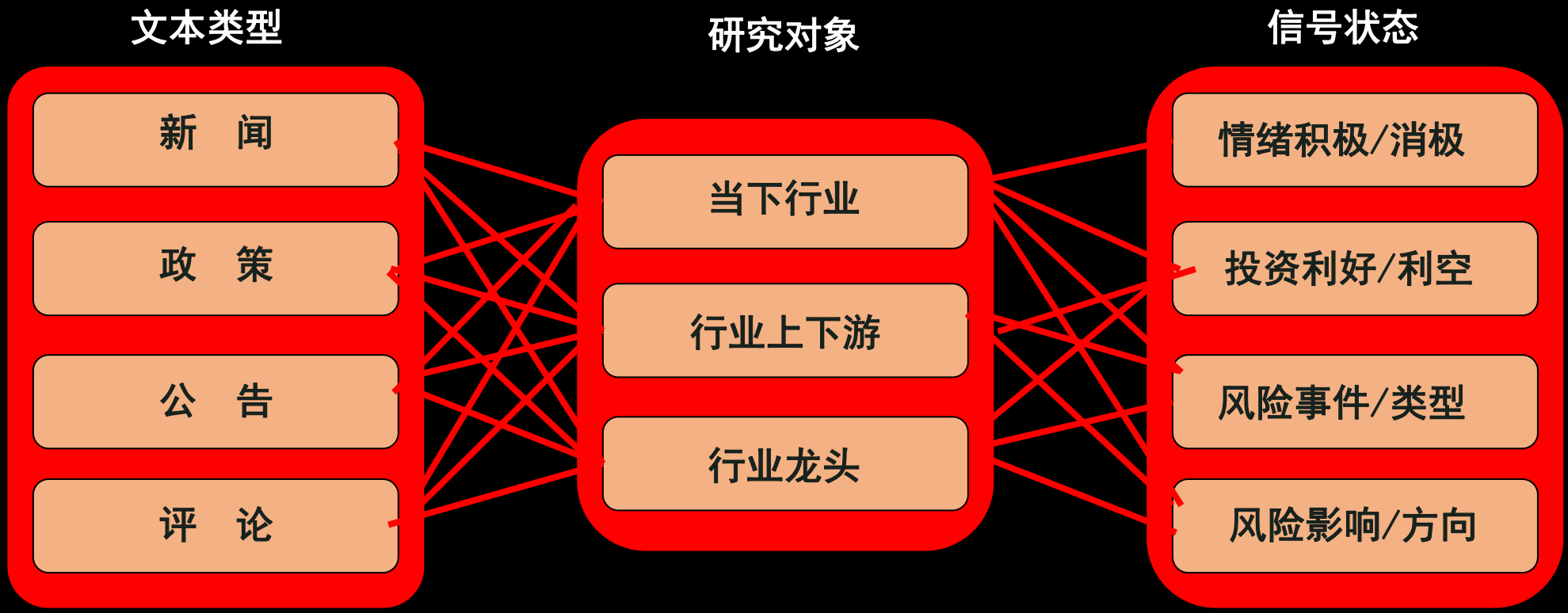
股海天天游：涨这么多了来说，不怀好意”“散户不愿来接盘，估计只到半山腰”

→→公众 情绪对于目前上行态势存疑，怀疑可能有刻意套路的风险

2.3 构建不同文本类别的指标体系

不同种类的文本类别，可以被**挖掘出多少种**指标信号？

不信息指标体系的设定需要对于不同文本信号的洞察，而特征还面临**不同应用场景**（情绪 高低，投资价值，风险类别和方向）和**不同主体**（当下行业，行业上下游，行业龙头），进而衍生出**64种体系化的文本信号**。



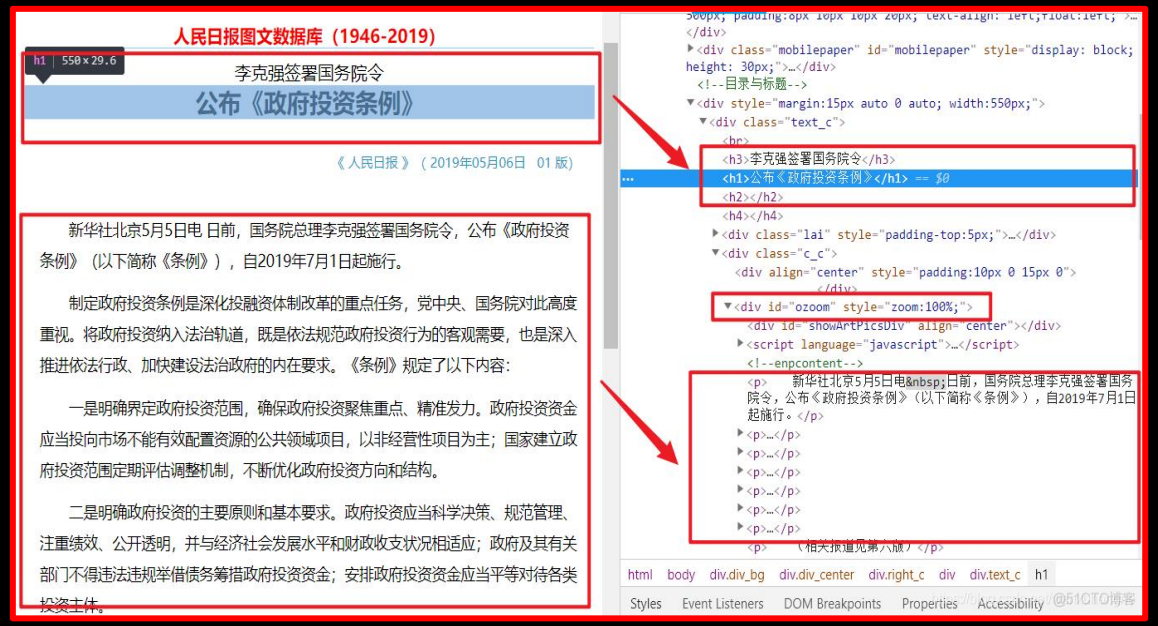
PART 03

技术路线

3.1 爬取文本数据源（以政策爬取为例）

爬取目标页面的Html分析

进入文章内容页面之后，我们可以知道，文章标题存放在 h1 ，h2 ，h3 标签中（有的文章标题只用到了 h1 标签，而有的文章有副标题可能会用到 h2 或 h3 标签）， 正文部分存放在 id = “ozoom” 的 div 标签下的 p 标签里。



使用同样的方法，我们可以知道，文章目录存放在一个 id = “titleList” 的 div 标签下的 ul 标签中，其中每一个 li 标签表示一篇文章。

制定、执行爬取策略

第一遍，先爬取版面目录，将每一个版面的链接保存下来；
第二遍，依次访问每个版面链接，将文章链接保存下来；
第三遍，依次访问每一个链接，将文章标题正文保存本地。

```
    # 输入起止日期，爬取之间的新闻
    beginDate = input('请输入开始日期:')
    endDate = input('请输入结束日期:')
    data = get_date_list(beginDate, endDate)

    for d in data:
        year = str(d.year)
        month = str(d.month) if d.month >=10 else '0' + str(d.month)
        day = str(d.day) if d.day >=10 else '0' + str(d.day)
        download_rmrb(year, month, day, 'data')
        print("爬取完成：" + year + month + day)
#         time.sleep(3)        # 怕被封 IP 爬一爬缓一缓，爬的少的话可以注释掉
```

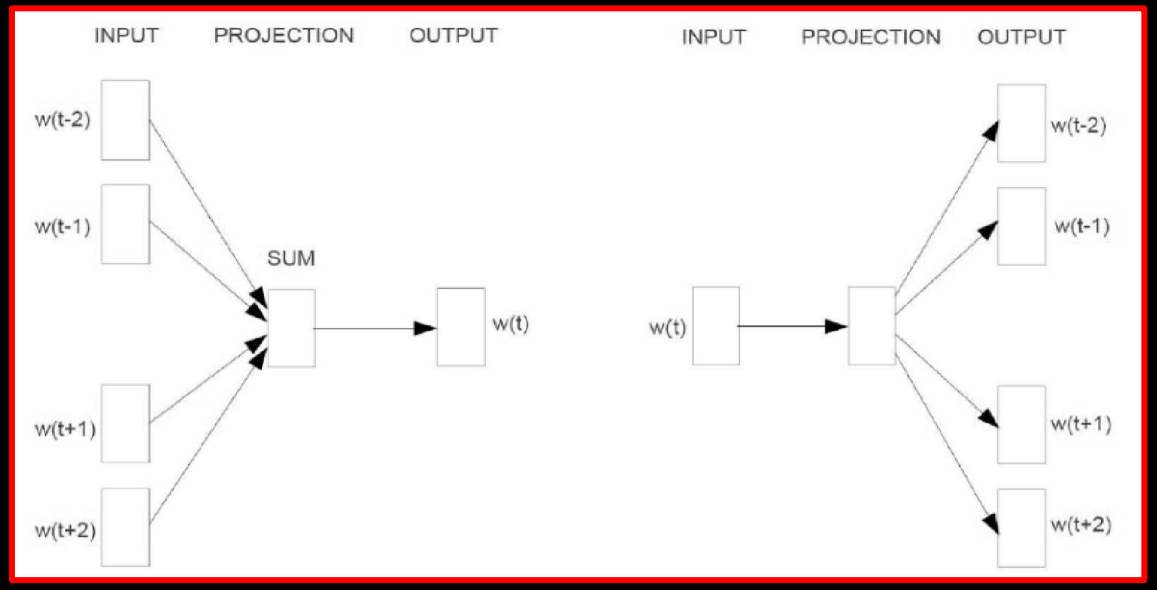
```
请输入开始日期:20190401
请输入结束日期:20190501
爬取完成：20190401
爬取完成：20190402
爬取完成：20190403
爬取完成：20190404
爬取完成：20190405
爬取完成：20190406
爬取完成：20190407
爬取完成：20190408
爬取完成：20190409
爬取完成：20190410
爬取完成：20190411
爬取完成：20190412
爬取完成：20190413
```

fetchUrl 函数用于发起网络请求
getPageList 函数，用于爬取当天报纸的各版面的链接
etPageList 函数，用于爬取当天报纸的某一版面的所有文章的链接

3.2 编码文本向量的特征工程

基于Word2vec的词向量编码

词向量(word embedding) 是词的一种表示，是为了让计算机能够处理的一种表示。 因为目前的计算机只能处理数值， 诸英文，汉字等等它是理解不了的， 最简单地让计算机处理自然语言的方式就是为每个词编号



word2vec的建模方式，有两个版本， 一个是Continuous bag of words (cbow)，另一个是 skip-gram. 实现方式也有两个方式： 一个是 Hierarchical softmax 另一个是Negative Sampling.

基于文本预处理的特征工程

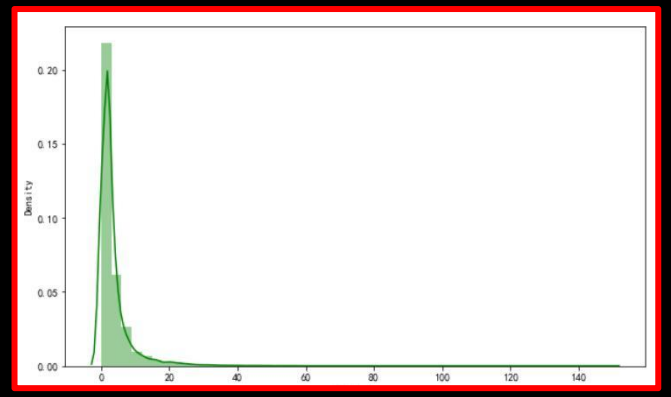
根据前文指标体系，构建对应特征工程

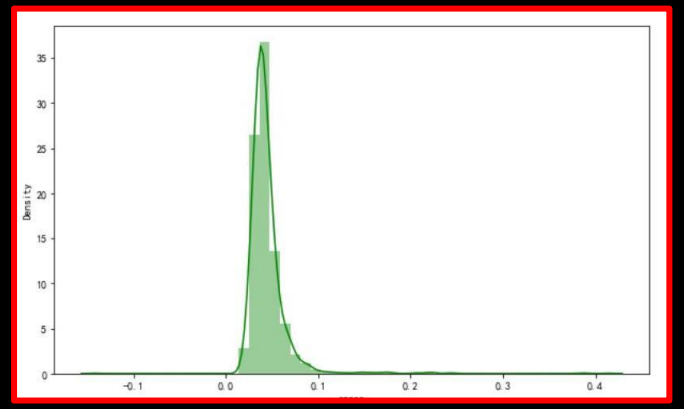
根据字段属性和类型，选用数据库并存储入语料库

id	nation_id	type	search_c1	search_c2	event	direction	grade	relevant_entity	create
2	1	301	服务业	假发	就业... 12B	1	0	政治稳定 12B	2022-0
3	2	301	化工业	凉茶	政治... 12B	1	1	就业情况 12B	2022-0
4	1	301	服务业	假发	政治... 12B	1	0	就业情况 12B	2022-0
5	2	301	化工业	假发	政治... 12B	1	0	政治稳定 12B	2022-0
6	1	301	化工业	假发	就业... 12B	1	1	就业情况 12B	2022-0
7	2	301	旅游业	凉茶	就业... 12B	1	1	政治稳定 12B	2022-0
8	1	301	食品业	凉茶	就业... 12B	1	2	就业情况 12B	2022-0
9	2	301	摩托车业	头盔	就业... 12B	1	0	政治稳定 12B	2022-0
10	1	301	食品业	假发	政治... 12B	0	2	政治稳定 12B	2022-0
11	2	302	旅游业	椰子	就业... 12B	0	1	就业情况 12B	2022-0
12	1	302	旅游业	椰子	就业... 12B	0	0	政治稳定 12B	2022-0
13	2	301	摩托车业	凉茶	就业... 12B	0	1	就业情况 12B	2022-0
14	1	301	旅游业	椰子	政治... 12B	0	0	政治稳定 12B	2022-0
15	2	302	摩托车业	假发	政治... 12B	0	2	就业情况 12B	2022-0
16	1	302	服务业	凉茶	政治... 12B	1	2	政治稳定 12B	2022-0
17	2	302	旅游业	凉茶	政治... 12B	0	0	政治稳定 12B	2022-0
18	1	302	旅游业	假发	政治... 12B	1	1	就业情况 12B	2022-0
19	2	302	化工业	椰子	就业... 12B	0	1	政治稳定 12B	2022-0
20	1	302	化工业	凉茶	就业... 12B	1	0	就业情况 12B	2022-0
21	2	302	旅游业	椰子	政治... 12B	0	0	政治稳定 12B	2022-0

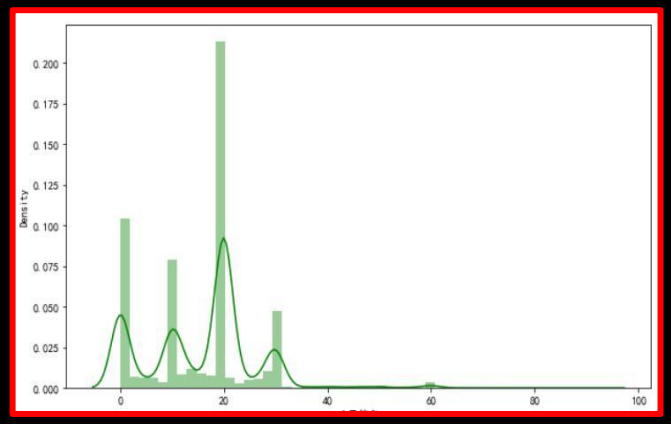
3.3 基于相关性与分布态的数据预处理

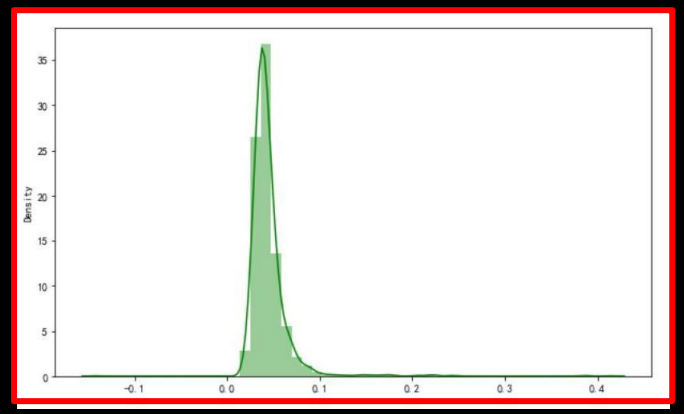
数值型特征的偏态处理





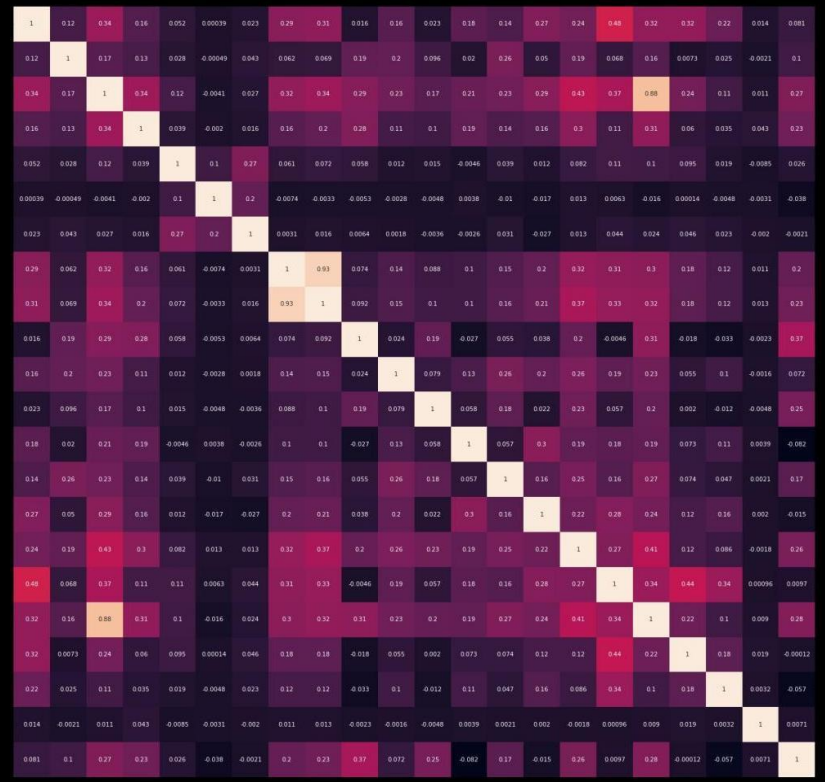
对数变换：处理相乘关系，高度偏态的数据





平方根变换：处理泊松分布（方差与均值近似相等），轻度偏度的数据

特征组合高度相关性处理



如果变量与变量之间存在较强的相关性，则代表着变量之间可以相互解释，不需要两个高相关 性的变量同时出现。

3.4 安装编译Google的Tensorflow

建立tensorflow的计算环境

先激活tensorflow环境，界面变化如下：

```
Fetching package metadata .......
Solving package specifications: ..........

Package plan for installation in environment /Users/lei.wang/anaconda/envs/tensorflow:

The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    openssl-1.0.2j             |                0         3.0 MB
    setuptools-27.2.0          |            py27_0         522 KB
    ------------------------------------------------------------
                                           Total:         3.5 MB

The following NEW packages will be INSTALLED:

    openssl:    1.0.2j-0
    pip:        8.1.2-py27_0
    python:     2.7.12-1
    readline:   6.2-2
    setuptools: 27.2.0-py27_0
    sqlite:     3.13.0-0
    tk:         8.5.18-0
    wheel:      0.29.0-py27_0
    zlib:       1.2.8-3

Proceed ([y]/n)? y

Fetching packages ...
^Copenssl-1.0.2j   0% |
Traceback (most recent call last):
```

激活tensorflow环境，然后用pip安装

先激活tensorflow环境，界面变化如下：

```
lei.wang ~/code/python/pdb $ source activate tensorflow
(tensorflow) lei.wang ~/code/python/pdb $
```

命令提交以后，等待即可，呈现如下：

```
(tensorflow) lei.wang ~/code/python/pdb $ pip install --ignore-installed --upgrade https://storage.googleapis.c
Collecting tensorflow==0.8.0rc0 from https://storage.googleapis.com/tensorflow/mac/tensorflow-0.8.0rc0-py2-none
  Retrying (Retry(total=4, connect=None, read=None, redirect=None)) after connection broken by 'NewConnectionEr
tion object at 0x10381a410>: Failed to establish a new connection: [Errno 65] No route to host',)': /tensorflow
  Downloading https://storage.googleapis.com/tensorflow/mac/tensorflow-0.8.0rc0-py2-none-any.whl (19.3MB)
    100% |████████████████████████████████| 19.3MB 21kB/s
Collecting six>=1.10.0 (from tensorflow==0.8.0rc0)
  Downloading six-1.10.0-py2.py3-none-any.whl
Collecting protobuf==3.0.0b2 (from tensorflow==0.8.0rc0)
  Downloading protobuf-3.0.0b2-py2.py3-none-any.whl (326kB)
    100% |████████████████████████████████| 327kB 709bytes/s
Collecting numpy>=1.10.1 (from tensorflow==0.8.0rc0)
  Downloading numpy-1.11.2-cp27-cp27m-macosx_10_6_intel.macosx_10_9_intel.macosx_10_9_x86_64.macosx_10_10_intel
    100% |████████████████████████████████| 3.9MB 2.6kB/s
Collecting wheel (from tensorflow==0.8.0rc0)
  Downloading wheel-0.29.0-py2.py3-none-any.whl (66kB)
```

简单测试tensorflow是否安装成功

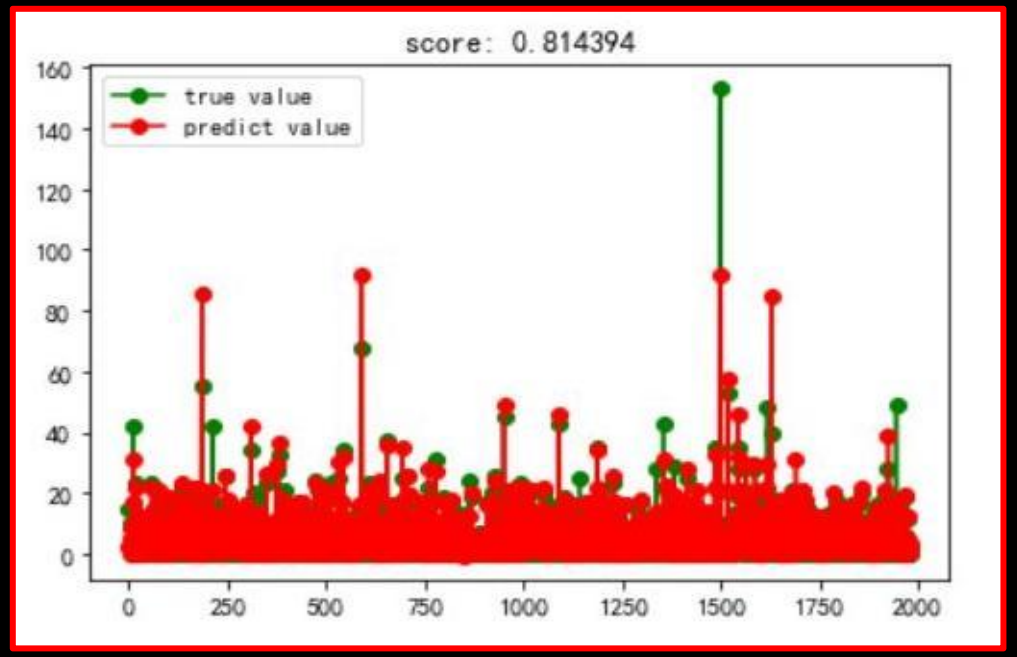
```
(tensorflow) lei.wang ~/code/python/pdb $ python
Python 2.7.12 |Continuum Analytics, Inc.| (default, Jul  2 2016, 17:43:17)
[GCC 4.2.1 (Based on Apple Inc. build 5658) (LLVM build 2336.11.00)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io/thanks and https://anaconda.org
>>> import tensorflow as tf
>>> hello = tf.constant('Hello, TensorFlow!')
>>> sess = tf.Session()
>>> print sess.run(hello)
Hello, TensorFlow!
```

备注：将tensorflow集成到IDE里，步骤也很简单。以IntelliJ为例，跟创建普通项目唯一的区别就是，创建普通项目的时候我们的Module SDK选项是系统默认的python解释器。如果我们想要使用tensorflow的相关代码，将Module SDK换为刚刚我们新建的tensorflow计算环境即可

3.5 基于长短期记忆神经网络（LSTM）的企业风险预测

搭建LSTM进行企业风险预测

以企业是否退市，即是否为ST，作为label，结合64种经过清洗预处理之后的文本信号和基本财务信号，进行LSTM的训练。

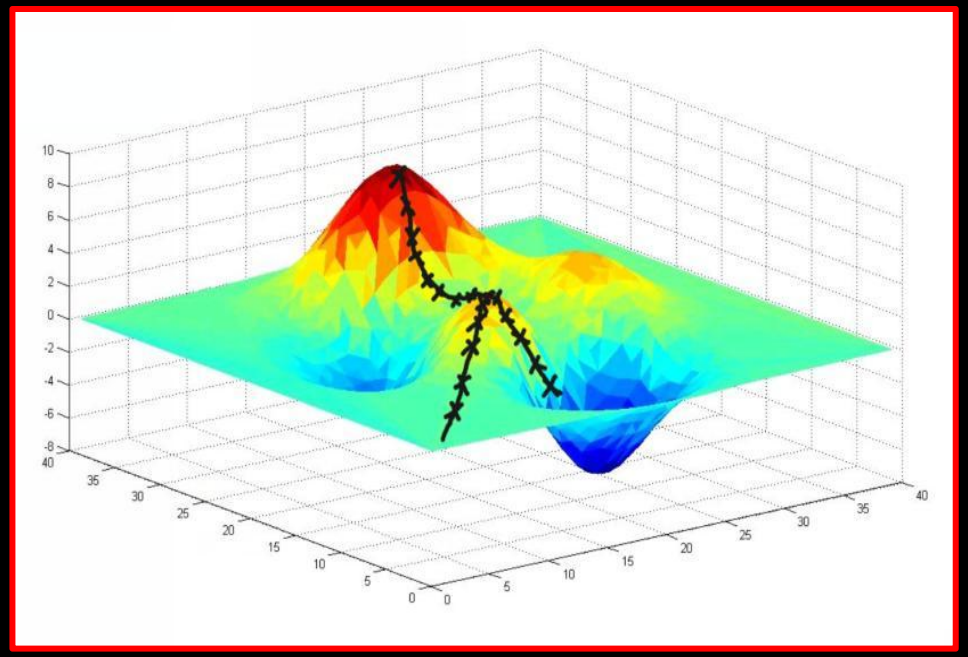



```
learning_rate = 0.001
training_steps = 10000
batch_size = 128
display_step = 200

num_inputs = 28
timesteps = 28
num_hiddens = 64
num_classes = 10
num_layers = 2
```

利用遗传算法进行神经网络参数调优

我们选择决策树的主要三个参数进行空间参数搜索的调优，由于逐步遍历耗费的时间和 空间复杂度太高，所以此处我们选择遗传算法进行启发式搜索来帮助我们进行参数优化。



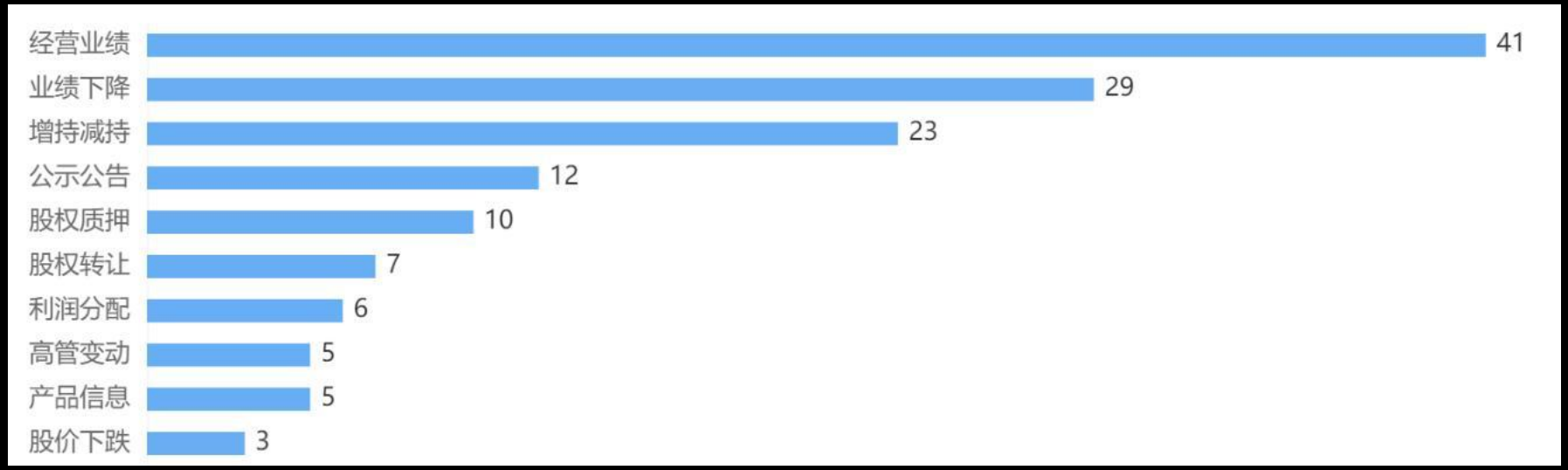
```
min_samples_split:  内部节点再划分所需最小样本数
min_samples_leaf:  叶子节点（即分类）最少样本数。
min_weight_fraction_leaf:  叶子节点最小的样本权重和
```

PART 04

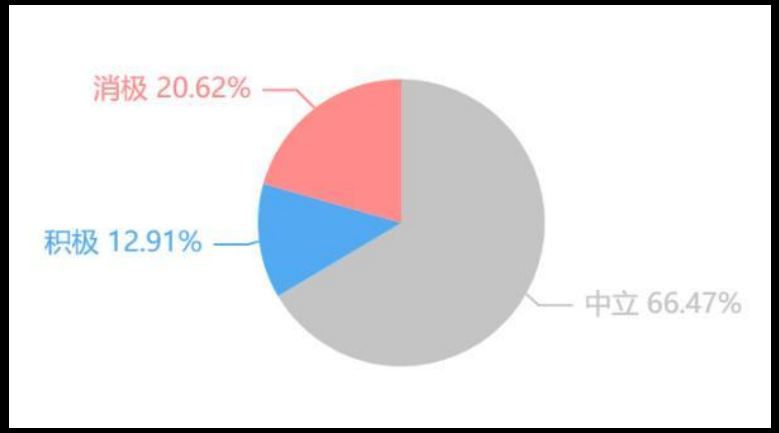
产品设计

4.1 企业风险评估的解释性

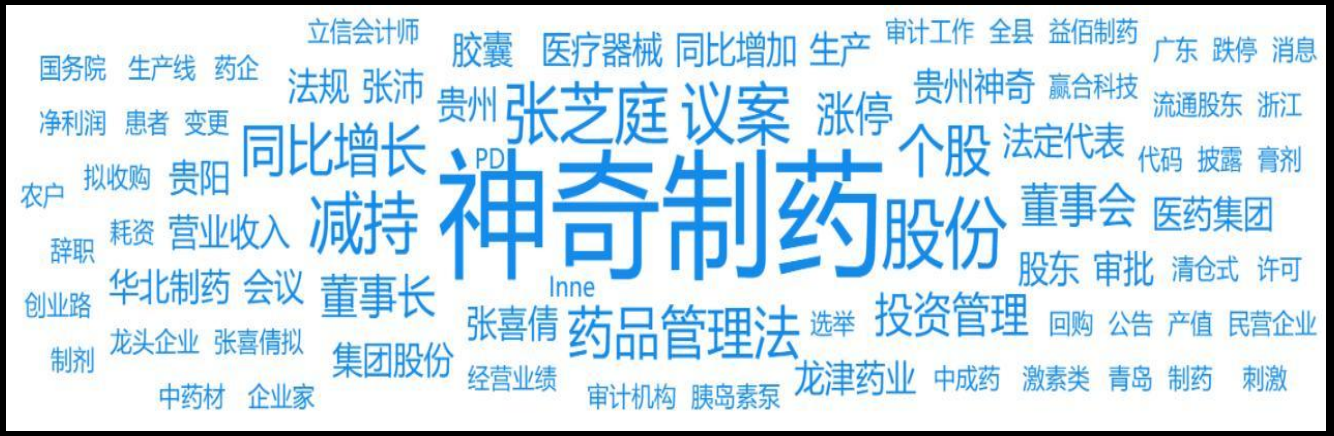
上海神奇制药投资管理股份有限公司的Top10公告事件类型



上海神奇制药投资有限公司的评论情绪分布图



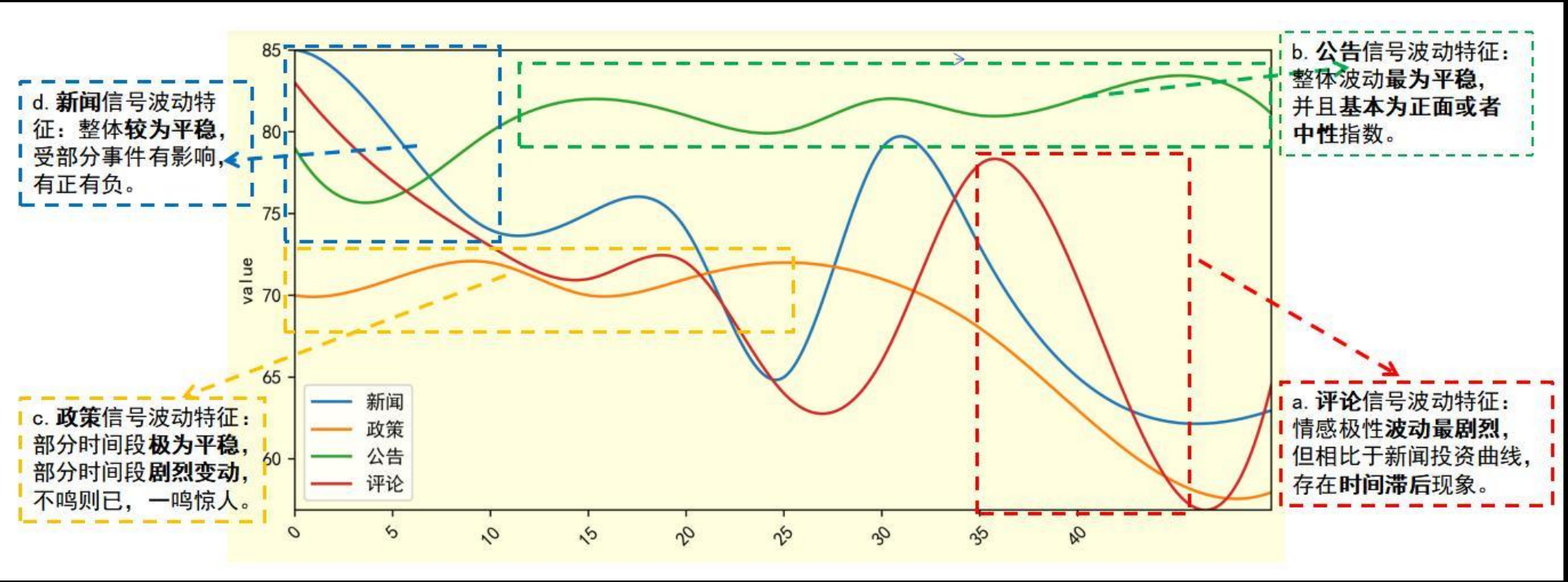
上海神奇制药投资管理股份有限公司的新闻资讯词云图



4.2 四类文本信号的波动特征总结

a. 评论信号波动特征：情感极性波动最剧烈，但相比于新闻投资曲线，存在时间滞后现象。 评论特征原因解读：股民发言情绪激烈，并且一般在看到明显事件之后才发生。

b. 公告信号波动特征：整体波动最为平稳，并且基本为正面或者中性指数。 公告特征原因解读：企业选择性披露重大事件公告，一般为正面或者中性事件。



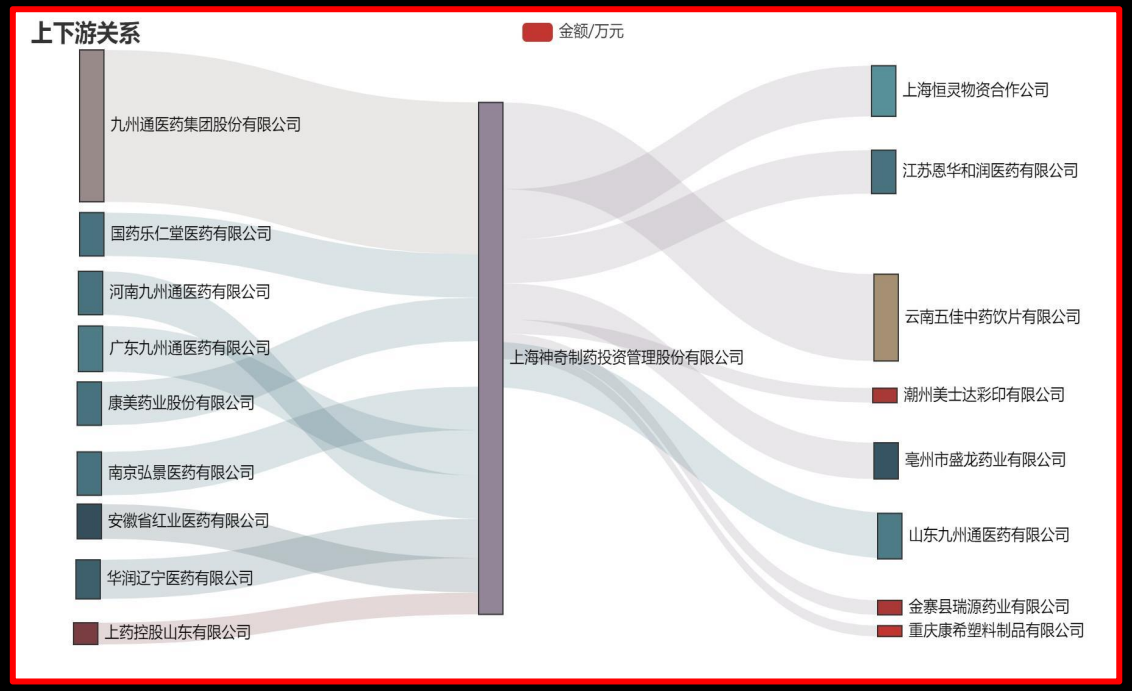
c. 政策信号波动特征：部分时间段极为平稳，部分时间段剧烈变动，不鸣则已，一鸣惊人。 政策特征原因解读：政策发布周期较长，一旦发布则影响覆盖面较广，影响较大。

d. 新闻信号波动特征：整体较为平稳，受部分事件有影响，有正有负。 新闻特征原因解读：新闻受媒体发布，不论好坏，受事件影响进行披露报道。

4.3 企业主体之间的风险传递总结

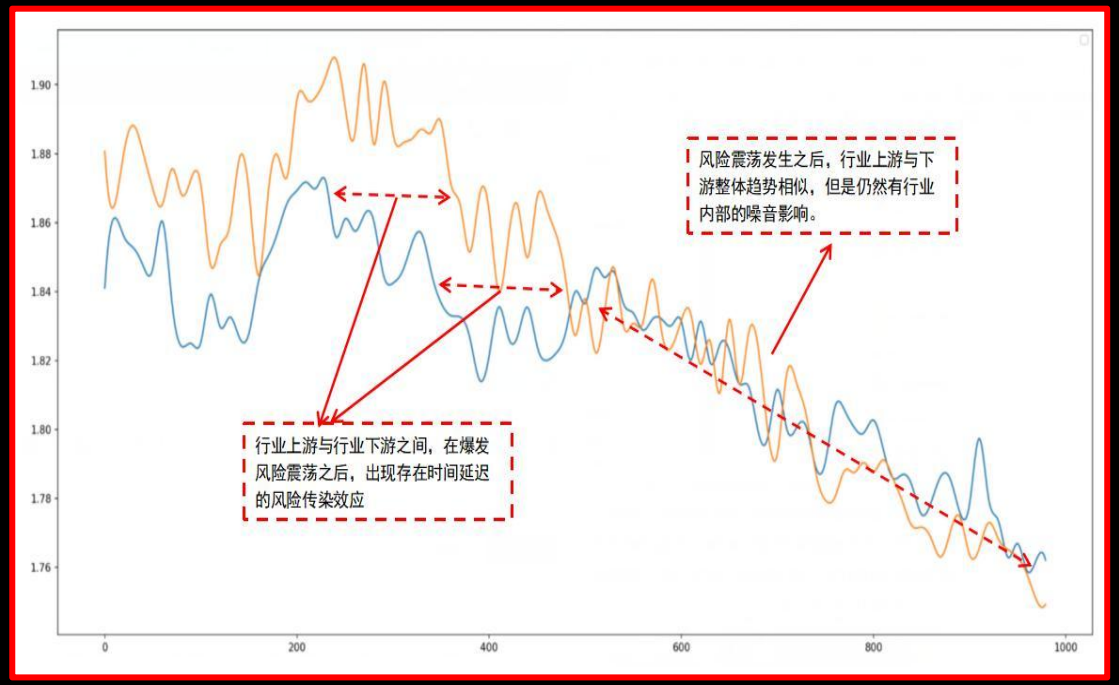
供应链风险上下游走势图

根据前文对于行业上下游的文本风险信号的衡量，我们首先梳理出上海神奇制药公司的 行业上下游分布图。



风险预警信号时序图

接下来针对上述行业上下游所在细分行业，进而根据时间戳划分得出数种信号的时间序列



利用互相关检测，来测算数个不同时间序列之间指标的相关性，并且可以测算出可 能的时间延迟区间，也就是滞后项

PART 05

团 队 协 作

5.1 团队组成：来自全国第1所Alternative Data高校实验室PraMet

PraMet实验室指导教授王闻

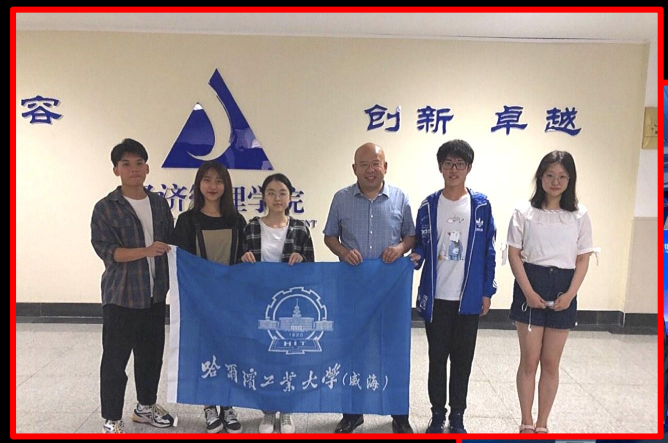
王闻，哈尔滨工业大学应用经济系主任，哈工大数量金融专业创办人出版**中国第1本Alternative Data**投资书籍的著作人

清华大学金融学博士，**浙江大学**博士后，在他“金融经济学”和“资产定价”、“金融工程”和“数理金融”这几个研究方向都有出色的研究成果

研究方向为机器学习(自编码器，决策树，LSTM, CNN, RNN, GNN. . .）在量化投资的应用，另类数据（文本数据，卫星数据，网络数据）在金融领域的应用

PraMet实验室团队成果

本团队成员均为**哈尔滨工业大学**的PraMet金融科技**实验室核心骨干成员**，我们致力于各类Alternative Data的挖掘，如本次课题研究的文本类数据



本团队成员已**合力斩获国内多项赛事**奖项，如**花旗杯金融科技全国一等奖**（作品：智能营销系统），**宁波银行金融科技挑战赛全国第1名**（作品：智慧信贷平台），**美团商业分析全国前十强**（作品：电商平台虚假评论监测系统），且均由本团队开发完成，合作经验丰富

5.2 团队协作：成员之间分工明确，整体规划紧密详实

成员之间分工明确

特征挖掘工程师

孙健坤，19级哈工大信管，负责数据预处理和特征工程环节

产品设计分析师

祝雨晴，19级哈工大信管，负责产品界面的UI设计

自然语言处理工程师

张伟杰，20级哈工大信管，负责自然语言处理环节

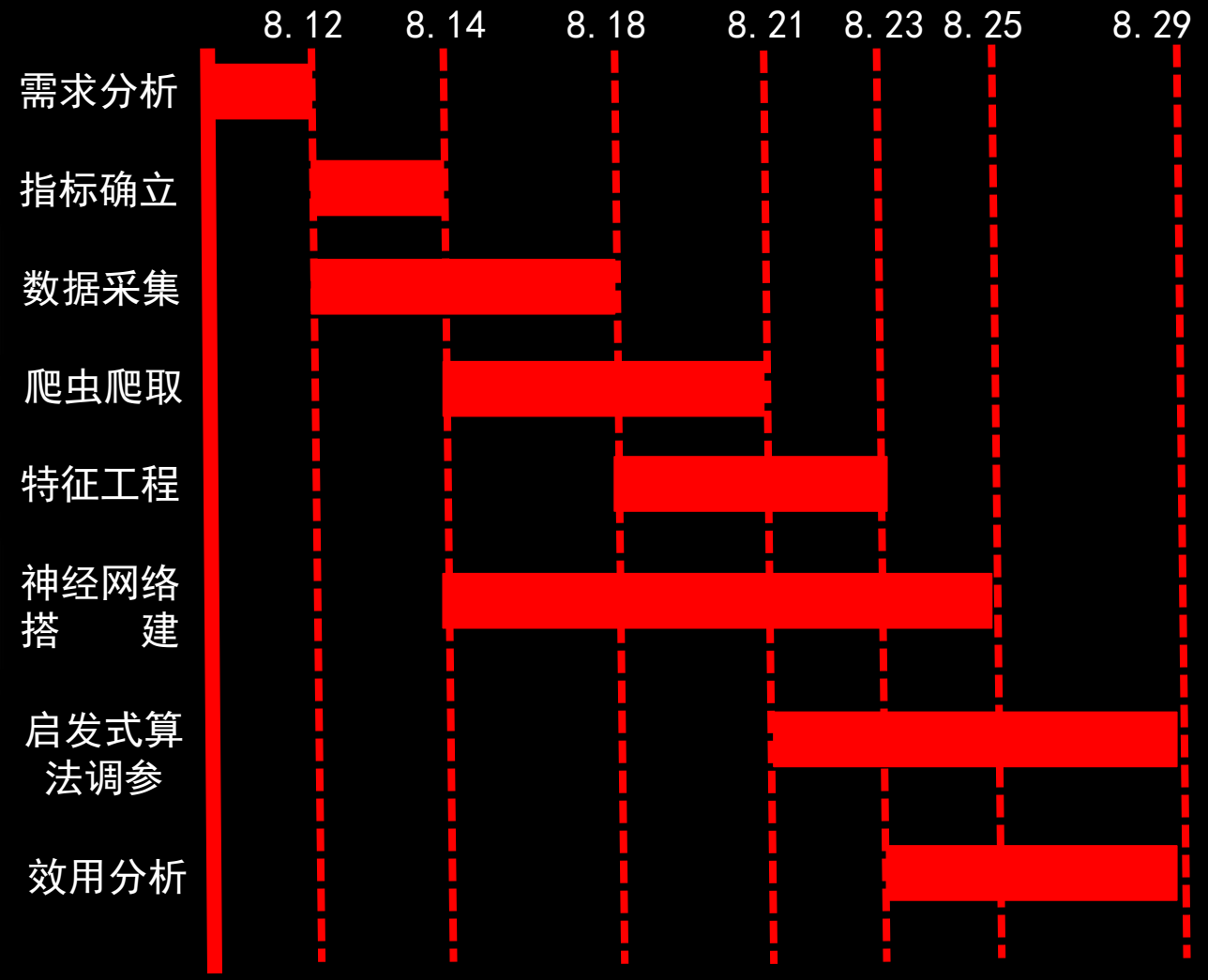
业务需求分析师

陈溪晗，20级哈工大智能会计，负责业务需求的对接

深度学习工程师

陆庆丰，20级哈工大信管，负责LSTM,RNN等深度学习网络的搭建

整体规划紧密详实



谢谢观看

风险侦察者