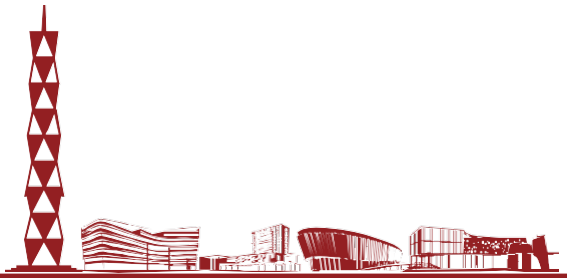
运用新闻和公司信息预测医疗行业股市波动

——沈昭正 井文可 钱丽玥

- Introduction
- Aims
- Preprocessing
- Training
- Result
- Issues and Challenges
- Conclusion
- Reference



Introduction

- 传统股票价格预测
  ---银行板块数据系列
  ---股票收盘价

- 新闻、社交媒体对金融市场产生了很大的影响

- 已有研究包括用支持向量机[1]、贝叶斯网络[2]、深度学习[3]

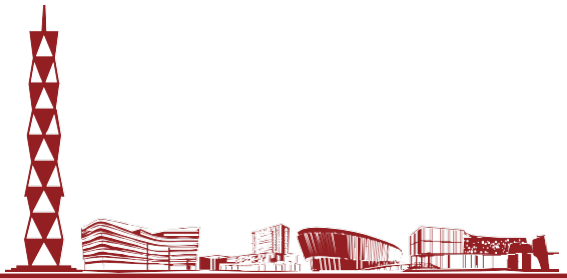
Aims

- 新闻情绪分析（NLP）

  ---预测新闻标题情绪

- 股价预测

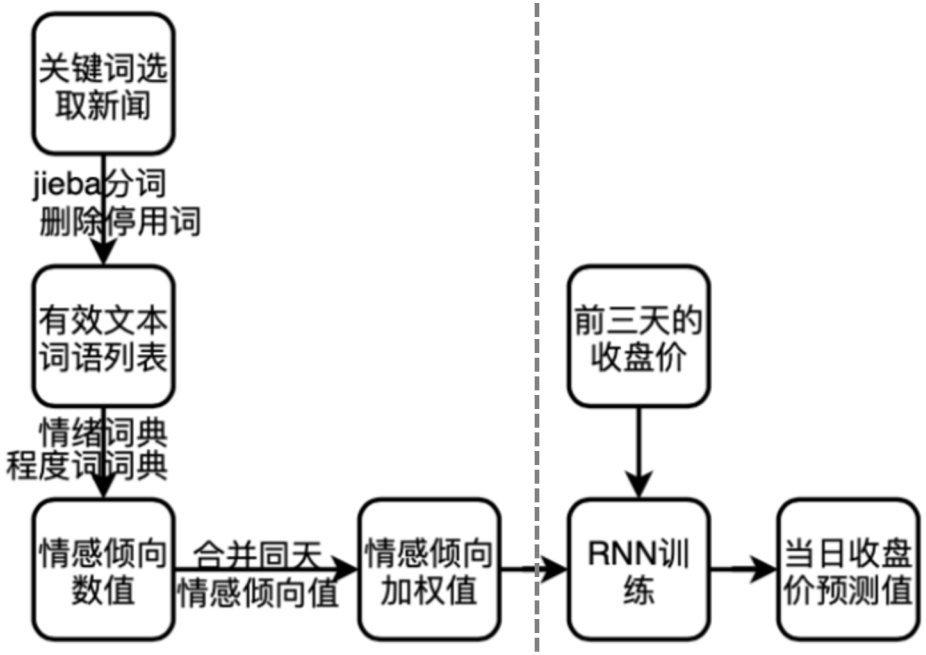
  ---RNN



Preprocessing

情感分析

收盘价预测



Preprocessing

- 新闻数据
  ---东方财富上公司的热帖和资讯信息
  --- 2020年5月25日至2021年5月23日
  ---复星医药、沃森生物、恒瑞医药

阅读	评论	标题	作者	最后更新
66949	555	复星医药，这一波底部极限在37-39元（39.01，37.50区	何然静	05-24 15:36
14096	36	新冠疫苗获批在即 复星医药“赶进度”：工厂已落地 3.	寻100倍股	05-24 15:17
31853	228	朋友们，下午好。这次再次压中复星医药阶段性高点和低	何然静	05-24 13:13

阅读	评论	标题	作者	发帖时间
2656	1	复星医药本周融资净买入1.26亿元，居医药制造板块第九	两融追踪	05-23 15:05
22508	135	复星医药：愿意将疫苗服务于台湾同胞	复星医药资讯	05-22 16:09
637	0	复星医药(600196)融资融券信息(05-21)	复星医药资讯	05-22 07:40

Preprocessing

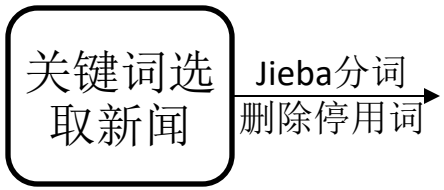
- 新闻数据量

公司名称 \ 新闻种类	热帖	资讯
复星医药	946	4471
沃森生物	1048	2511
恒瑞医药	286	3603

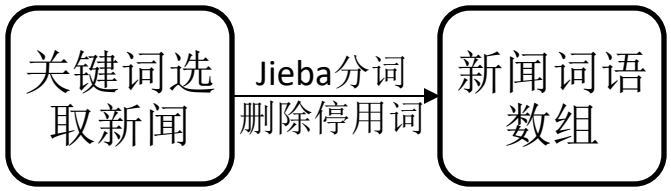
Preprocessing

关键词选取新闻

Preprocessing



Preprocessing



Preprocessing

- 情感词典

  ---我们自己建立了财经新闻相关的积极情感词典与消极情感词典。程度词词典与否定词词典来自知网情感分用词语集beta版。

情感极性	数量/个	示例
most	69	非常、极端
very	42	格外、很
more	37	较、还要
ish	29	略微、一些
insufficiently	12	轻度、不怎么
pos	162	上涨、增加
neg	421	下滑、损失
inverse	58	不、没有

Preprocessing

Preprocessing

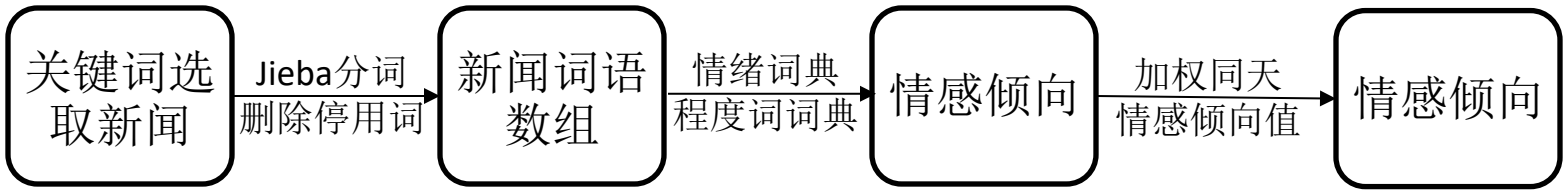
- 加权同天情感倾向值（以复星医药为例）

date	pos	neg	read
2020/7/2	2	0	6073
2020/7/2	9	0	3769
2020/7/2	4	5	6710
2020/7/2	2	0	8631

↓

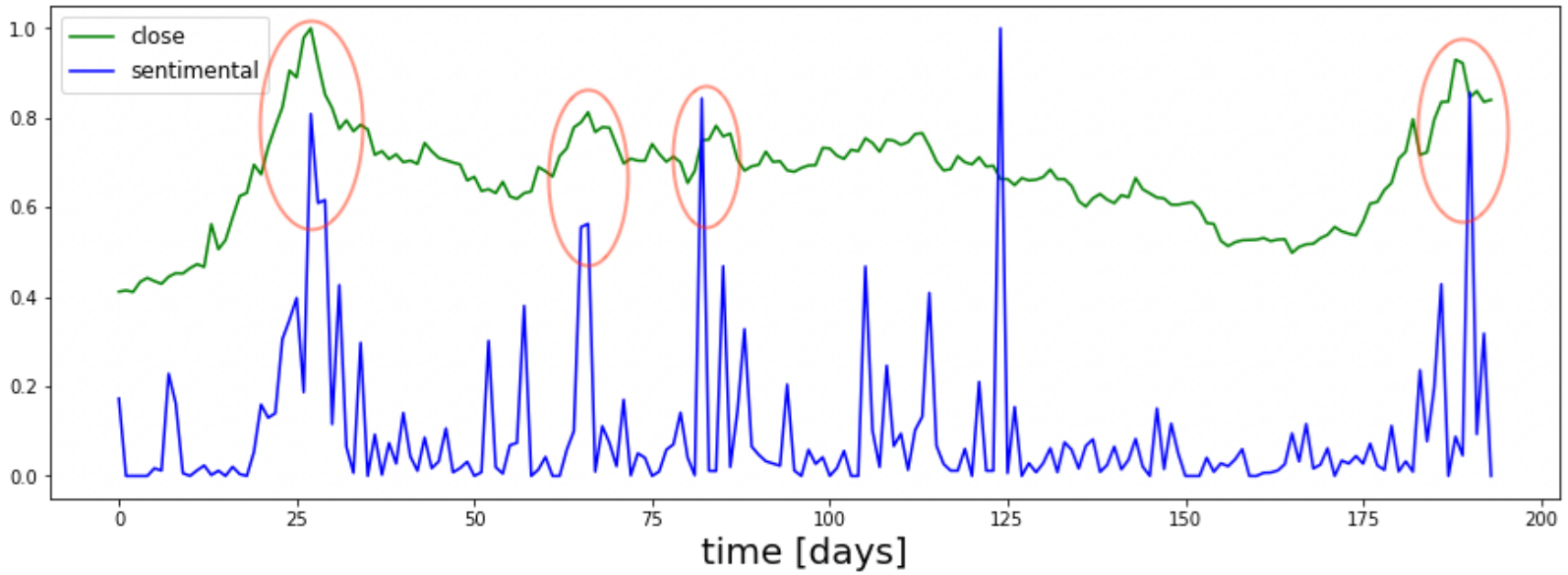
date	pos	neg
2020/7/2	90169	33550

Preprocessing



Preprocessing

- 验证情感分析准确性



Preprocessing

- 股票数据

	date	open	close	high	low	volume	money
0	2020-05-25	30.91	31.12	31.52	30.78	20599861.0	6.412297e+08
1	2020-05-26	31.33	31.33	31.66	31.16	18963313.0	5.943644e+08
2	2020-05-27	31.37	30.80	31.47	30.37	27555024.0	8.486539e+08
3	2020-05-28	30.68	30.11	31.03	29.75	30655938.0	9.267066e+08
4	2020-05-29	30.15	30.54	30.74	30.12	23890232.0	7.284473e+08

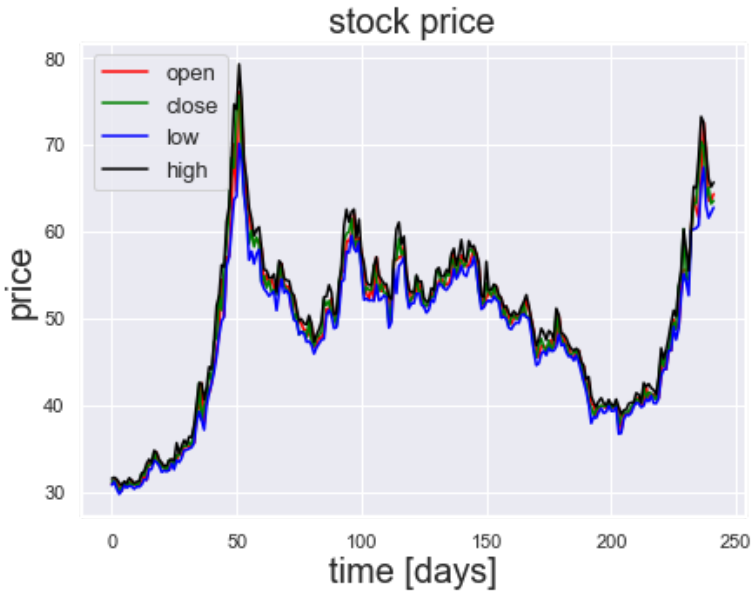
Training

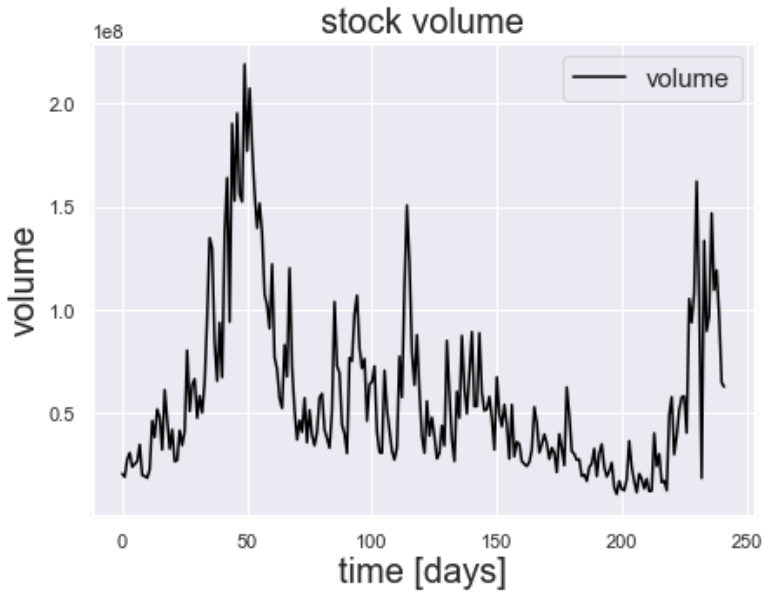
- 股票信息、新闻情感倾向整合

	date	pos	neg	open	close	high	low	volume	money
0	2020-05-25	0.451054	0.025495	30.91	31.12	31.52	30.78	20599861.0	6.412297e+08
1	2020-05-26	0.052651	0.193771	31.33	31.33	31.66	31.16	18963313.0	5.943644e+08
2	2020-05-27	0.239771	0.046422	31.37	30.80	31.47	30.37	27555024.0	8.486539e+08
3	2020-05-28	0.092130	0.100642	30.68	30.11	31.03	29.75	30655938.0	9.267066e+08
4	2020-05-29	0.094521	0.053147	30.15	30.54	30.74	30.12	23890232.0	7.284473e+08

Training

- 合成数据集的可视化分析

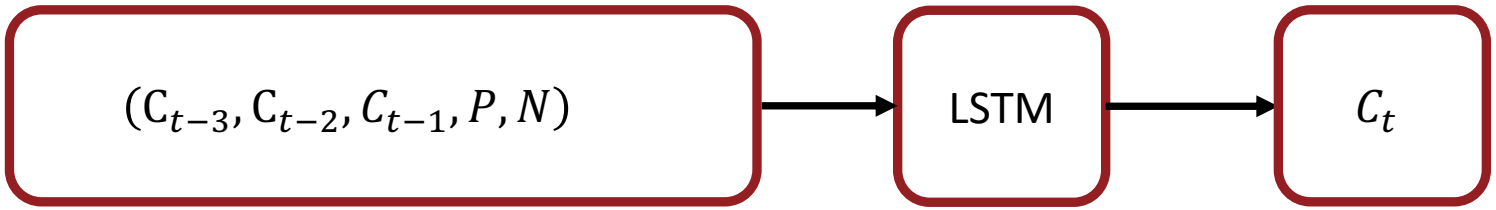




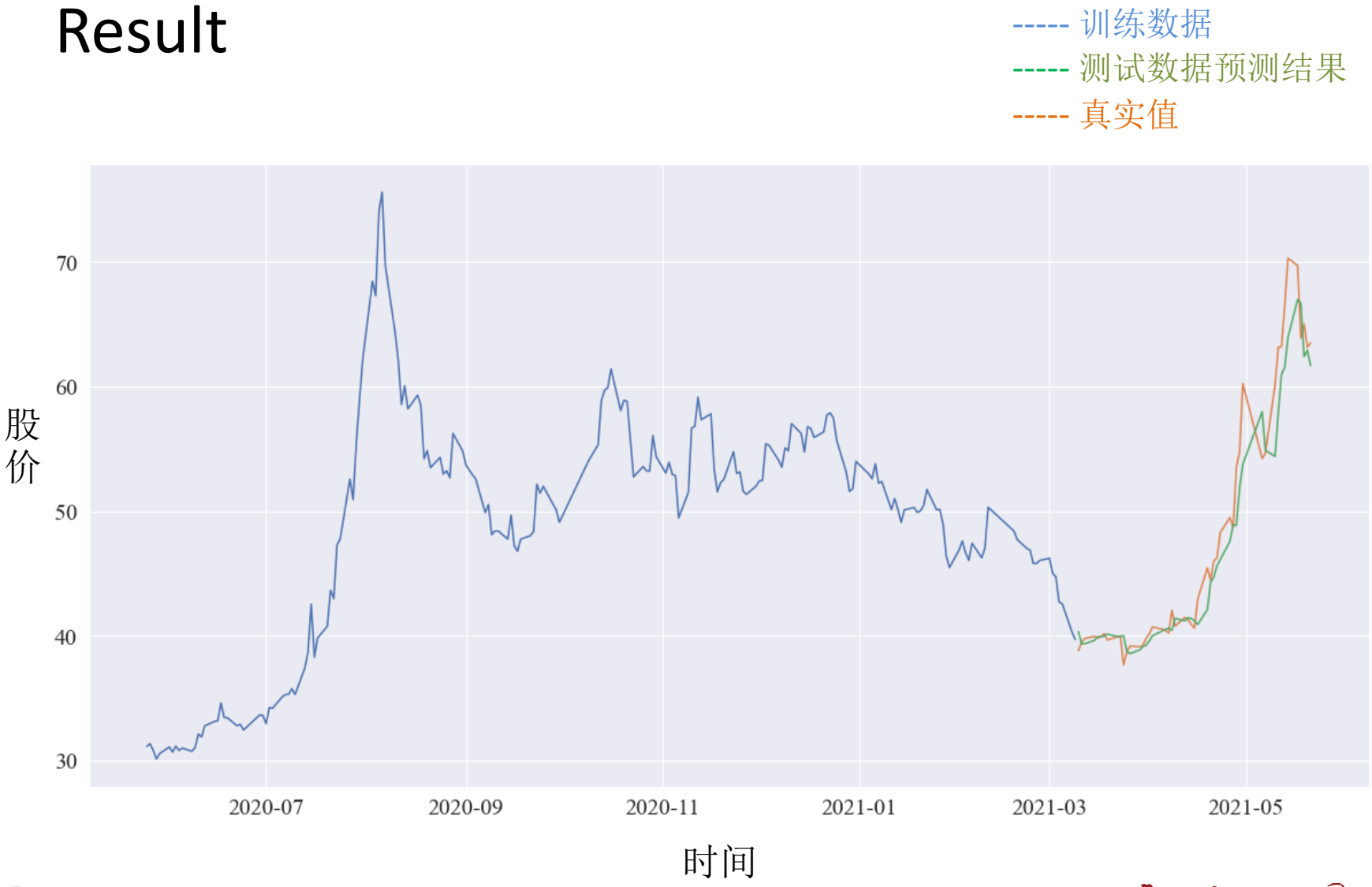
Training

- 收盘价预测

C: 收盘价
P: 正向情感
N: 负向情感



Result


----- 训练数据
----- 测试数据预测结果
----- 真实值
70
60
50
40
30
股价
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
时间

Result

- ACC

$$\text{ACC} = \frac{\text{涨跌情况预测正确的天数}}{\text{总天数}}$$

- MSE

$$\text{MSE} = \frac{1}{n}\sum_{i}(p_i - c_i)^2$$

Result

- 有新闻无股价：删除数据
- 有股价无新闻
  - 删除数据（内积）
  - 用前一天的情感倾向值作为当天的情感值

新闻种类
- 热帖
- 资讯
- 热帖+资讯

预测当天股价
- 前一天的新闻
- 当天新闻
- 不使用新闻

Result

- 股票数据与新闻数据内积的ACC

ACC	新闻类型	用第i-1天新闻	用第i天新闻	不加新闻
复星医药	热点	55.263%	**57.895%**	52.632%
	资讯	**54.348%**	50.000%	47.826%
	热点+资讯	52.174%	**54.348%**	47.826%
沃森生物	热点	45.833%	**53.846%**	51.282%
	资讯	**44.681%**	**44.681%**	**44.681%**
	热点+资讯	**44.681%**	**44.681%**	**44.681%**
恒瑞医药	热点	39.130%	**43.478%**	**43.478%**
	资讯	44.681%	**46.809%**	44.681%
	热点+资讯	44.681%	**48.936%**	44.681%

Result

- 股票数据与新闻数据内积的MSE

MSE	新闻类型	用第i-1天新闻	用第i天新闻	不加新闻
复星医药	热点	**8.011**	8.221	11.356
	资讯	6.859	**6.698**	7.723
	热点+资讯	**6.383**	6.983	7.723
沃森生物	热点	**5.602**	10.002	8.733
	资讯	**4.888**	5.715	5.628
	热点+资讯	**5.104**	5.447	5.628
恒瑞医药	热点	17.308	**14.998**	18.905
	资讯	**5.665**	5.745	5.668
	热点+资讯	5.651	**5.370**	5.668

Result

- 填补为前一天的新闻数据的ACC

ACC	新闻类型	用第i-1天新闻	用第i天新闻	不加新闻
复星医药	热点	41.667%	45.833%	**47.917%**
	资讯	**47.917%**	45.833%	**47.917%**
	热点+资讯	45.833%	45.833%	**47.917%**
沃森生物	热点	46.154%	**47.917%**	**47.917%**
	资讯	41.667%	41.667%	**47.917%**
	热点+资讯	43.750%	43.750%	**47.917%**
恒瑞医药	热点	**50.000%**	43.750%	41.667%
	资讯	**43.750%**	**43.750%**	41.667%
	热点+资讯	43.750%	**45.833%**	41.667%

Result

- 填补为前一天的新闻数据的MSE

MSE	新闻类型	用第i-1天新闻	用第i天新闻	不加新闻
复星医药	热点	5.107	**4.932**	5.826
	资讯	5.029	**4.882**	5.826
	热点+资讯	**4.520**	5.004	5.826
沃森生物	热点	9.905	**5.037**	6.682
	资讯	**5.348**	6.024	6.682
	热点+资讯	5.077	**4.985**	6.682
恒瑞医药	热点	**3.562**	4.121	4.521
	资讯	3.354	**3.192**	4.521
	热点+资讯	3.435	**3.161**	4.521

Issues and Challenges

- 采用的jiaba分词器不够完备，例如不能将双重否定词组（例如“难道不是”）保留，而是分成了“难道”与“不是”，导致最后的情感分析出错。
- 情感分析词库不够全面，有待扩充和修正。
- 根据公司的新闻去预测可能存在的股票涨跌情况，但是事实上可能存在新闻标题是反向预测股票真正的走势，这会给预测结果带来一些偏差。
- 同时也存在有些股市新闻标题中并不包含任何利好或者利空的关键词，对股票的预测带来了一定难度。
- 新闻的渠道挖掘的还不全面，仅考虑了股吧中人们的反应和资讯新闻，短视频等新兴渠道还未涉及。

Conclusion

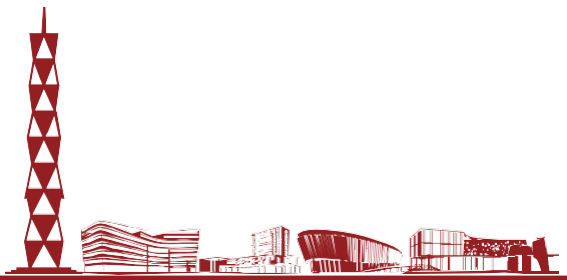
- 新闻作为信息传播的一大媒体，它对股市波动有着强烈的影响
- 点击量越多的新闻影响力也越大，对股票波动的程度影响也越大
- 我们的情感分析结果经验证可知是基本可靠的
- 将股票数据与新闻数据按照时间取内积作为输入数据时，使用当天新闻对预测趋势更有效，而使用前一天的新闻对预测价格更有效
- 使用前一天的新闻数据填补当天缺失的新闻数据时，不使用新闻的ACC普遍上优于使用新闻，使用当天新闻的MSE普遍上优于使用前一天的新闻和不使用新闻

Reference

[1] C.-Y. Yeh, C.-W. Huang, and S.-J. Lee, "A multiple-kernel support vector regression approach for stock market price forecasting," Expert Systems with Applications, vol. 38, no. 3, pp. 2177-2186, 2011.

[2] L. S. Malagrino, N. T. Roman, and A. M. Monteiro, "Forecasting stock market index daily direction: A Bayesia Network approach," Expert Systems with Applications, vol. 105, pp. 11-22, 2018.

[3] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," Expert Systems with Applications, vol. 83, pp. 187- 205, 2017.



Q&A