运用新闻和公司信息预测医疗行业股市波动

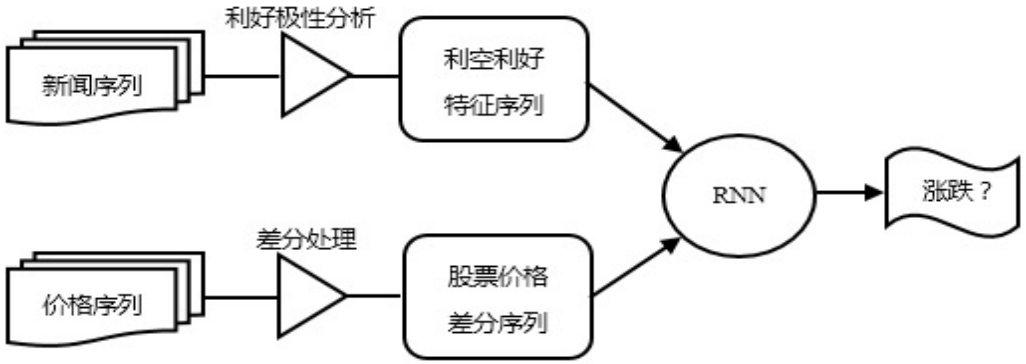
——沈昭正 井文可 钱丽玥

Introduction

- Fama于1965年提出有效市场假说----->股票预测
- 股票相关的新闻信息影响股市
- 疫情之下医药行业澎湃发展

Related Work

- 《股票预测:一种基于新闻特征抽取和循环神经网络的方法》
- 特征提取
  - 选取利好关键词和利空关键词的种子集合
  - 使用最优化方法选出利好与利空的标准关键词集合，并用标准关键词集合计算其他单词的利好极性
  - 构造出新闻文本的利好利空特征
- 股票预测模型
  - 使用循环神经网络



Related Work

- 《股票预测:一种基于新闻特征抽取和循环神经网络的方法》
  - 某只股票会产生一个收盘价格的序列 $p_1, p_2, \ldots p_t$
  - 该股票相关的新闻序列 $d_1, d_2, \ldots d_t$
  - 新闻文本可以看作单词的序列，对于中文文本而言，即新闻内容进行分词后的序列

$$d_t = w_1 w_2 \ldots w_{l_t}$$

$w_i \in V$表示单词表中的一个单词，l_t表示第t个交易日新闻的长度

$$c_t = \begin{cases} 1, & if\ p_t > p_{t-1} \\ 0, & if\ p_t \leq p_{t-1} \end{cases}$$

c_t：*第t个交易日股票价格的涨跌情况*

Related Work

- 《股票预测:一种基于新闻特征抽取和循环神经网络的方法》
  - 假设我们有一个足够大的与股票相关的新闻文档集合：

$$\mathrm{D} = \{doc_1, doc_2, \ldots, doc_N\}$$

  - 每个新闻看作单词序列

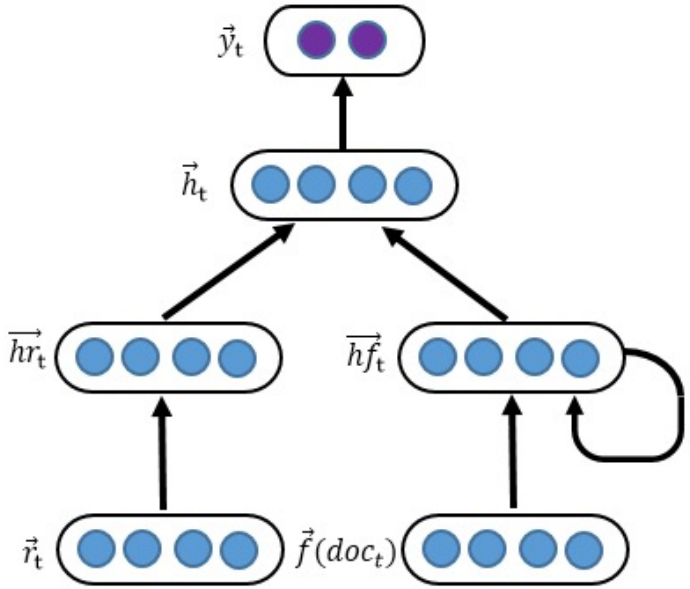
$$doc_i = w_1 w_2 \ldots w_{n_i}$$

  - 单词 w 在文档集中出现的概率记为 $p(w)$，两个单词 w 和 v 之间的点互信息记为 $pmi(w, v)$：

⇨ 最优标准集P*和N*

Related Work

- 《股票预测:一种基于新闻特征抽取和循环神经网络的方法》
  - 输入层1:股价差分序列 $r_1, r_2, \ldots r_t$
  - 输入层2:新闻相关的特征序列



Related Work

- 《基于社交情感数据挖掘的股票市场预测研究》
  - 利用jieba分词工具，统计每个词出现的频率，人工筛选标注出不同情感极性的词语。考虑到中文文字语法的复杂性，需要建立3类词典：积极情感词典、消极情感词典、否定词词典。据此，获得的3个词典中含积极词汇3212个，消极词汇3230个，否定词汇27个。

情感极性	数量/个	示例
积极	3 212	涨、涨停、反弹、买、好等
消极	3 230	跌、垃圾、跑、下跌、砸等
否定词	27	不、没有、没、不要、不能等

Related Work

- 《基于社交情感数据挖掘的股票市场预测研究》
  - 由于股吧评论中股民的情绪不定，无法事先得知股民当下的情绪，因此假设先验概率都为50%，即不影响情感判断的可以舍去。另外，在评论文本中，在分词之后去除与情绪判断无关的如连词、量词等避免影响判断。并且由于网评的特殊性，词与词之间的关系并不大，也就是说，句子 D 可以表示为 $D = \{word_i, \dots, word_n\}$。

$$P(D \mid h) \propto \sum_{j=1}^{n} P(w_j \mid h) = \begin{cases} P(w_j \mid h), \text{if word}_{j-1} \text{ is not nagation word}, \\ P(w_{j-1} + w_j \mid h), \text{otherwise}; \end{cases}$$

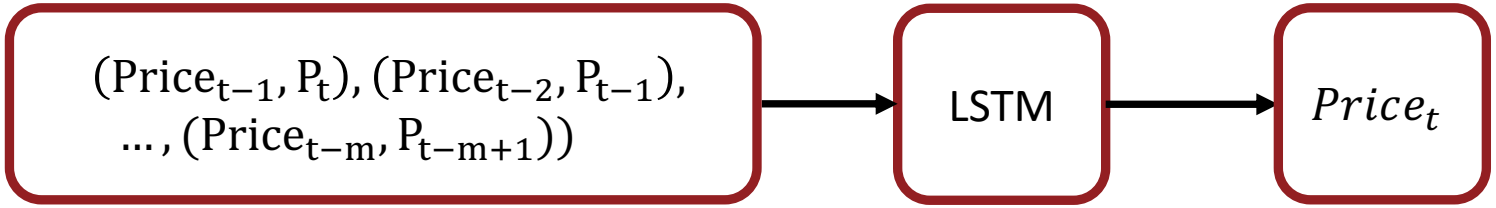
  - 把最终的结果转换为到[0, 1]之间的数，越接近0表示情绪越消极，越接近1表示情绪越积极。

Related Work

- 《Stock Price Prediction Using News Sentiment Analysis》
  - *RNN LSTM with Stock Prices and Textual Polarity:*

$$Polarity_i^c = (+/-)\max(\text{abs}(N_i^c, P_i^c)),$$

$$Polarity^c = \frac{1}{k}\sum_{i=1}^{k} Polarity_i^c,$$



- N_i^c and P_i^c are negative and positive values corresponding to words in the i_{th} of k documents about company c.
- m is the window size of past stock prices.

Aims and Objectives

- 新闻：投资者对金融新闻和日常事件的反应会影响股价
- 点击量：新闻点击量能够反映新闻的热度，与新闻发布的平台相关
- 公司信息：公司的风险信息和对外投资详细信息会对股票价格的走势产生影响

- 通过**新闻文本**的关键信息提取对所选股票的涨跌幅预测
- 考虑**新闻发布平台**对股票波动幅度的影响
- 利用**公司信息**找到股票之间的波动关系

Proposed Methodology

- **股票数据提取**
- 数据来源：新浪财经 https://finance.sina.com.cn

Proposed Methodology

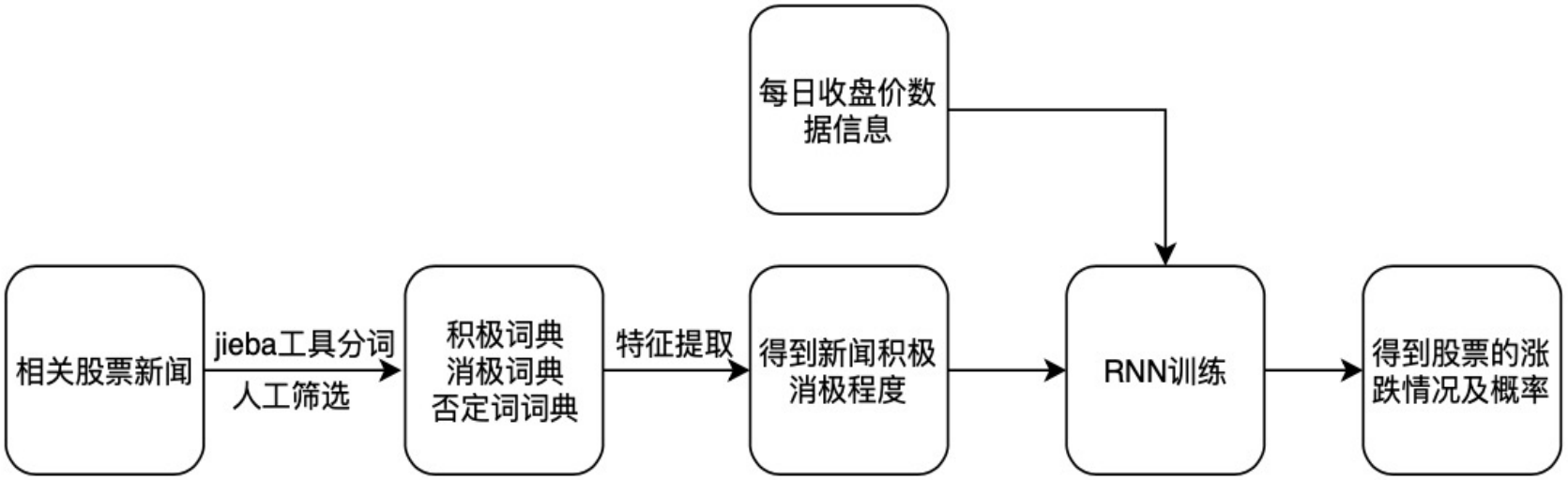
- 财经新闻数据提取

Chinese financial news

- 来源：东方财富和新浪财经；
- 时间：2020年5月-2020年12月；
- 数量：141423条；
- 数据链接：[上科大云盘]

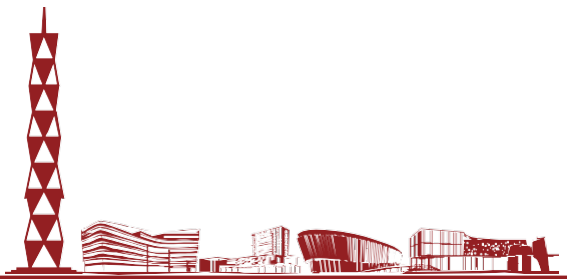
tid	Datetime	Source	Titile	Text
c81b09334d2f38d4e831f549b51998db	2020-08-27 22:22:00	东方财富网	核心产品“集采”失标后 华东医药业绩保持增长	华东医药(000963)核心产品阿卡波糖片在“集采”的失标，曾让市场担忧公司今年的业绩。不过，实际情况却是公司上半年业绩依旧保持逆势增长……

Proposed Methodology



Feasibility

- 数据可行性分析
  - 数据本身的可信度较强，不会存在较大噪声
- 算法可行性分析
  - 使用财经新闻文章作为输入的模型表现很好，而仅根据历史股价预测未来股价的模型会导致较高的百分比误差
  - Saloni Mohan等人利用 MAPE 度量不同方法的效果评测，其中 RNN 在 2.03 和 2.17 之间，然而 ARIMA 和 Facebook Prophet 等经典时间序列模型的得分要低得多，介于 7.39 和 7.98 之间
- 运算能力分析
  - 由于我们需要通过 RNN 进行训练，需要运算能力比较强的电脑



Timetable and Workload Division

Timetable	Week1-Week4		Week5-Week8		Week9-Week10		Week11-Week14	
阅读文献								
讨论实现方法								
爬取、筛选、训练股票新闻数据								
测试								
分析实验结果								
完成报告、ppt								

Q&A