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

Summary

- 新闻情绪分析（NLP）

  ---预测整条新闻情绪

  ---从新闻中提取关键词

  名词+动词，根据点互信息得到标准关键词集合，NLTK库

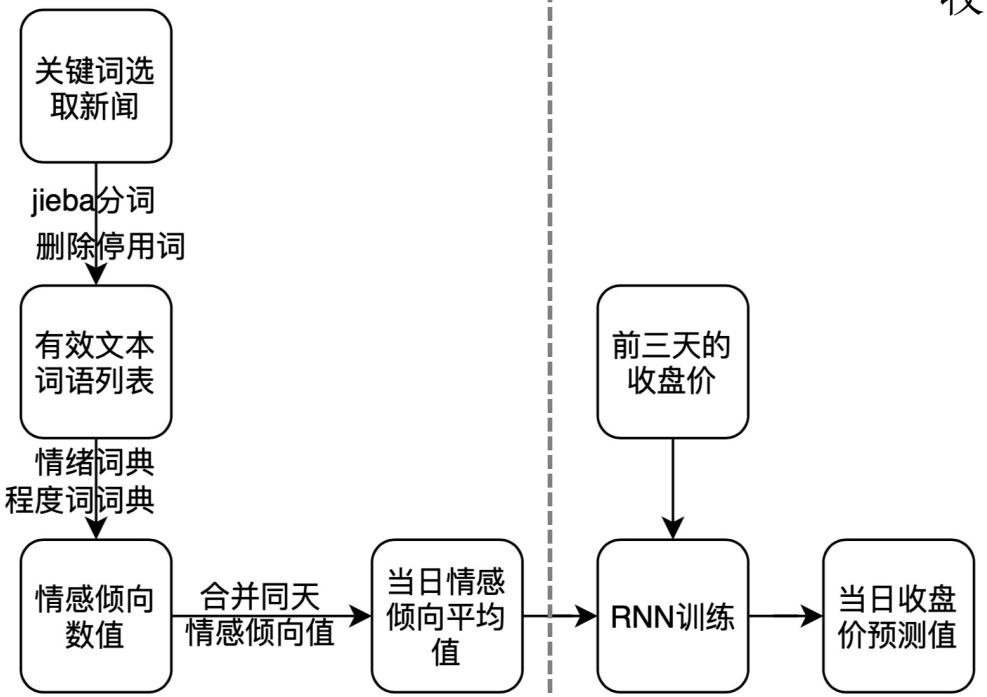
- 股价预测

  ---RNN

  ---离散值挖掘算法

Progress

情感分析



收盘价预测

Progress

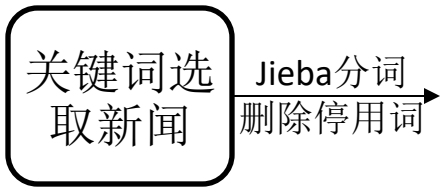
- 新闻数据
  - 来自上科大云盘。2020年5月到2020年9月的来自东方财富与新浪财经的新闻csv文件，共141423条新闻。每条新闻都有五个特征，包括新闻ID、时间、来源、标题和内容，如下图所示。

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……

Progress

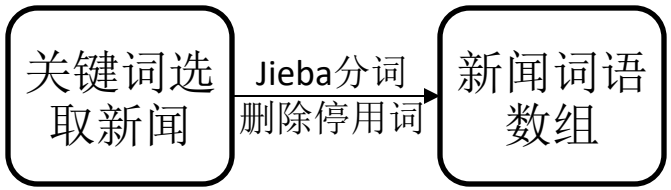
关键词选取新闻

Progress



42
新闻文本

Progress



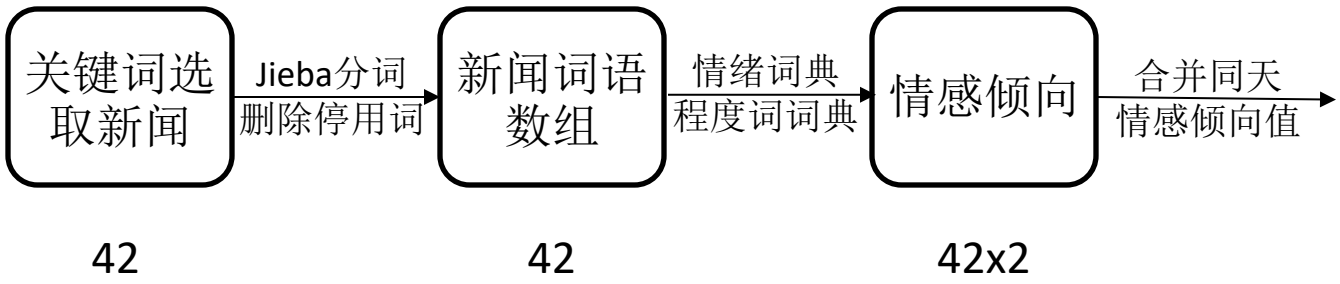
42 42

Progress

- 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、绝等
very	42	多么、格外、很等
more	37	更加、较、还要等
ish	29	略微、一点、一些等
insufficiently	12	不怎么、轻度、不大等
pos	146	上涨、增加、疫苗等
neg	168	损失、下滑、减少等
inverse	58	不、别、没有等

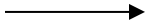
Progress


关键词选取新闻
Jieba分词
删除停用词
新闻词语数组
情绪词典
程度词词典
情感倾向
合并同天
情感倾向值
42
42
42x2

Progress

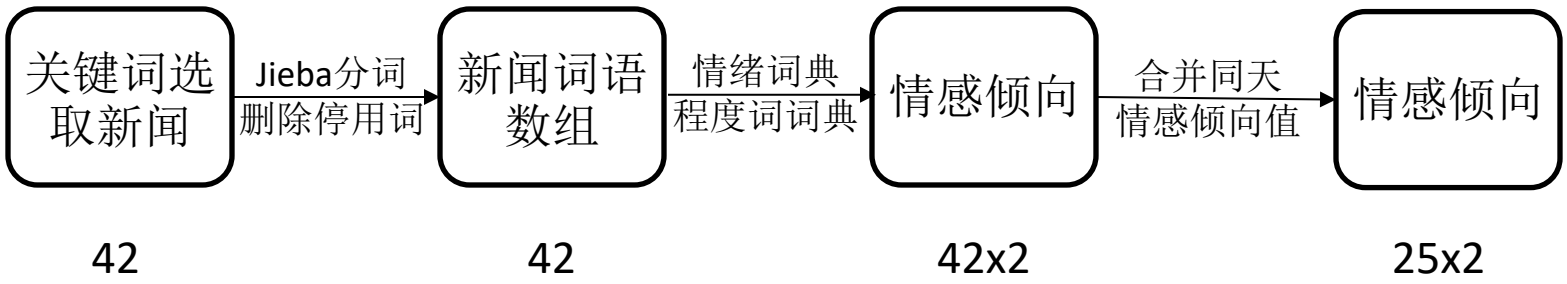
- 合并同天情感倾向值

date	pos	neg
2020-05-14	0	6
2020-05-19	7	12
2020-05-19	20	43
2020-05-28	37	0
2020-06-02	3	9



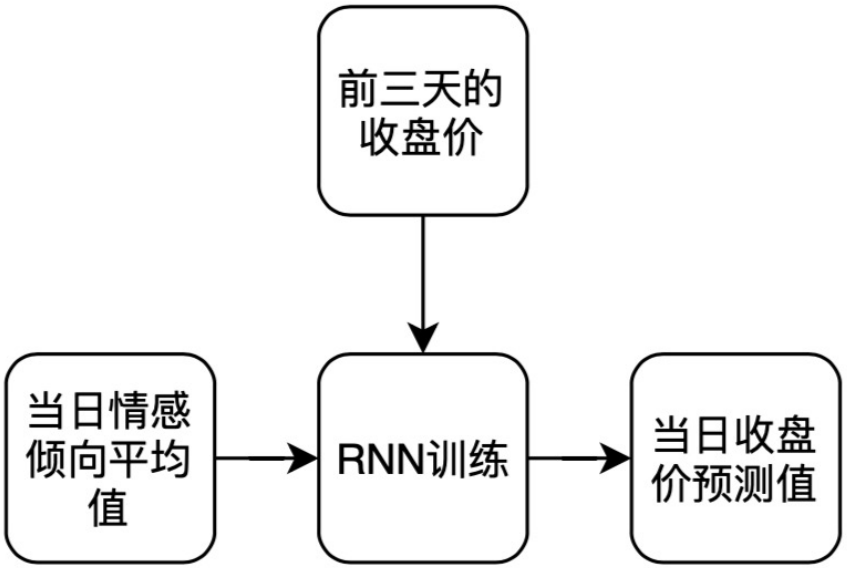
date	pos	neg
2020-05-14	0.0	6.0
2020-05-19	13.5	27.5
2020-05-28	37.0	0.0
2020-06-02	3.0	9.0

Progress


关键词选取新闻
Jieba分词
删除停用词
新闻词语数组
情绪词典
程度词词典
情感倾向
合并同天
情感倾向值
情感倾向
42
42
42x2
25x2

Progress

- 收盘价预测



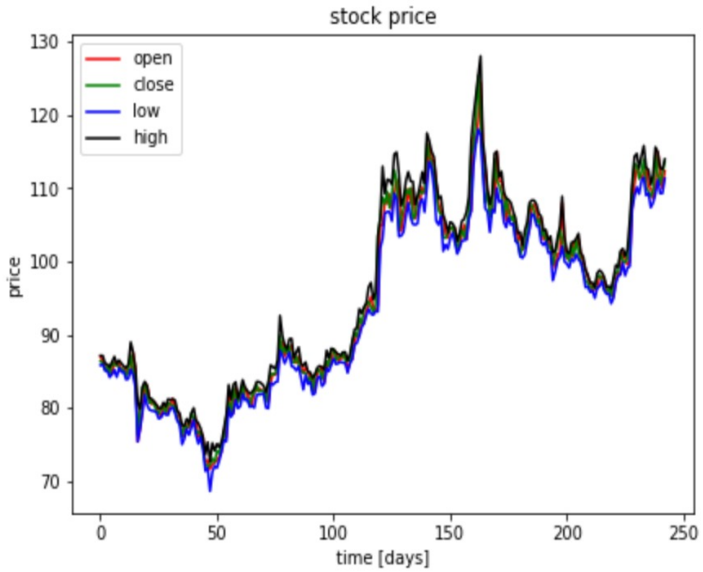
Progress

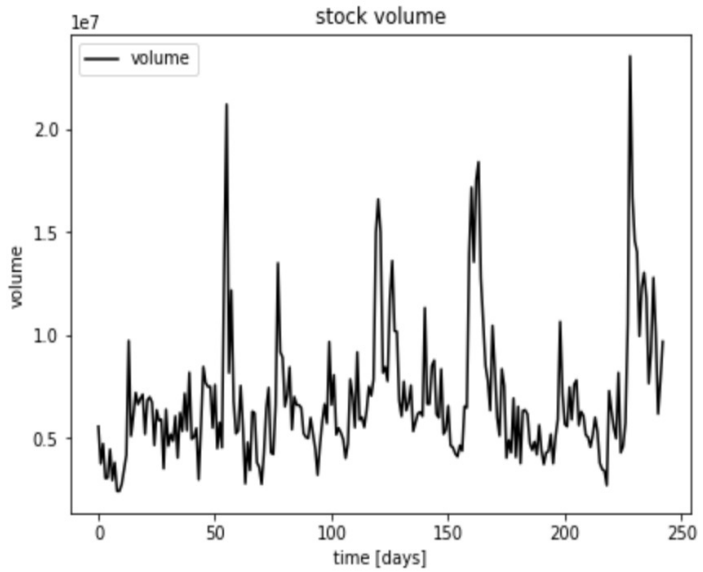
- 股票数据

	date	open	close	high	low	volume	money
0	2020-01-02	87.03	86.37	87.13	85.83	5570437.0	4.804153e+08
1	2020-01-03	86.41	86.16	87.13	85.95	3771176.0	3.262361e+08
2	2020-01-06	85.55	85.16	86.07	85.10	4734226.0	4.046714e+08
3	2020-01-07	85.04	85.64	85.95	85.03	3061176.0	2.618692e+08
4	2020-01-08	85.58	84.64	85.58	84.19	3085713.0	2.617135e+08

Progress

- 股票数据





Progress

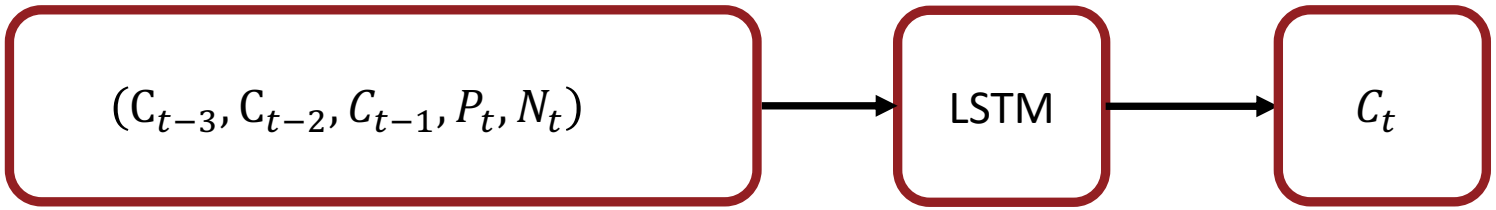
- 股票信息、新闻情感倾向整合

	date	pos	neg	open	close	high	low	volume	money
0	2020-05-06	39.312500	27.687500	87.03	88.29	88.67	86.08	8919331.0	7.801134e+08
1	2020-05-07	27.322581	34.516129	87.83	87.47	87.99	86.84	6539936.0	5.703585e+08
2	2020-05-08	118.615385	10.923077	87.86	88.47	89.41	87.58	7096194.0	6.294167e+08
3	2020-05-11	175.387097	13.959677	88.58	86.26	89.45	85.68	8450599.0	7.337356e+08

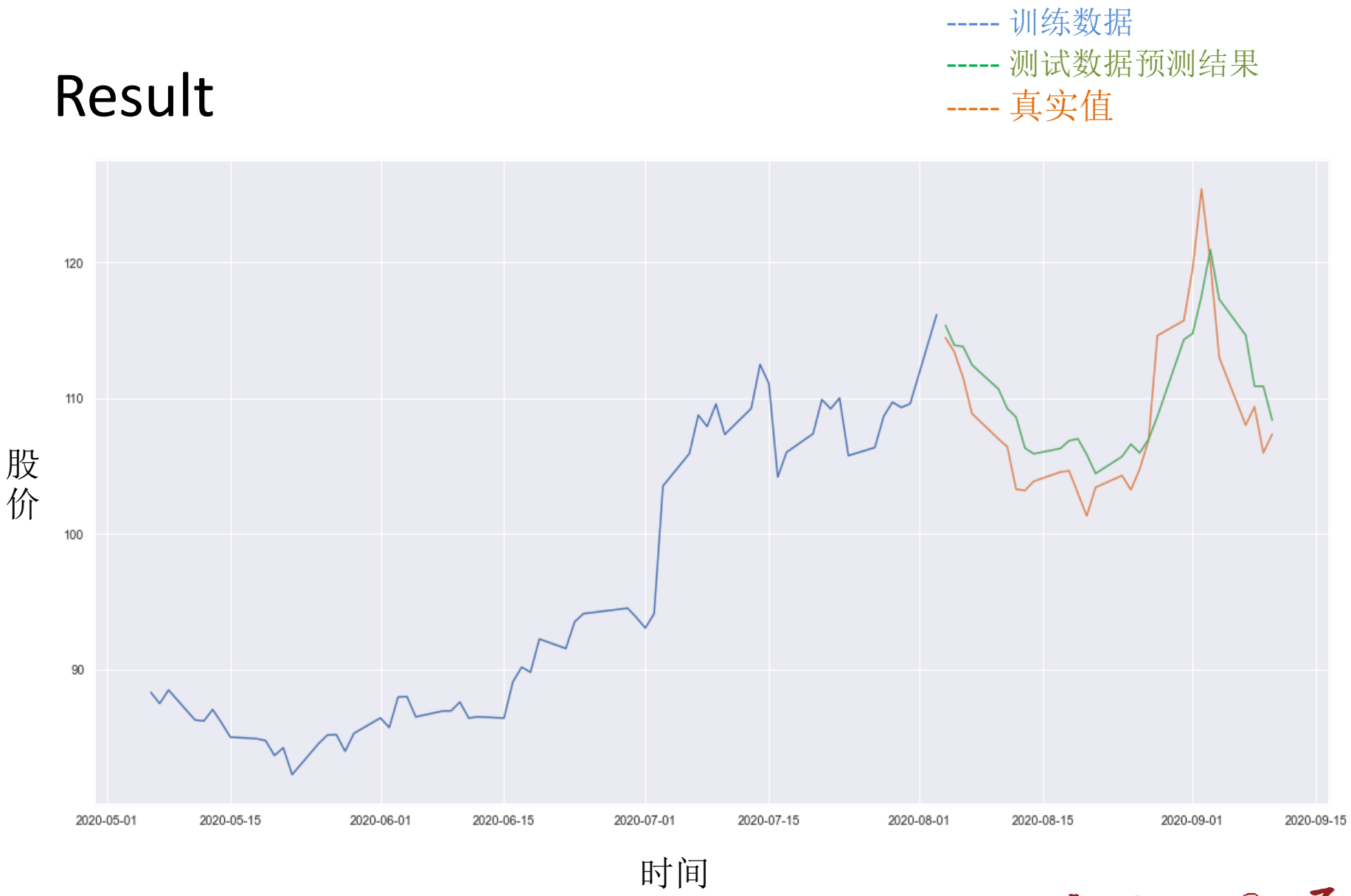
Progress

- 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
股价
120
110
100
90
2020-05-01
2020-05-15
2020-06-01
2020-06-15
2020-07-01
2020-07-15
2020-08-01
2020-08-15
2020-09-01
2020-09-15
时间

Result

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

- MAPE：平均绝对百分误差

	无新闻	有新闻
以岭药业	0.02195	0.01763
复星医药	0.19965	0.10607
恒瑞医药	0.01953	0.01683
片仔癀	0.03936	0.03703
云南白药	0.02769	0.02045

Recall 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
- 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**新闻点击量**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

Issues and Challenges

- 财经新闻数据
  - 财经新闻数据只包括2020年5月3日至9月10日的新闻。每日针对单只股票的新闻数据太少，无法对股票进行很好的预测。
  - 在大数据、智能化的时代，股民已经不单单从财经新闻这一个渠道获得信息，手机短视频等也在很大程度上影响了股民的选择，我们考虑从其他渠道获取更全面的新闻信息。
  - 由于财经新闻数据没有新闻的点击量的信息，我们只是将每日行业相关的所有新闻每个情感相加，而非根据点击数量加权求和。
- 情感词典
  - 我们选用的词典并不是针对医药行业的词典，如果考虑更有针对性的词典我们的准确率可能更高。

Timetable and Workload Division

任务列表	week1-4	week5-8	week9-10	week11-13	week13-14
阅读文献	井文可、沈昭正、钱丽玥				
讨论实现方法	井文可、沈昭正、钱丽玥				
爬取股票数据预处理		沈昭正			
新闻数据情感倾向		井文可、钱丽玥			
训练、测试			井文可、沈昭正、钱丽玥		
修改方法，进行实验				井文可、沈昭正、钱丽玥	
完成报告、ppt					井文可、沈昭正、钱丽玥

Reference

- E. Guresen, G. Kayakutlu, and T. U. Daim, "Using artificial neural network models in stock market index prediction," Expert Systems with Applications, vol. 38, no. 8, pp. 10389-10397, 2011.Conference Short Name:WOODSTOCK'18
- D. Shah, H. Isah and F. Zulkernine, "Predicting the Effects of News Sentiments on the Stock Market," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4705-4708, doi: 10.1109/BigData.2018.8621884.
- 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.
- ChenCC, HuangHH, ChenHH.NTUSD-Fin:AMarket Sentiment Dictionary for Financial Social Media Data Applications.

Q&A