MM5425 商业分析

WEEK 10 LECTURE – TEXT MINING

WK10 目录 Contents

- 情感分析： Sentiment Analysis
- 处理文本：自然语言处理
- 推荐： Recommendation

情感分析
Sentiment Analysis

FEELINGS, EMOTIONS

情感分析 Sentiment Analysis

情感是指感受、观点、情绪、喜欢/不喜欢、好/坏。

情感分析是一项自然语言处理和信息抽取任务，旨在通过分析大量文档，获取作者在正面或负面评论、问题和请求中表达的感受。

最简单的任务：

- 判断这段文本的态度是积极还是消极？

更复杂的任务：

- 将这段文本的态度从1到5进行评分

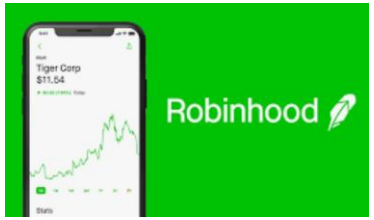
高级任务：

- 检测目标、来源或复杂的态度类型

应用 Applications

- 企业利用情感分析进行品牌分析、新产品认知、产品和服务的基准比较。
- 例子 Examples:
  - 电影：这条评论是正面的还是负面的？
  - 产品：人们对新款 iPhone 有什么看法？
  - 公众情绪：消费者信心如何？绝望情绪是否在增加？
  - 政治：人们对这个候选人或议题有什么看法？
  - 预测：根据情感预测选举结果或市场趋势

商业案例：Robinhood 对 GameStop 事件的回应



Robinhood：

Robinhood 由两位斯坦福大学本科生 **Vlad Tenev** 和 **Baiju Bhatt** 于 **2013** 年创立。

公司名称来源于其使命：“为所有人提供进入金融市场的机会，而不仅仅是富人。”（~维基百科）

公司于 **2021** 年 **7** 月 **29** 日在纳斯达克上市。

截至 **2022** 年，**Robinhood** 拥有 **2280** 万个有资金账户和 **1590** 万月活跃用户。

（~维基百科）

https://techcrunch.com/2022/06/27/robinhood-report-meme-stock-gamestop/

GameStop 做空逼仓

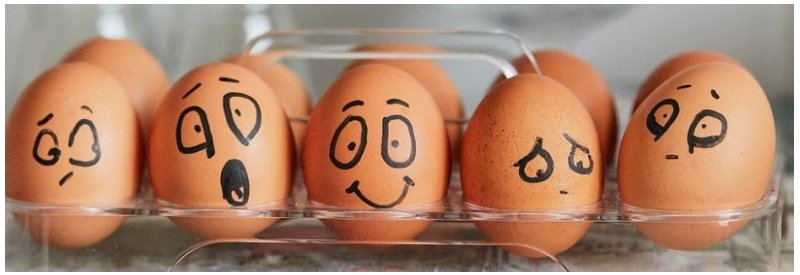
从2021年1月13日开始，GameStop的股价出现了突然且剧烈的上涨，随之而来的是波动性的增加。

据报道，这一上涨主要由使用Robinhood金融平台的散户投资者推动，他们通过社交媒体，特别是Reddit上的WallStreetBets论坛组织行动。

Robinhood于2021年1月28日至29日对GameStop实施了交易限制，禁止用户开设新的多头仓位。这一举措引发了客户和媒体的强烈愤怒，许多政界人士也对此表示愤慨。

https://www.cato.org/cato-journal/fall-2021/gamestop-episode-what-happened-what-does-it-mean#background-the-rise-of-low-cost-trading-platforms

更多的挑战



- 2021年2月8日，Robinhood被一名20岁交易者的家属起诉，该交易者因误以为自己亏损了76万美元而自杀。
- 诉状称，该交易者曾三次尝试联系Robinhood客服，寻求关于巨额负债的帮助，但据投诉所述，他收到的都是自动回复。
- 据报道，家属表示Robinhood应用程序针对年轻人，并鼓励他们参与高风险交易。
- 2021年2月18日，Robinhood首席执行官Vlad Tenev在国会听证会上就公司在GameStop一月逼仓事件中的角色向家属道歉。
- 2021年7月，此案以双方达成和解而撤诉。

https://www.cnbc.com/2021/02/08/robinhood-sued-by-family-of-alex-kearns-20-year-old-trader-who-killed-himself-.html

利用推特进行情感分析

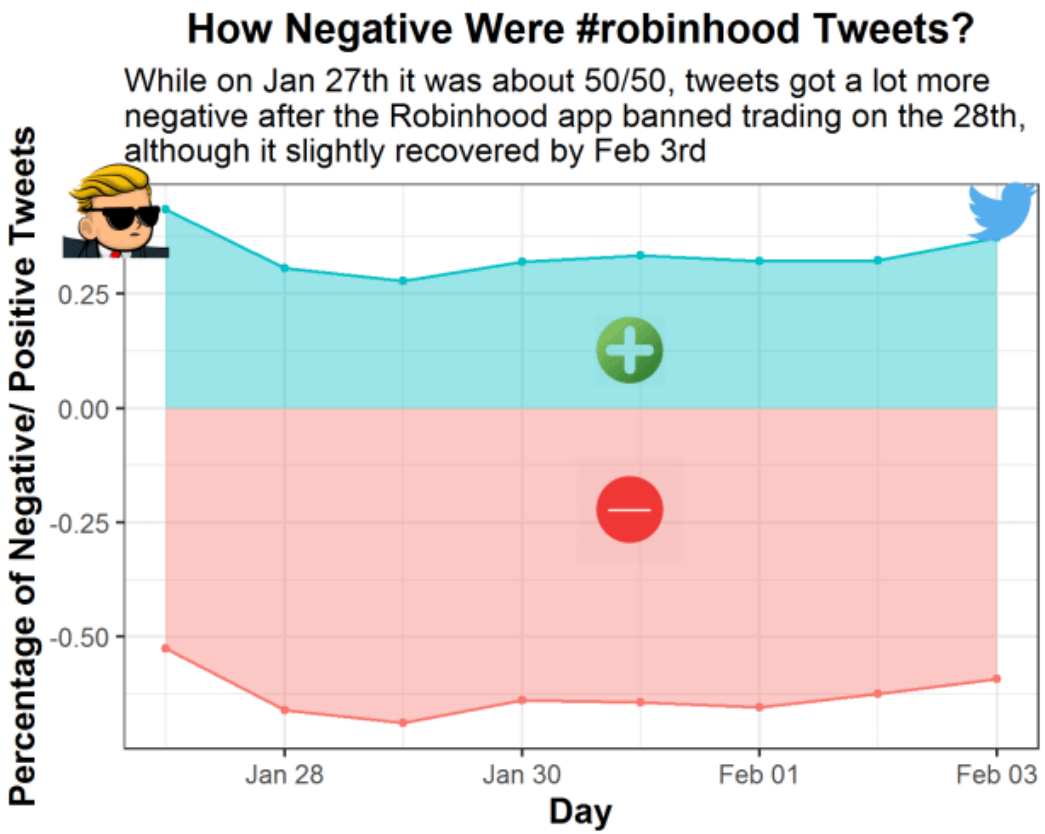
从Twitter上收集了提及“Robinhood”的推文数据集，分为以下三个时间段：

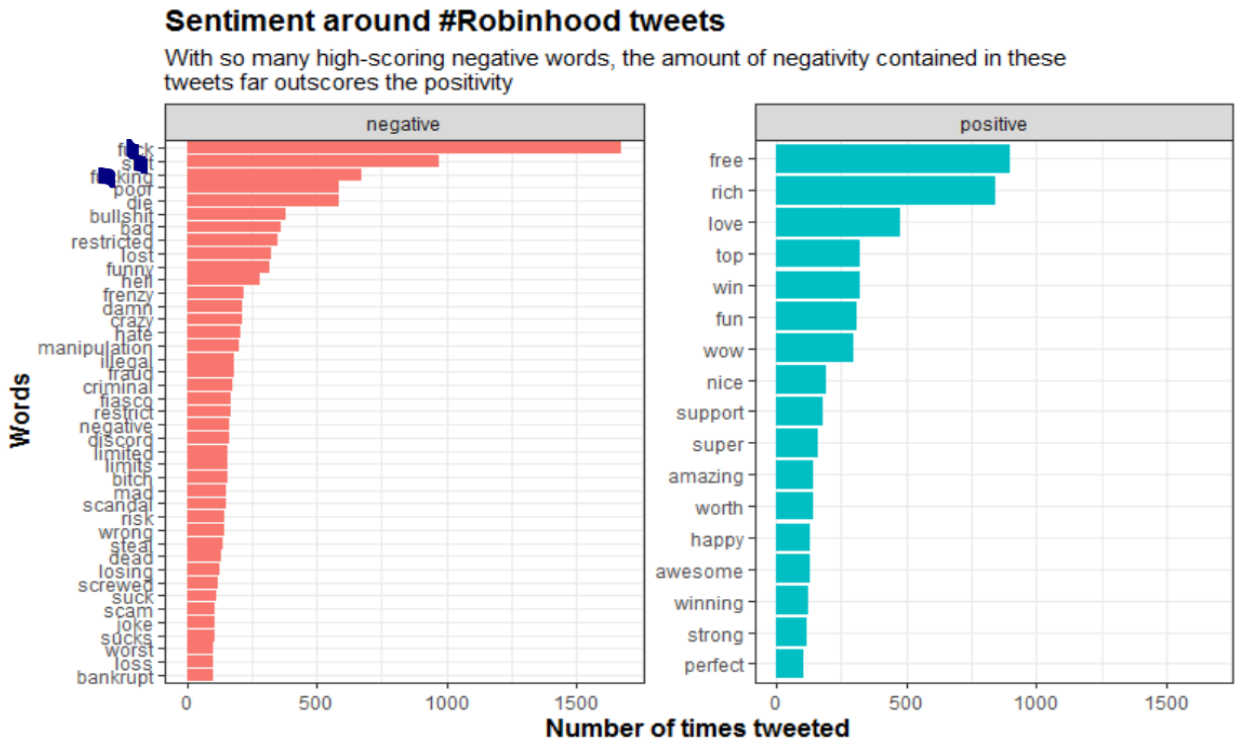
1. 事件发生前
2. GameStop做空逼仓高峰期间（即2021年1月22日至2月1日）
3. 事件发生后

需要解决的问题

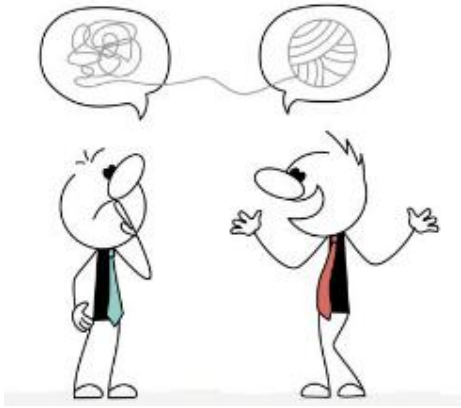
- - 这是否只是一个可以通过精心策划的广告和公关活动来应对的问题？
- - 这场危机是否真的有可能侵蚀用户对Robinhood的信任，并促使他们转向竞争对手？
- - 是否需要对应用进行改进，以减少新用户的不当使用？
- - 是否应该减少应用中的“游戏化”元素？

这些问题都需要Robinhood认真考虑和应对，以维护其品牌形象和用户基础。





https://www.r-bloggers.com/2021/04/analyzing-brand-sentiment-with-robinhood-gamestop-r/



自然语言处理 (NLP)

UNSTRUCTURED DATA

处理文本 Dealing with Text

自然语言处理（NLP）是计算机科学、人工智能和语言学的一个领域，关注计算机与人类语言之间的交互。

- 它能够处理大量文本，分析和理解“自然”语言。
- NLP的目标是识别单词、句子、文本和对话的结构与含义。

语言处理的主要方面包括：

- 句法（Syntax）：分析语言的结构和语法规则。
- 语义（Semantics）：理解词语和句子的意义。
- 语篇分析（Discourse Analysis）：理解更大范围的语言单元，如段落、对话和上下文关系。

NLP在商业中的一些应用

文本分析：

- NLP可用于分析文本的情感、关键词、实体等。
- 监控社交媒体提及，及时处理负面评论。
- 评估客户对市场活动或产品发布的反应。

聊天机器人：

- NLP为聊天机器人的对话界面提供支持，使其能够理解并回应人类语言。

语音识别：

- NLP应用于Siri、小爱同学和Alexa等语音助手，理解并响应语音指令。

机器翻译：

- NLP实现了不同语言之间的自动翻译。

文本分析很困难。

- 文本是“非结构化”的
  - 语言结构是为人类交流设计的，而不是为计算机设计的。
- 有时词序很重要
- 文本数据可能很杂乱
  - 人们写作时常常不遵循语法规则，拼写错误，缩写不规律，标点随意。存在同义词、同形异义词、缩写等问题。
- 上下文很重要
  - 例如：“I ran to the store because we ran out of milk.”（我跑去商店，因为我们家牛奶用完了。）

大规模带来的挑战

- 圣经（钦定版）：约70万词
- 新闻稿合集：超过5亿词
- 维基百科（英文）：29亿词
- 网络：数十亿词

由于词语歧义带来的挑战

"The chicken is ready to eat."

She saw the man with the telescope.

文本的预处理

应执行以下步骤：

- 统一大小写

所有词语都转换为小写。

- 词干提取

去除词语的后缀，例如将名词复数形式转换为单数形式。

- 去除停用词

停用词是指英语（或其他语言）中非常常见的词。

通常会去除如the、and、of、on等词。

词频 Term Frequency

词频（Term Frequency）：

- 词频指的是在文档中某个词出现的次数（而不仅仅是出现与否），用来区分一个词在文档中出现的频繁程度。

归一化词频（Normalized Frequency）：

- 不同文档长度不同，词语出现的频率也不同。
- 因此，原始的词频通常需要进行归一化处理，使得不同长度的文档或不同频率的词可以进行比较。

常见的归一化方法包括：
- 用词频除以文档总词数（即相对频率）。
- 使用TF-IDF等加权方法，进一步考虑词语在整个语料库中的重要性。

词云 Word Cloud

- 词云是一种词语的可视化表示方式，其中每个词语的大小和粗细反映了它在特定文本中的出现频率或重要性。
- 词云常用于对大型文档（如报告和演讲）进行摘要、突出某一主题，或将表格中的数据进行可视化展示。

常用文本分析工具

Python packages:

- NLTK (Natural Language Tookkit)
- spaCy
  - Entity recognition, text classification, dependency parsing
- TextBlob
  - Noun phrase extraction, sentiment analysis, translation
- Gensim
  - Topic modelling and document similarity analysis

Websites/apps:

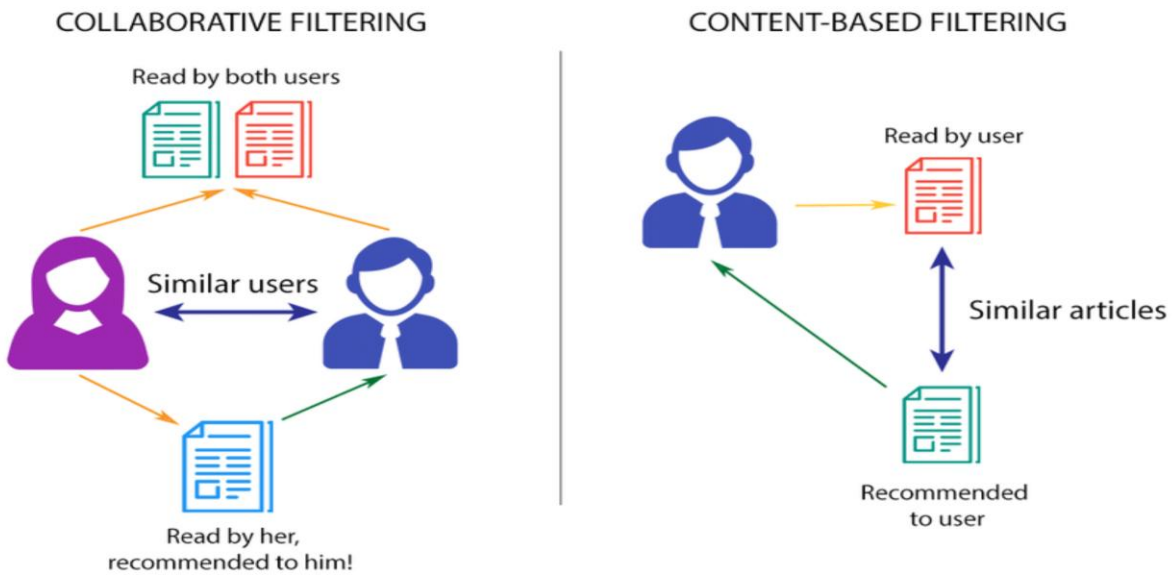
- **Genai**.polyu.edu.hk
- **Copilot**: copilot.Microsoft.com
- **Galaxy AI** Text Analyzer: https://galaxy.ai/ai-text-analyzer
- **Poe** or Poe.com
- **DeepSeek**

推荐Recommendations

WIN WITH OPTIMAL RESULTS

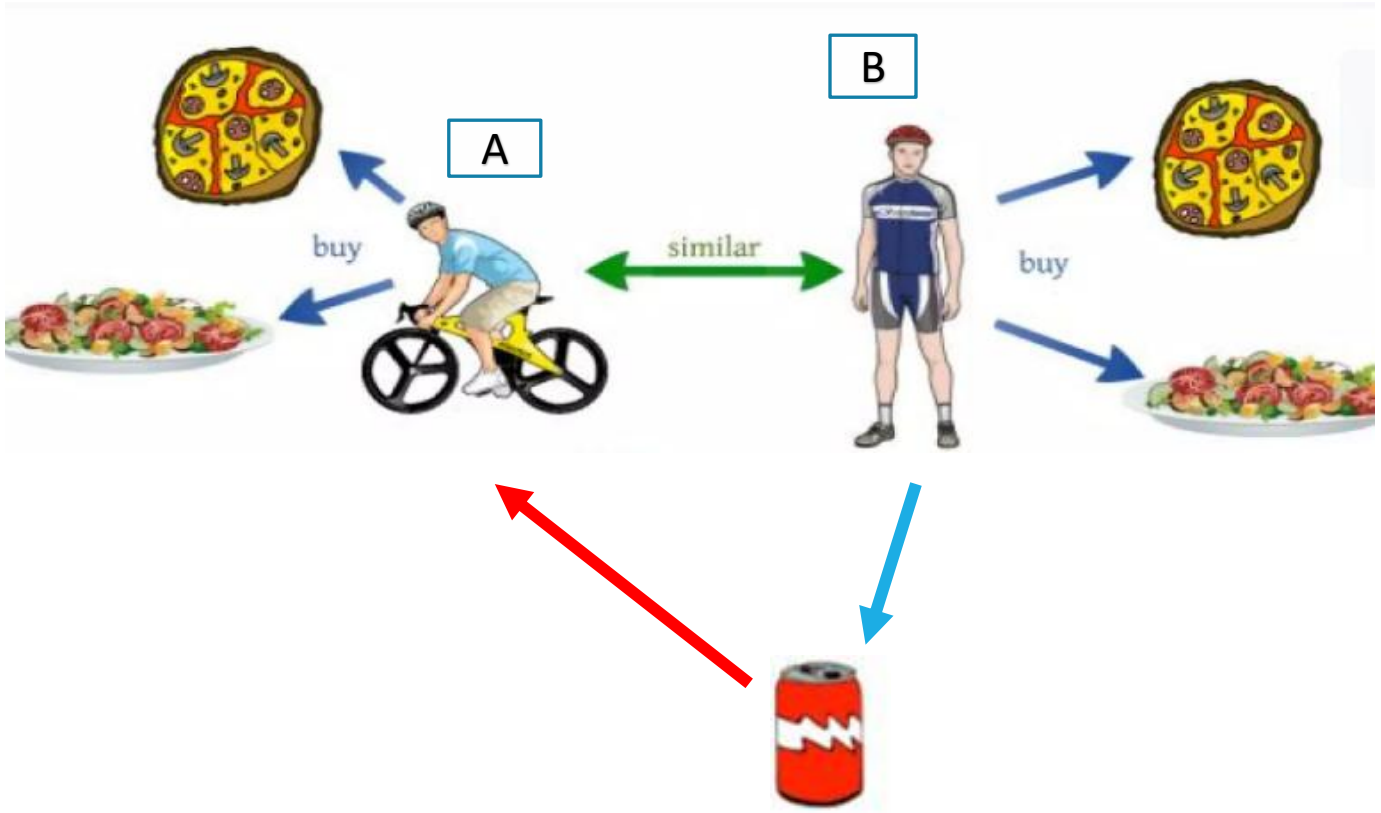
推荐的种类Types of Recommendations

1. 基于流行度的推荐引擎
2. 基于内容的推荐引擎
3. 基于协同过滤的推荐引擎



Source: https://www.themarketingtechnologist.co/building-a-recommendation-engine-for-geeksetting-up-the-prerequisites-13/

基于用户的协同过滤

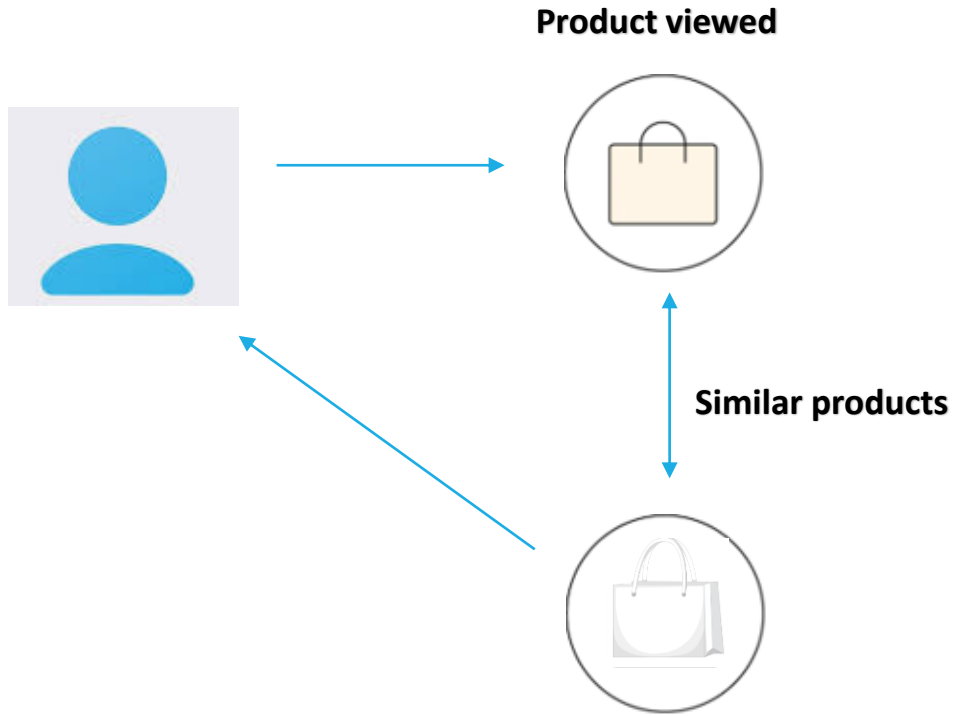


皮尔逊相关系数

欧氏距离

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

内容过滤



	Item 1	Item 2	Item 3	Item 4
User A	1			1
User B		1	1	
User C	1			1
User D		1	1	
User E	1	1		1