Data Science & AI for Economists

Lecture 12: Text Analysis(II)

Zhaopeng Qu
Business School, Nanjing University
December 11 2025

Roadmap

Today's Agenda

Part I: Basic Concepts

- What is Text Analysis?
- Why Text Analysis for Economists?
- Applications in Economics
- Basic Terminology
- Text Preprocessing
- Chinese Text Analysis Challenges
- Bag of Words & DTM
- TF-IDF

Part II: Advanced Topics

- Sentiment Analysis
- Topic Modeling (LDA)
- Text Classification
- Word Embeddings
- Information Extraction

Text Similarity and Semantic Analysis

Part I: Four Paradigms of Text Vectorization

Why Convert Text to Numbers?

Core Question

> How to transform text into numbers for computers to understand?

Four Paradigms of Method Evolution

Paradigm 1: Statistical Representation

Core Idea: Count Word Frequencies

Bag of Words (BoW) Model

将文档视为词的"袋子"，忽略词序，只统计出现次数

Example:

- Doc1: "央行提高利率抑制通胀"
- Doc2: "通胀压力促使央行加息"

文档	央行	提高	利率	抑制	通胀	压力	促使	加息
Doc1	1	1	1	1	1	0	0	0
Doc2	1	0	0	0	1	1	1	1

Paradigm 1: Statistical Representation

TF-IDF: Highlight Important Words

- 假设文档总数为1000篇，平均每篇文档总词数为5000个
- TF = 词在文档中的出现次数 / 文档总词数(5000)
- IDF = log(总文档数1000 / 包含该词的文档数)
- TF-IDF = TF × IDF

词	出现文档数(共1000篇)	在单篇文档中出现次数	TF(词频)	IDF(逆文档频率)	TF-IDF	含义
的	1000	500	500/5000=0.10	log(1000/1000)=0.00	0.10×0.00=0.000	无信息量
经济	800	250	250/5000=0.05	log(1000/800)=0.22	0.05×0.22=0.011	太常见
量化宽松	15	100	100/5000=0.02	log(1000/15)=1.82	0.02×1.82=0.036	高区分度

Paradigm 1: Statistical Representation

Disadvantages

- 忽略词序："狗咬人"="人咬狗"
- 同义词无法区分："汽车"≠"轿车"
- 高维稀疏（维度=词汇量）：TF-IDF矩阵通常非常稀疏
- 无法处理一词多义：TF-IDF无法区分一词多义

Advantages

- 简单直观：易于理解
- 计算快速：计算速度快
- 可解释性强：易于解释

Applications

- 经济学期刊的主流方法：可用于文本分类、文本聚类、文本相似度计算等。

Paradigm 2: Static Word Embeddings (Word2Vec)

Core Idea: Distributional Hypothesis

> "一个词的含义由它的上下文决定" — J.R. Firth, 1957

- 所以我们可以让经常出现在**相似上下文中的词，拥有相似的向量**，这就是Word2Vec的核心思想。

Example

> "The Fed raised interest rates..."
>
> "The central bank increased borrowing costs..."

- 下列词组可能会拥有**相似的向量**:
  - "The Fed"和"The central bank"
  - "raised"和"increased"
  - "interest rates"和"borrowing costs"

Paradigm 2: Static Word Embeddings (Word2Vec)

- Then the vectors of "The Fed" and "The central bank" are:

> "The Fed" → [0.12, -0.45, 0.78, 0.33, -0.21, ..., 0.56] (300个连续值)
>
> "The central bank" → [0.15, -0.41, 0.82, 0.29, -0.18, ..., 0.51] (300个连续值)

- 向量的维度通常为50-500维，通常远小于词汇量，最常用的维度为300维。
- Question: How to get the actual values of vectors?
- Answer: Use ML algorithms normally Neural Network to get the values of vectors.

- CBOW（连续词袋）：用周围词预测中心词

> 输入: ["The", "central", "_", "raised", "rates"]
>
> 目标: 预测 "bank"

- Skip-gram：用中心词预测周围词

> 输入: "bank"
>
> 目标: 预测 ["The", "central", "raised", "rates"]

Paradigm 2: Static Word Embeddings (Word2Vec)

Pre-trained Chinese Resources

资源	语言	说明
Google News	英文	300万词，Word2Vec经典模型
腾讯AI Lab	中文	800万词，推荐
哈工大HIT词向量	中文	哈工大贡献的中文词向量
百度百科词向量	中文	基于百科大规模训练

- **核心思想**：无需从零训练，直接加载已训练好的词向量。

Paradigm 3: Contextual Dynamic Embeddings

Limitations of Word2Vec

- In Word2Vec,

> "bank" → [0.12, -0.45, 0.78, ...] ← 永远是这个向量

- In reality,

> "I deposited money in the bank" → bank = 银行
>
> "I sat on the river bank" → bank = 河岸
>
> "You can bank on me" → bank = 依靠

Paradigm 3: Contextual Dynamic Embeddings

BERT's Breakthrough:

- 情景动态嵌入：Same Word, Different Context = Different Vector
- 核心机制：Transformer的自注意力机制(Self-Attention Mechanism)
  - BERT在生成某个词的向量时，会同时"看"整个句子，让每个词根据上下文调整自己的表示

> "The central bank raised interest rates"

- When processing "bank", BERT will ask:
  - "central" is nearby → "central" might be the central bank
  - "raised" is after → the action is "raising"
  - "rates" is after → it involves interest rates
- Comprehensive judgment: "central" means "central bank"
- Then generate the vector of "bank" by the context of the sentence.

Paradigm 4: Generative Large Language Models

Paradigm Shift: From "Training Models" to "Writing Prompts"

传统流程（BERT时代）	LLM流程
收集数据	设计Prompt
人工标注	API调用
训练模型	获取结果
预测	
需要：编程、GPU、标注数据、数周	需要：清晰描述、几分钟

- **LLM让文本分析重新民主化**：不需要深度学习专业知识，不需要编程，不需要GPU，不需要标注数据，不需要数周，只需要清晰描述任务，即可完成分类。

Core Capabilities of LLM Text Analysis

1. Zero-shot Classification

任务：判断央行声明是鸽派还是鹰派

- 传统方法：需要数千条人工标注的训练数据
- LLM方法：直接描述任务，即可完成分类

2. Complex Information Extraction

任务：从年报中提取风险因素、管理层态度、关键指标

- 传统方法：设计复杂的正则表达式和NER模型
- LLM方法：用自然语言描述需要提取的信息

3. Fine-grained Sentiment Analysis

- 识别讽刺、反语
- 理解隐含立场
- 区分事实 vs 观点

4. Topic Classification

- 自动发现文档的潜在主题
- 无需预定义主题
- 可解释性强

Paradigm 4: Generative Large Language Models

LLM vs Traditional Methods

任务	传统方法	LLM方法	优势
情感分析	训练分类器	Zero-shot prompt	无需标注
主题分类	LDA	直接分类+解释	可解释性强
信息抽取	正则+NER	结构化输出	灵活准确
文本标注	众包平台	API批量处理	成本低、一致性高
翻译/摘要	专业服务	API调用	即时、便宜

Embedding Word and Sentence Method Selection

- Question: Which method should I use to embed words and sentences?
- Answer: It depends on your task.
- Simple Word Frequency / Keyword Statistics → TF-IDF (Simple, Explainable)
- Measure Document Similarity / Clustering → Word2Vec + Average
- Classification / Sentiment Analysis → BERT Fine-tuning / LLM
- Complex Information Extraction / Reasoning → LLM

Applications of Text Analysis

Applications Overview

Core Applications of Text Analysis

应用	核心问题	是否需要标注	输出
文本相似度	这些文档有多相似?	不需要	相似度分数
文本分类	这篇文档属于哪个类别?	需要	预定义类别
文本聚类	这些文档可以分成几组?	不需要	自动发现的组
情感分析	这段文字表达什么情感/态度?	通常需要	情感标签/分数
主题模型	这些文档在讨论什么话题?	不需要	潜在主题分布

Relationships

- 直接计算：文本相似度
- 有监督的学习：文本分类、情感分析
- 无监督的学习：文本聚类、主题模型

Integration with Four Paradigms

每种应用都可以使用不同范式的文本表示方法：

- **传统方法**：TF-IDF + 机器学习分类器
- **嵌入方法**：Word2Vec/BERT + 深度学习
- **LLM方法**：直接用Prompt完成任务

Similarity Measurement Methods

1. Cosine Similarity (余弦相似度)

$$\text{CosSim}(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{||\vec{A}|| \times ||\vec{B}||}$$

- $\vec{A} \cdot \vec{B}$，即向量A和向量B的点积
- $||\vec{A}|| \times ||\vec{B}|$，即向量A和向量B的模长的乘积

几何直觉：测量两个向量的夹角

- 1 = 方向相同 → 非常相似
- 0 = 垂直 → 不相关
- 优点：对文档长度不敏感

bank 和 finance 的向量表示

$$\text{bank} \rightarrow [0.12, -0.45, 0.78, 0.33, -0.21, \ldots, 0.56]$$
$$\text{finance} \rightarrow [0.15, -0.41, 0.82, 0.29, -0.18, \ldots, 0.51]$$

$$\text{Cosine Similarity} = \frac{\text{bank} \cdot \text{finance}}{||\text{bank}|| \times ||\text{finance}||} = 0.92 \quad \leftarrow \textbf{非常相似}$$

Similarity Measurement Methods

2. Jaccard Similarity

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

集合直觉：两个文档共享了多少词汇?

> Doc1词集: {经济, 增长, 技术, 创新} Doc2词集: {经济, 技术, 创新, 发展}

```
交集：{经济，技术，创新} = 3
并集：{经济，增长，技术，创新，发展} = 5
```

- the Jaccard Similarity is $\frac{3}{5} = 0.6$

Similarity Measurement Methods

3. Method Selection

场景	推荐方法
比较语义内容	余弦相似度 + 嵌入向量
检测词汇重叠/抄袭	Jaccard相似度
考虑词频权重	余弦相似度 + TF-IDF

- 经济学研究中，我们更关注语义内容，因此推荐使用余弦相似度 + 嵌入向量。

Applications: Measuring Technological Innovation

Kelly et al. (2021, AER: Insights): Patent Text and Innovation

研究问题：如何客观测量一项专利的创新程度?

方法：

1. 将每项专利表示为TF-IDF向量
2. 计算新专利与过去5年所有专利的余弦相似度
3. **创新度** = 1 - max(相似度)

核心指标：

- Backward Similarity：与过去专利的最大相似度 | 低 = 更具突破性 |
- Forward Similarity：被未来专利引用时的相似度 | 高 = 更有影响力 |

发现：

- 能识别出真正的突破性创新
- 与传统引用指标互补
- 可追溯200年技术史

Text Classification

Definition

> 将文档自动归入**预先定义**的类别中

Typical Applications

研究场景	分类任务	类别示例
央行沟通	政策立场分类	鸽派/中性/鹰派
企业年报	风险披露分类	财务风险/运营风险/法律风险
新闻分析	经济新闻分类	宏观/行业/公司/政策
政治文本	意识形态分类	左/中/右

Method Selection

Task	Method	Characteristics
传统	TF-IDF + SVM/朴素贝叶斯	简单、可解释、需要特征工程
深度学习	Word2Vec + CNN/LSTM	自动学习特征、需要大量数据
预训练	BERT微调	效果最好、需要GPU
LLM	Zero-shot/Few-shot Prompt	无需训练、灵活、成本高

Application: Media Bias Measurement

Gentzkow & Shapiro (2010, Econometrica)

问题：如何客观测量媒体的政治倾向？

方法：

1. 收集国会议员演讲，标注其党派（民主党/共和党）
2. 识别各党派的"标志性短语"
3. 用这些短语对媒体报道进行分类
4. 计算媒体的"政治倾斜指数"

发现：媒体偏见与受众偏好高度相关（迎合读者）

对同一件事情，不同党派的用语不同：

民主党倾向用语	共和党倾向用语
estate tax (遗产税)	death tax (死亡税)
undocumented workers	illegal aliens
war in Iraq	war on terror

Text Clustering

Definition

> 将文档自动分成若干**事先未知**的组，使组内文档相似、组间文档不同

Classification vs Clustering

维度	文本分类	文本聚类
类别	预先定义	算法发现
标注数据	需要	不需要
学习方式	监督学习	无监督学习
适用场景	类别已知且稳定	探索性分析

Common Methods

方法	原理	优缺点
K-Means	最小化组内距离	简单快速，需指定K
层次聚类	自底向上合并	可视化好，计算慢
谱聚类	基于图论	效果好，计算复杂

Application: Clustering Analyst Report

Analyst Report Clustering

问题：不同分析师的研究风格有何差异?

方法：

1. 收集某行业所有分析师报告
2. 用Sentence-BERT编码每份报告
3. 聚类发现不同的"报告类型"

可能发现的簇：

簇	特征	代表词
簇1	基本面分析	营收、利润率、增长
簇2	技术分析	支撑位、压力位、趋势
簇3	行业展望	政策、竞争格局、行业趋势
簇4	事件驱动	并购、管理层变动、诉讼

Sentiment Analysis

Definition

> 自动识别文本中表达的情感、态度或观点

Task Types

类型	输出	示例
极性分类	正面/中性/负面	产品评论
情感强度	连续分数(-1到+1)	情感程度
细粒度情感	具体情感类型	愤怒/恐惧/喜悦/
立场检测	支持/反对/中立	对议题的态度

方法演进

方法	原理	适用场景
词典方法	统计正负面词数量	简单任务、需要可解释性
机器学习	训练分类器	有标注数据时
深度学习	BERT等预训练模型	追求最高精度
LLM	Zero-shot分析	无标注、复杂情感

Sentiment Analysis Cases

Case 1: Tetlock (2007, JF) - Media Pessimism and Stock Market

问题：媒体情绪能预测股市吗？

方法：

1. 收集《华尔街日报》"Abreast of the Market"专栏
2. 使用Harvard-IV词典计算悲观情绪指数
3. 检验与次日股市收益的关系

Harvard-IV词典示例：

负面词	正面词
loss, decline, risk	gain, growth, profit
concern, fear, weak	strong, improve, success

发现：

- 高悲观情绪 → 次日市场下跌压力
- 异常高/低情绪 → 随后反转
- 媒体情绪包含噪音交易者信息

Challenges in Sentiment Analysis

Core Challenges

挑战	示例	难点
反语/讽刺	"这个政策真是太'棒'了"	字面与实际意思相反
隐含情感	"该公司连续三年亏损"	无明确情感词但含负面信息
条件句	"如果经济衰退，利润将下降"	假设性陈述
否定	"不会出现大幅下跌"	双重否定
比较	"比竞争对手更差"	需要理解比较对象

Advantages of LLM in Sentiment Analysis

> 传统方法的困境: "虽然一季度业绩不及预期，但管理层对全年展望保持乐观"

- 词典方法: 检测到"不及预期"(负面) + "乐观"(正面) → 混淆
- LLM方法: 理解整句语义 → 判断为"谨慎乐观"

Deepening Topic Model Applications

Topic Models vs Other Methods

维度	主题模型	文本分类	聚类
类别/主题	自动发现	预定义	自动发现
文档归属	可属于多主题	只属于一个类	只属于一个簇
输出	概率分布	类别标签	簇标签
可解释性	高（主题词）	取决于方法	需人工解读

Deepening Topic Model Applications

Unique Value of Topic Models in Economics

1. 发现未知结构

- 不需要预设研究假设
- 让数据"说话"

2. 时间序列分析

- 追踪主题随时间的变化
- 构建"注意力指数"

3. 降维与特征工程

- 将高维文本压缩为低维主题向量
- 作为回归分析的控制变量

Extended Topic Model Cases

Case 1: Hanley & Hoberg (2010, RFS) - IPO Information Content

问题：IPO招股书的信息含量如何影响定价？

方法：

1. 对IPO招股书应用LDA提取主题
2. 计算每份招股书的"标准化程度"
3. 检验与IPO抑价的关系

发现：

- 主题分布越独特（信息越丰富）→ 抑价越低
- 标准化的"模板"语言 → 信息不对称更严重

Extended Topic Model Cases

Case 2: Mueller & Rauh (2018, QJE) - Linguistic Signals of Political Violence

问题：能否从领导人演讲预测政治暴力？

方法：

1. 收集多国领导人演讲文本
2. 用主题模型提取演讲主题
3. 构建"好战性"主题指标

发现：

- 特定主题强度上升 → 后续暴力冲突概率增加
- 文本信号具有预测价值

Combining Methods

Combining Methods

组合	应用示例
聚类 + 分类	先聚类发现类别，再训练分类器
主题 + 情感	分析每个主题的情感倾向
聚类 + 情感	识别不同情感群体
主题 + 分类	用主题分布作为分类特征

References

- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*
- Ash, E., & Hansen, S. (2023). Text Algorithms in Economics. *Annual Review of Economics*