文本分析 (一)：词频法与向量空间

汪小圈

2025-05-12

文本数据与 NLP 初步

- 文本数据是一种非常常见但又极其复杂的数据类型
- 自然语言处理 (NLP) 是人工智能和语言学的交叉领域
- 主要任务包括：文本分类、情感分析、机器翻译、问答系统等

文本数据的特点

- **高维稀疏性**：文本可以表示为向量空间中的点，但这个空间往往有数万维（对应词汇量），而每个文档只使用其中很少的词
- **顺序性**：词的顺序对语义至关重要（“狗咬人”和”人咬狗”含义完全不同）
- **语义性**：文本承载复杂的语义信息，存在歧义、隐喻、引用等多种语言现象

文本处理流程概览

1. **采集**：网络爬虫、API 接口、数据库、PDF 解析等
2. **清洗**：去除 HTML 标签、特殊字符、错别字纠正等
3. **表示**：将文本转换为机器可理解的形式（向量化）
4. **建模**：应用机器学习算法执行分类、聚类、主题提取等任务

文本预处理：中文分词

- 与英文不同，中文没有明显的词语分隔符，需要专门的分词工具
- **jieba** 是目前最流行的中文分词工具之一：
  - 支持精确模式、全模式和搜索引擎模式
  - 允许添加自定义词典
  - 具有词性标注功能
- **常见问题**：
  - 专业术语、新词识别困难（需添加自定义词典）
  - 歧义分词（例如”结合成分子”可能被分为”结合/成/分子”或”结合/成分/子”）

文本预处理：停用词过滤

- 停用词是在文本处理中经常被过滤掉的常见词
- 主要包括：
  - 冠词、介词、连词等功能词
  - 代词、数词等
  - 高频但低信息量的词
- 过滤步骤：
  1. 准备停用词表（中文常用停用词表包含几百个词）
  2. 对分词结果进行过滤，去除停用词

词袋模型详解

- **词袋模型 (Bag of Words)** 是自然语言处理中最基础的文本表示方法
- 该模型将文本视为一组**无序的词语集合**，完全忽略语法和词序
- 仅关注词语在文档中是否出现或出现次数

词袋模型详解

数学表示：

对于文档集合 $D = \{d_1, d_2, ..., d_n\}$，构建词汇表 $V = \{w_1, w_2, ..., w_m\}$

文档 d_i 表示为向量 $X_i = [x_{i1}, x_{i2}, ..., x_{im}]$，其中：

$$x_{ij} = \text{词语} w_j \text{ 在文档} d_i \text{ 中的出现次数}$$

形成文档-词项矩阵（Document-Term Matrix, DTM）：

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

N 元语法模型

N 元语法模型（N-gram Model）：不仅考虑单个词，还考虑连续的 N 个词组合，可以部分保留词序信息

- **一元语法（Unigram）**：单个词，如“发展”、“经济”
- **二元语法（Bigram）**：两个相邻词，如“经济发展”、“促进增长”
- **三元语法（Trigram）**：三个相邻词，如“促进经济发展”

特点：随着 N 的增加，可以捕捉更多的上下文信息，但也会导致维度爆炸、数据稀疏性增加

词袋模型的向量化过程

1. **构建词汇表**：
   - 从所有文档中收集唯一的词语
   - 可去除停用词、低频词
   - 可能限制词汇表大小（最频繁的 K 个词）
2. **统计词频**：计算每个文档中每个词语出现的次数
3. **创建文档向量**：每个文档表示为一个向量，向量长度等于词汇表大小

词袋模型的优缺点

优点：

- 简单直观，易于理解和实现
- 计算高效，适用于大规模文本数据
- 维度确定，便于应用各种机器学习算法
- 捕捉文档的主题关键词

缺点：

- 完全忽略词序和语法，无法捕捉上下文信息
- 无法处理词语的多义性和同义词关系
- 高维稀疏表示，导致”维度灾难”
- 新词问题：测试文档中可能出现训练集中未见过的词

特征稀疏性分析

文本向量的稀疏性是文本分析中的重要特性。在词袋模型表示下，文本向量具有高度稀疏性，即绝大多数元素为零。

稀疏性源于以下事实：

1. **词汇量巨大**：自然语言的词汇量通常非常大（中文常用词汇约有几万个）
2. **单个文档用词有限**：任何一篇文档通常只使用全部词汇的很小一部分
3. **齐普夫定律 (Zipf's Law)**：自然语言中，词频与词频排名成反比

稀疏表示的优势和挑战

优势：

- 存储效率高：只需存储非零元素及其位置
- 计算效率高：只需处理非零元素
- 降低过拟合风险：稀疏性可视为一种正则化

挑战：

- 信息密度低：需要更多样本学习有效特征
- 难以直接应用某些算法，如神经网络
- “维度灾难”问题：高维空间中数据点趋于疏远

TF-IDF 理论基础

TF-IDF（Term Frequency-Inverse Document Frequency）是对词袋模型的重要改进，它不仅考虑词频 (TF)，还考虑词语的区分度 (IDF)。

核心思想：**如果一个词在某篇文档中出现次数多，但在整个文档集合中出现次数少，那么这个词很可能对该文档的主题具有较高的区分度。**

TF-IDF 数学定义

TF-IDF 由两部分组成：

1. **词频 (Term Frequency, TF)**：衡量词语在文档中的重要性
   常见计算方式：
   - **原始频率**：$TF(t,d) = f_{t,d}$
   - **归一化频率**：$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$
   - **对数归一化**：$TF(t,d) = \log(1 + f_{t,d})$
2. **逆文档频率 (Inverse Document Frequency, IDF)**：衡量词语提供信息的程度

$$IDF(t) = \log \frac{N}{DF(t)}$$

其中 N 是文档总数，$DF(t)$ 是包含词语 t 的文档数量。通常加平滑项：

$$IDF(t) = \log \frac{N}{DF(t) + 1} + 1$$

TF-IDF 权重

TF-IDF 权重：将 TF 和 IDF 相乘

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

TF-IDF 的理论性质：

1. 词频越高，TF 值越大：反映了词在文档中的重要性
2. 文档频率越高，IDF 值越小：惩罚了常见词
3. 独特性：对于罕见但在特定文档中高频的词，给予最高权重

文本相似度的理论基础

文本相似度是衡量两篇文档内容相似程度的度量，广泛应用于信息检索、文档聚类和分类、推荐系统等领域。

主要相似度度量：

- **欧氏距离 (Euclidean Distance)**：向量空间中两点之间的直线距离
- **曼哈顿距离 (Manhattan Distance)**：向量各维度差的绝对值之和
- **余弦相似度 (Cosine Similarity)**：向量夹角的余弦值
- **杰卡德相似系数 (Jaccard Similarity)**：集合交集与并集的比值

余弦相似度详解

余弦相似度 (Cosine Similarity)：计算两个向量夹角的余弦值，是文本相似度计算中最常用的度量之一

$$similarity(X, Y) = \cos(\theta) = \frac{X \cdot Y}{||X|| \times ||Y||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

- 取值范围：[-1, 1]，越接近 1 表示越相似
- 不受向量长度影响，只关注向量方向
- 适合于高维稀疏向量的比较

为什么余弦相似度适合文本分析?

余弦相似度在文本分析中特别常用，原因在于：

1. **文档长度不敏感**：长文档和短文档可以直接比较
2. **方向敏感**：关注的是词汇分布的模式而非绝对频率
3. **高效计算**：特别适合稀疏向量计算
4. **范围明确**：值域为 [-1,1]，便于理解和比较

词云图的理论基础

词云图 (Word Cloud) 是文本数据可视化的常用方法，通过调整词语的大小和颜色来表示其在文本中的重要性。

词重要性的数学表示：

1. **词频 (TF)**：词语在文档中出现的次数
2. **TF-IDF 权重**：结合词频和逆文档频率
3. **其他自定义权重**：如情感分析中的极性强度等

视觉编码原则：

1. **大小编码**：词语的大小与其重要性成正比
2. **颜色编码**：可表示词语的类别、情感极性等
3. **方向编码**：词语的方向可增加视觉多样性
4. **位置编码**：中心位置通常放置最重要的词

情感分析的理论基础

情感分析 (Sentiment Analysis)，也称为意见挖掘 (Opinion Mining)，是通过自然语言处理、文本分析和计算语言学等方法来识别、提取和量化文本中主观信息的过程。

情感分析的主要任务：

1. **文档级情感分析**：确定整个文档的情感倾向（积极、消极或中性）
2. **句子级情感分析**：确定单个句子的情感倾向
3. **方面级情感分析**：识别文本中提到的特定方面/属性及其相关情感
4. **比较情感分析**：比较不同实体之间的情感差异

情感分析的主要方法

情感分析的方法大致可分为三类：

1. **基于词典的方法**：
   - 依赖预定义的情感词典
   - 通过计算情感词的出现频率和强度来评估整体情感
   - 数学表示：

$$Score(d) = \frac{\sum_{t \in d} s_t \times w_t}{\sum_{t \in d} w_t}$$

   其中 s_t 是词 t 的情感得分，w_t 是权重
2. **基于机器学习的方法**：
   - 监督学习：使用标注数据训练分类器
   - 深度学习：使用 CNN、RNN/LSTM、Transformer 等神经网络模型
3. **混合方法**：结合词典和机器学习方法的优点

主题建模理论

主题建模 (Topic Modeling) 是一类无监督机器学习技术，旨在从文档集合中发现抽象” 主题”。

核心思想：**每篇文档可以看作是多个主题的混合，而每个主题又是词语上的概率分布。**

常见主题模型:

- **潜在语义分析 (LSA)**：基于奇异值分解 (SVD)
- **概率潜在语义分析 (PLSA)**：引入概率框架的主题模型
- **潜在狄利克雷分配 (LDA)**：最流行的主题建模方法，是 PLSA 的贝叶斯版本

潜在狄利克雷分配 (LDA)

LDA 的生成过程:

1. 对于每个文档 d:
   - 从狄利克雷分布 $Dir(\alpha)$ 中抽取主题比例向量 θ_d
2. 对于每个主题 k:
   - 从狄利克雷分布 $Dir(\beta)$ 中抽取词语分布 ϕ_k
3. 对于文档 d 中的每个词位置 i:
   - 从多项式分布 $Mult(\theta_d)$ 中抽取主题 z_{di}
   - 从多项式分布 $Mult(\phi_{z_{di}})$ 中抽取词语 w_{di}

LDA 的优缺点

优点：

- 完整的生成概率模型，理论基础扎实
- 解释性强，主题和词语分布有明确的语义
- 可扩展性好，适用于大规模文档集合
- 有效避免过拟合

缺点：

- 需要预先指定主题数量
- 不考虑词序和语法
- 对短文本效果较差
- 计算复杂度较高

主题一致性评估

评估主题模型质量的主要指标包括：

1. **困惑度 (Perplexity)**：衡量模型对未见文档的预测能力

$$Perplexity(D_{test}) = \exp\left(-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right)$$

2. **主题一致性 (Topic Coherence)**：衡量主题内部词语的语义相关性
   - **PMI(Pointwise Mutual Information)**
   - **NPMI(Normalized PMI)**
   - **UCI coherence**
   - **UMass coherence**

实践案例：政府工作报告分析

基于我们所学知识，可以对政府工作报告进行系统分析：

- **文本预处理**：分词、去停用词、清洗
- **词频分析**：识别高频词汇，跟踪政策关键词变化
- **TF-IDF 分析**：发现每年报告的独特关键词
- **相似度分析**：计算不同年份报告间的相似程度
- **情感分析**：分析政策语言的情感倾向
- **主题建模**：发现潜在政策主题及其变化趋势

总结

在本课中，我们学习了：

1. 文本数据的特点与预处理技术
2. 词袋模型与文本向量化方法
3. TF-IDF 加权与文本特征提取
4. 文本相似度计算与应用
5. 文本可视化技术（词云图）
6. 情感分析的基本原理
7. 主题建模的理论与方法（重点是 LDA）

这些基础知识为后续深入学习更复杂的 NLP 技术（如词嵌入、深度学习模型等）奠定了基础。