文本分析 (二)：词向量与深度学习基础

汪小圈

2025-05-19

本讲内容

- 从稀疏到密集表示
- Word2Vec 原理讲解
- 词向量的语义特性
- 金融文本中应用词向量
- 预训练词向量模型比较
- 词频法与词向量对比分析

Bag of Words 模型的局限性

- **丢失词序信息**：“政府调控房价” vs “房价调控政府”
- **语义鸿沟问题**：无法捕捉词与词之间的语义关系
- **维度灾难**：高维稀疏向量（维度 = 词汇量大小）
- **未登录词问题**：无法处理训练集中未出现的词语

词向量的直觉理解

词向量（Word Embedding）：将词语映射到低维稠密实数向量空间

- **稠密表示**：向量的每个维度都有非零值
- **语义编码**：不同维度隐含地编码了语义特征
- **相似性可计算**：语义相近的词在向量空间中距离较近

例如：“银行”和”金融”在向量空间中距离较近，而与”蔬菜”距离较远

分布式假设：词向量的理论基础

> *"You shall know a word by the company it keeps."*
> —— *J.R. Firth (1957)*

分布式假设：上下文相似的词，其语义也相似

例如："银行" 和" 金融机构" 经常出现在相似的上下文中，因此它们语义相似

词向量学习的核心任务：学习一个映射函数，使得上下文相似的词在向量空间中位置相近

稀疏向量 vs 密集向量

稀疏向量：$\mathbf{v} = [0, 0, 1, 0, ..., 0, 2, 0]$

- 大多数元素为 0
- 维度 = 词汇表大小
- 通常数十万维

密集向量：$\mathbf{v} = [0.2, -0.6, 0.5, ...]$

- 大多数元素非 0
- 维度通常 50-300
- 包含语义信息

密集表示的优势：

- **降维性**：从高维降至低维
- **连续性**：支持向量代数运算
- **泛化能力**：更好地泛化到未见例子

Word2Vec 简介

Word2Vec：Mikolov 等人于 2013 年提出的高效学习词向量的方法

核心思想：通过预测上下文中的词来学习词语的向量表示

两种模型：

- **Skip-gram 模型**：给定中心词，预测上下文词
- **CBOW 模型**：给定上下文词，预测中心词

Skip-gram 模型详解

目标：给定中心词，预测其上下文词

网络结构：

- **输入层**：中心词的 one-hot 编码
- **隐藏层**：无激活函数的全连接层
- **输出层**：预测上下文词的 softmax 层

数学表示：

$$p(w_o|w_i) = \frac{\exp(v'^T_{w_o} \cdot v_{w_i})}{\sum_{w=1}^{|V|} \exp(v'^T_w \cdot v_{w_i})}$$

CBOW 模型详解

目标：给定上下文词，预测中心词

网络结构：

- **输入层**：多个上下文词的 one-hot 编码
- **隐藏层**：无激活函数的全连接层
- **输出层**：预测中心词的 softmax 层

数学表示：

$$\hat{v} = \frac{1}{2c} \sum_{-c \le j \le c, j \neq 0} v_{w_{t+j}}$$

$$p(w_t|\hat{v}) = \frac{\exp(v'^T_{w_t} \cdot \hat{v})}{\sum_{w=1}^{|V|} \exp(v'^T_w \cdot \hat{v})}$$

Skip-gram 与 CBOW 对比

Skip-gram:

- 更适合小型语料库
- 对低频词表现更好
- 计算复杂度较高

CBOW:

- 训练速度更快
- 对高频词表现更好
- 在大型语料库上更稳定

负采样（Negative Sampling）技术

问题：softmax 计算复杂度与词汇量成正比，计算效率低

解决方案：负采样技术，将多分类问题转化为二分类问题

负采样原理：

1. 对真实词对 (w_i, w_o)，标记为正样本 (1)
2. 随机采样 k 个负样本 (w_i, w_n)，标记为负样本 (0)
3. 使用逻辑回归判断词对是否真实共现

优化目标：

$$J(\theta) = \log \sigma(v'^T_{w_o} \cdot v_{w_i}) + \sum_{j=1}^{k} \mathbb{E}_{w_j \sim P_n(w)} [\log \sigma(-v'^T_{w_j} \cdot v_{w_i})]$$

词向量空间的语义特性

语义相关性：相似概念在向量空间中距离较近

- “银行” 和” 金融” 距离近
- “苹果”(水果) 和” 橙子” 距离近
- “苹果”(公司) 和” 微软” 距离近

余弦相似度：

$$similarity(w_1, w_2) = \cos(\theta) = \frac{v_{w_1} \cdot v_{w_2}}{||v_{w_1}|| \cdot ||v_{w_2}||}$$

语义计算：向量代数运算

词向量支持向量代数运算，可进行” 语义计算”：

$$v(\text{”king”}) - v(\text{”man”}) + v(\text{”woman”}) \approx v(\text{”queen”})$$

更多例子：

- v(” 北京”) $-$ v(” 中国”) $+$ v(” 法国”) $\approx$ v(” 巴黎”)
- v(” 比特币”) $-$ v(” 数字”) $+$ v(” 实物”) $\approx$ v(” 黄金”)

这些语义运算表明词向量确实捕获了复杂的语义关系

中文预训练词向量模型

1. 腾讯 AI Lab 词向量

- 训练语料：8 亿 + 句子，200 亿 + 词汇
- 词汇量：800 万词、词组和实体
- 向量维度：200 维

2. 哈工大/讯飞词向量

- 训练语料：人民日报等新闻语料
- 词汇量：约 100 万词
- 向量维度：300 维

3. 北师大/人大词向量

- 训练语料：百度百科、维基百科、人民日报 1947-2017、知乎、微博、文学、金融、古汉语等
- 词汇量：约 200 万词
- 向量维度：300 维

英文预训练词向量模型

1. Word2Vec 词向量

- 训练语料：Google News (1000 亿词)
- 词汇量：约 300 万词和短语
- 向量维度：300 维

2. GloVe 词向量

- 训练语料：CommonCrawl/Wikipedia
- 词汇量：40 万-200 万词
- 向量维度：50-300 维

3. FastText 词向量

- 训练语料：维基百科等
- 词汇量：约 200 万词
- 向量维度：300 维

预训练模型的表现比较

1. 语义捕捉能力

测试语义关系准确度:

- 国家-首都关系
- 性别关系
- 形容词-比较级关系

2. 领域适应性

- 金融领域
- 医疗领域
- 法律领域

3. 处理未登录词能力

FastText > GloVe > Word2Vec

预训练模型选择指南

应用场景	推荐中文模型	推荐英文模型	理由
通用文本分类	腾讯 AI Lab	GloVe 300d	覆盖面广，维度适中
命名实体识别	哈工大词向量	FastText	对实体和罕见词好
情感分析	腾讯 AI Lab	Word2Vec	语义细微差别好
金融领域	领域特定模型	领域特定模型	专业术语需求高

词向量在实际应用中的评估

1. 内在评估 (Intrinsic Evaluation)

- **词相似度任务**：WordSim-353、SimLex-999
- **类比关系任务**：国家-首都、动词时态等
- **多语言对齐质量**：跨语言词语语义一致性

2. 外在评估 (Extrinsic Evaluation)

- **下游任务性能**：分类准确率、F1 得分、BLEU 分数
- **迁移能力**：从一个领域到另一个领域的泛化能力
- **资源效率**：训练和推理时的计算和内存消耗

3. 案例分析：金融领域词向量与通用词向量在市场情绪分析中的性能比较可提升 8-15% 准确率

选择或训练自己的词向量模型

使用预训练模型的情况：

- 数据量有限，无法支持训练
- 任务是通用领域
- 计算资源有限
- 需要快速开发原型

训练自己的模型的情况：

- 有大量特定领域的文本数据
- 应用领域有特殊术语
- 现有预训练模型表现不佳
- 有足够的计算资源

折中方案：微调预训练模型

- 从预训练模型开始，用领域数据继续训练
- 保留通用语言知识，学习领域特定表示

词频法与词向量对比

比较维度	词频法 (Bag of Words/TF-IDF)	词向量 (Word2Vec)
数据表示	高维稀疏向量 (~ 万维)	低维稠密向量 (~ 百维)
语义捕捉	基于表面词频统计，无语义	基于分布式假设，有语义
计算复杂度	低，适合大规模文档	中等，训练需要时间
新词处理	无法处理未见词	也无法直接处理 (FastText 可以)
应用场景	文档分类、信息检索	语义搜索、推荐系统

词向量的优缺点

优点

- **语义丰富**：捕获了词语间语义关系
- **维度可控**：典型为 50-300 维
- **泛化能力**：处理未见过的词组合
- **通用性**：适用于各种 NLP 任务

局限性

- **多义词问题**：无法区分同一词的不同含义
- **上下文依赖**：固定向量无法根据上下文调整
- **预训练依赖**：需要大量语料预训练
- **领域专一性**：通用词向量在专业领域可能不佳

本讲小结

1. 词向量通过低维稠密向量表示词语，克服了传统方法的局限
2. Word2Vec 通过 Skip-gram 和 CBOW 两种模型高效学习词向量
3. 负采样等技术大幅提高了训练效率
4. 词向量空间具有丰富的语义特性，支持相似性计算和向量代数运算
5. 在金融文本分析中，词向量可以发现政策热点、分析语义变化等