语料之舞：从数据采集到词向量深度解析之演武

潘宇轩

2025 年 10 月 29 日

目 录

一、开场之辞

此份文书，乃是为了一场从网络语料的采集直至词向量模型幽微之处的深度剖析的完整自然语言处理（NLP）实践所做的系统性记述。此番实践的流程，始于数据的自动化采集，终于词向量模型内在行为的深度分析。我等首先实现了指定的爬虫与预处理流水线，分别从古登堡计划与新浪新闻获取了英文文学与中文新闻两类语料。

在对语料进行探索性分析，并验证其符合自然语言的统计特性之后，我们以 Skip-gram 算法为基准，训练了初始的 Word2Vec 模型。以词向量空间中的相似词检索为评估手段，初步的评估揭示了模型在捕捉常规语义上的有效性。

为深入探究模型行为，我设计并执行了三个维度的拓展对比实验，宛如在舞台上上演三幕精彩的对决。**领域对比实验**有力地证明了词汇的向量表示会随语料环境发生显著的“语义漂移”，其变化之剧烈，不亚于一场华丽的舞步；**算法对比实验**清晰地揭示了 Skip-gram 与 CBOW 在学习具体共现关系与通用聚合关系上的不同倾向，二者各有其优雅之处；**超参数对比实验**则验证了增加负采样数能够引导模型学习到更泛化、更鲁棒的语义表示。

二、缘起之章

甲、演武之背景与动机

自然语言处理（NLP）的核心基石，在于如何将人类的语言——这一充满模糊、抽象与上下文依赖的符号系统——转化为机器可解、可算的数学表示，这本身就是一场绝妙的挑战。词向量（Word Embedding）技术，特别是以 Word2Vec 为代表的分布式表示方法，正是此番挑战中的一座里程碑。它通过将词语从高维稀疏的独热编码，映射到低维、稠密的连续向量空间，成功地赋予了机器在向量层面理解语义相似度的能力，为机器翻译、情感分析等众多下游任务的突破提供了坚实支撑。

本项目的核心动机，或者说整个课程的最终理想，便是搭建一条完整的 NLP 流水线，从而不仅在工程层面获得宝贵的实践经验，更能在理论层面，通过亲手设计和观测控制变量实验，深度洞察在数据驱动下，模型行为内在的复杂逻辑与规律。

乙、Word2Vec 之奥义

Word2Vec 乃是 2013 年提出的经典之作，其核心思想根植于语言学的“分布式假设”：一个词的意义，由其频繁共同出现的上下文所决定。它包含两种主要的模型架构：

- **CBOW（Continuous Bag-of-Words）**：使用一个词的上下文作为输入，预测该中心词；更擅长高频词的表示，训练速度较快。

- **Skip-gram**：使用中心词作为输入，预测其上下文词；对低频词与精细共现更敏感，但训练开销更大。本项目主实验采用 Skip-gram。

丙、演武之目标与阵法

本实验的核心目标，是完成一次从数据到模型的 NLP 全流程，并系统性地探究关键因素对模型行为的影响，实验分为以下几个阶段：获取数据、预处理与清洗、基础分析与建模、拓展实验设计与执行

三、粮草之筹备与精炼

甲、宝库之所在

语料来源涵盖两类：

- **英文文学语料（Project Gutenberg）**：古登堡计划（Project Gutenberg）是一个历史悠久的数字图书馆项目，致力于免费提供公共版权的文学作品。本项目从中采集了经典文学作品，这些文本具有叙事丰富、词汇多样、文学性强的特点，涵盖小说、诗歌、戏剧等多种体裁，为英文词向量模型提供了充足的语义上下文与语言表达的多样性。

- **中文新闻语料（新浪新闻）**：新浪新闻作为国内主流新闻门户网站之一，提供了大量实时更新的新闻报道。本项目采集的中文语料主要来自财经、科技、社会等新闻板块，这些文本具有主题广泛、信息密度高、时效性强的特点，词汇涵盖了当代社会热点与专业术语，为中文词向量模型提供了现代汉语在新闻领域的真实使用场景。

两类语料在文体、主题与词汇分布上形成鲜明对比，为后续的跨领域对比实验奠定了数据基础。

乙、神速采集之策略

为高效采集新闻，设计了基于线程池与队列的并发爬虫，采用广度优先遍历，设定文件大小阈值实现终止

然在实践中，常规之单线程爬取，速度很慢，不知道什么原因，在短时间很难完成数据读取。为此，设计了并发爬虫阵法，以提升采集效率。

此阵法之精髓，在于启动多个并发工作线程（Worker），各线程独立地从任务队列中获取待采之网址，并同时执行网络请求与页面解析。此举极大缩短了因网络延迟而产生的等待时间，令数据采集之速，有如神助，效率倍增。其核心调度逻辑如代码清单 1 所示。

其核心调度逻辑如下：

```
def worker():
    while not stop_event.is_set():
        try:
            current_url = urls_to_visit.get(timeout=1)
        except Exception:
            continue

        article_text, new_links = get_text_and_links(
           current_url)

        if article_text:
            with file_lock:
                with open(ZH_FILENAME, 'a', encoding='utf-8')
                   as f: f.write(article_text + "\n\n")
                current_size_mb = os.path.getsize(ZH_FILENAME)
                    / (1024 * 1024)
                pbar.n = current_size_mb
                pbar.refresh()
                if current_size_mb >= TARGET_SIZE_MB:
                   stop_event.set()

        with file_lock:
            for link in new_links:
                if link not in visited_urls:
                    visited_urls.add(link)
                    urls_to_visit.put(link)
        urls_to_visit.task_done()
```

Listing 1: 并发工作线程与优雅终止

丙、玉石之雕琢

预处理是整个流程的关键基础步骤。针对中英文语料的不同特点，采用了不同的处理策略：

英文预处理：将文本转为小写，去除标点与数字，过滤停用词（如 ”the”、”and” 等），并进行词形还原（Lemmatization），将词汇统一为基本形式。

中文预处理：使用 jieba 分词工具进行分词，去除标点与数字，同样过滤停用词（如”的”、”了”等）。

这两步处理既保证了数据的清洁度，又保留了足够的语义信息。

丁、铸模之操练

表 1 列出了本次实验中各模型与训练的具体设定（包含模型架构、向量维度、窗口大小、最小词频、训练轮数、负采样数与并行线程数等）。后续章节中呈现的训练结果与相似词检索均以该配置为基准；若需对比其他配置（如 CBOW、不同的 min_count 或 negative 值），可在脚本级别进行参数替换并复现实验以获得可比结果。

表 1: 模型与训练设定（中文与英文分别训练）

项目	设定
语料	中文与英文语料分别独立训练
工具	Gensim `Word2Vec`
模型架构	Skip-gram（`sg=1`）
向量维度	150
上下文窗口	5
最小词频	5
训练轮数	10
负采样数	5
并行线程	4

四、沙场之勘察与初阵

甲、军情之洞察

对预处理后的中英文语料进行了探索性数据分析（EDA）。规模统计显示：英文语料总词数达 1,378,438，独立词汇量约 42,615；中文语料总词数为 796,648，独立词汇量约 45,575。尽管中文语料总词数较少，但独立词汇量略高于英文，这与中文分词粒度及新闻语料主题多样性相关。

表 2 总结了中英文语料的基础统计信息（词数、独立词汇量及高频词），为下文的语料对比与模型行为分析提供量化依据。

表 2: 中英文语料基础统计对比

统计项	英文语料 (文学)	中文语料 (新闻)
总词数	1,378,438	796,648
独立词汇量	42,615	45,575
Top 1	the (80,874)	的 (35,659)
Top 2	and (47,436)	在 (9,333)
Top 3	of (40,340)	和 (6,692)
Top 4	to (36,175)	了 (5,923)
Top 5	a (27,982)	是 (5,477)

高频词分布体现出两类语料的鲜明主题特征：英文语料以文学叙事为主，高频词如 ”the”、”and”、”of” 等功能词占据主导，反映出叙事性文本的语法特点；中文语料则以财经新闻为主，高频词如”的”、”在”、”公司”、”市场”等既包含功能词，也涵盖领域相关词汇，体现了新闻语料的信息密度与主题聚焦度。

词云可视化进一步印证了这一差异，如图1所示。中文词云中”公司”、”市场”、”技术”等财经与科技词汇突出；英文词云则以人物、地点等文学叙事要素为主，词汇分布更为发散。

(a) 中文语料高频词云　　(b) 英文语料高频词云

图 1: 中英文语料高频词云

齐夫定律的引入与验证

齐夫定律（Zipf’s law）是自然语言中广泛观察到的一条经验规律：若将语料中所有词按频率从高到低排序，词频 f 与其秩 r 之间近似满足幂律关系 $f(r) \propto r^{-s}$（通常 $s \approx 1$）。该规律反映了自然语言中少数高频功能词与大量低频内容词并存的分布特性，对语料质量与模型训练结果有直接影响。

在本报告中，我们对中英文预处理后语料分别计算词频并按秩排序，随后在双对数坐标上绘制词频-秩图以检验幂律线性趋势（见图 2）。图中中文与英文曲线在中等至高秩区间均呈现近似线性关系，斜率接近 -1，说明两套语料均符合齐夫定律的统计特

性注意到，在极高频处存在偏离，可能由于高频处被功能词（如中文的“的/在/和”或英文的“the/and/of”）主导，所以出现偏差。

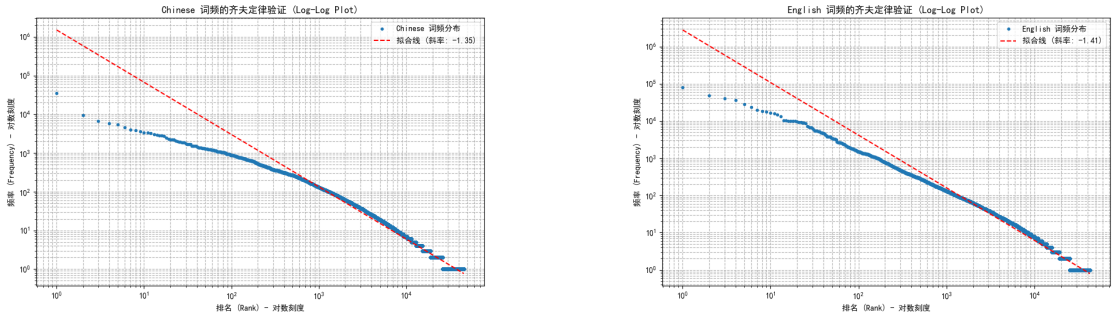

图 2: 齐夫定律验证（中文与英文，双对数坐标）

乙、寻邻之鉴

基于此语料与已训练模型，对二十个英文与二十个中文检索词逐一求取 Top-10 相似词.

以数据言志：英语检索揭示明显的语料偏斜与结构噪声——例：查询词“city” 的近邻中出现了大量人名与地名碎片（如 ”boylan”、”irish”），而查询词“night” 的近邻被罗马数字（”ii”、”iii”）占据，显现该英文语料以文学古籍为主，含章节编号与专有名，故模型在若干查询上被这些格式化或专名信号带偏

反观中文检索，连贯性更强且多为领域相关词汇：例如“芯片”的近邻包含“触控”“散热”“内核”等硬件与性能词条；“公司”与“市场”则被金融语境词（如“减持”“股份”“资本”）所包围。此异同印证：模型表现并非单纯算法优劣，而深植于语料之分布特征与文本结构。

表 3 给出了若干代表性查询及其 Top-3 相似词，用以具体说明英文模型在专名与格式化信息下的偏差以及中文模型的领域连贯性。

表 3: 中英文代表性词语相似词对比

查询词	Top 1	Top 2	Top 3
英文模型 (文学语料)			
city	richie (0.9989)	places (0.9988)	irish (0.9988)
night	ii (0.9831)	iii (0.9828)	iv (0.9818)
中文模型 (新闻语料)			
芯片 (chip)	触控 (0.9920)	散热 (0.9897)	165 (0.9866)
公司 (company)	减持 (0.9511)	股份 (0.9502)	万股 (0.9502)

五、演武之三幕对决

下述三组控制变量对比实验用于解析“领域、算法、超参”如何塑造语义空间。

第一幕：领域之舞

领域对比实验设计

额外采集 2MB 科技新闻，训练 `word2vec_chinese_tech.model`，与通用新闻模型对比观察“苹果/智能/芯片”等词的语义漂移。

领域对比结果呈现

对比结果见表 4。

表 4: 第一幕：领域之舞——相似词对比

对比词	模型	Top-10 相似词（及相似度）
苹果	通用新闻	大姐(0.9868), 自信(0.9851), 下沉(0.9843), 看来(0.9840), 找到(0.9832), 经历(0.9828), 安全(0.9827), 元素(0.9820), 举例(0.9812), 格力电器(0.9805)
苹果	科技新闻	MacBook(0.9336), Share(0.9227), 太高(0.9185), 是不是(0.9180), Air(0.9161), 一台(0.9124), 极致(0.9115), 高品质(0.9109), 素质(0.9104), 换上(0.9096)
智能	通用新闻	硬件(0.9804), 系统(0.9748), 架构(0.9713), 应用(0.9707), G2(0.9705), 极致(0.9702), 电竞(0.9700), 流畅(0.9694), 冰河(0.9691), 时代(0.9689)
智能	科技新闻	IDC(0.9961), 合同(0.9928), 最为(0.9923), 无线(0.9918), 合作伙伴(0.9918), 投入(0.9916), 崛起(0.9915), 自给自足(0.9913), 推进(0.9913), 首批(0.9913)
芯片	通用新闻	触控(0.9920), 散热(0.9897), 165(0.9866), 玩家(0.9860), 内核(0.9844), 流畅(0.9840), 设计(0.9832), 超高(0.9827), 同档(0.9819), 硬件(0.9813)
芯片	科技新闻	海思(0.9618), 禁止(0.9576), 技术(0.9395), 出售(0.9393), 实现(0.9376), 全球(0.9370), 参与(0.9301), 终端(0.9262), 厂商(0.9230), 设备(0.9227)

领域对比深度分析

“苹果”在通用模型中近邻发散，科技模型中则收敛到品牌与产品；“芯片”在通用模型更偏消费端语义，在科技模型聚焦于半导体产业链。**词向量语义是随领域而动的相对位置**。

为便于直观比较，本节的领域对比结果在图 3 中以条形图形式汇总，图 4 展示了领域对比的代表性词云示例，读者可据此快速把握不同语料下相似度与语义分布的差异。

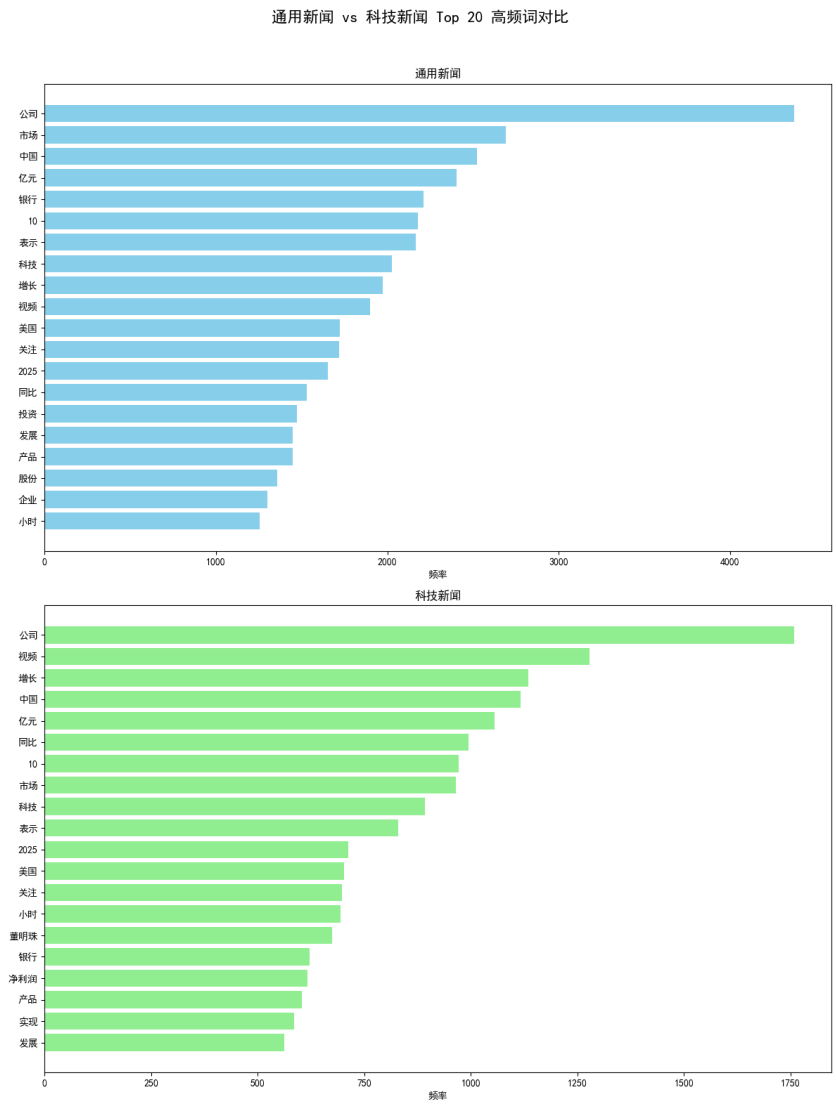

图 3: 第一幕：领域之舞——统计汇总条形图（通用 vs 科技）

图 4: 第一幕：领域对比的代表性词云对比

第二幕：招式之辨

实验设计

在同一通用语料上训练 CBOW 与 Skip-gram 两个模型，对比“公司/技术”等词的相似词风格差异。

分析与结论

表 5 明确展示了在相同通用语料上的 CBOW 与 Skip-gram 两类模型在相似词返回结果上的差异。

表 5: 第二幕：招式之辨——CBOW vs Skip-gram 对比

对比词	模型	Top-10 相似词（及相似度）
公司	Skip-gram	减持(0.9511), 股份(0.9502), 万股(0.9502), 股本(0.9487), 不超(0.9448), 股东(0.9428), 承诺(0.9420), 转让(0.9408), 锁定(0.9386), 制药(0.9377)
公司	CBOW	减持(0.9990), 万股(0.9988), 竞价(0.9988), 交易方式(0.9983), 持股(0.9982), 股份(0.9979), 股本(0.9976), 有限公司(0.9976), 回购(0.9972), 股东(0.9968)
技术	Skip-gram	硬件(0.9623), 领先(0.9610), 护眼(0.9563), 165(0.9559), 流畅(0.9557), 系统(0.9548), 引领(0.9546), 智能(0.9541), 应用(0.9498), 首发(0.9496)
技术	CBOW	系统(0.9998), 金融(0.9997), 空间(0.9997), 正在(0.9997), 厂商(0.9997), 货币(0.9997), 全面(0.9997), 护眼(0.9997), 方面(0.9997), 退税(0.9997)

CBOW 在示例中呈现出异常偏高的相似度数值（表中数值接近 0.99 的项），且近邻词往往更平滑、聚合，反映模型在平均化上下文信息时对常见搭配的强化；而 Skip-gram 返回的相似词相对分散但更贴近具体共现（如行业术语或短期搭配），数值更具判别性。

CBOW 与 Skip-gram 各有优势：前者更平滑、稳定，适合对泛化性有更高要求的场景；后者更贴近共现细节，适合需要保留语料精细搭配信息的分析或应用。模型选型应以实际任务目标与定量验证结果为准。

参数更新策略的选择

在模型训练中，参数更新策略的选择对效率和效果有着重要影响。全批量更新（Batch Gradient Descent）和单样本更新（Stochastic Gradient Descent, SGD）虽然各有优点，但并未被本实验采用，原因如下：

全批量更新需要在每次迭代中计算整个数据集的梯度。这种方法虽然能够提供精确的梯度估计，但计算开销极大，尤其在处理大规模数据时，容易导致内存不足或计算效率低下。此外，全批量更新的收敛速度较慢，缺乏随机性带来的探索能力。

单样本更新每次仅基于一个数据点计算梯度，虽然计算开销较小，但梯度估计的波动性较大，可能导致收敛过程不稳定。此外，这种方法对现代硬件的并行计算能力利用不足，难以充分发挥计算资源的优势。

本实验采用的小批量更新（Mini-batch Gradient Descent）在每次迭代中使用一个小规模的数据子集计算梯度，综合了全批量和单样本更新的优点。它既能提高计算效率，又能在一定程度上平滑梯度波动，同时充分利用硬件的并行计算能力，是一种平衡效率与效果的合理选择。

第三幕：心法之炼

实验设计

对比 Skip-gram 的负采样数量（`negative=5` vs. `15`），观察“市场”一词的相似词变化。

分析与结论

从表 6 可见：当负采样数由 5 增加到 15 时，Top-10 相似词的绝对相似度值整体下降（例如若干条相似度由 0.95 降至 0.91），且近邻词的组成发生了可观察的调整——高相似度且紧密的搭配项在低负采样（neg=5）下更易被凸显，而在高负采样（neg=15）下，模型倾向于给出更为稳健、主题性更强但相似度数值更保守的近邻。

这说明增大负采样数会在一定程度上抑制噪声信号（如偶发的格式化标记或非常规搭配）对相似度排序的影响，从而使语义邻居更聚焦于语料的主题相关词。但代价是相似度分数整体下降，意味着顶端相似度的置信度变得更为保守。

表 6: 第三幕：心法之炼——负采样数量对比

对比词	模型	Top-10 相似词（及相似度）
市场	neg=5	成为(0.9526), 资本(0.9425), 点心(0.9370), 预期(0.9356), 节奏(0.9312), 保持(0.9290), 提高(0.9257), 持续(0.9221), 智能手机(0.9216), 科创债(0.9207)
市场	neg=15	节奏(0.9219), 成为(0.9208), 预期(0.9187), 点心(0.9144), 显著(0.9139), 海外(0.9085), 整体(0.9071), 逐步(0.9041), 扩张(0.9028), 有所(0.9012)

六、谢幕之言

本项目完成了从数据采集、预处理、分析到词向量训练与深度评估的端到端实践。主要结论如下：本文构建并验证了一条端到端的自然语言处理实验流水线，覆盖网络语料的采集（中、英语料均在数 MB 量级）、系统化的文本预处理（中文分词、英文标记化）、统计分析与可视化，以及基于 Word2Vec 的词向量训练与相似词检索。实验结果通过词频分布、词云与齐夫定律验证了语料的统计特性，同时揭示了语料中存在的结构噪声（如章节编号与专有名格式化）对模型行为的干扰。通过领域（通用 vs 科技）、算法（CBOW vs Skip-gram）与超参数（负采样数）三组对比实验，我们观察到语料领域与超参设置会显著改变若干目标词的近邻分布，从而影响语义判定的稳定性。