北京郵電大學
Beijing University of Posts and Telecommunications

自然语言处理

基于LangChain的本地知识库智能检索与问答增强

小组序号：05　组长（汇报人）：王博远　组员：池瀚

汇报时间：2025 年 6 月 20 日

一、工作概览

北京郵電大學 Beijing University of Posts and Telecommunications

NO.1 语料收集与处理和知识库构建

基于 **BeautifulSoup4** 和 **Selenium** 框架搭建自动化爬虫程序，爬取**CSDN**上**1500+**机器学习相关语料，以 **JSON 元数据 + TXT 正文内容**存储，为检索模块构建高质量底层知识库。

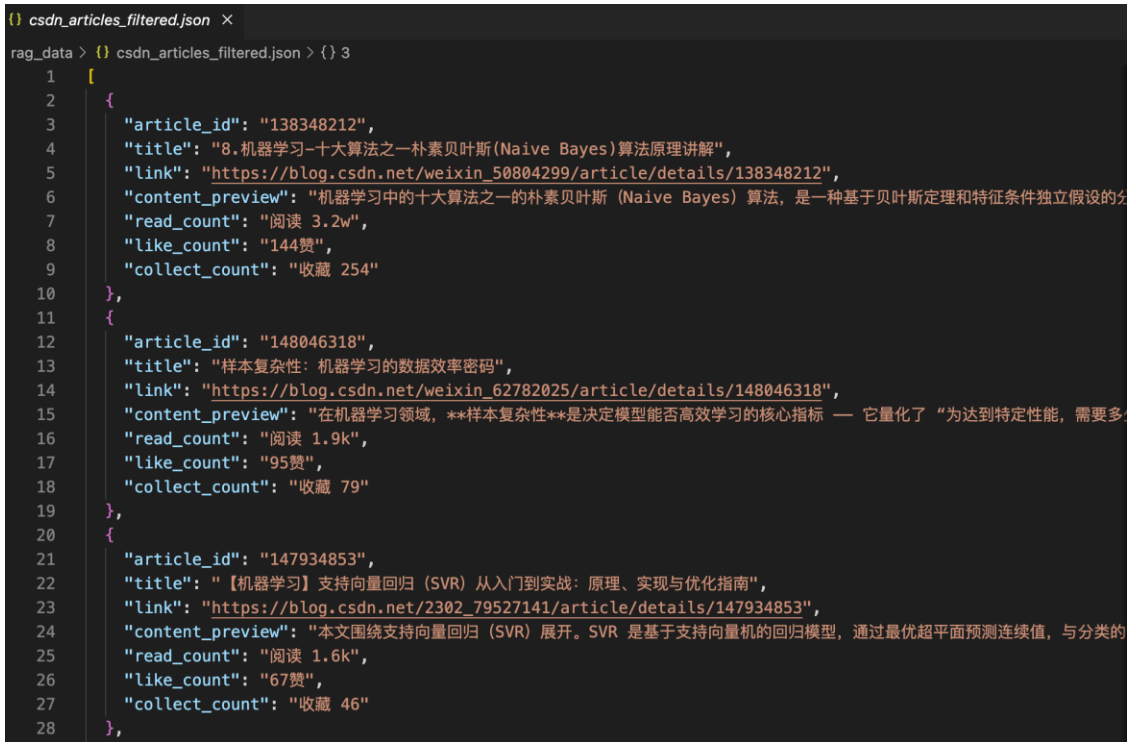
NO.2 基于LoRA方法对BGE模型微调与优化

构建有效**QA对**和**正负样本**，并采用**关键词重叠**和**TF-IDF相似度**等策略挖掘困难负样本，使用 **LoRA** 方法对 **BAAI/bge-large-zh-v1.5** 模型进行微调，提高模型检索层面的性能指标。

NO.3 RAG知识库问答系统集成与应用

面向**知识问答**和**教学资源推荐**等应用领域搭建微服务系统，集成**多轮对话、PDF文档问答和知识库浏览**三大核心功能，采用**混合检索**和**智能重排序**等策略优化系统性能和使用体验。

二、CSDN语料收集与处理



CSDN语料爬取

1. **内容质量高：** 文章多由一线人员撰写，技术深度有保障
2. **数据丰富度：** 覆盖面广，从基础概念到前沿算法均有涉及
3. **获取便利性：** 大部分文章公开发布，便于合法获取

- **语料结构与组成：** 文章ID+标题+链接+预览+阅读量+点赞数+收藏量
- **数据过滤与筛选：** 1500篇 → 827篇高质量技术博客 (55%精选率)
- **存储格式标准化：** JSON结构化元数据 + TXT非结构化正文内容

二、QA对构造


```
{
"question": "朴素贝叶斯算法的核心思想是什么？",
"answer": "朴素贝叶斯算法的核心思想是通过考虑特征概率来预测分类，即对于给出的待分类样本，求解在此样本出现的条件下各个类别出现的概率，哪个最大，就认为此待分类样本属于哪个类别。",
"source_article_id": "138348212",
"source_title": "8.机器学习-十大算法之一朴素贝叶斯(Naive Bayes)算法原理讲解",
"source_link": "https://blog.csdn.net/weixin_50804299/article/details/138348212"
},
```

```
{
"question": "什么是样本复杂性在机器学习中的核心问题？",
"answer": "样本复杂性是指机器学习算法为实现目标性能（如准确率≥90%）所需的最小数据量。",
"source_article_id": "148046318",
"source_title": "样本复杂性：机器学习的数据效率密码",
"source_link": "https://blog.csdn.net/weixin_62782025/article/details/148046318"
},
```


.....


```
{
"question": "SVR与SVM的区别是什么？",
"answer": "SVR是回归模型，用于预测连续型变量；而SVM是分类模型，用于预测离散型变量。SVR允许数据点存在误差，而SVM不允许。",
"source_article_id": "147934853",
"source_title": "【机器学习】支持向量回归（SVR）从入门到实战：原理、实现与优化指南",
"source_link": "https://blog.csdn.net/2302_79527141/article/details/147934853"
},
```


多阶段爬虫框架

1. **csdn_spider： 文章导航列表智能爬取**
   - 反爬检测规避技术
   - 动态滚动加载机制
   - 断点续爬功能
2. **single_article： 单篇文章内容深度抓取**
   - 智能内容等待机制
   - 自动处理"阅读全文"限制
   - 多重重试错误处理
3. **qa_generation： AI驱动的QA对生成**
   - 并发API调用优化
   - 智能提示词设计
   - 质量过滤机制

三、正负样本对构造

正样本对构建 基于两种主要策略构建了**4148个**正样本对

1. 基于QA对的直接构建

- 构建策略：直接将原始的**1830个**question-answer作为正样本对
- 质量保证：每个QA对都基于具体的技术文章内容得到，确保准确性

2. 基于文档内容的多层次构建：

- 标题-内容预览对： 利用文章标题与摘要的天然对应关系
- 标题-完整内容片段对： 标题与文章正文前500字符的匹配
- 内容预览-完整内容对： 摘要与正文中间部分的语义关联


```
"text1": "什么是批量学习和在线学习？",
"text2": "批量学习是一次性训练模型，适用于静态数据；在线学习逐条或小批量更新模型，适用于动态环境。",
"label": 1.0
```

```
"text1": "机器学习有监督学习sklearn实战二：六种算法对鸢尾花(Iris)数据集进行分类和特征可视化",
"text2": "项目的主要环节：从数据探索、预处理、模型训练与比较，到结果分析和可视化，是一个标准的分类问题解决方案模板。针对鸢尾花数据集的特点，通过多种可视化手段和模型比较方法，全面评估了不同算法的性能表现。",
"label": 1.0
```

```
"text1": "链式法则是微积分中的一个基本法则，用于计算复合函数的导数。在神经网络中，它允许我们计算损失函数相对于网络中任何参数的梯度。",
"text2": "……BP神经网络中的链式法则……反向传播（Backpropagation，简称BP）算法是神经网络训练中的核心技术，而链式法则则是BP算法的基础。本文将深入探讨……",
"label": 1.0
```

三、正负样本对构造

正样本对构建 基于两种主要策略构建了**4148个**正样本对

1. 基于QA对的直接构建

- 构建策略：直接将原始的**1830个**question-answer作为正样本对
- 质量保证：每个QA对都基于具体的技术文章内容得到，确保准确性

2. 基于文档内容的多层次构建：

- 标题-内容预览对： 利用文章标题与摘要的天然对应关系
- 标题-完整内容片段对： 标题与文章正文前500字符的匹配
- 内容预览-完整内容对： 摘要与正文中间部分的语义关联

"text1": "什么是批量学习和在线学习？",
"text2": "批量学习是一次性训练模型，适用于静态数据；在线学习逐条或小批量更新模型，适用于动态环境。",
"label": 1.0

"text1": "机器学习有监督学习sklearn实战二：六种算法对鸢尾花(Iris)数据集进行分类和特征可视化",
"text2": "项目的主要环节：从数据探索、预处理、模型训练与比较，到结果分析和可视化，是一个标准的分类问题解决方案模板。针对鸢尾花数据集的特点，通过多种可视化手段和模型比较方法，全面评估了不同算法的性能表现。",
"label": 1.0

"text1": "链式法则是微积分中的一个基本法则，用于计算复合函数的导数。在神经网络中，它允许我们计算损失函数相对于网络中任何参数的梯度。",
"text2": "”.....BP神经网络中的链式法则.....反向传播（Backpropagation，简称BP）算法是神经网络训练中的核心技术，而链式法则则是BP算法的基础。本文将深入探讨.....",
"label": 1.0

负样本对构建 共计构建**2903个**负样本，包括**871个**基础负样本和经过多种策略组合深度挖掘得到的**2032个**困难负样本。

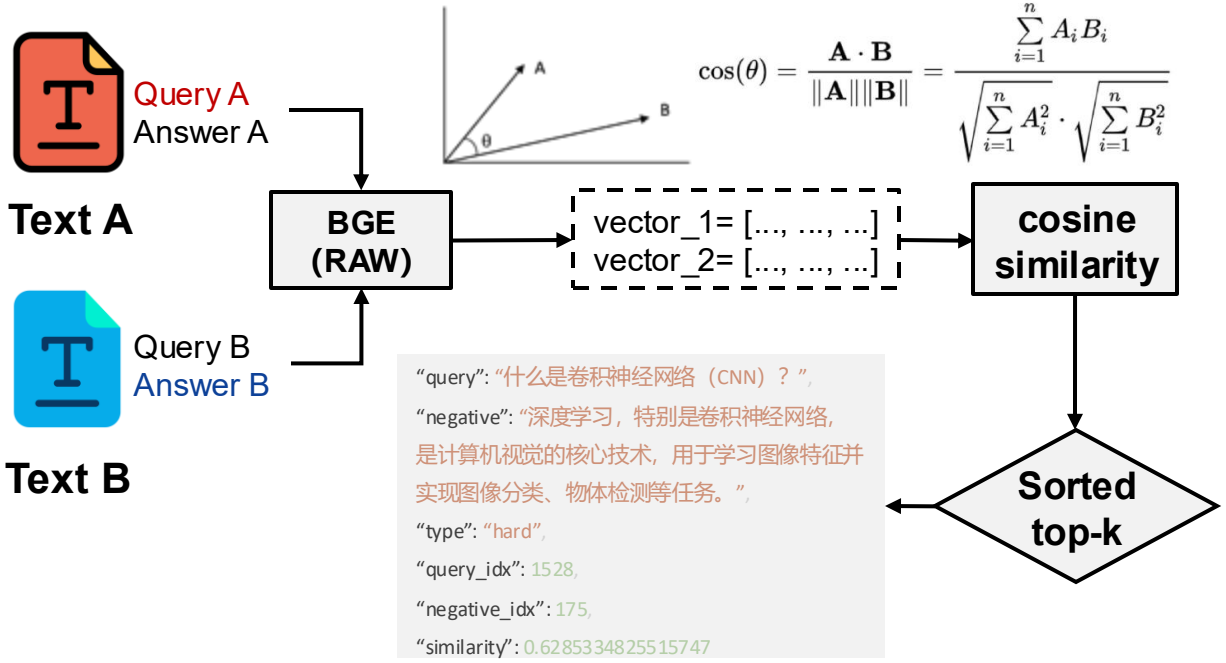
- **基础负样本构建：** 随机采样非正样本对，确保基本正负样本区分能力

"text1": "机器学习作为人工智能的重要分支，能够让计算机系统从数据中自动学习模式和规律，并利用这些知识进行预测和决策。在工业4.0的背景下，机器学习可以处理和分析海量的生产数据，为生产过程优化、质量控制、设备维护等提供智能支持。",
"text2": "无监督学习中的聚类问题是什么？",
"label": 0.0

三、困难负样本挖掘

① 语义相似度挖掘

- **原理：**使用原始的BGE预训练模型计算文本向量，选择语义空间中接近但标签不同的样本
- **目标：**训练模型区分语义相似但实际不匹配的文本对
- **实现：**对每个查询选择余弦相似度较高但非正样本文本

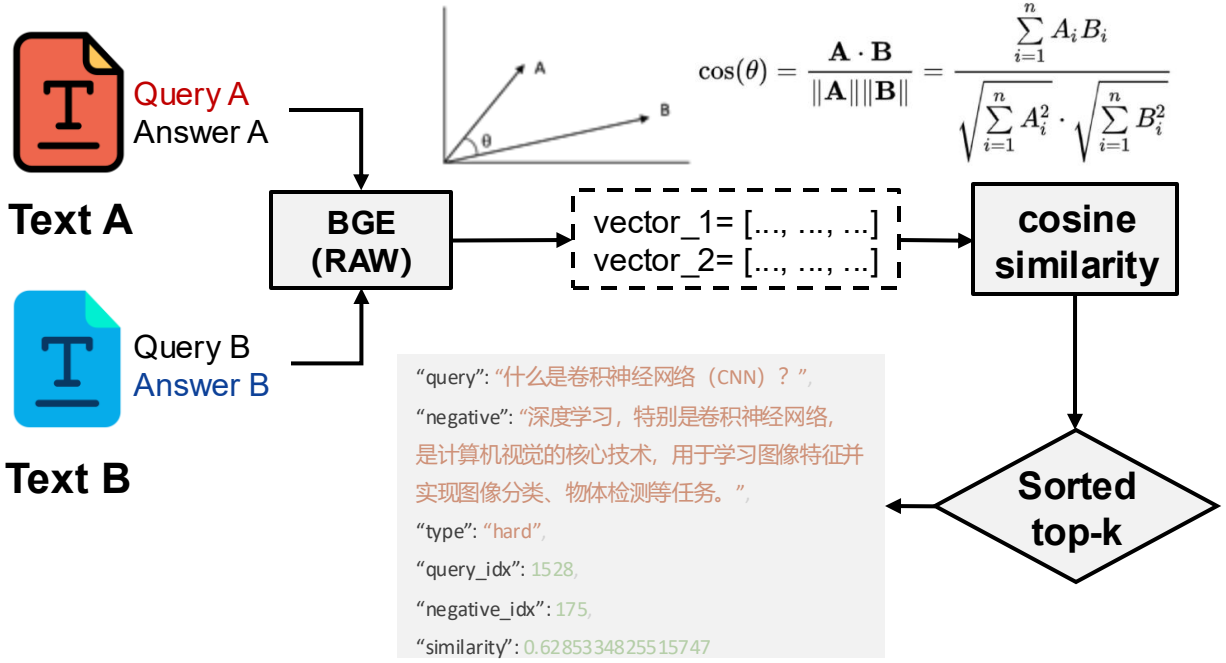


"query": "什么是卷积神经网络（CNN）？",
"negative": "深度学习，特别是卷积神经网络，是计算机视觉的核心技术，用于学习图像特征并实现图像分类、物体检测等任务。",
"type": "hard",
"query_idx": 1528,
"negative_idx": 175,
"similarity": 0.6285334825515747

三、困难负样本挖掘

① 语义相似度挖掘

- **原理：**使用原始的BGE预训练模型计算文本向量，选择语义空间中接近但标签不同的样本
- **目标：**训练模型区分语义相似但实际不匹配的文本对
- **实现：**对每个查询选择余弦相似度较高但非正样本文本



② 关键词重叠挖掘

- **原理：**基于jieba分词和术语词典，计算关键词重叠度
- **策略：**选择重叠度在0.1-0.4之间的文本对
- **价值：**防止模型过度依赖关键词匹配，提升语义理解能力

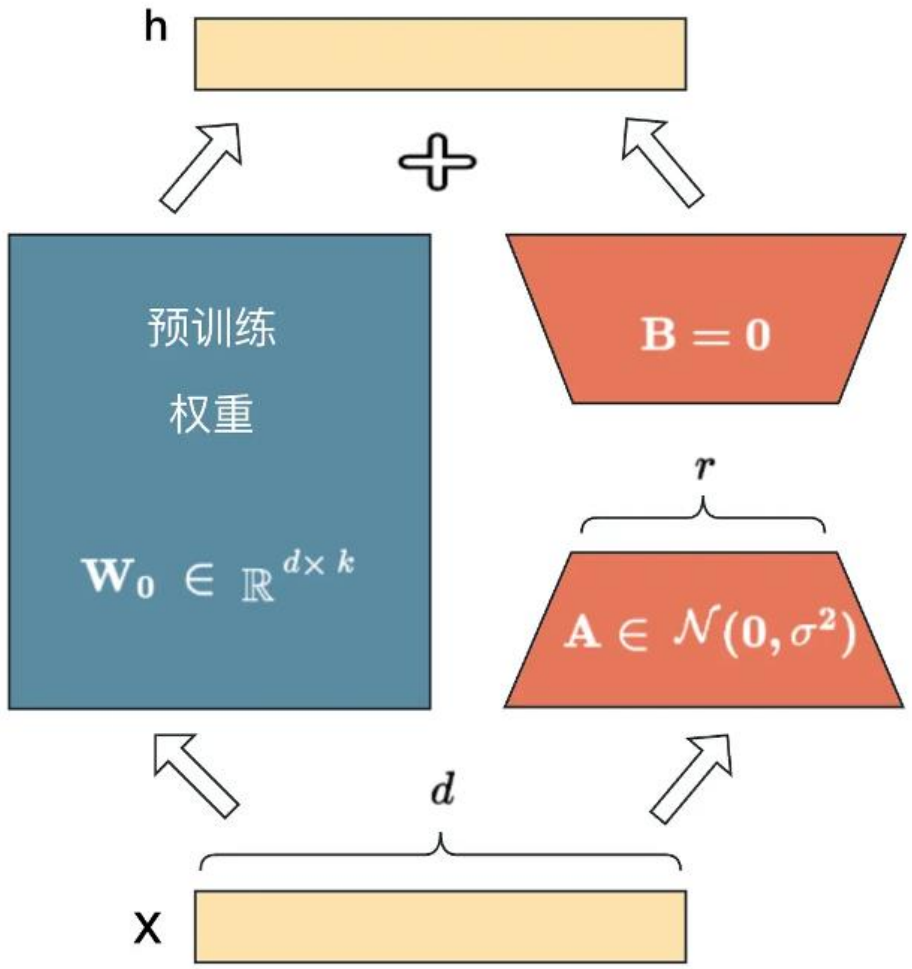
$$overlap = |\ key_A \cap key_B\ | \ /\ |\ key_A \cup key_B\ |$$
$$hard\_negative = \{doc | 0.1 \le overlap \le 0.4\}$$

③ TF-IDF相似度挖掘

- **方法：**构建TF-IDF向量化器，计算词汇级别的相似度
- **筛选标准：**选择TF-IDF相似度在0.2-0.6之间中等相似文本
- **作用：**挖掘词汇相似但语义不同的困难样本

$$TF-IDF(t,d) = tf(t,d) \times log(N/df(t))$$
$$similarity = cosine\_similarity(tfidf\_query, tfidf\_doc)$$
$$hard\_negative = \{doc | 0.2 \le similarity \le 0.6\}$$

四、LoRA微调BGE模型



基础模型： BGE-large-zh-v1.5 (3.2B参数)

hidden_size=1024,layers=24,attention_head=16,intermediate_size=4096

微调方法： LoRA (Low-Rank Adaptation)

训练参数： batch_size=8, epochs=3, lr=2e-5, evaluation_steps=500

损失函数： CosineSimilarityLoss

优化策略： warmup + 梯度裁剪（max_grad_norm）

假设要在下游任务微调预训练模型，则需要更新预训练模型参数：

$$W_0 + \Delta W$$

其中，W_0是预训练模型初始化的参数，ΔW就是需要更新的参数。

具体来看，假设预训练的权重矩阵$W_0 \in \mathbb{R}^{d\times k}$，它的更新可表示为：

$$W_0 + \Delta W = W_0 + BA, B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times k}$$

其中，秩$r \ll \min(d, k)$。

在 LoRA 的训练过程中，W_0是固定不变的，只有A和B是训练参数。

四、数据集划分与样本统计

训练数据分布

数据类型	数量	占比
Train	4,935	70.0%
Validate	1,058	15.0%
Test	1,058	15.0%
Positives	4,148	58.8%
Negatives	2,903	41.2%
Hard Negatives	2,032	70% (among Negatives)
Total	**7,051**	**100%**

困难负样本挖掘统计

挖掘策略	样本数量	占比
Cross Domain	1,999	93.4%
Semantic Similarity	200	9.4%
Keyword Overlap	138	6.4%
TF-IDF	23	1.1%
Total (filtered)	**2,137**	**100%**

五、检索性能指标

Metrics	Epoch_1	Epoch_2	Epoch_3
Accuracy@1	66.29%	**69.82%**	67.74%
Accuracy@3	82.66%	**84.59%**	80.58%
Accuracy@5	86.36%	**87.64%**	84.11%
Precision@1	66.29%	**69.82%**	67.74%
Precision@3	27.55%	**28.20%**	26.86%
Precision@5	17.27%	**17.53%**	16.82%
Recall@1	66.29%	**69.82%**	67.74%
Recall@3	82.66%	**84.59%**	80.58%
Recall@5	86.36%	**87.64%**	84.11%
MRR@10	0.7538	**0.7793**	0.7491
NDCG@10	0.7913	**0.8116**	0.7788
MAP@100	0.7572	**0.7824**	0.7534

MRR@k： 平均倒数排名，衡量正确答案的平均排名

$$MRR@k = (1/|Q|) \times \Sigma(1/rank_i)$$

其中，$rank_i$ 是查询 i 的第一个相关文档的排名位置

NDCG@k： 归一化折损累积增益，综合考虑排名和相关性

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i}-1}{log_2(i+1)} \qquad IDCG_p = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i}-1}{log_2(i+1)}$$

其中，REl_p 表示语料库中相关性最高的 p 个文档列表。

MAP@k： 平均精确率均值，所有查询的平均精确率的均值

$$MAP@k = (1/|Q|) \times \Sigma(AP@k_i)$$

$$AP@k_i = (1/min(m,k)) \times 2(Precision@j \times rel_j)$$

五、模型微调性能对比

Metrics	Vanilla	Finetuned	Absolute	Relative
accuracy	88.85%	95.18%	+6.33%	**+7.13%**
best_accuracy	89.79%	96.22%	+6.43%	**+7.16%**
auc_approx	94.10%	98.86%	+4.76%	**+5.06%**
similarity_separation	0.2787	**0.7817**	+0.5030	**+180.45%**
avg_positive_similarity	0.7062	0.8977	+0.1915	**+27.12%**
avg_negative_similarity	0.4274	0.1160	-0.3115	**-72.87%**

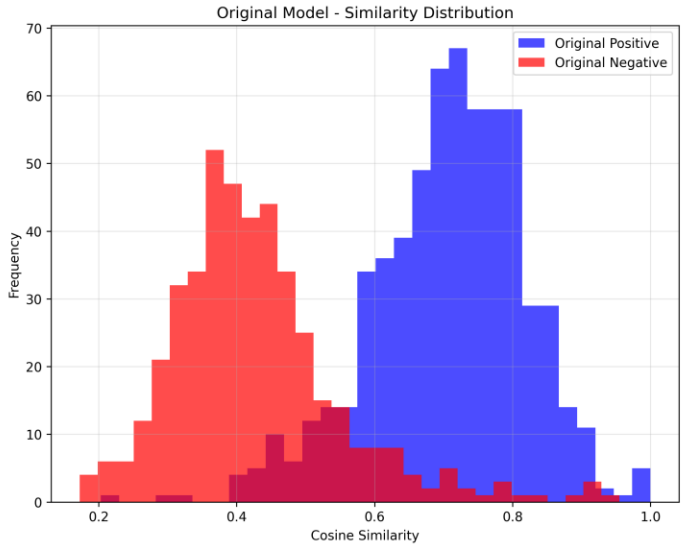
➢ 核心分类指标

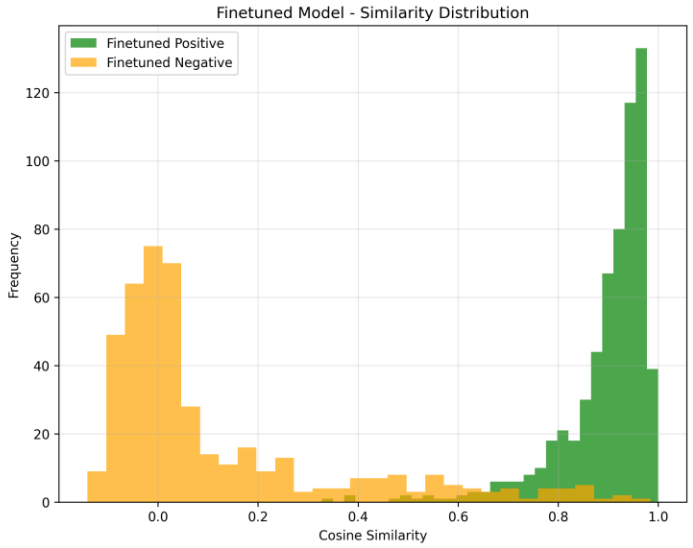
- **准确率 (Accuracy)**：正确分类的样本占总样本的比例
- **最佳准确率**：通过优化阈值得到的最高准确率
- **AUC近似值**：衡量模型区分正负样本的排序能力，越接近1.0越好

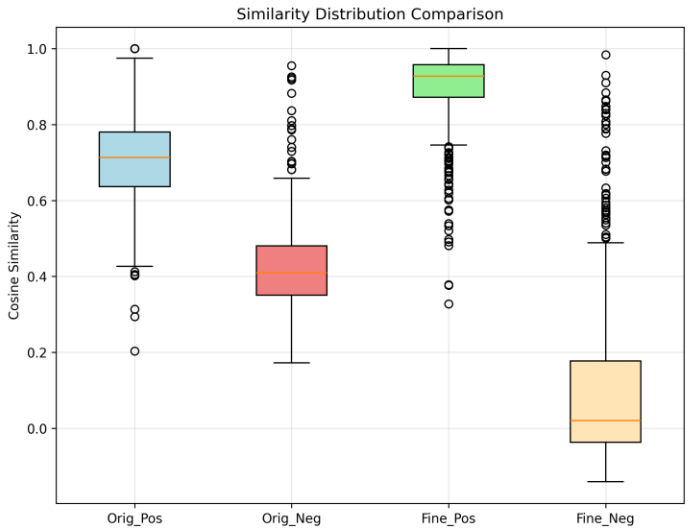
➢ 相似度分析指标

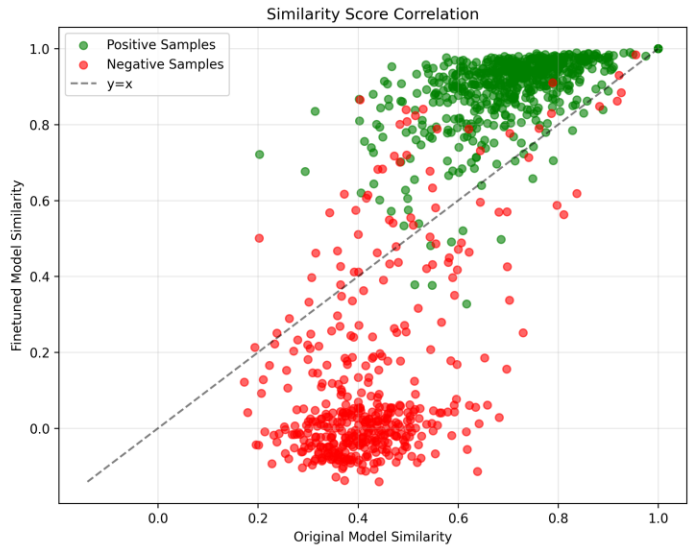
- **相似度分离度**：正样本平均相似度 - 负样本平均相似度
  (数值越大说明模型区分能力越强)
- **正样本平均相似度**：所有正样本对的余弦相似度均值，较高
- **负样本平均相似度**：所有负样本对的余弦相似度均值，较低

五、可视化对比









原始模型的问题 (左上图)

- **正负样本重叠严重**: 蓝色(正样本)和红色(负样本)分布大量重叠在0.4-0.6区间
- **区分界限模糊**: 两个分布峰值过于接近，模型难以准确判断
- **负样本相似度偏高**: 红色分布集中在0.3-0.5，说明错误匹配相似度较高

微调模型的改进 (右上图)

- **较好的分离效果**: 绿色(正样本)集中在0.8-1.0高相似度区间
- **负样本相似度降低**: 橙色分布主要在0.0-0.2区间，接近完全不相关
- **零重叠现象**: 正负样本分布几乎没有重叠，分离度达到最优

六、系统集成与应用

1. 混合检索策略设计——双粒度双策略检索

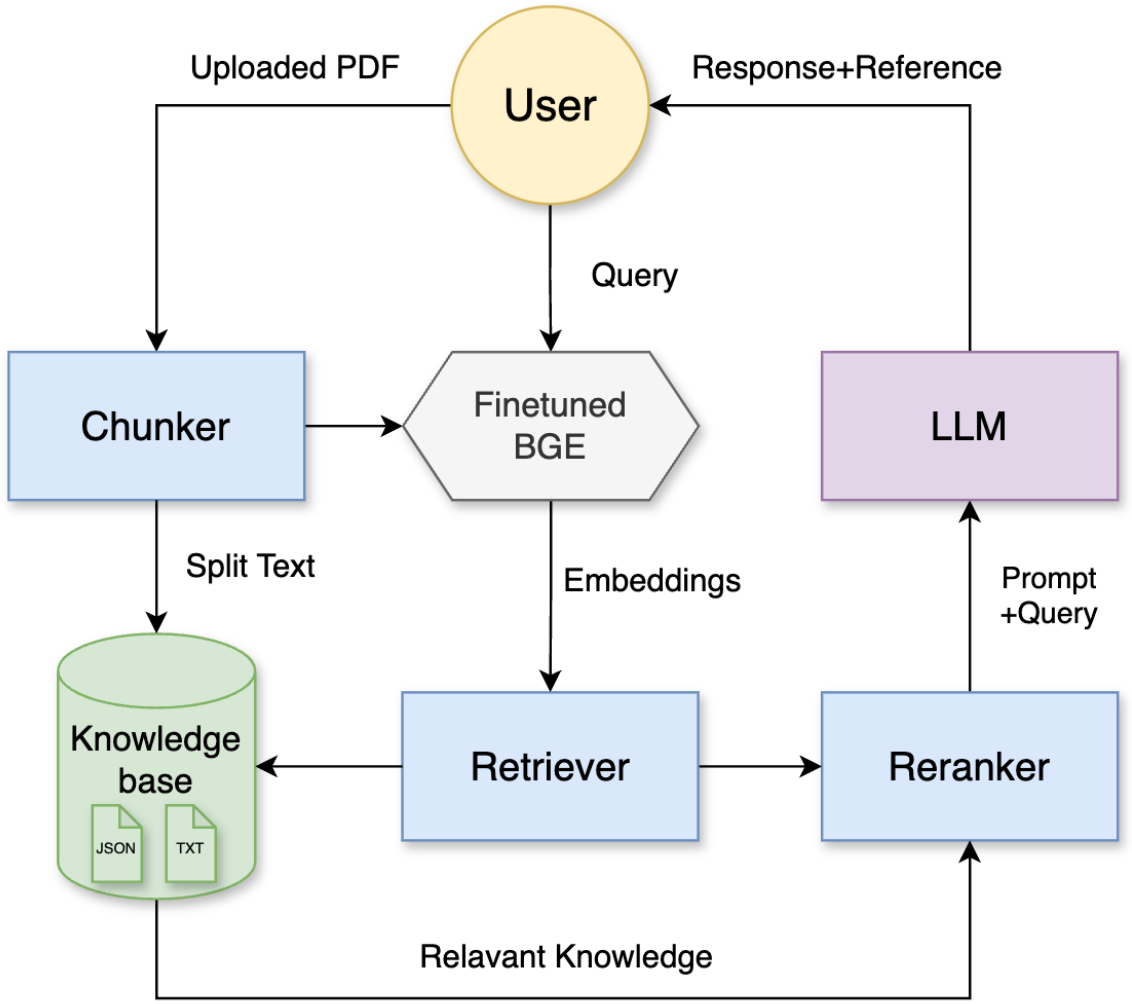
- **段落级 + 文档级**：既能精确定位细节，又能获取全局理解
- **向量检索 + BM25**：语义匹配与关键词匹配的最优融合
- **动态权重调节**：alpha参数让用户可根据需求调整检索策略

2. 专业化的领域问答系统

- **智能领域过滤**：自动识别机器学习相关问题
- **权重增强的文章匹配**：对ML专业术语给予更高权重
- **引用参考资料集成**：深度理解和引用而非简单检索

3. 微服务架构的系统性设计

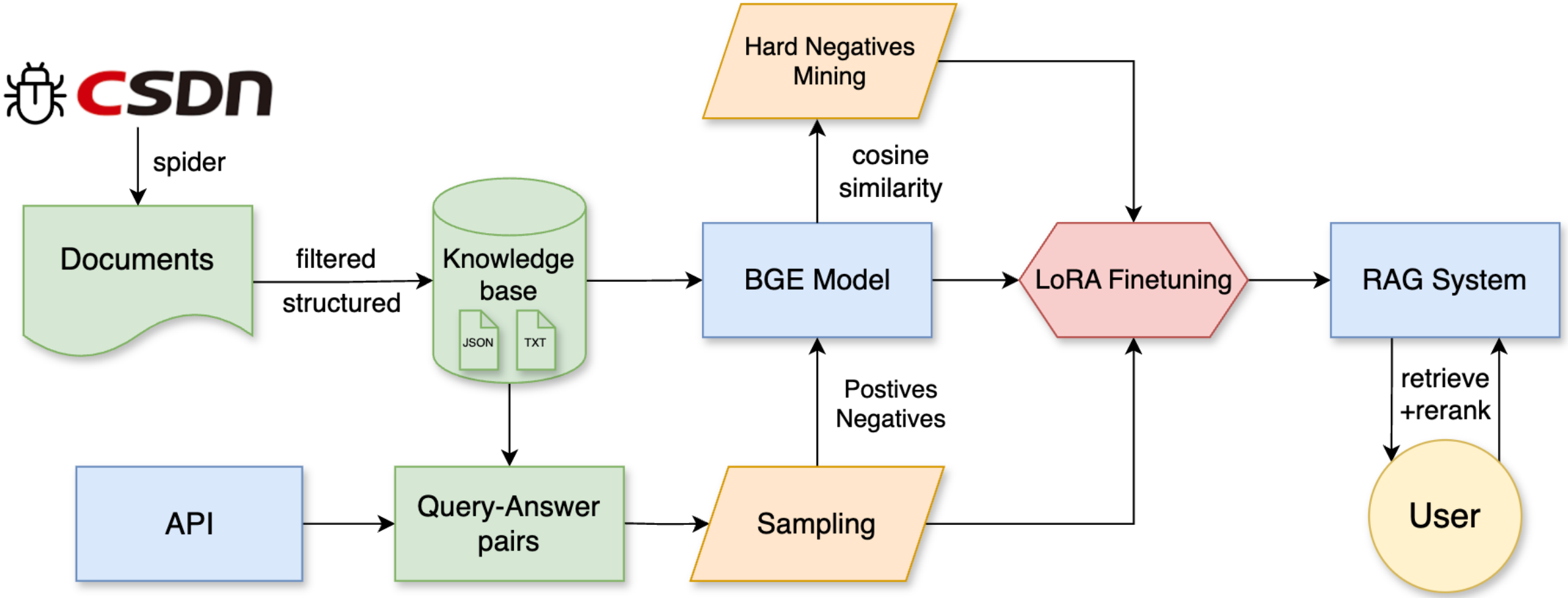
- **服务解耦**：检索、重排序、生成三个独立服务
- **性能监控**：实时追踪各服务调用次数和耗时
- **容错机制**：服务失败时采用降级策略



六、系统演示


机器学习知识问答系统
localhost:8501
Deploy
ML问答系统
选择功能
多轮对话
系统说明
多轮对话: 与机器学习助手进行对话
PDF文档问答: 上传PDF并进行问答
知识库浏览: 浏览机器学习文章库
机器学习助手
我是专业的机器学习助手，可以回答机器学习、深度学习、数据科学相关的问题。
清除对话历史
请输入您的问题...

七、总结框图


CSDN
spider
Documents
filtered
structured
Knowledge base
JSON
TXT
Hard Negatives Mining
cosine similarity
BGE Model
LoRA Finetuning
RAG System
retrieve +rerank
User
Postives Negatives
API
Query-Answer pairs
Sampling

工作概览

北京郵電大學 Beijing University of Posts and Telecommunications

NO.1 语料收集与处理和知识库构建

基于 **BeautifulSoup4** 和 **Selenium** 框架搭建自动化爬虫程序，爬取**CSDN**上**1500+**机器学习相关语料，以 **JSON 元数据 + TXT 正文内容**存储，为检索模块构建高质量底层知识库。

NO.2 基于LoRA方法对BGE模型微调与优化

构建有效**QA对**和**正负样本**，并采用**关键词重叠**和**TF-IDF相似度**等策略挖掘困难负样本，使用 **LoRA** 方法对 **BAAI/bge-large-zh-v1.5** 模型进行微调，提高模型检索层面的性能指标。

NO.3 RAG知识库问答系统集成与应用

面向**知识问答**和**教学资源推荐**等应用领域搭建微服务系统，集成**多轮对话、PDF文档问答和知识库浏览**三大核心功能，采用**混合检索**和**智能重排序**等策略优化系统性能和使用体验。