**二、基于动态采样策略的Few-Shot实体关系抽取实验——以PHM领域文献为例**

**1. 实验概述**

本实验旨在评估“动态Few-Shot示例采样”策略在大模型实体关系抽取中的效果。通过语义向量表示与相似度计算，从人工标注数据中选取与当前待抽取文献语义最相关的示例，引导大模型更好理解目标文本语境，提高抽取质量。

**2. 实验步骤**

**步骤1：构建语义嵌入库（使用GTE模型）**

本研究第一步针对训练集中的三篇飞机预测与健康管理（PHM）领域文献，使用中文通用语义嵌入模型 BAAI/bge-large-zh-v1.5，对文档进行分段处理后生成语义向量。具体流程如下：

首先，将每篇文档按照空行划分为若干段落，保证每个段落语义相对完整且文本长度适中。随后，利用预训练的 BGE 模型对每个段落文本进行编码，采用模型输出中 CLS token 的向量表示作为段落的语义嵌入。通过该方式，获得了包含三篇文档所有段落的语义向量集合。

最后，将所有段落的向量及对应文本信息保存为 JSON 格式文件，为后续的主题聚类与动态示例采样模块提供高质量的语义基础。

此步骤保证了动态采样模块能基于丰富且粒度适中的文本语义信息，准确筛选与待抽取文本语义高度相关的示例，从而提升大模型在实体关系抽取任务中的泛化能力和效果。

**步骤2： 主题聚类**

在第二步中，我们将第一步中构建的文本语义向量与人工标注的实体关系信息相结合，完成了降维、聚类与数据合并的流程。首先，脚本从 embedding\_vectors.json 加载每条记录，其中包含段落文本及其在高维语义空间中的向量表示；与此同时，依据文件名读取对应的 JSON 标注文件，将实体与关系信息映射到每条文本记录中。随后，使用 UMAP 算法将高维向量压缩至五维空间，以保留语义邻域结构并去除冗余特征；训练完成后，UMAP 模型被持久化保存，以便后续快速加载与降维。接着，基于降维后的向量执行 HDBSCAN 聚类，自动识别出语义主题簇并标记噪声。最后，脚本将每条记录的聚类标签与降维向量一起附加回原始数据，并将完整的文本、向量、降维结果、主题簇号及实体关系标注以 JSON 格式输出到 embedding\_clusters\_with\_annotations.json 中。该文件即为后续动态 Few‑Shot 示例选取模块提供了既包含语义分布信息、又携带结构化标注的、可直接检索的训练示例库。

**步骤3：构建S模块**

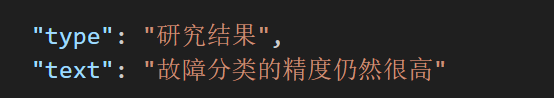
在本研究中，针对每篇待抽取的论文文本，我们设计并实现了动态构建 Few-Shot 动态采样模块（以下简称 S 模块）的流程。该模块旨在从预先聚类并标注好的示例库中，筛选出与目标文本语义最为接近的若干段落及其对应的实体与关系标注，作为大语言模型抽取任务的辅助提示。

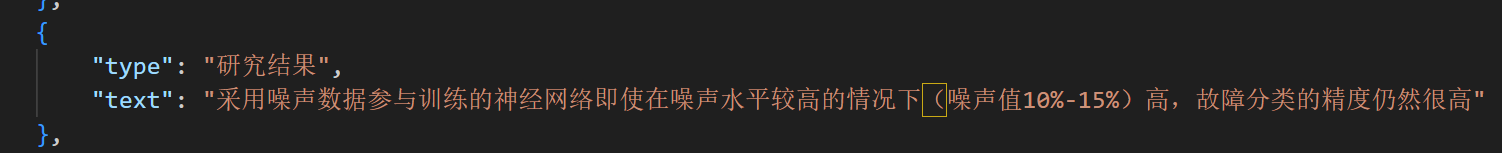
具体而言，首先利用预训练的文本编码模型对论文全文进行向量化表示，并借助已训练的降维模型将高维嵌入映射到低维空间。随后，通过计算论文向量与示例库中各聚类段落向量之间的余弦相似度，确定最匹配的聚类簇。再从该簇中选取若干最相似的示例段落。最终，将这些示例段落及其人工标注的实体与关系信息按固定格式拼接成 S 模块文本，存储于指定目录，供后续抽取环节调用。

**步骤4： 用更新过后的动态S模块+llm抽取实体和关系**

将构建好的Prompt（保存在prompt存储文件夹中）按顺序提交给大模型，得到结果。

**3.实验结果**

从实验结果来看，llm对于**研究结果**等实体的理解有很大提高，例如paper1中，若未提供针对性的S模块，其输出结果为:

图一：仅使用llm抽取的部分示例

图二：使用主题聚类+llm模型抽取的部分示例

如图所示，图二的可读性和信息丰富程度较1高了很多。