【哥飞解答】为什么 RAG 要用到 Embedding

原创　我是哥飞　哥飞　2024-03-23 18:24:44　广东

大家好，我是哥飞。

上一篇文章《以月访问量2058万的16型性格测试网站为例说说搜索流量的品牌和非品牌区别暨链接好坏判断方式》本来昨天就要发，后来写着写着超过0点了，最后就占用掉了今天的推送额度。

所以今天这篇文章没有通知到所有人，只有打开哥飞公众号才能看到。

欢迎大家关注哥飞公众号，二百多篇原创文章，让你学会出海赚美元。

很多人对于让AI回答问题时，为什么要用到 RAG，以及如何用 Embedding 实现 RAG 有疑问。

在今天的这篇文章中，哥飞想用非技术人也能看懂的方式给大家讲清楚这个问题。

我们知道目前大模型的上下文 Token 长度是有限的，像GPT模型，目前最大的也才 128K 。

MODEL	DESCRIPTION	CONTEXT WINDOW	TRAINING DATA
gpt-4-0125-preview	New **GPT-4 Turbo** The latest GPT-4 model intended to reduce cases of “laziness” where the model doesn’t complete a task. Returns a maximum of 4,096 output tokens. Learn more.	128,000 tokens	Up to Dec 2023
gpt-4-turbo-preview	Currently points to gpt-4-0125-preview.	128,000 tokens	Up to Dec 2023

假设我们想让 GPT 基于指定的一本 30万字的 PDF 来回答某个问题，我们无法一次把全部内容放进去，那就只能挑选一部分放进去。

这个时候就需要去挑选最相关的内容，也就是最有可能跟问题的答案相关的，最有可能能够回答问题的内容。

这个挑选就可以用 RAG 技术实现，RAG 是 Retrieval augmented generation 的简称，中文含义是“检索增强生成”。

传统的挑选可能是用搜索引擎，关键词匹配。

但是关键词匹配无法匹配到语义上相似的。

这时候就需要用到 Embedding，把内容向量化之后，就能够实现语义上的匹配。

内容向量化又是什么意思呢？

拿 OpenAI 的 text-embedding 系列模型来说， text-embedding-3-small 能够把内容变成 1536 维的向量， text-embedding-3-large 则可以变成 3072 维的向量。

维度又是什么意思呢？

我们拿“苹果”这个词来解释，会有各种维度，如：
1. 水果
2. 吃的
3. 树上长的
4. 红色的
5. 青色的
6. 甜的
7. 酸的
8. 名词
9. 植物
……

当然上面的维度只是哥飞按照我们人类能够看懂的方式写出来的，实际是算法根据各种信息计算出来的。

总共会有几千个维度来解释每一个内容，这些维度放在一起，就是一个向量，来描述“苹果”这个信息。

相似的信息，在向量空间里的距离更相近。

如“苹果”和“水果”的距离就比较近，而“苹果”和“猫”的距离就比较远了。

距离越接近，语义上也就越接近。

我们在用户问问题之后，把问题向量化，然后拿着问题的向量去找所有跟问题向量距离很近的内容片段向量。

这样通过匹配，就能找到最有可能能够回答问题的内容，放到上下文里去，提供给大模型作为参考资料，让大模型来回答问题。

要对 30 万字的 PDF 做 RAG，估计有人会以为直接把 30万字扔进去 Embedding 就完事了。

不是这样的，我们需要做预处理，先分为多个片段，之后将每一个片段内容向量化。

分段这里也有一些讲究，有可能还需要让每个片段有重叠部分。

另外具体每一个片段要放多少字也有讲究，不能太长，不能太短。

这里需要根据内容，以及你的使用场景去调整。

然后每次基于问题的向量，找出最可能能够回答问题的多个片段，提交给大模型来总结回答。

为了达到“找出最可能能够回答问题的段落”，我们可以多种方法一起使用，如把传统的搜索也用上。

所以 RAG 并不仅仅只可以用 Embedding，其实可以多种方法组合起来使用。

以上就是关于RAG和 Embedding 向量化的一点小分享，希望对大家有帮助。

不知道讲清楚了没有，欢迎评论区反馈。

关于哥飞社群的介绍，请看《是时候给大家好好介绍一下哥飞的社群了，毕竟刚被二十年站长大佬夸过》。

#AI