

一篇让你搞懂向量 Embeddings 如何使用

原创 我是哥飞 哥飞 2023-10-03 08:00:12 广东

大家好，我是哥飞。

这几天在给社群配套的网站增加搜索能力，目前来讲，最好用的模糊搜索就是将内容向量化之后进行搜索，可以做到语义化搜索。

哥飞之前做了一个 ChatGPT Plugin Store 里的插件中文网站，里边的搜索就用上了向量化搜索。

如下面这个例子，我输入的需求是“制作表情包”，出来的结果是“Meme Generator”，结果里没有出现“制作表情包”五个字中的任何一个字，但依然搜索出来了正确的我们需要的结果。

制作表情包

Search

All

Games

Shopping

Travel

Stock Market

Education

Entertainment

Health

Tools

Law

Marketing

News

Cryptocurrency

Real Estate

Programming

Charity

Medical

Recruitment

Home

Safety

Workplace

Academic

Beauty


Recipes

Religion

Wedding

Entertainment

80%



Meme Generator

一个AI模因生成器，帮助你快速生成有趣的模因图片。

人工智能

表情包

生成器

就是因为讲内容向量化之后，就可以跨语言进行语义化搜索。

那么原理是什么呢？具体如何实现呢？

今天哥飞就给大家讲清楚。

要将内容向量化，就需要用到 OpenAI 的 Embeddings 接口，文档地址和接口介绍网址如下：

- <https://platform.openai.com/docs/guides/embeddings>
- <https://platform.openai.com/docs/api-reference/embeddings>

具体使用也很简单，我们假设有100段要被搜索的文字，那么我们每一段文字都调用一次 Embeddings 接口，每调用一次就得到一组向量，100段文字就得到了100组向量。

```
• curl https://api.openai.com/v1/embeddings \
• -H "Authorization: Bearer $OPENAI_API_KEY" \
• -H "Content-Type: application/json" \
• -d '{
•   "input": "这里是要被搜索的文字",
•   "model": "text-embedding-ada-002"
• }'
```

input 中放入文字，model 目前能用的只有一个，就是 text-embedding-ada-002，价格是 \$0.0001 / 1K tokens，还是很便宜的，几乎不要钱。

返回的结果格式如下：

```
• {
•   "object": "list",
•   "data": [
•     {
•       "object": "embedding",
•       "embedding": [
•         0.0023064255,
•         -0.009327292,
•         .... (1536 floats total for ada-002)
•         -0.0028842222,
•       ],
•       "index": 0
•     }
•   ],
•   "model": "text-embedding-ada-002",
•   "usage": {
•     "prompt_tokens": 8,
•     "total_tokens": 8
•   }
• }
```

data 数组中的第一个数据里的 embedding 数组就是我们得到的向量数组，这个数组长度是1536，也就是目前 OpenAI 的向量维度是 1536 维。

得到向量之后，最简单的我们可以把向量数组存储为一个一个的文本文件，也即是纯文本保存。

更复杂一点的，你可以存储到专门的向量数据库里，OpenAI 官方推荐了一些：

1. Chroma: Chroma 是一个开源的嵌入式存储库。它主要用于存储和检索向量嵌入。
2. Elasticsearch: Elasticsearch 是一个非常受欢迎的搜索/分析引擎，同时也是一个向量数据库。它可以用于全文搜索、结构化搜索和分析，并且支持向量数据的存储和相似性搜索。
3. Milvus: Milvus 是为可扩展的相似性搜索而构建的向量数据库。它提供了高效的大规模向量检索能力。
4. Pinecone: Pinecone 是一个完全托管的向量数据库，用户无需关心底层的维护和管理，可以专注于其应用的开发。
5. Qdrant: Qdrant 是一个向量搜索引擎，专门为高效的向量检索而设计。

6. Redis: 虽然 Redis 主要是一个内存数据结构存储, 但它也可以作为一个向量数据库来使用, 存储和检索向量数据。
7. Typesense: Typesense 是一个快速的开源向量搜索工具, 它提供了简单易用的 API 来进行向量数据的存储和检索。
8. Weaviate: Weaviate 是一个开源的向量搜索引擎, 它支持语义搜索和自然语言查询。
9. Zilliz: Zilliz 是一个数据基础设施, 由 Milvus 提供支持。它提供了一系列的数据解决方案, 包括向量搜索和分析。

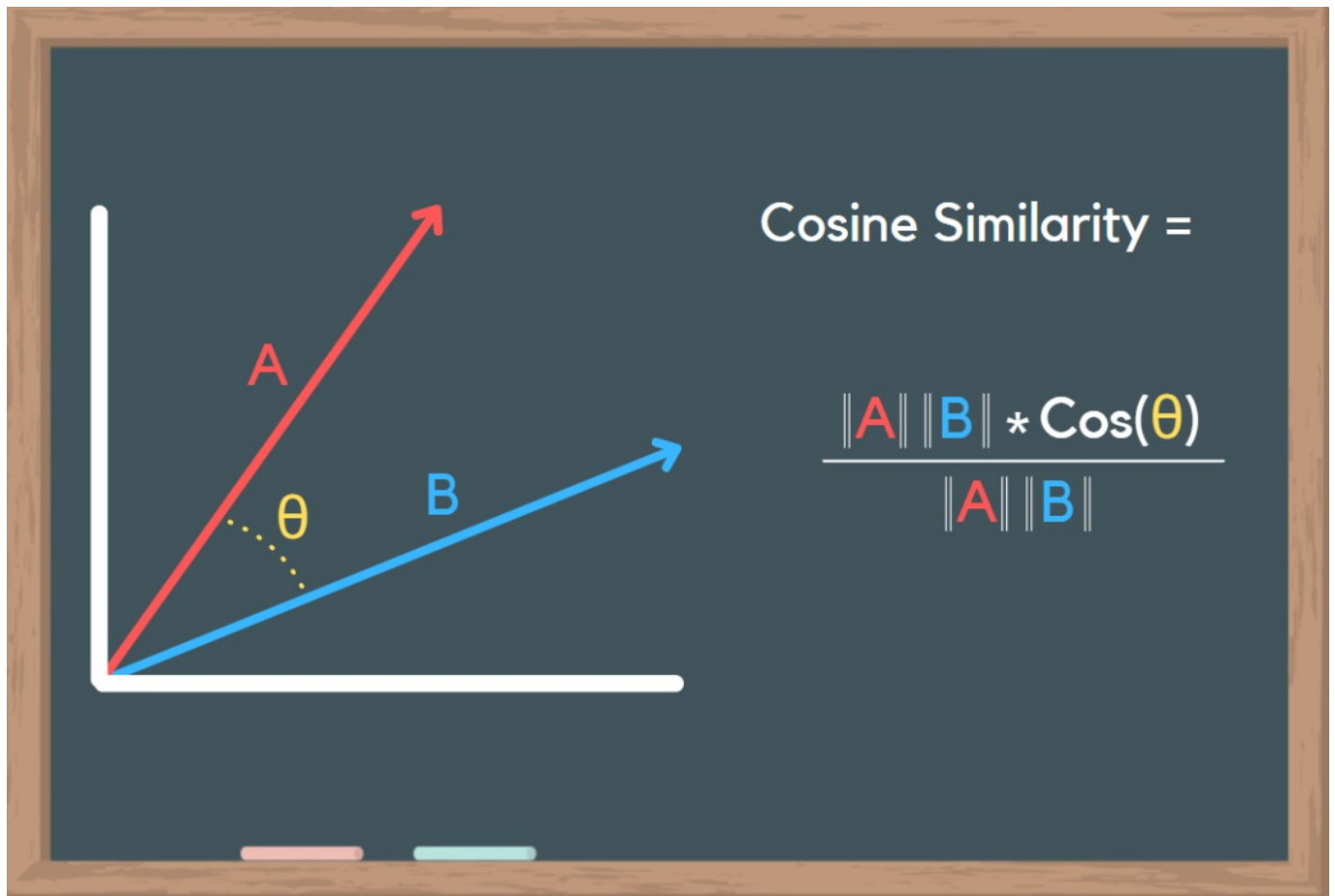
这些向量数据库在不同的应用场景中都有其独特的优势, 大家可以根据自己的需求选择合适的数据库进行使用。

但其实哥飞不推荐新手一上来还没搞懂原理就用这些向量数据库, 大家可以先直接存为文本文件形式, 每次要搜索时, 把向量数组载入内存中进行搜索。

搜索第一步, 先把用户输入的内容也调用 Embeddings 接口得到一个搜索向量数组, 我们命名为向量数组A。

搜索第二步, 把所有待搜索的向量数组都载入内存, 来一个循环, 我们把循环到的每一个待搜索向量数组命名为Bn, 第1个就是B0, 第2个就是B1,, 第100个就是B99。

用A与B0到B99都进行一次余弦相似度计算, 然后得到余弦距离D0到D99, 再从小到大排序, 取出前10作为搜索结果, 余弦距离越小表示越相似。



具体到代码, 哥飞给大家几个PHP函数就知道了:

```
• /**
•  * 计算两个向量的点积。
•  * @param array $vec1 第一个向量
•  * @param array $vec2 第二个向量
•  * @return float 向量的点积
•  */
• function dotProduct($vec1, $vec2) {
•     $result = 0;
•     foreach ($vec1 as $key => $value) {
•         if (isset($vec2[$key])) {
•             $result += $value * $vec2[$key];
•         }
•     }
• }
```

```

    }
    return $result;
}

/**
 * 计算向量的幅度（或长度）。
 * @param array $vec 向量
 * @return float 向量的幅度
 */
function magnitude($vec) {
    return sqrt(dotProduct($vec, $vec));
}

/**
 * 计算两个向量之间的余弦相似度。
 * @param array $vec1 第一个向量
 * @param array $vec2 第二个向量
 * @return float 两个向量之间的余弦相似度
 */
function cosineSimilarity($vec1, $vec2) {
    $v2 = magnitude($vec1) * magnitude($vec2);
    if($v2==0){
        return -1;
    }
    return round(dotProduct($vec1, $vec2) / (magnitude($vec1) * magnitude($vec2)), 2);
}

/**
 * 计算两个向量之间的余弦距离。
 * @param array $vec1 第一个向量
 * @param array $vec2 第二个向量
 * @return float 两个向量之间的余弦距离
 */
function cosineDistance($vec1, $vec2) {
    return round(1 - cosineSimilarity($vec1, $vec2), 2);
}

```

如果大家想要其它语言代码，可以把上面PHP代码给GPT4，让AI帮你生成别的语言代码。

好了，今天的文章就到这里了，大家想看更多文章，可以查看哥飞公众号9月文章一览：

坚持写作三个月，哥飞公众号涨了6000+关注；社群朋友9月份新上的网站从谷歌获得了1万个点击。

从7月2日开始，哥飞还同时运营着一个付费社群，其实聊的内容跟公众号差不多，但会比公众号更细，更深入，并且哥飞作为出海鼓励师，一直在陪伴大家走向成功。

到今天为止，这个价格666元/365天的付费社群已经有了368人加入了。

是什么原因让这么多人选择付费加入哥飞这个付费社群呢？

因为社群干货多，值！

跟着哥飞做海外网站，能够快速拿到结果。

如@Banbri 9月做的一个新网站，仅靠SEO，从谷歌搜索就获得了1万个点击。



哥飞的朋友们(368)



昨天 23:58



Banbri

还有 3 分钟就进入 10 月了，分享一下这个月上的第一个站的数据



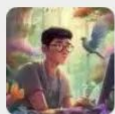
Banbri



Banbri

9 月份新站新词点击量过万 🎉

昨天 23:58



Can

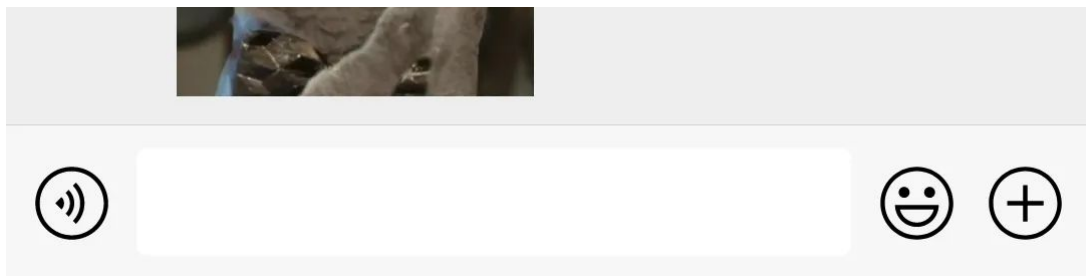


总之，牛



dream Qi





方法就是新词新站，关于新词，其实之前哥飞分享过《找新词：一个永远有效的建站策略，让你快速拿到搜索引擎流量》，大家可以复习一下。

同时社群还有一个配套网站，把群聊历史精华整理成了一个一个的话题，大家可以按话题来浏览群聊内容。

哥飞的朋友们 > 群聊

我发现我的流量都是直接访问，很不健康

#16038 2023.07.31 19:16 16条讨论

@Console.log() 大佬你的 semrush 在哪买的

#16028 2023.07.31 19:15 14条讨论

semrush 也有关键字SEO难度分析工具

#16020 2023.07.31 19:09 8条讨论

开源几款我转的iOS离线模型 中英多语言句子向量 <https://github.com/sinnuswong/Sentence-transformer-Core>

#16014 2023.07.31 18:58 2条讨论



#16013 2023.07.31 18:57 5条讨论



#16010 2023.07.31 18:51 2条讨论

App在桌面的显示名字可以是单字母，但在App Store的上架名字是不可以单字母的

#15998 2023.07.31 18:48 8条讨论



#15993 2023.07.31 18:45 10条讨论

<https://www.similarweb.com/zh/website/w>

#15988 2023.07.31 18:42 12条讨论

DemoChen 07-31 19:16:49

我发现我的流量都是直接访问，很不健康

Console.log() 07-31 19:17:30

说明被传播出去的

AUDI 07-31 19:17:33

我觉得最好的就是先冷启动直接流量没然后能有留存访问，然后在自发传播

AUDI 07-31 19:18:07

对，域名好记，基本就直接输入了

DemoChen 07-31 19:20:47

 redian.news

<https://redian.news> : wxnews

这个网站进去容易，出来难... - Redian新闻

11小时前 — 是时候请出今天的主角了，一个专门收录优质文档的网站：DocHub。DocHub 地址：<https://www.dochub.wiki>。这是上周偶然间在一个帖子里的留言区发现的 ...

DemoChen 07-31 19:20:48

哈哈哈

DemoChen 07-31 19:21:35

AUDI 07-31 19:22:42

自发传播，被动增加外链，这个真好

DemoChen 07-31 19:26:50

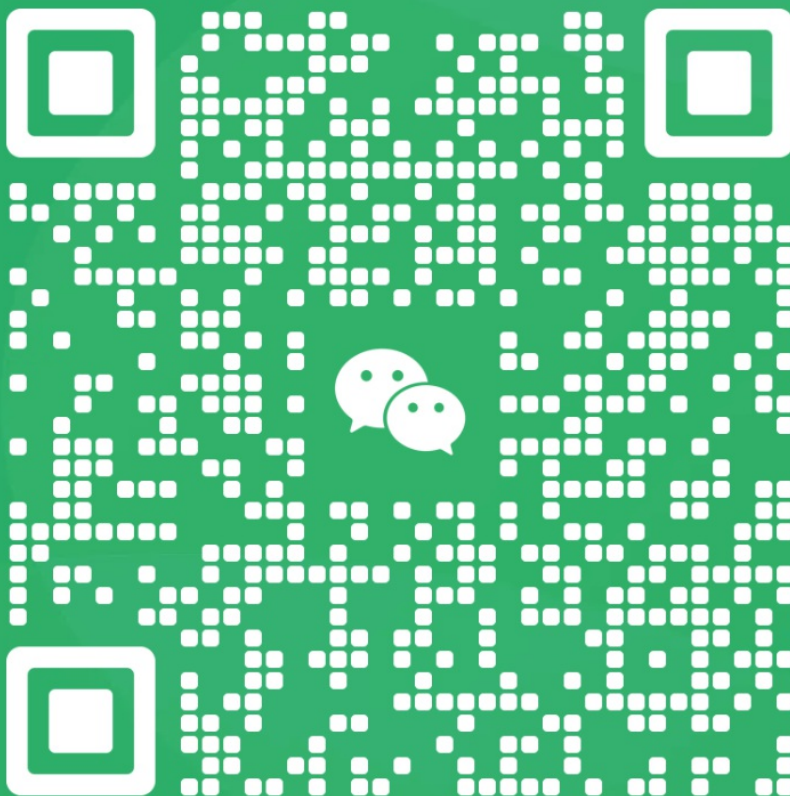
这个网站进去容易，出来难...

如果对社群感兴趣，欢迎加哥飞微信 [qiayue](#) 咨询了解。



Console.log()

广东 深圳



扫一扫上面的二维码图案，加我为朋友。