AFAN的金融科技

# 从word2vec到大模型RAG——基于向量数据库faiss和深度学习的实践

24/09/01 19:00 UTC+8

直播进入/回放见星球→

会员权益如下，快来加入吧：
1、每月至少1次的线上群体直播交流
2、不定期的金融科技专业话题分享

AFAN的金融科技

微信扫码加入星球

Aug

知识星球会员直播

fintech

AFAN的金融科技

# 从零到一搭建
# 金融大模型对话机器人
# 基于Llama3-8B的微调

24/05/17 20:00 UTC+8

直播进入/回放见星球→

会员权益如下，快来加入吧：
1、每月至少1次的线上群体直播交流
2、不定期的金融科技专业话题分享

May

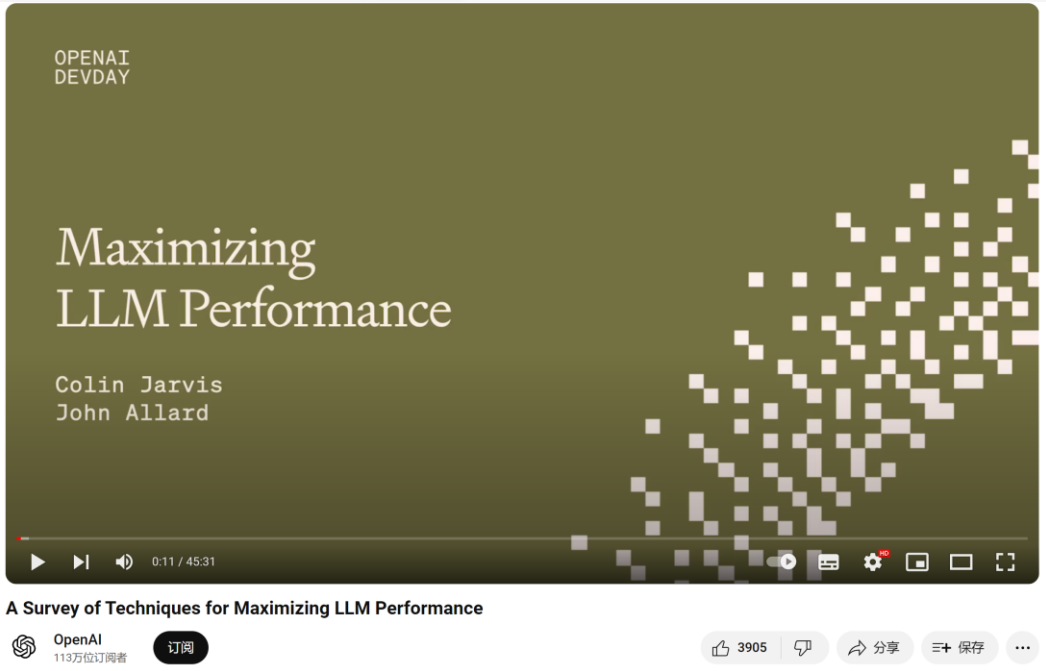
知识星球会员直播

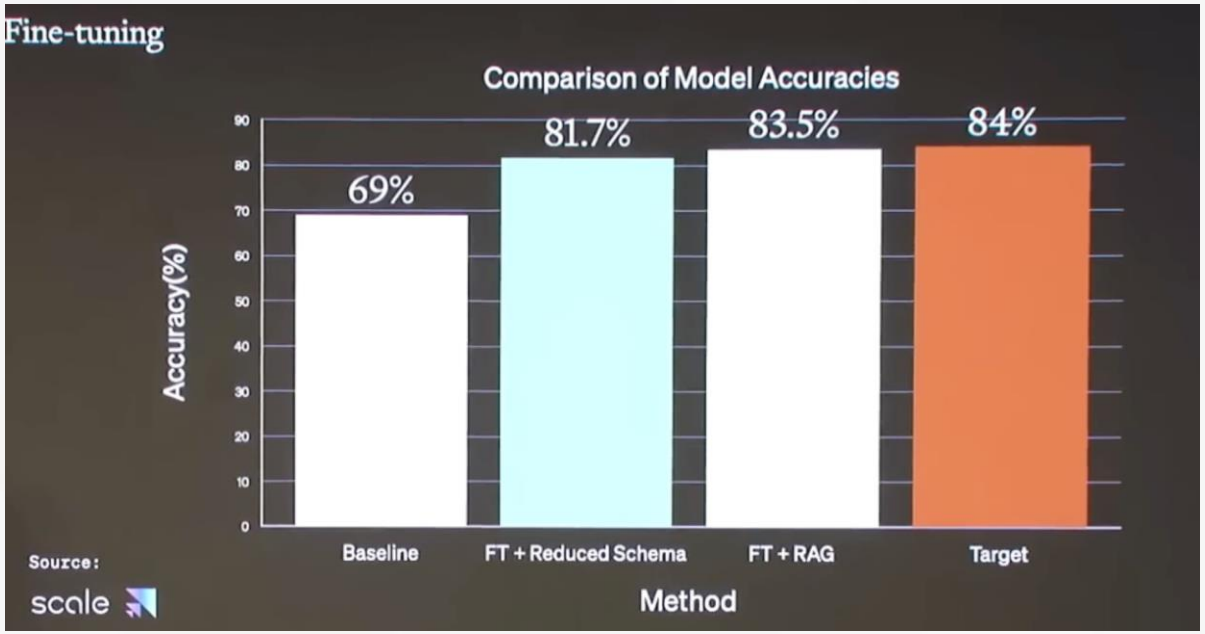
AFAN的金融科技

微信扫码加入星球

# 大模型的几种应用路线

- 23.11.14 OpenAI DevDay分享




原视频：https://www.youtube.com/watch?v=ahnGLM-RC1Y
笔记参考：https://www.53ai.com/news/qianyanjishu/493.html

# 大模型的几种应用路线

AFAN的金融科技

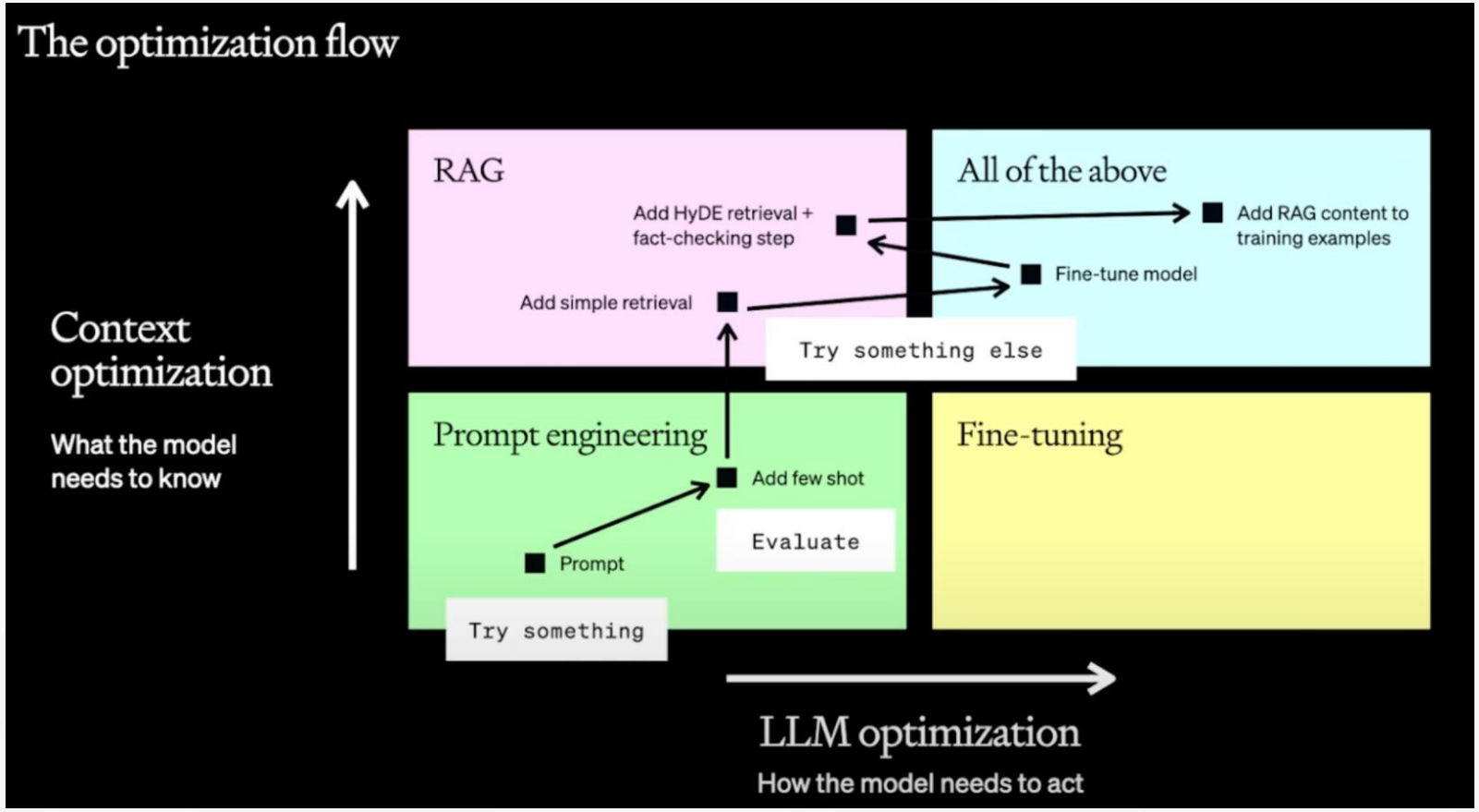
- 大模型优化的2个维度

维度1：模型增强
维度2：知识增强

RAG检索增强生成(Retrieval Augmented Generation)：通过自有垂域数据库检索相关信息，然后合并成为提示模板，给大模型生成漂亮的回答。

VS

Fine-tuning微调指的是为一个已经经过预训练的模型进行额外训练，以使其更好的适应某些特定的任务或数据

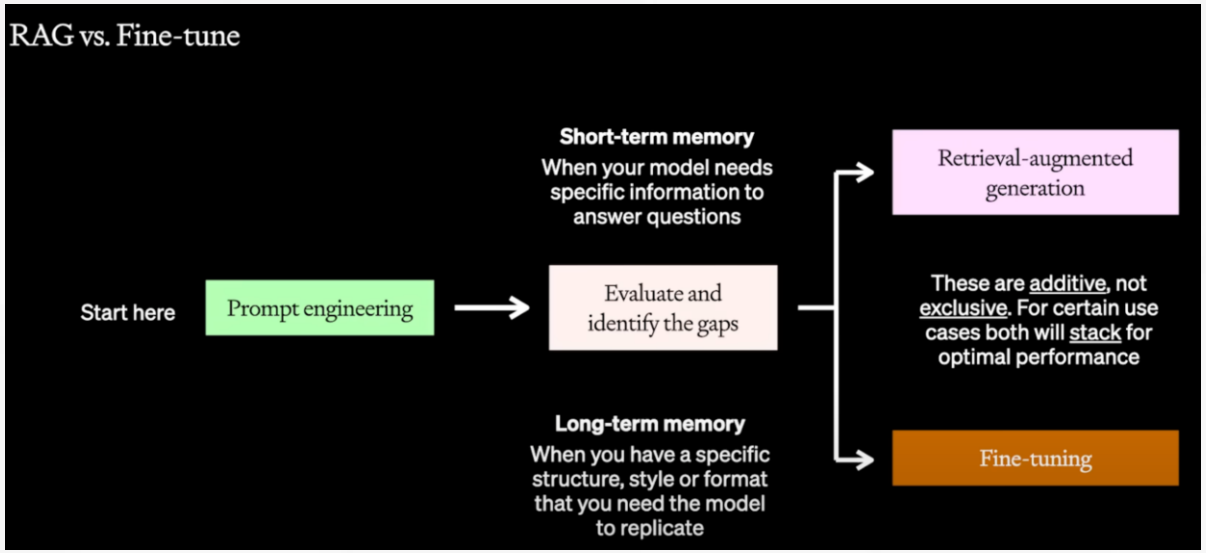


# 大模型的几种应用路线

AFAN的金融科技

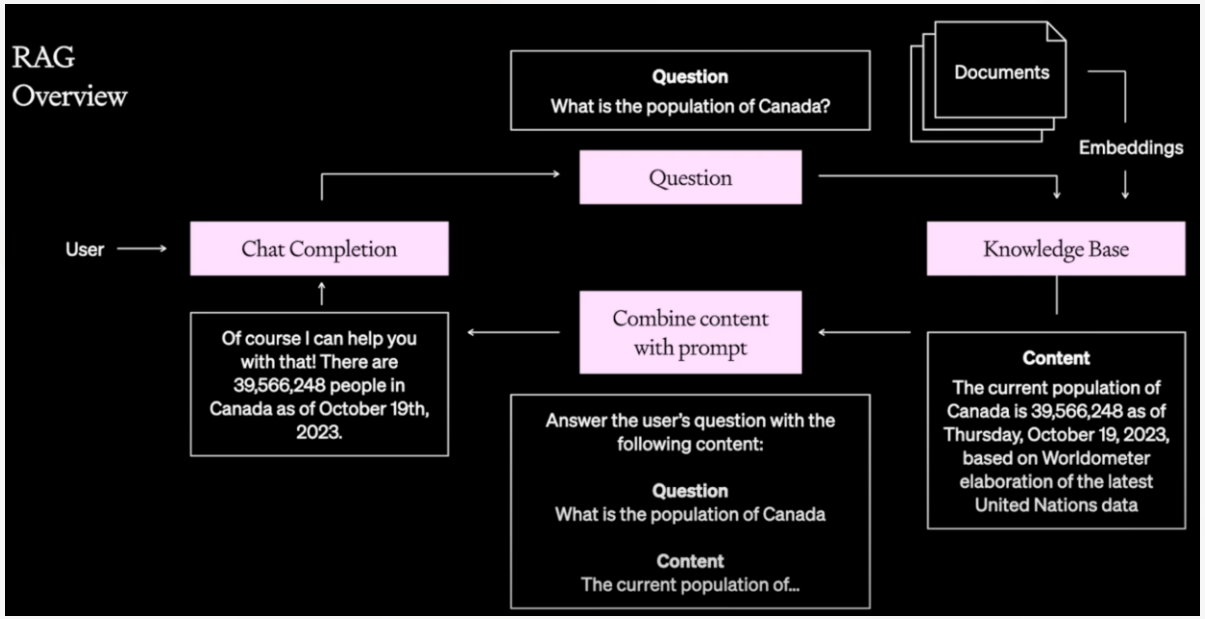
## ➢ FT和RAG的对比

- Prompt：提问和Baseline对比有gap
- RAG：短期记忆（开卷考试）
- FT：长期记忆（知识融会贯通）




## ➢ RAG的大体过程

- 用户先进行交流提问
- 拿问题和知识库先进行匹配
- 将匹配到的结果和回答融合
- 将结果最后进行返回

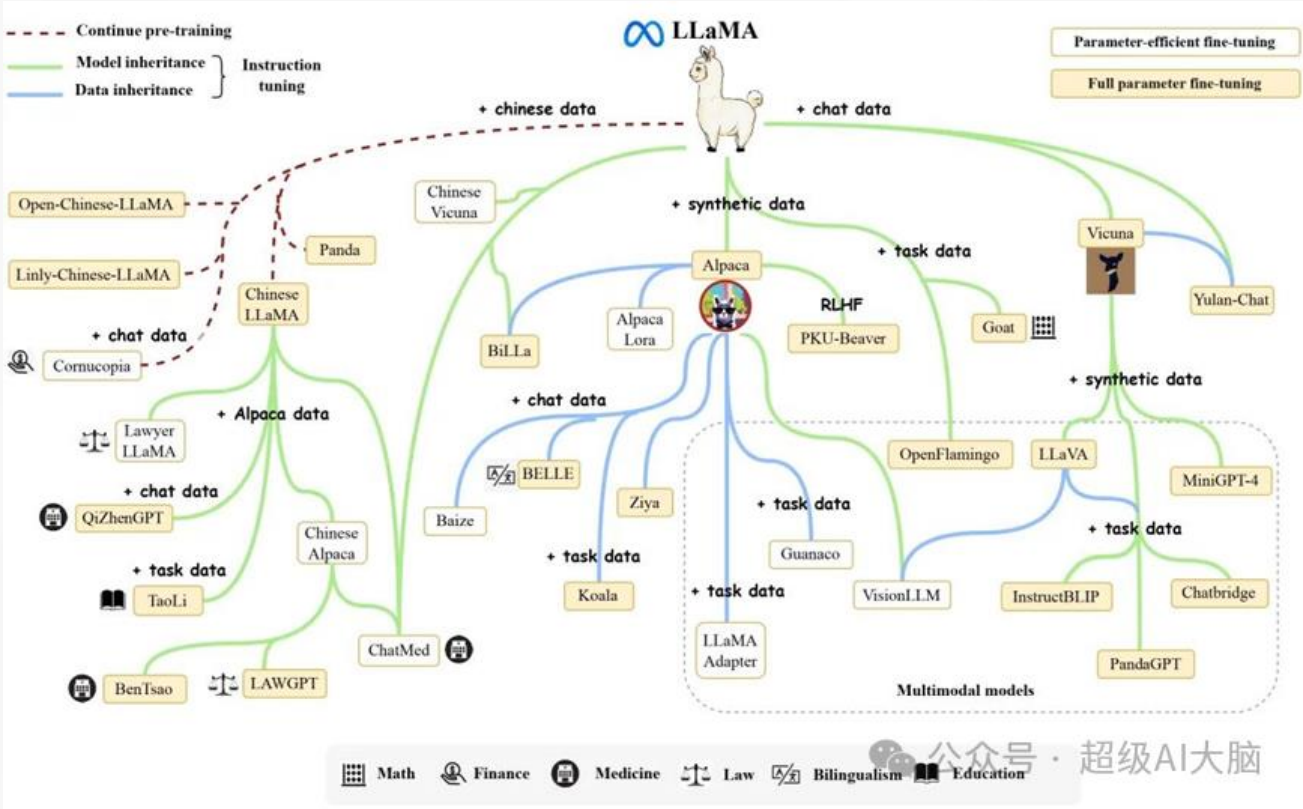


# Llama3介绍

➢ Llama3历史

- Llama-1，作为Meta在2023年2月推出的重磅大语言模型，其性能卓越，在当时便跻身顶尖开源模型之列。该模型提供了7B、13B、30B和65B四个不同参数规模的版本。众多研究者纷纷将Llama作为基座模型，进行了深入的继续预训练或微调工作，由此衍生出了一系列**各具特色的变体模型**
- Meta在2023年7月推出了可供商业使用的免费版本Llama-2，这个版本包含了四个不同参数量的模型，分别是7B、13B、34B和70B。
- 2024年4月，Meta公司宣布了其开源的大型语言模型Llama 3的正式亮相，该模型包括8亿和70亿参数的版本。此外，Meta还透露了正在开发中的400亿参数版本Llama-3的进展。

万字长文梳理Llama开源家族：从Llama-1到Llama-3
https://mp.weixin.qq.com/s/riB8kA_JKkzNepEDVO8xQA



# Llama3介绍

## ➢ Llama3排名

大模型评测社区LMSYS(https://chat.lmsys.org)发布了一份大模型排行榜单，Llama 3位列第五，英文单项与GPT-4并列第一。但中文榜排到了第20

| Rank | Model | Arena Elo | 95% CI | Votes | Organization | License |
|---|---|---|---|---|---|---|
| 1 | GPT-4-Turbo-2024-04-09 | 1259 | +4/-5 | 23823 | OpenAI | Proprietary |
| 1 | GPT-4-1106-preview | 1254 | +3/-3 | 67933 | OpenAI | Proprietary |
| 1 | Claude 3 Opus | 1252 | +3/-3 | 68656 | Anthropic | Proprietary |
| 2 | GPT-4-0125-preview | 1249 | +3/-3 | 56475 | OpenAI | Proprietary |
| 5 | Meta Llama 3 70b Instruct | 1210 | +5/-5 | 12719 | Meta | Llama 3 Community |
| 5 | Bard (Gemini Pro) | 1208 | +6/-6 | 12435 | Google | Proprietary |
| 5 | Claude 3 Sonnet | 1202 | +2/-3 | 70952 | Anthropic | Proprietary |
| 8 | Command R+ | 1192 | +3/-4 | 39243 | Cohere | CC-BY-NC-4.0 |
| 8 | GPT-4-0314 | 1189 | +3/-3 | 46299 | OpenAI | Proprietary |
| 10 | Claude 3 Haiku | 1181 | +3/-3 | 64106 | Anthropic | Proprietary |

Meta Llama 3 is now top-5 in Arena!

中文通用大模型综合性基准SuperCLUE（https://www.superclueai.com/）评估Llama3-70B排名第10，7B量级第一

| 排名 | 模型名称 | 机构 | 总分 |
|---|---|---|---|
| - | GPT-4-Turbo-0125 | OpenAI | 79.13 |
| - | GPT-4-Turbo-0409 | OpenAI | 77.02 |
| - | GPT-4(官网) | OpenAI | 75.32 |
| - | Claude3-Opus | Anthropic | 74.47 |
| 🥇 | Baichuan3 | 百川智能 | 73.32 |
| 🥈 | GLM-4 | 清华&智谱AI | 72.58 |
| 🥈 | 通义千问2.1 | 阿里巴巴 | 72.45 |
| 🥉 | 腾讯Hunyuan-pro | 腾讯 | 72.12 |
| 🥉 | 文心一言4.0 | 百度 | 71.9 |
| 6 | MoonShot(Kimichat) | 月之暗面 | 70.42 |
| 6 | 从容大模型V1.5 | 云从科技 | 70.35 |
| 6 | MiniMax-abab6.1 | 稀宇科技 | 70.18 |
| 9 | 山海大模型 | 云知声 | 69.51 |
| 9 | 讯飞星火V3.5 | 科大讯飞 | 69.43 |
| - | Llama-3-70B-Instruct(poe) | Meta | 68.77 |

SuperCLUE7B量级排行榜（2024年4月）

| 排名 | 模型名称 | 机构 | 总分 |
|---|---|---|---|
| - | Llama-3-8B-Instruct | Meta | 57.44 |
| 🥇 | qwen-1.5-7B-Chat | 阿里巴巴 | 55.49 |
| 🥈 | ChatGLM3-6B | 清华&智谱AI | 50.6 |
| - | Gemma-7b-it | Google | 44.46 |
| - | Llama2-7B-Chat | Meta | 40.17 |

# Llama3介绍

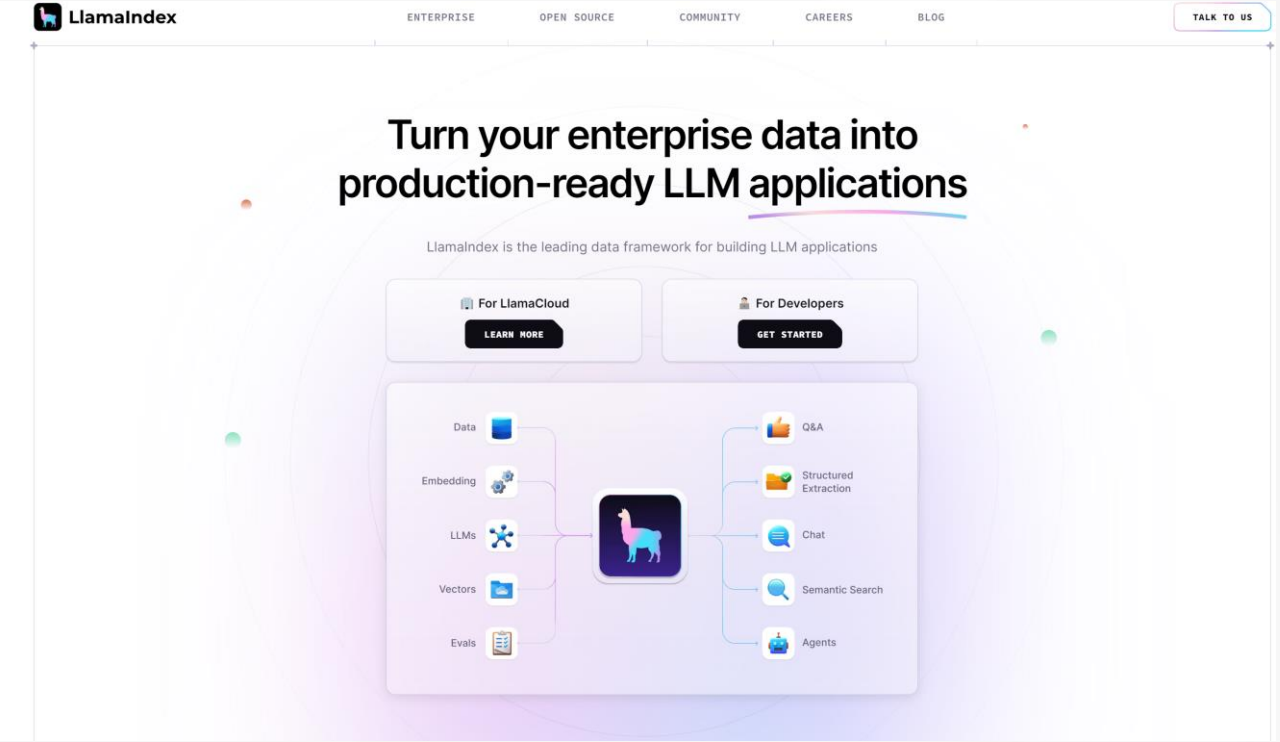
AFAN的金融科技

- 部署参考


Windows下中文微调Llama3，单卡8G显存只需5分钟，可接入GPT4All、Ollam...

UP AI百晓生 · 4-27

- 下一步RAG尝试



AFAN的金融科技

# 从word2vec到大模型RAG——基于向量数据库faiss和深度学习的实践

24/09/01 19:00 UTC+8

直播进入/回放见星球→

会员权益如下，快来加入吧：
1、每月至少1次的线上群体直播交流
2、不定期的金融科技专业话题分享

AFAN的金融科技

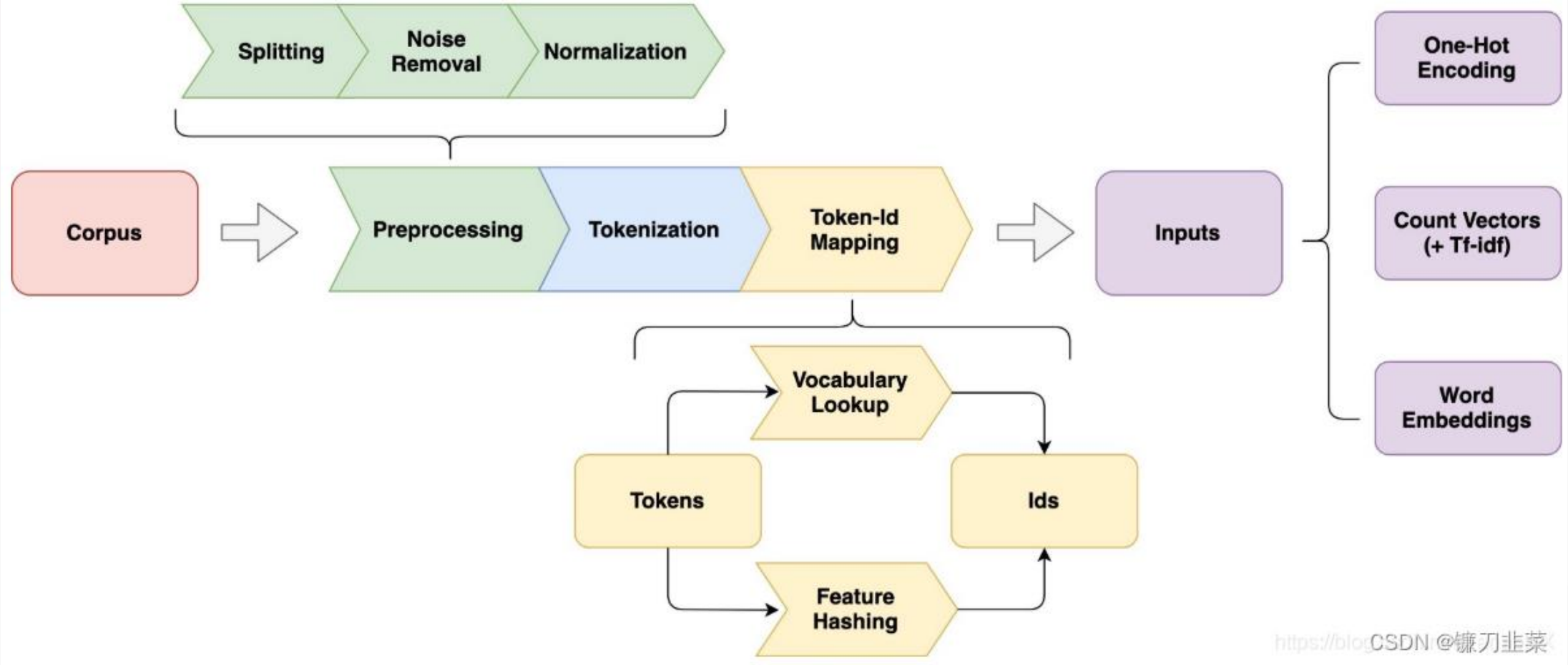
微信扫码加入星球

Aug

知识星球会员直播

# 什么是词向量?

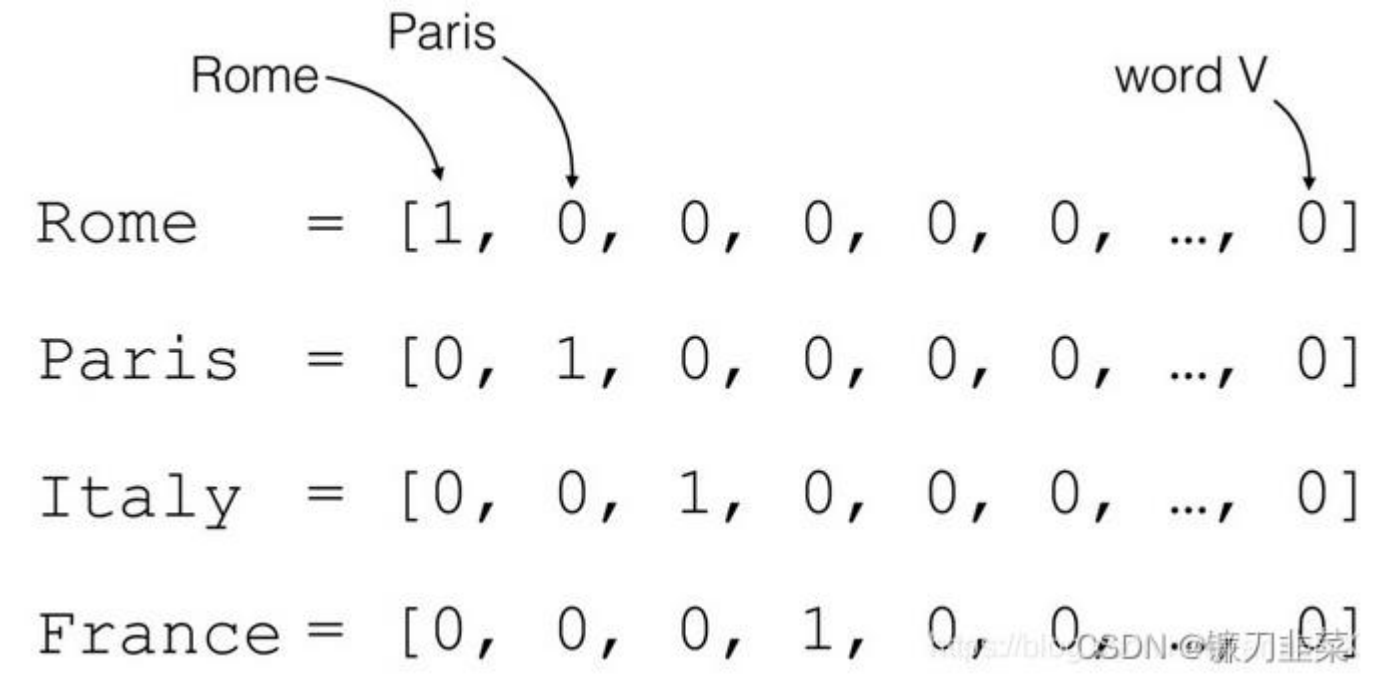
- 计算机理解单词的方式



# 什么是词向量?

- **One-hot编码**

  - 单词的上下文丢失了。
  - 没有考虑频率信息。
  - 词汇量大的情况下，向量维度高且稀疏，占用内存。



# 什么是词向量?



- **TF（Term Frequency）**

核心思想：**拥有相似的Context 上下文的词的语义是相似的，对应的词向量也是相似的**

- Couting-based Approach 如果单词在相同的上下文出现的次数越多，那就说明它们越相似。

| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ |
|---|---|---|---|---|---|
| ekonomi | 0 | 1 | 40 | 38 | 1 |
| pusing | 4 | 5 | 1 | 3 | 30 |
| keuangan | 1 | 2 | 30 | 25 | 2 |
| sakit | 4 | 6 | 0 | 4 | 25 |
| Inflasi | 8 | 1 | 15 | 14 | 1 |



水平表示文档，垂直表示单词，表示某单词在某文档中的出现次数

# 什么是词向量?

- **TF-IDF（Term Frequency - Inverse Document Frequency）**

  给了TF中频繁词的权重惩罚

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

**TF-IDF**

Term $x$ within document $y$

$tf_{x,y}$ = frequency of $x$ in $y$

$df_x$ = number of documents containing $x$

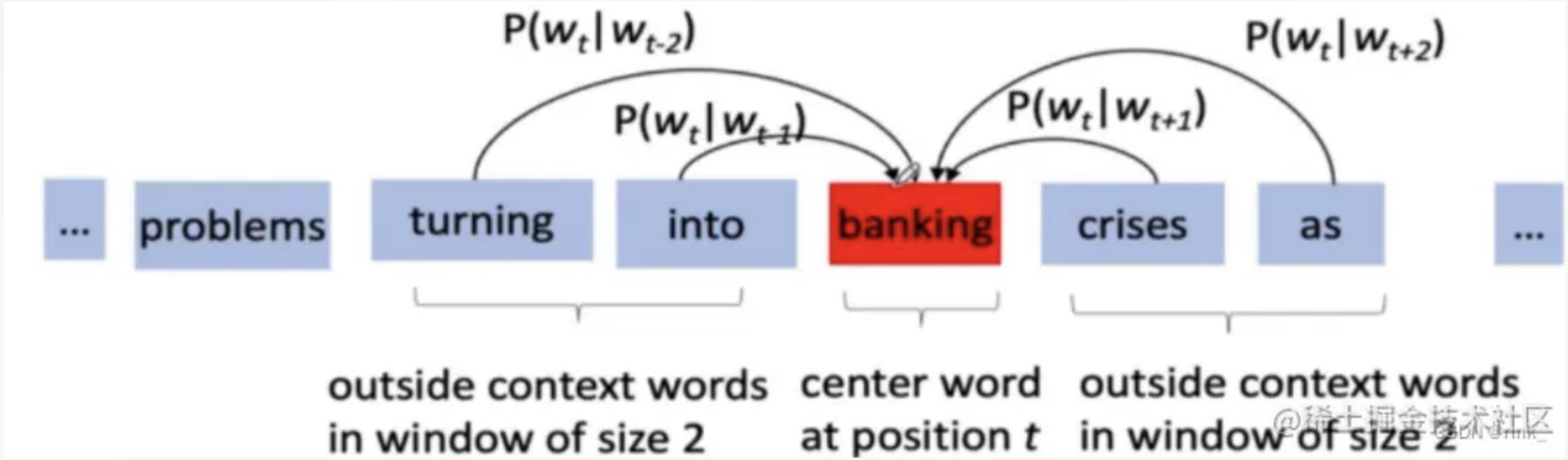
$N$ = total number of documents

# 什么是词向量?

- **word2vec：word to vector**

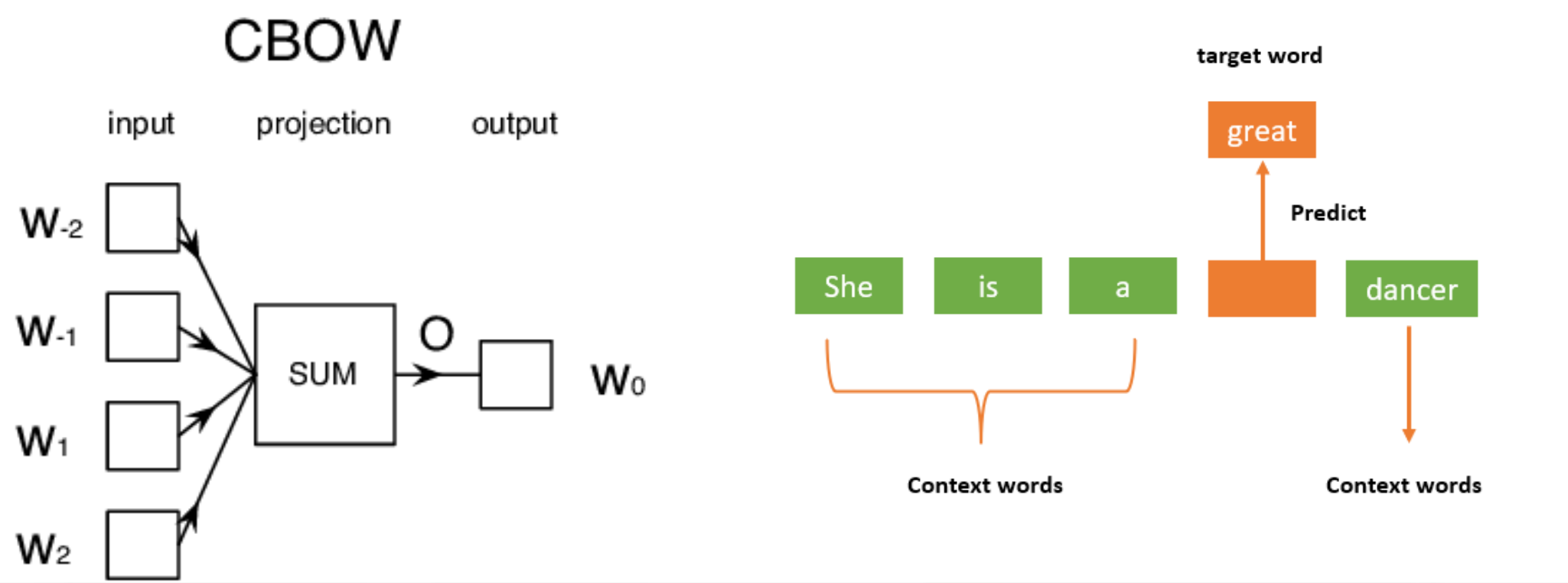
核心思想：**拥有相似的Context 上下文的词的语义是相似的，对应的词向量也是相似的**

- Prediction-based Approach 输入上下文，预测一个单词，预测的这个单词肯定是符合这个Context的。



# 什么是词向量?

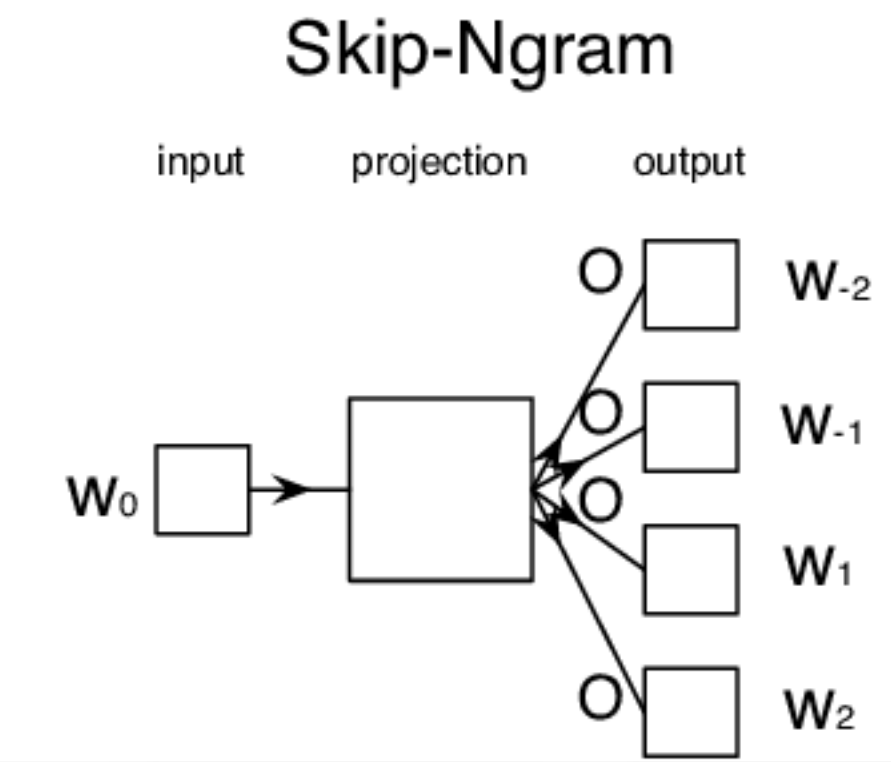
- 两种训练方式——CBOW

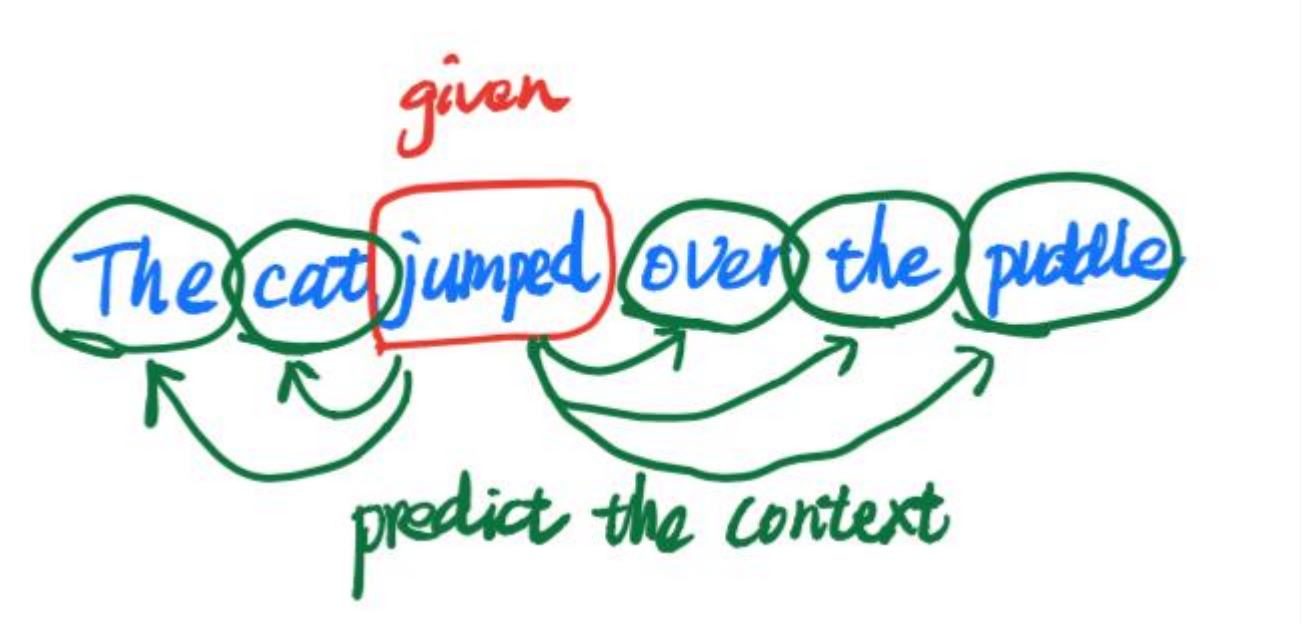



基于周围词预测中间

# 什么是词向量?

- 两种训练方式——Skip-gram



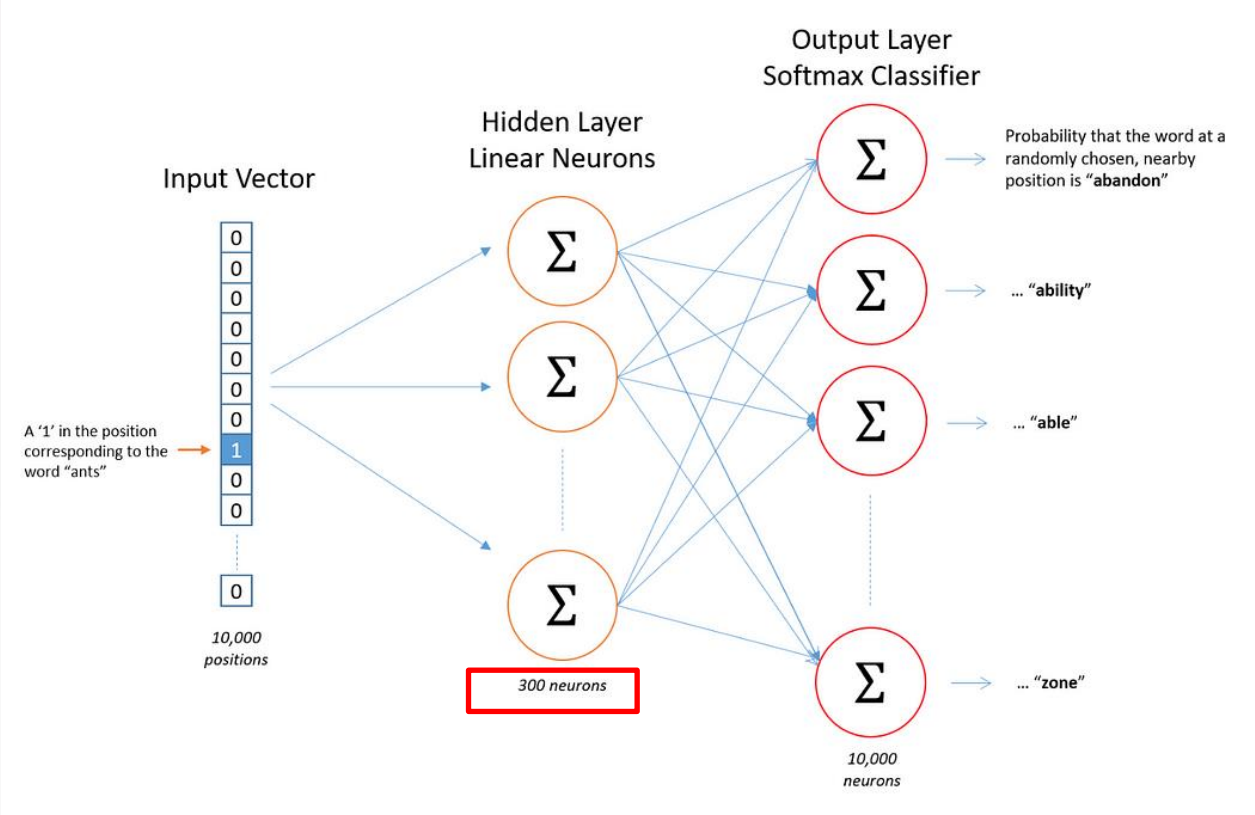



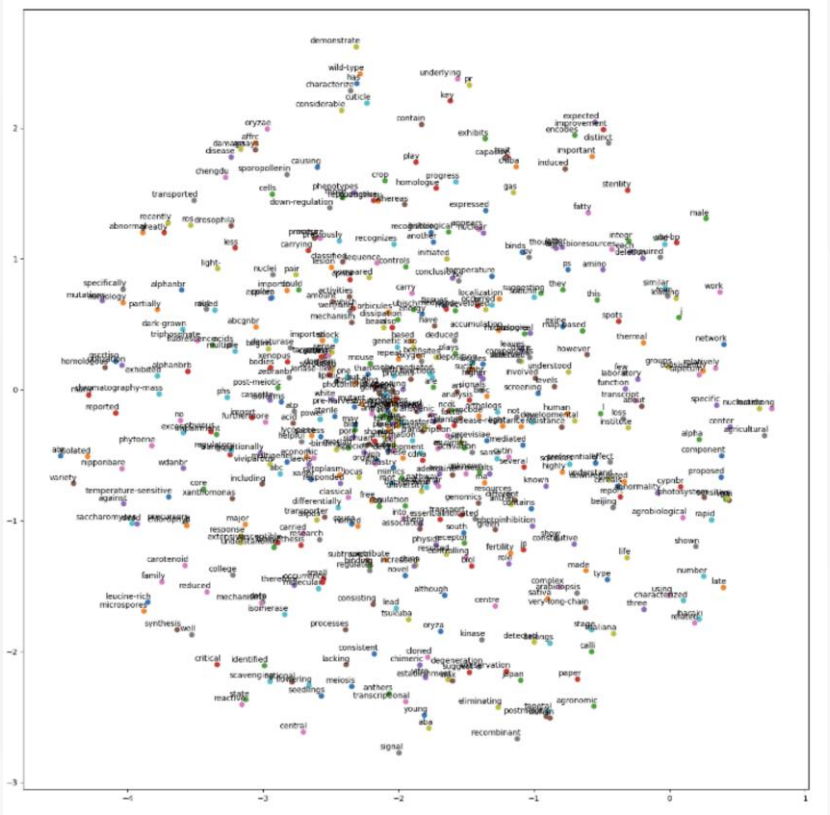

基于中间词预测周围

# 什么是词向量?

AFAN的金融科技

## ➢ 高维向量产生原理



中间层有300个神经元组成了300维向量

关联性更高的单词会映射到更近的位置

# 什么是词向量?

- **总结：让计算机理解词**

**人类眼中**：有真实世界映射的文字符号

| i | want | to | eat | chinese | food | lunch | spend |
|---|---|---|---|---|---|---|---|

**计算机眼中**：数字序号+字符串字典

0 1 2 3 4 5 6 7

{

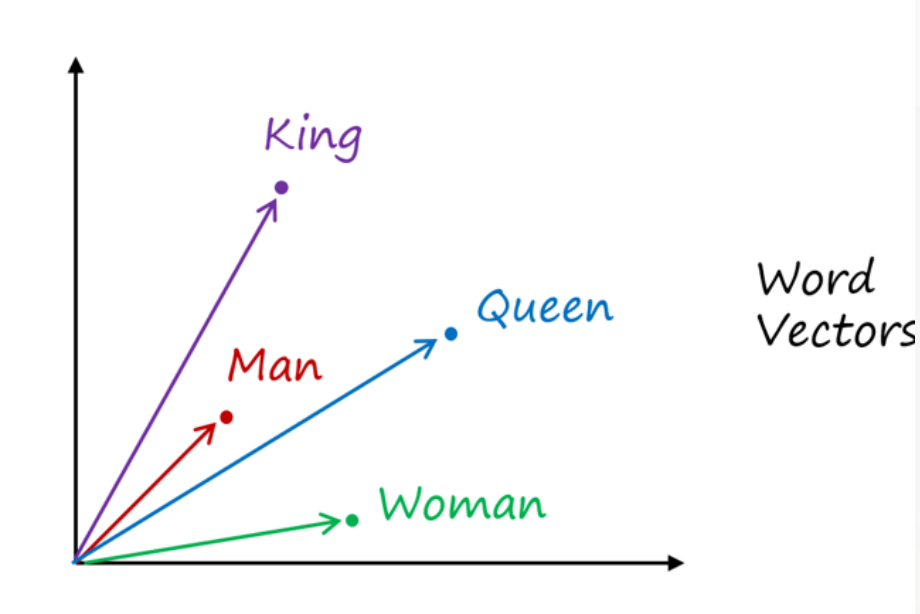
0 : I, 1: want, 2: to, 3: eat, 4: Chinese …

}

**传统统计编码**

**人类眼中**：King和Queen都很有权势，最大区别是性别

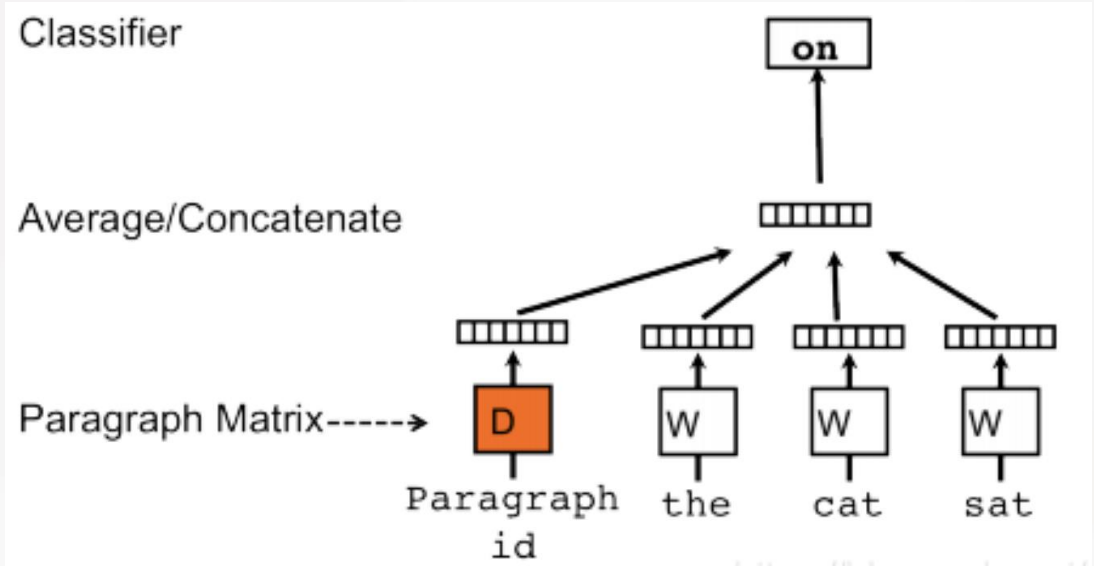
**计算机眼中**：4个有多维度的数据点，King – Queen = Man – Woman




**Word2Vec编码**

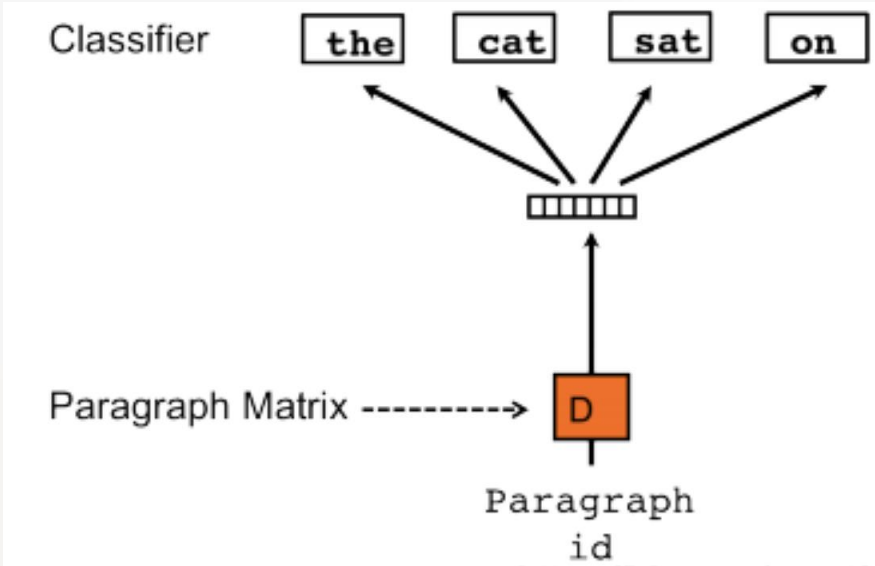
# 什么是句向量?

- 两种训练方式




**句向量的分布记忆模型（PV-DM）**

在给定上下文和文档向量的情况下预测单词的概率




**句向量的分布词袋（PV-DBOW）**

在给定文档向量的情况下预测文档中一组随机抽样的单词的概率。

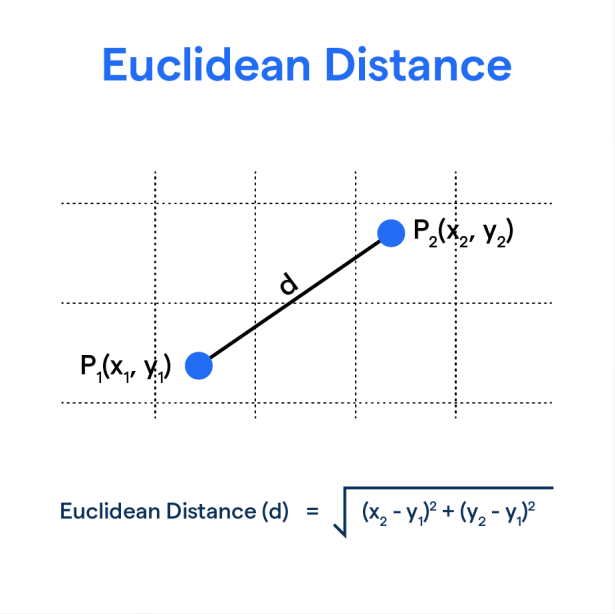
# 什么是句向量匹配

AFAN的金融科技

- Faiss向量库匹配


Faiss的全称是Facebook AI Similarity Search,是FaceBook的AI团队针对大规模相似度检索问题开发的一个工具,使用C++编写,有python接口,对10亿量级的索引可以做到毫秒级检索的性能。

**Faiss的相似度搜索是在多维向量空间中寻找**




**本质上是欧几里得空间距离**

# 什么是句向量匹配

- Demo案例匹配效果展示

```
query = "农业保险"
```

```
docs = [
    # 句子1，句意有农业保险
    '实施金融支持农业纾难解困政策。鼓励市县积极开展特色渔业养殖保险，在参保农户自缴保费比例不低于20%的前提下，对险种绩效评价结果达标的市县，省级财政按'
    '照25%的比例给予保费补贴。对2022年8月至12月到期的农民小额贷款和新发生的农民小额贷款贴息由5%提升至6%。对脱贫人口小额信贷，允许其调整还本计划或办'
    '理贷款展期、续贷。对受疫情影响暂时出现还贷困难的涉农企业及农户(包括脱贫户、监测户)，支持银行机构按市场化原则予以降息、减息或免息扶持，开展征信保护等。',
    # 句子2，句意无农业保险
    '分级分类开展社会化服务。针对中高风险区农业生产人员无法外出生产问题，组织有关企业和社会化服务组织提供托管、代耕、代收服务。各村委会统计当地需要种植'
    '或收获的作物品种、面积、产量，乡镇政府商请当地农业农村局统一协调专业化服务组织提供托管服务。当地力量不足时，市县农业农村局向省农业农村厅申请统一协调安排',
    # 句子3，句子中有农业保险
    '财政支持。各级财政部门履行牵头主责，会同有关部门从发展方向、制度设计、政策制定、资金保障等方面推进农业保险发展，通过保费补贴、机构遴选等多种政策手'
    '段，发挥农业保险机制性工具作用，督促承保机构依法合规展业，充分调动各参与方积极性，推动农业保险高质量发展。'
]
```

有明确提到农业保险的效果最好

句意次之

```
page_content='财政支持。各级财政部门履行牵头主责，会同有关部门从发展方向、制度设计、政策制定、资金保障等方面推进农业保险发展，通过保费补贴、机构遴选等多种政策手段，发挥农业保险机制性工具作用，督促承保机构依法合规展业，充分调动各参与方积极性，推动农业保险高质量发展。'
0.34435654
page_content='实施金融支持农业纾难解困政策。鼓励市县积极开展特色渔业养殖保险，在参保农户自缴保费比例不低于20%的前提下，对险种绩效评价结果达标的市县，省级财政按照25%的比例给予保费补贴。对2022年8月至12月到期的农民小额贷款和新发生的农民小额贷款贴息由5%提升至6%。对脱贫人口小额信贷，允许其调整还本计划或办理贷款展期、续贷。对受疫情影响暂时出现还贷困难的涉农企业及农户(包括脱贫户、监测户)，支持银行机构按市场化原则予以降息、减息或免息扶持，开展征信保护等。'
0.4557104
page_content='分级分类开展社会化服务。针对中高风险区农业生产人员无法外出生产问题，组织有关企业和社会化服务组织提供托管、代耕、代收服务。各村委会统计当地需要种植或收获的作物品种、面积、产量，乡镇政府商请当地农业农村局统一协调专业化服务组织提供托管服务。当地力量不足时，市县农业农村局向省农业农村厅申请统一协调安排'
0.520805
```