|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 原理 | 优点 | 缺点 |
| 正则表达式 (Regular Expression, Regex) | 精确控制：对字符串的表面模式（如电话号码、邮箱地址、特定关键词的出现）可以进行极高精度的匹配。  高效快速：匹配过程非常高效，尤其在处理大规模文本时，速度远胜于复杂的模型。  通用性强：几乎所有编程语言和文本编辑器都支持，是工程师的必备工具。 | 精确控制：对字符串的表面模式（如电话号码、邮箱地址、特定关键词的出现）可以进行极高精度的匹配。  高效快速：匹配过程非常高效，尤其在处理大规模文本时，速度远胜于复杂的模型。  通用性强：几乎所有编程语言和文本编辑器都支持，是工程师的必备工具。 | 僵硬、无语义理解：最大的缺点。它只能匹配它“看到”的字符模式，完全无法理解文本背后的含义。例如，它无法知道“苹果”公司和一个水果“苹果”之间的区别。  难以维护：复杂的正则表达式公式非常难以编写和阅读，被称为“天书”。  规则穷举困难：一种含义可能有多种表达方式（如“你好”和“您好”），需要手动编写所有可能的情况，容易遗漏。。 |
| TF-IDF (Term Frequency-Inverse Document Frequency) | 一种用于信息检索和文本挖掘的统计方法。它将文本转换为一个数值向量（一串数字），以便计算机能够处理和比较。  TF (词频)：一个词在当前文档中出现的频率。频率越高，说明该词对当前文档越重要。  IDF (逆文档频率)：一个词在所有文档（整个文档集/语料库）中出现的频率。频率越高，说明该词越常见、越普通（如“的”、“是”），区分能力就越差。  TF-IDF = TF \* IDF：最终值越高，表示该词对当前文档越具有代表性，越能区分该文档和其他文档。 | 统计直观：原理简单，易于理解和实现。  一定程度上的重要性衡量：能够过滤掉常见停用词，突出文档中的关键词。  可解释性：可以清楚地看到是哪些关键词的权重高导致了文档的相似性或检索结果。 | 词袋模型 (Bag-of-Words)：完全忽略词语的顺序、语法和上下文信息。“狗咬人”和“人咬狗”的向量表示是完全一样的。  语义稀疏：无法捕捉词语之间的语义关系（如同义词、多义词）。“电脑”和“计算机”被视为两个完全不同的词。  依赖语料库：IDF权重高度依赖于预先选择的语料库，语料库选择不当会导致效果变差。 |
| BERT (Bidirectional Encoder Representations from Transformers) | 基于Transformer架构的深度双向预训练语言模型。它的核心思想是：  预训练 (Pre-training)：在海量无标注文本上，通过两个自监督任务（掩码语言模型 MLM 和下一句预测 NSP）进行训练，学习语言的内在规律和知识。这个过程让它获得了强大的语言理解能力。  微调 (Fine-tuning)：在预训练好的模型基础上，针对特定的下游任务（如情感分析、问答系统）用少量标注数据进行微调，即可获得极好的性能。  双向上下文：与传统模型从左到右或从右到左分析文本不同，BERT能同时利用一个词左右两侧的上下文信息来生成该词的表示，从而更深刻地理解词义。 | 深度上下文语义：彻底解决了多义词问题，能够根据上下文生成动态的词向量表示。例如，“苹果很甜”和“苹果发布了新手机”中的“苹果”会得到不同的向量。  强大的泛化能力：预训练模型包含了丰富的语言知识，通过微调可以轻松适配多种NLP任务，并大幅提升各项基准测试的纪录。  特征提取能力强：即使不微调，仅使用BERT生成的词向量作为特征输入给其他模型，也能取得很好的效果。 | 计算资源消耗大：模型参数量巨大（数亿甚至数亿），训练和推理需要大量的GPU计算资源和时间。  黑盒模型：模型决策过程难以解释，不像TF-IDF那样直观。  静态微调：微调后的模型参数是固定的，无法在推理时动态适应新的信息。 |
| 大语言模型 (Large Language Model, LLM) | 如GPT系列、LLaMA等，是比BERT规模更大（参数达千亿甚至万亿）、训练数据更多（覆盖整个互联网文本、代码、书籍等）的自回归或自编码语言模型。它们通常也基于Transformer架构，但训练目标和规模发生了质变。  核心任务是生成：根据给定的上文（提示），以概率预测下一个最可能的词/ token，并循环往复地生成连贯的文本。  涌现能力 (Emergent Ability)：当模型规模超过某个阈值后，会产生小模型不具备的能力，如复杂的推理、思维链、代码生成等。  情境学习 (In-Context Learning)：无需更新模型参数，仅通过几个示例（Few-shot）或任务描述（Zero-shot）就能让模型理解并执行新任务。 | 通用性与生成能力：不再是单一功能的模型，而是一个“万能”的文本理解和生成基础。可以对话、翻译、总结、创作、编程等，几乎无所不能。  强大的推理和泛化能力：能够进行一定程度的逻辑推理、知识运用和类比，处理未知任务的能力极强。  自然的人机交互：通过提示词（Prompt）即可与之交互，门槛极低。 | 巨大的计算和资源消耗：训练成本高达数百万美元，需要庞大的算力基础设施，碳排放高。  “幻觉” (Hallucination)：会生成看似合理但实际不正确或不存在的信息，这是目前最棘手的问题。  偏见与毒性：会放大训练数据中存在的社会偏见、歧视性言论和错误观点。  知识截止：模型的知识仅限于其训练数据截止的日期，无法获取新知识（除非通过后续技术如RAG）。 |