1. 标题：向量化技术在文本相似度评分中的应用与比较研究

此段落应该属于上面模版的“结果与讨论”部分。

改写后的文章：

在处理导师信息长文本和用户短句输入的向量化过程中，我们可以探索使用不同的技术。对于长篇导师信息，可应用词袋模型（如CountVectorizer或TF-IDF），该技术将文本转化为基于词频或TF-IDF权重的向量形式。此向量化方式适用于长篇文本，虽然不考虑词序和语义关系，但可有效捕获文档内词语的重要性。而对于用户的短句输入，我们可以考虑使用词嵌入（如Word2Vec、GloVe或FastText）进行向量化，预训练的词嵌入模型可将词语映射为低维度的连续向量，这种向量的位置和方向编码了单词的语义和上下文信息。

尽管在初次尝试使用CountVectorizer时，固定导师如Sharib Ali和Noorhan Abbas得出的分数较高，这是由于CountVectorizer基于词频进行向量化，文本内容丰富的导师关键词较多，因此他们常会被优先推荐。这在某种程度上合理，因为当某项技术或背景被导师频繁提及，用户在搜索此项技术时应获得更高的匹配分数。

然而，使用预训练的词向量模型（如 Word2Vec 或 FastText）可能更能捕获词语间的语义关系，这些模型通过大规模文本数据训练得到，能将相似语义的词语映射到向量空间中的相似位置，从而在计算相似度时更准确地捕获词语间的语义相似性。

我们的实验结果表明，使用预训练的词嵌入模型可能面临词汇不匹配问题，例如你的数据集可能包含一些预训练模型的词汇表中不存在的词汇，尤其是一些专有名词、短语或新词。此外，Word2Vec模型可能由于上下文的理解导致了对一些计算机领域的单词误解，因此可能推荐出与用户输入不符的导师。

通过实验发现，使用Word2Vec的模型有时可能会将一些描述简短的导师如amy lowe作为software engineering and web application的首选推荐，原因是她在个人描述中使用了大量的与这些技术相关的词语。然而，与此同时，这个模型可能会忽视一些专门研究这些技术的导师。这一现象凸显出词嵌入模型在处理特定领域专业词汇时可能存在的挑战。

此外，我们发现在某些情况下，TF-IDF模型和CountVectorizer模型可能会过度优先考虑在长篇文本中出现频率较高的词汇，从而忽视了用户输入中一些关键词的重要性。例如，当用户查询“机器学习”时，这些模型可能会偏向于推荐在描述中频繁提及“学习”这个词的导师，而忽视了“机器学习”作为一个整体的重要性。

最后，虽然词嵌入模型如Word2Vec和FastText在捕获词汇间的语义关系上具有优势，但这些模型对文本的理解是基于单词级别的，对于理解整个句子或段落的语义信息仍然存在局限。而BERT等基于Transformer的模型则可以更好地理解句子和段落的整体语义，从而在某些情况下可能提供更准确的推荐结果。

总的来说，不同的向量化技术都有其独特的优点和局限，选择哪种技术取决于具体的应用场景和需求。在我们的研究中，我们发现结合使用不同的技术可能是一个有效的策略，例如在处理长篇文本时使用基于词频的向量化技术，而在处理用户的短句输入时使用词嵌入模型。同时，基于Transformer的模型在理解整体语义信息方面的优势也值得我们进一步探索和利用。

1. 标题：文本清理方法及其在导师推荐系统中的应用

本研究探讨了两种文本清理策略的设计与应用，特别是在导师推荐系统中的效果。首先，我们将这两种策略简述如下：

策略一：基础文本清理

策略二：在策略一的基础上，进一步使用spacy工具对文本进行过滤，仅保留名词、形容词、副词和专有名词。

我们最初预设的假设是，文本清理应对CountVectorizer有所影响，因为CountVectorizer主要关注词语的出现频率，而对词语在各文档中的分布并不敏感。然而，实验结果显示，即使清理了主语、动词和连词等，其结果与未进行清理的差异并不显著。通过详细分析，我们发现，当导师信息和用户输入都采用CountVectorizer时，可以获得最佳效果，这表明CountVectorizer更适合处理具有特定主题的文档。

另一方面，我们也考虑了TF-IDF（词频-逆文档频率）的向量化方法。TF-IDF通过惩罚在大量文档中频繁出现的词语，有助于突出那些更具区分能力的词语。在清理过程中，我们并未直接删除词语，因此TF-IDF提供了一种更纯净的方法。实验结果显示，TF-IDF方法的效果与CountVectorizer相当。

此外，我们设定了一项实验来验证各种推荐算法的有效性，其中包括自训练的Word2Vec、Doc2Vec，以及Google Word2Vec等。在此实验中，我们以Dr. Ban Al Jassani的相关内容为输入，希望推荐算法能将其作为首选推荐项。实验结果显示，Jaccard相似度方法，CountVectorizer和TF-IDF都成功将Dr. Ban Al Jassani列为首选，而自训练的Word2Vec和Doc2Vec并未成功做到，Google Word2Vec将其排在第三。

此外，我们也发现，以词频为基础的向量化推荐系统在识别同一领域的具体词语时存在困难。例如，无法分辨"Visual Basic"和"Visual C++"之间的差异。尽管有语义向量化的推荐系统进行辅助，但这类问题仍无法完全解决。我们发现最优秀的Word2Vec模型是由Google使用新闻文本进行训练的，而如果能获得基于计算机领域专门训

练的模型，我们就可能能解决这个问题。

综上，本研究表明，各种向量化技术在导师推荐系统中的应用有其优点和限制，选择哪种技术可能取决于特定的应用场景和数据集。

这段文本应该属于研究模板的"实验设计与结果"部分。

1. 标题：比较文本生成技术与主题建模策略在导师推荐系统中的应用与性能

在导师推荐系统中，我们探索了各种文本生成和主题建模方法，其中包括brat生成、T5生成、TextRank、LDA主题分类以及BERTopic等。

1. 文本生成方法：我们比较了brat生成、T5生成和TextRank三种主要的文本摘要方法。每种方法都有其优点和局限性，如brat生成具有高度的灵活性，但受规则精确性和覆盖范围的限制；T5生成能够适应不同类型的文本和任务，但需要更多的计算资源和时间进行模型训练；而TextRank简单且有效，但可能无法捕捉到文本的语义信息，对于较长的文本可能存在信息损失。我们还发现T5模型生成的摘要通常比Bart模型生成的摘要短，而且在内容质量上也稍逊一筹。

2. 主题建模策略：我们首选使用主题模型，如Latent Dirichlet Allocation（LDA）对导师进行分类。在这个过程中，我们必须先删除可能造成干扰的词语，如主语、动词、连接词等，尽量只保留关键词。然后，我们使用LDA算法总结出10-30个主题，并为每个导师打上对应的主题标签。我们发现，使用原始文本的LDA推荐效果优于使用bart生成摘要的文本。

我们也尝试了另一种主题建模方法，名为BERTopic。虽然BERTopic有着强大的功能，包括BERT编码器的能力、UMAP的降维能力和HDBSCAN的聚类能力，但我们发现在处理我们的导师文本数据时，其结果并不理想。主要的问题在于，我们的数据量相对较小且质量参差不齐，这导致BERTopic难以生成有效的分类结果。

综上，我们发现不同的方法在实现和效果上存在差异，选择合适的方法取决于具体的需求、数据和任务要求。在我们的案例中，LDA主题分类在大多数情况下提供了更好的结果。

这段文本应该属于研究模板的"实验设计与结果"部分。

1. 标题：OpenAI API与Ada嵌入模型在主题分类与导师推荐中的应用

摘要：在本研究中，我们利用OpenAI的API来整合导师的主题分类和摘要生成，通过一个统一的接口来实现上述功能。在这个过程中，我们使用了一种名为text-davinci-003的收费API，它来自于OpenAI，并使用了特定的提示词来确定每个导师信息的主题。我们的实验结果显示，利用TFIDF向量化这些主题以及使用NMF和LDA模型对导师进行进一步的分类，我们能够对导师的推荐系统做出改进。此外，OpenAI也提供了一种名为text-embedding-ada-002的嵌入API，它是OpenAI最新和最强大的嵌入API。我们利用这个API对清洗后的导师文本进行向量化，再对用户输入进行向量化，并计算它们的余弦相似度。结果显示，这种方法和使用OpenAI进行总结后再进行TFIDF向量化的方法在性能上具有相当的优势。然而，这种嵌入API的一个缺点是，每次用户输入时都需要通过API进行向量化，这将导致每次使用都需要向OpenAI发出请求并支付费用，尽管费用很小，但这对于面向用户的系统来说可能是不现实的。

根据我们的实验模板，这段内容可以被归类为"实验结果与讨论"部分。