## 1. Background Research for the project

### 1.1 Context

在计算机科学领域，文本分析和处理已经成为一个重要的研究领域。自然语言处理（NLP）和机器学习的技术使我们能够从大量的文本数据中提取有价值的信息。在这个背景下，我们计划开发一个导师推荐系统，该系统旨在帮助利兹大学的计算机科学系的学生找到最适合他们的导师。这个系统应用了一系列的文本处理和分析的技术，包括文本向量化技术、相似度度量算法和文本摘要技术。

### 1.2 Problem statement

目前，计算机系学生寻找导师的方法只能是阅读每个导师的个人网页，导师大概有100多个，选择导师的这个过程可能是低效的。我们的目标是开发一个基于学生输入的文本（如他们的背景、期望的导师背景或期望完成的项目类型等）来匹配并推荐合适导师的系统。这个系统需要能够理解和处理自然语言，比较文本的相似度，并对大量的导师信息进行有效的摘要，方便学生选择导师。

### 1.3 Possible solution

我们的解决方案是开发一个基于自然语言处理（NLP）和机器学习的导师推荐系统。我们将利用一系列的文本向量化技术，包括CountVectorizer，TF-IDF，Word2Vec和Doc2Vec，将文本转化为数值数据，以便于机器学习算法处理。然后，我们将使用余弦相似度和Jaccard相似度算法，来比较学生输入的文本和导师信息的相似度。最后，我们将使用T5，TextRank和BART等模型对导师的信息进行总结，以便于用户查看。

### 1.4 How to demonstrate the quality of the solution

我们将通过实验评估来展示我们的解决方案的质量。我们将设计一系列的实验，比较不同的文本向量化技术、相似度度量算法和文本摘要技术的性能。我们的目标是找到最优化的组合，使得推荐结果的准确性最高。此外，我们还将进行用户满意度调查，以评估系统的实

## 3. Literature Review

### 3.1 Text Vectorization Techniques

文本向量化是自然语言处理中的一个重要步骤，它将文本数据转化为数值数据，以便于机器学习算法处理。有多种不同的文本向量化技术，包括CountVectorizer，TF-IDF，Word2Vec和Doc2Vec等。

CountVectorizer是一种用于向量化文本数据的工具，它会将文本转化为数值数据，形成一个矩阵，其中每一行代表一个文档（句子），每一列代表一个词语【14†source】。TF-IDF（term frequency–inverse document frequency）是一种反映一个词语对文档的重要性的数值统计指标，它常常被用作信息检索、文本挖掘和用户模型的权重因子【18†source】。

Word2Vec是一种在2013年发布的自然语言处理技术，它使用神经网络模型从大量的文本中学习词语的关联。一旦训练完成，这样的模型可以检测同义词，或者为一个部分句子建议额外的词语【22†source】。Doc2Vec是一种将每个文档表示为一个向量的模型【26†source】。

### 3.2 Similarity Measures

在我们的系统中，我们需要一种方法来比较学生输入的文本和导师信息的相似度。余弦相似度和Jaccard相似度是两种常用的相似度度量算法。

余弦相似度是一个衡量两个非零向量之间相似度的度量，它是两个向量之间角度的余弦值，也就是两个向量的点积除以它们长度的乘积【6†source】。Jaccard相似度，也被称为Jaccard相似系数，是一种用于评估样本集相似度和多样性的统计量【10†source】。

### 3.3 Text Summarization Techniques

为了让用户更容易地理解和查看导师的信息，我们需要一种方法来对导师的信息进行总结。T5，TextRank和BART是三种常用的文本摘要技术。

T5是一个模型，它将所有的NLP任务转化为一个统一的文本到文本的格式，这使得一个T5模型可以适应多种任务【30†source】。TextRank是一种基于PageRank的算法，它常常被用于关键词提取和文本摘要【34†source】。BART是一种用于预训练序列到序列模型的去噪自编码器，BART通过破坏文本，然后学习一个模型来重构原始文本

## 5. Conclusion

本文提出了一个基于自然语言处理和机器学习的导师推荐系统。我们的系统将能够根据学生的研究兴趣和需求来推荐合适的导师。为了实现这个目标，我们将使用文本向量化技术，相似度计算算法和文本摘要技术。我们的系统将进行实验性的评估，并会进行持续的优化和改进，以提供更高质量的推荐。

## 1. Background Research for the project

### 1.1 Context

The recent technological advancements in Natural Language Processing (NLP) and Machine Learning (ML) have paved the way for intelligent systems capable of understanding, interpreting, and responding to human language in meaningful ways. These advancements hold significant potential for enhancing a variety of fields, including education. One such application is in the realm of matching students with suitable academic advisors, a critical factor in the academic success and overall student experience at universities. The University of Leeds, renowned for its strong emphasis on research and innovation, is an ideal setting for such an application. This project seeks to harness these advancements to develop an intelligent tutor recommendation system for the University of Leeds' Computer Science department.

### 1.2 Problem statement

The process of matching students with suitable advisors is often a challenging and time-consuming task for both students and university administrators. The current systems in place often rely on manual processes and lack the ability to consider a student's specific interests, aspirations, and background comprehensively. Moreover, the process can be overwhelming for students due to the vast number of potential advisors and the breadth of their expertise. This project aims to address this problem by developing an intelligent tutor recommendation system that matches students with suitable advisors based on their input, which could include their background, desired advisor's background, and project ideas.

### 1.3 Possible solution

The proposed solution is a tutor recommendation system that employs various NLP and ML techniques to understand and match the student's input with the information of potential advisors. These techniques include text vectorization (CountVectorizer, TF-IDF, Word2Vec, Doc2Vec), text summarization (T5, TextRank, BART), and similarity measures (Cosine similarity, Jaccard similarity). Additionally, topic modeling techniques such as BERTopic, LDA, and NMF will be explored for their potential in enhancing the recommendation process. The use of OpenAI's text generation and embedding APIs will also be explored for their potential in summarizing tutor information and transforming the tutor information and student input into vectors, respectively.

### 1.4 How to demonstrate the quality of the solution

The quality of the solution will be demonstrated through systematic testing and evaluation, ensuring that the system effectively matches students with suitable tutors. The performance of the system will be evaluated using a variety of metrics such as precision, recall, and F1-score. The system's usability and user satisfaction will also be gauged through user testing and feedback. Furthermore, the performance of the different NLP and ML techniques used in the system will be compared to select the most effective combination for the final solution.

## 2. Scope for this project

This section specifies what the project will deliver. It should be written in a concise manner, to be used as a basis for assessment.

### 2.1 Aim

The aim of the project is the overall top-level goal. It might be helpful to consider this in conjunction with the project title.

分析，试验，并构建一个导师推荐系统--为利兹大学计算机学院的同学推荐毕业设计的导师，需要的算法和技术还有可行性。旨在在未来为学校构建此系统时提供算法和原理上的帮助。

该系统功能是可以通过用户输入的文本，告诉用户潜在的满足用户需求的导师，并为导师的档案进行总结，方便用户查看。同时当用户选择了某一导师时，系统还可以为用户推荐同类导师。

本系统目前在python3.8环境下的jupyter notebook上运行，有未来升级为完整的web application的基础。

### 2.2 Objectives

List up to five objectives. When you phrase an objective, think about how you can demonstrate its achievement.

To summarise, characteristics of suitable objectives are:

* Deliverable -- you will hand them in! 一个再ipynb下运行导师推荐系统，包含多种算法和分析结果。

1.使用向量化技术和使用余弦相似度计算推荐分数推荐

Word2Vec (self-trained/google-trained), Doc2Vec, CountVectorizer, TF-IDF, 使用jaccard相似度计算分数推荐

2. 导师文本生成摘要（总结）

brat生成、T5生成、TextRank

3.主题分类推荐

LDA, NMF, BERTTopic

4. 使用OpenAI api的主题分类和向量化推荐

使用text-davinci-003总结各个导师的主题，然后使用tf-idf向量化文本然后使用余弦相似度进行分数计算。或者使用NMF或LDA进行再分类，之后推荐给用户。

使用text-embedding-ada-002对导师文本和用户输入进行向量化，拥有最好的效果，缺点是花钱并且必须远程访问openai（必须联网）

5. 使用KMeans聚类推荐同类导师

对text-embedding-ada-002的向量进行预处理（标准化，PCA），之后使用Silhouette Coefficient进行聚类分析，找出最合适的聚类数目（导师分成多少类）。然后使用KMeans对导师进行clustering，最终当用户选择某一导师后，系统可以为他找出同一类的导师。

* Measurable -- examiners are able to judge/quantify if you have done a good job.

可展示推荐结果，总结结果。

* Appropriate -- they should solve a sufficiently difficult problem.

不同向量化技术的分析和实验，不同文本生成模型的分析，对新兴的强势的openai的技术的应用。对聚类算法的分析和结果。

* Agreed – by your supervisor, assessor, other members of the School (where appropriate).

### 2.3 Deliverables

These are items for assessment under ‘delivery’. These could be written up as sections in the final project report (e.g. comparison of algorithms, or feasibility assessment, or design documentation, etc.) or be handed in separately (e.g. code, user manual or installation guide etc.). It is important to have the agreement from the supervisor at this stage that these deliverables are suitable as delivery for the type of project.

To ensure that these deliverables are within the scope of the project, cross-referencing to the objectives may be a helpful check.

Some examples of deliverables: an analysis of current infrastructure, a comparative study of techniques or tools, a recommendation to the client, requirements specification, design documentation, algorithms, software functionality, a qualitative or quantitative evaluation study, and so on, as appropriate for the type of project.

## 3. Project schedule

The schedule for completion of the project should relate the activities (or tasks) to objectives or deliverables. A few milestones should be identified for self monitoring of progress.

### 3.1 Methodology

Outline the underpinning project approach that is appropriate for the chosen type of project. This should help to plan for the order of the activities /tasks.

### 3.2 Tasks, milestones and timeline

Any appropriate method of presentation is acceptable. A common method is the use of Gantt chart.

Timeline

6.21 向量化和余弦相似度

6.27 算法分析，文本生成，聚类推荐等

### 3.3 Risk assessment (if appropriate)

If there is any risk identified at this stage (e.g. availability of stakeholders, technical issues or suitable test data etc.), mitigating strategy should be discussed.

使用openai的api需要缴费，向量化的费用很低，但文本生成的费用不低。同时使用openai的技术会让数据进入他们的服务器，安全问题。