# Abstract

This project proposes an approach to help students in computer school of the University of Leeds in finding suitable supervisors for their final projects. Leveraging multiple machine learning and NLP techniques, this project offers a system that recommends supervisors based on user input, generates summaries for each potential supervisor and recommend similar supervisors to users. The methodologies encompassed include various vectorization techniques, recommendation algorithms, text generation models, topic modeling, and clustering recommendation. Using the OpenAI API for vector generation and text generation further enhanced the system's performance. The study illustrates the strengths and weaknesses of each approach and indicates the most effective techniques for the specific context of the project.

# Introduction (12 marks)

## 1.1. Problem statement

The process of finding an appropriate supervisor of their final projects can be confusing and time-consuming for students. Currently students have to read through tutor profiles one by one, and given that the School of Computing now has approximate 100 tutors, this process takes at least an hour. If the number of supervisors is increased in the future, they will spend more time on selecting supervisors. Moreover, the process of selecting supervisors was complex. After reading through a large number of profiles, students also need to compare and rank supervisors manually, then they can submit their rank of 6 selected supervisors. As a result, the current supervisor selection process is inefficient and results in many students submitted supervisors who do not match their academic background or the needs of their proposed project. Therefore, the necessity of building an effective system that can alleviate this burden and optimize the supervisor selection process is significant.

## 1.2. Project Aim

This project aims to address this issue by investigating and evaluate various machine learning and natural language processing approaches to determine the most optimal approach for developing a supervisor recommendation system. Through massive experimentation and analysis, the project seeks to identify the most effective recommendation approaches, vectorization techniques, and text generation models that can effectively address the requirements of a supervisor recommendation system.

By conducting a thorough analysis of different recommendation approaches, such as Jaccard similarity and cosine similarity, then use cosine similarity along with various vectorization methods like Word2Vec, Doc2Vec, CountVectorizer, and TF-IDF, this project will assess their respective performances and suitability in the context of the supervisor recommendation task. Furthermore, the project will explore text generation models, such as T5, Bart, and TextRank, to provide concise and informative summaries of each supervisor to students.

Another part of the project is the investigation of topic modelling techniques, including LDA, NMF, and BERTopic, to determine their potential for recommendation and compare with classic recommendation approaches. By utilizing these techniques, the system can find latent topics and identify supervisors with similar backgrounds to recommend supervisors have same topics with student’s demand.

Additionally, the project will leverage the capabilities of OpenAI APIs, specifically text-davinci-003 for text generation and text-embedding-ada-002 for vector generation align with classic recommendation approaches to augment the recommendation system's performance and accuracy. The project will assess the capabilities of these APIs, exploring their potential to enhance the quality of generated topics and embeddings.

At last, this project tested and evaluated clustering recommendation based on KMeans clustering. The system will allocate supervisors into different clusters based on their background and research area, then If the user has already selected certain supervisors, cluster recommendation will recommend similar supervisors to user. This approach includes standardization, PCA dimension reduction and silhouette analysis techniques to enhance the final clusters outcome and performance of recommendation.

Overall, by conducting extensive experiments and evaluations, this project aims to provide valuable insights into the effectiveness of various vectorization, recommendation, topic modelling approaches and clustering recommendation for developing a supervisor recommendation system. The findings and conclusions of this project will contribute to the broader field of recommendation systems and assist in the future development of efficient and accurate recommendation mechanisms designed specifically to the supervisor selection system in academic institutions.

## 1.3. Objectives

- Obtain personal profiles of supervisors in the School of Computer from the website of University of Leeds by Using a web scraping approach, beautiful soup.

- Clean the dataset: expand abbreviations, remove generic nouns, remove html expressions and special characters. Prepare another dataset based on the former but without subjects and connectives.

- To analyse and compare different recommendation algorithms, cosine similarity and Jaccard similarity, in order to identify the most effective one for recommendation system.

- Evaluate various vectorization techniques such as Word2Vec, Doc2Vec, CountVectorizer, and TF-IDF for their efficiency and applicability to the dataset and recommendation algorithm.

- Generate brief introductions for each supervisors using text generation models like T5, Bart, or TextRank.

- Explore topic modelling techniques (LDA, NMF, BERTopic) and OpenAI APIs (text-davinci-003 and text-embedding-ada-002) to compare with classic recommendation approaches.

- Design a cluster-based recommendation approach that recommends supervisors with similar research topics or backgrounds.

## 1.4. Deliverables

1. A cleaned and processed dataset of faculty members in the School of Computer at the University of Leeds, including their profile, Areas of expertise, and published works.

2. A comparative study of different vectorization techniques and recommendation methods, along with their strengths, weaknesses, and the rationale for the selected method.

3. Providing a brief introduction for each supervisor.

4. A comprehensive evaluation of different topic modeling techniques and OpenAI APIs, compare with the former tested techniques and justify the final chosen one.

5. A functional recommendation system that takes user input and outputs a list of recommended supervisors.

6. Implementation of a clustering recommendation system, providing clusters of similar supervisors to students based on their inputs.

7. A GitLab repository that contains the source code of experiments and the recommendation system.

8. The MSc project report”

# Background Research（15marks）

## 1.1. Literature Review

The project is grounded in a variety of research fields extensively investigated in the literature, each contributing valuable methodologies and insights to the proposed system.

Vectorization Methods: Vectorization is a crucial process in Natural Language Processing (NLP), transforming textual data into numerical vectors that machine learning algorithms can interpret and analyze. The cosine similarity approach mentioned above would require vectorised text for the calculation. Various vectorization techniques have been proposed and used over the years, each with unique properties and advantages.

CountVectorizer, for instance, offers a simple method to tokenize a collection of text documents and build a vocabulary of known words, but it treats each word as a discrete entity, ignoring any semantic relationship between words (scikit-learn developers, 2021).

On the other hand, TF-IDF, or Term Frequency-Inverse Document Frequency, is a statistical method that reflects how important a word is to a document in a collection (Ramos, 2003). Unlike CountVectorizer, TF-IDF gives more importance to words that are less frequent in the corpus, reducing the importance of common words that are less informative.

Lastly, Word2Vec and Doc2Vec are prediction-based methods that go beyond simple tokenization and frequency counts. These techniques learn continuous word representations using shallow neural networks, capturing semantic relationships between words and even entire documents (Mikolov et al., 2013) (Le & Mikolov, 2014).

Recommendation Systems: Recommendation systems have become a critical component in numerous applications related to information retrieval and personalized user services. The core principle behind recommendation systems is to provide users with personalized suggestions based on their preferences and behaviour. In the context of this project, the system will take students as users and find the right supervisors for them based on their needs

Various approaches are utilized in recommendation systems, including Collaborative Filtering, Content-Based approach, Hybrid approach, Social approach, and Demographic Approach (Kumar et al., 2020). Collaborative Filtering fulfils the needs of this project. It usually uses for analysing historical user data (like purchase and browsing history) to identify similar users or items, then recommends based on these similarities. In this project, it uses for analysis supervisors’ data and recommend them for students.

Two common techniques used in Collaborative Filtering systems are cosine similarity and Jaccard similarity. They usually serve as measures to calculate the similarity between different items in recommendation system, thus aiding in suggesting the most appropriate items to users.

Cosine similarity, a measure of similarity between two non-zero vectors in an inner product space, gauges the cosine of the angle between these vectors. Cosine similarity is particularly beneficial when managing high-dimensional data, like text data (Verma and Aggarwal, 2020) (Huang et al., 2021). It's insensitive to the number of dimensions, making it an excellent choice for handling high-dimensional data. Each document can be represented as a vector in this context, with each dimension representing a specific word's frequency or TF-IDF weight. Cosine similarity can then be used to ascertain the similarity of these documents.

On the other hand, Jaccard similarity measures the similarity between two nominal attributes by taking the intersection of both and dividing it by their union. It's notably effective when dealing with binary or discrete data (Verma and Aggarwal, 2020). For example, in a recommendation system, Jaccard similarity can be used to compare the set of items liked by two users (Zahrotun, 2017), or in bioinformatics, it can be used to compare the similarity of two gene samples (Huang et al., 2021).

In conclusion, both cosine similarity and Jaccard similarity have their strengths in different scenarios and data types, and the choice of method depends on the specific application scenario and requirements (Verma and Aggarwal, 2020). These methods play a pivotal role in constructing Collaborative Filtering recommendation systems.

Text Generation Models: Text generation models and automatic summarization methods have gained considerable attention in NLP research. These models provide a way to generate human-like text, which is a critical requirement for the proposed system.

TextRank, an extractive summarization technique, brings order into text by ranking the relevance and significance of sentences in a document, similar to how PageRank ranks web pages (Mihalcea & Tarau, 2004). Transformer-based models like T5 (Raffel et al., 2019) and Bart (Lewis et al., 2019) have achieved state-of-the-art performance on multiple NLP tasks. These models adopt the transformer architecture (Vaswani et al., 2017) which is based on self-attention mechanisms, allowing the models to generate more coherent and contextually relevant sentences.

Topic Modeling: Topic modeling techniques such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Non-negative Matrix Factorization (NMF) (Lee & Seung, 1999) are valuable tools for discovering abstract topics from document collections. These techniques can identify the latent semantic structure in the data, providing a means to group similar documents together.

Recently, a new approach, BERTopic (de Vries et al., 2020), has emerged, which combines BERT embeddings (Devlin et al., 2018) and c-TF-IDF to generate more coherent and interpretable topics. The use of transformer-based embeddings allows this method to capture deeper semantic relationships between words, improving the quality of the discovered topics.

OpenAI: The rise of large transformer-based architectures has dramatically reshaped the landscape of NLP. OpenAI's text-davinci-003 and text-embedding-ada-002 are two such examples that have been developed to allow easy access to these powerful models (OpenAI, 2021). The former is capable of generating human-like text based on a prompt, while the latter produces embeddings for the given input text, which can be beneficial for downstream tasks like clustering or similarity estimation.

Clustering Recommendation: Clustering algorithms, such as KMeans, have been employed in various applications for grouping similar instances together. KMeans works well with high dimensional data, especially when dimension reduction techniques like Principal Component Analysis (PCA) are used to capture the essential structure of the data (Jolliffe & Cadima, 2016). Before applying KMeans, it's often advisable to standardize the data to ensure all features have the same scale (Hastie et.al, 2001). To determine the optimal number of clusters, silhouette analysis can be used, which provides a succinct graphical representation of how well each object lies within its cluster (Rousseeuw, 1987).

## 1.2. Methods and Techniques

The project will utilize a range of methods and techniques derived from the aforementioned areas of research. The choice of these methods is guided by theirrelevance to the problem at hand and their performance in respective contexts as suggested by the literature.

In Vectorization, various methods such as CountVectorizer, Word2Vec, Doc2Vec, TF-IDF and OpenAI’s text-embedding-ada-002 will be evaluated for their efficacy and compatibility with the dataset and the recommendation algorithm. These methods have been chosen as they offer a balance between computational efficiency and the ability to effectively capture the semantics of the text data (scikit-learn developers, 2021) (Mikolov et al., 2013) (Le & Mikolov, 2014) (Ramos, 2003) (OpenAI, 2021).

For recommender system, as this system needs to make recommendation based on text content, and cosine similarity and Jaccard similarity has ability to effectively compute similarity between different text content, which is vital for the recommendation task (Adomavicius & Tuzhilin, 2005). So, they will be evaluated separately and their performance and compatibility will be compared. The cosine similarity will be evaluated using different vectorisation methods as mentioned above.

Text generation for brief introductions of each supervisor will be carried out using models like T5, Bart, or TextRank. The selection of these models is based on their ability to generate high-quality, human-like text and their performance in recent NLP research (Mihalcea & Tarau, 2004) (Raffel et al., 2019) (Lewis et al., 2019).

Topic modeling techniques including LDA, NMF, BERTopic will be explored. These techniques, known for their effectiveness in discovering latent topics and generating high-quality embeddings respectively, are chosen to compare their performance with classic vector-calculation-based recommendation approaches (Blei et al., 2003) (de Vries et al., 2020) (Devlin et al., 2018).

Finally, a cluster-based recommendation approach will be designed that leverages KMeans clustering, standardization, PCA dimension reduction, and silhouette analysis techniques. The selection of these methods is based on their proven effectiveness in grouping similar instances together and handling high dimensional data (Jolliffe & Cadima, 2016) (Hastie et.al, 2001) (Rousseeuw, 1987).

## 1.3. Choice of methods

The selection of methods for this project is driven by the specific requirements and objectives of the supervisor recommendation task faced to students. The chosen methods aim to optimize the supervisor selection process by reducing the time and improving the final accuracy for matching teachers to students compared to manual selection.

Therefore, the first step is selecting the most optimal recommendation algorithm from cosine similarity and Jaccard similarity, based on their performance and scalability. This is followed by testing to select the most suitable vectorization methods for various recommendation algorithms and other techniques. All these methods are analyzed considering their performance and matching in two specific contexts: supervisor information text and user input text.

Subsequently, the potential of topic modelling recommendation methods is investigated. These methods can allocate the topics for supervisors and users based on their contents and inputs, and match topics for users. From LDA, NMF, and BERTopic, the highest performing ones will be chosen, compared with vector-calculation-based recommendation methods, and then analysed to conclude which recommendation approach is most effective for this system.

Inspired by the topic modelling approaches, the project will also assess the performance of OpenAI's text-davinci-003 for generating topics from supervisors' contents. These topics, along with vector-based approaches, will be utilized to recommend supervisors to users. OpenAI's text-embedding-ada-002 will be used for vectorising the supervisors’ contents and users’ inputs, and compare with the traditional vectorisation approaches. Exploring the applicability and performance of OpenAI's API and finally summarising the best performing recommendation approaches and vectorisation approaches for this system.

Lastly, due to their potential to identify latent topics, cluster similar supervisors, and recommend supervisors with similar research interests or backgrounds to students, clustering recommendation approaches were selected for testing. KMeans clustering will be implemented, with standardization and PCA dimension reduction employed as data pre-processing tools. Silhouette analysis will be used to find the most appropriate number of clusters.

The selection of these approaches is also considered by the need for the system to be scalable, economic efficiency and capable of handling a growing number of supervisors and different background of supervisors in the future.

# Design of experiments （8marks）

# Results of the Empirical Investigation (implementation) （15marks）

# Evaluation （5marks）

# Conclusion & Future Work （3marks）

## Conclusion

## Future Work