**COMP4003 / COMP4031 / COMP4026 Project Plan**

**Student:** Ting-Chun Chen

**Student ID:** 20498346

**Course:** MSc Projects COMP4096

**1. Project title.**

This project is about building relations between sentences and scientific articles in FinTech based on word embedding and document summarisation techniques. We would like to develop a word-embedding method that can best describe sentences in FinTech research articles and yield the best result of relating citation sentences and scientific articles.

**2. Statement of the research problem.**

This project aims to help academics find proper references for their statement when writing literature reviews and introductions of scientific reports. Citations are commonly used in writing scientific articles to acknowledge the source of a statement in our work. Though we usually file and organize papers well while doing literature reviews, it happens that some statements are based on months or years of experience or accumulated knowledge in a particular subject. It could be relatively hard work to find an appropriate source for this idea. Hence, we are proposing a method to find the best math articles for our writing.

We want to look for articles that mainly describe similar thoughts to the statement we write down in a certain domain in scientific writing. Among the processes within NLP (natural language processing) pipeline, word-embedding techniques would be mostly focused on in this research, which is to convert texts and words into meaningful numbers for algorithms to compute, summarise and compare. We decide to apply this research in the fintech field and answer the question “What word embedding techniques can best represent a scientific article sentence to derive an academic writing citation-hinting system in fintech field?”

**3. Related work.**

Document summarisation methods are widely used in scientific articles to generate a representative conclusion for the whole research. Citation-based summarisation uses citation sentences as the prediction target of document summarisation. Citation sentences are thought to be more precise and represent main ideas and contribution of the research[[Elkiss et al., 2008](https://onlinelibrary.wiley.com/doi/epdf/10.1002/asi.20707?src=getftr)], making it another object of document summarisation instead of manually generated summaries. [Qazvinian et al., 2013](https://www.sciencedirect.com/science/article/pii/S1319157820303554?ref=pdf_download&fr=RR-2&rr=7e07e010aa4c7729#b0395) proposed a single-article citation-based summarisation with ANN dataset and [Saini et al., 2023](https://link.springer.com/article/10.1007/s10489-022-04166-z) proposed a multi-objective clustering for citation-based scientific document summarisation.

Word embedding can be used in a wide range of tasks, including visualisation in medical field [[Oubenali et al., 2022](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-022-01822-9)], topic modelling on social networks of Twitter and Reddit[[Curiskis et al., 2020](https://www.sciencedirect.com/science/article/pii/S0306457318307805?via%3Dihub)] and sentiment analysis on product reviews[[Onan, 2020](https://onlinelibrary.wiley.com/doi/10.1002/cpe.5909)]. As for the application of word embedding to scientific sentences, [Tshitoyan et al., 2019](https://www.nature.com/articles/s41586-019-1335-8) used embedding to capture complex materials science concepts to extract knowledge and relationship from scientific literature. [Zhang et al., 2022](https://www.sciencedirect.com/science/article/pii/S1532046422001228?via%3Dihub) used 15 text representation models to construct a similar article recommendation system in biomedical field.

**4. Methodology.**

**5. Programme of work.**

**6. Time plan.**

**Enter word count out of /1000 words**

**Assessment criteria**

Is there a clear statement of what the project is about?

Is there an explanation of why this problem needs to be addressed / solved?

Is the novelty of the proposed research substantiated with reference to related work?

Is there a clear statement of how the problem is going to be addressed / solved?

Are the proposed methods for addressing / solving the problem appropriate?

Is there a clear statement of how the results are going to be evaluated?

Is the proposed evaluation appropriate?

Is there a breakdown of the research into work packages?

Is this breakdown appropriate?

Is there a time plan?

Is the time plan realistic?

*Affirmative answers to each of these questions will result in a first class mark.* *Markers should briefly describe how the project plan diverges from these desired outcomes to justify their mark.*