NPS2001B Group 1: Hirwan, Ruth, Yee Shien
Milestone 1

Which real world problem have you chosen to work on? Why is it an important problem?

Reading academic papers presents a challenge for students, particularly those in university. These individuals often struggle to balance their time across their different commitments. Having to spend extra time deciphering difficult jargon and abstract topics requiring prior knowledge can be disheartening. Furthermore, a majority of Singaporean students may find themselves unacquainted with the rigours of delving into such scholarly works. Hence, to be thrown head first into doing such intensive readings of unfamiliar topics without much guidance has proven to be a difficult task.

We have recognized these difficulties as a real-world issue, as it is not uncommon for students to approach academic papers more out of obligation than with confidence in their ability to grasp the core ideas. This situation has the potential to breed frustration and a waning enthusiasm for academic literature, ultimately compromising their own educational journey. Thus, it is important that there are resources to guide these students in their learning and navigation of academic papers.

What will your app do and how will it help solve or mitigate the problem?

Our app will have the capability to analyse entire academic texts and extract their primary points and crucial examples, effectively generating a comprehensive summary of the scanned paper. This feature aims to alleviate the time-consuming process of sifting through extensive pages of academic content, offering university students a more efficient means of comprehending the paper's core concepts.

By equipping students with a concise overview of the paper's content, our app not only enhances their understanding but also enriches their overall learning experience. Moreover, the time saved can be redirected towards tackling other assignments and aspects of their lives, ultimately promoting a healthier work-life balance for students.

What is the central algorithm or class of algorithms that will enable this app to work?

Our app relies on two key algorithms: a text summarization algorithm and a keyword definition algorithm. We elaborate on them below:

1. **Text Summarization Algorithm:** This algorithm serves to condense academic texts while preserving their essential elements, including key points, critical examples, and the underlying structure of the content. Its primary objective is to ensure that the summary provides a coherent representation of the original text's main insights.

   Our group has identified two existing potential options. The first one is Textrank, a graph-based ranking model for text processing. It has the capacity to identify the most relevant sentences in a given text and extract keywords. The second option is the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm, which assesses the importance of words based on their frequency within a document.

   A simple online search for both algorithms revealed numerous guides and papers explaining how they function and how to apply them. While delving into these materials might unveil some complex mathematical concepts involved in the algorithms' development, we have noted that actually using them does not necessitate a deep understanding of the underlying maths.

2. **Keyword Definition Algorithm:** This algorithm will elucidate the context-specific meanings of keywords within academic texts. It dissects the usage of these keywords within the given context, providing users with a nuanced understanding of their significance in relation to the

subject matter.

Our group has identified the existing Lesk algorithm to aid us in this. This algorithm compares the dictionary definitions of an ambiguous word with the words in its surrounding context to determine the most appropriate sense. Similar to the previous algorithms, there is a lot of material on this algorithm found through a quick google search and from a quick scan, it seems to be easier to implement and understand than the search algorithms.

These algorithms will empower our app to deliver precise, insightful, and context-aware summaries of academic papers, facilitating a more efficient and enriched learning experience for university students.

Under what circumstances do you expect users to use your app? What is (are) the target demographic(s) for your app?

This app primarily targets university students and anyone dealing with academic texts. University students, in particular, are expected to use the app regularly due to the substantial amount of reading required for their courses. Additionally, researchers can benefit significantly from the app as it allows for quick summaries of important sections, saving them valuable time in assessing the relevance of research papers for their work.

We have also identified a few issues users might face when using our app:

1. **Language recognition.** Academic texts often incorporate words and references from various languages. We'll need to explore whether our algorithms can accommodate this diversity by expanding their parameters to include other languages.

2. **Formatting concerns.** Our current algorithms may be unable to read papers with 2 columns of paragraphs, a format often used in scientific papers. If our algorithm struggles with this format, it could limit the app's usefulness for students encountering such papers frequently.

3. **Evolution of academic content.** Given that our target audience(s) would likely expect and require updated information and definitions, we have to keep in mind that there is a continuous influx of new content being introduced, while existing content undergoes constant revision. Thus staying abreast of these changes presents a challenge in our efforts to ensure that the App consistently offers the most up-to-date information.

4. **Accuracy.** Our users expect accurate summaries that aid their understanding. If the app fails to identify relevant keywords and key points within a paper, it could lead to inaccuracies, undermining the user's comprehension. While TextRank is a powerful algorithm, it may struggle to capture subtle and nuanced key points in complex readings, particularly in the arts and humanities, where there are many nuances. These limitations may present challenges that are not easily overcome.