CS 510 Project Proposal
Name: Sindhu, Kavvya, Srilekha
Team: 15

Track - Development Track

Team Members:

- Srilekha Kodavati(sk121) - Coordinator
- Sindhu Vydana(svydana2)
- Kavvya Ramarathnam(kavvyar2)

FUNCTIONS AND USERS

As part of this project, we are planning to implement a document summarization tool. The main objective of the document summarization tool is to help users quickly identify and understand the main points of a document without having to read through the entire document. The tool will be available as a web extension, which means users can easily access it through their web browser. The problem statement behind the application is that when users want to search about a topic, they stumble upon a lot of links and documents and they generally do not have the time to read and analyze each and every document. It would save the user's time and energy by providing a summary of each link.

The web extension would be implemented as a standalone software tool. In addition, It could also be incorporated into existing systems like Community digital library. We are also planning to generate tags that would help users associate the document with a list of topics. Generating topic tags would also enable users to search their desired topics faster.

Major functions of the tool would be

- **Summarization**: Taking a long article or text document linked to the web page and generating a summary that captures the main points and essential information in a shorter format. This helps to effectively reduce reading time, and make the document selection process easier.
- **Topic tagging**: Analyzing the document and generating relevant tags and keywords that reflect the main topics covered in the document. This can help users quickly identify the main themes and topics within the text.

Users of the tool would be

- **Students**: College or high school students could use the tool to quickly summarize articles and research papers for their coursework.
- **Professionals**: Business professionals, researchers, and journalists could use the tool to quickly analyze and understand long reports, market research, and news articles.

- **General audience**: Anyone who reads a lot of online content but has limited time could benefit from the tool's summarization and tagging functions.

SIGNIFICANCE

Many people start an internet search for a specific topic and end up reading a lengthy article or blog to find the information they need. Most of the time, they may not have the time to read lengthy documents, or they may have to read the entire article only to discover that it doesn't have the required information. The summarization tool will be useful for summarizing articles and documents, allowing people to read the summary to find the necessary information and then proceed to read the entire document if it appears relevant to what they need. This will save time and effort.The main pain point that the summarization tool addresses is the information overload on the internet. The summarization tool addresses this issue by providing users with a quick and efficient way to process large amounts of information.

The summarization tool will have the following impacts on the world:

1. People can quickly and easily identify the most important information, reducing information overload.
2. People can process more information in less time, resulting in increased efficiency.
3. People can save time and effort.
4. People can easily identify relevant documents by filtering using the topic tags our tool generates.

Yes, the summarization tool addresses the societal need for information accessibility. Everyone can get the information they require quickly and easily without having to read the entire document. In today's world, access to information is more important than ever. It is critical to address this need because it will save users time, increase their efficiency, and productivity. Users will be able to quickly and easily access information using the summarization tool without having to read the entire document. This will allow people to quickly go through and process a large number of documents, as well as help people with limited time to understand the important points in the document. It will also help people who have difficulty reading and comprehending long documents. In addition, users can also filter documents using the topic tags generated by the tool. This can help users quickly identify the main themes and content of the text.

CS 510 Project Proposal
Name: Sindhu, Kavvya, Srilekha
Team: 15

APPROACH

Our approach to building the document summarization tool will involve the following steps:

1. **Research and Algorithm Selection**: We will begin by researching existing summarization algorithms and techniques, including extractive and abstractive summarization methods. After careful consideration, we will select the most appropriate algorithm for our use case.
2. **Text Preprocessing:** We will preprocess the input document text to remove stop words, perform tokenization, and other relevant preprocessing tasks.
3. **Summarization Implementation**: Using the chosen algorithm, we will implement the summarization functionality, which will take the preprocessed text and generate a summary capturing the main points of the document.
4. **Topic Tagging**: We will generate relevant tags and keywords, which will help users quickly identify the main topics of the document.
5. **Web Extension Development**: The summarization tool will be available in the form of web extension allowing users to easily access and use the tool via their web browser.
6. **Testing and Refinement**: We will thoroughly test our tool on various types of documents and refine the summarization and tagging algorithms based on the test results.

As for the technologies we plan to leverage NLP libraries such as Gensim, NLTK, and spaCy for text processing and summarization. We can use web development technologies like HTML, CSS, JavaScript, and Chrome Extension API to create a web extension. Additionally, we will refer to research papers that offer valuable insights and best practices for enhancing our document summarization tool.(References included at the end)

One potential challenge we might face is the difficulty of accurately summarizing diverse and complex documents. To overcome this issue, we will thoroughly research and test different algorithms and approaches, and iterate on the implementation to improve its performance. Additionally, we will gather feedback from users to identify areas for improvement and further refine the tool.

EVALUATION

To demonstrate the usefulness of our tool and correctness of our implementation, we will do the following:

CS 510 Project Proposal
Name: Sindhu, Kavvya, Srilekha
Team: 15

1. **Evaluation metrics:** ROUGE for measuring Recall and BLEU for evaluating Precision. These metrics will help to evaluate the quality of the output generated by our tool.
2. **User testing and feedback:** We intend to conduct user testing by providing them with documents and the results from our tool and asking them to rate the tool's overall efficiency and correctness. We would also appreciate feedback on any missing features or errors so that we can make the necessary changes and improve the tool even further.

TIMELINE

- **Milestone 1**: Looking into existing tools and mechanisms to understand how summarization is implemented and choosing an efficient algorithm
- **Milestone 2**: Document text preprocessing to remove stop words and creating similarity matrix to find similarity between two sentences.
- **Milestone 3**: Applying a ranking algorithm to calculate the importance of each sentence and combining sentences to create a summary. Perform post-processing to remove redundant sentences
- **Milestone 4**: Integrating the summarization tool as a chrome extension and testing them on a website
- **Milestone 5**: Generating topic tag words from the summary by eliminating stop words using a background reference model. We would rank the words in the document and use top-5 high probability content words as tags for the given document.

TASK DIVISION

Sindhu Vydana (svydana2): Sindhu will focus on researching and selecting the appropriate summarization algorithm, as well as working on the implementation of the summarization functionality.

Kavvya Ramarathnam (kavvyar2): Kavvya will be responsible for implementing the topic tagging functionality and developing the Chrome web extension.

Srilekha Kodavati (sk121): Coordinator: Srilekha will oversee the project coordination, manage testing and refinement, and contribute to the implementation of summarization and topic tagging functionalities.

CS 510 Project Proposal
Name: Sindhu, Kavvya, Srilekha
Team: 15

REFERENCES

1. Liu, Yang, and Mirella Lapata. "Text summarization with pretrained encoders." *arXiv preprint arXiv:1908.08345 (2019).*
2. See, Abigail, Peter J. Liu, and Christopher D. Manning. "Get to the point: Summarization with pointer-generator networks." *arXiv preprint arXiv:1704.04368* (2017).
3. Nallapati, Ramesh, et al. "Abstractive text summarization using sequence-to-sequence rnns and beyond." *arXiv preprint arXiv:1602.06023* (2016).
4. Narayan, Shashi, Shay B. Cohen, and Mirella Lapata. "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization." *arXiv preprint arXiv:1808.08745* (2018).
5. Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).