Document Comparison and Topic Modeling Tool

Bohao Wu (bohaowu, *Coordinator*), Shusen Han (shusenh2), Wenjie Guo (wenjie6)

Introduction: In the realm of text analysis and information retrieval, there exists a pressing need for tools that can delve deeper into the content of documents, beyond mere surface-level similarities. Our proposed software tool aims to address this gap by offering a comprehensive solution for comparing multiple documents and extracting common topics and frequently occurring words through advanced natural language processing techniques.

Functions and Users: The envisioned software tool will be a standalone application, primarily intended for researchers, students, content creators, and professionals engaged in analyzing and comparing textual content. It is also designed for ordinary people to easily compare documents. The tool's major functions will include document comparison to identify similarities and differences, topic modeling using a mixture language model with Probabilistic Latent Semantic Analysis (PLSA) to cluster words related to common topics, semantic analysis to understand the context and meaning of these words, and a user-friendly visualization of the results.

Significance: The significance of this tool lies in its ability to provide a deeper analysis of document content, addressing the existing "pain point" of limited analytical depth in current document comparison tools. By uncovering underlying themes and contexts, our tool will enable users to gain more profound insights into the commonalities and differences between documents, thereby enhancing the understanding and interpretation of textual information. This capability is particularly important in academic research, content development, and information retrieval, where a nuanced comprehension of text is crucial. It also makes a difference to document workers or writers to keep track of the topic shift.

Novelty: Something unique about our project is that we are going to combine Probabilistic Latent Semantic Analysis (PLSA) with other documents comparing models to build a mixture model. Related work includes some probability theory models and some deep learning models for document comparison. Our model is able to combine the strength of related work and possibly will perform better.

Approach: To build this tool, we plan to leverage Python as the primary programming language, utilizing well-established libraries such as NLTK, spaCy, and Gensim for natural language processing and topic modeling. We will also explore existing open-source frameworks and resources to support the development process. One potential risk in this endeavor is the accuracy of the topic modeling component. To mitigate this risk, we will conduct thorough testing with various parameters and models to optimize the accuracy of topic clustering and ensure the reliability of the tool.

Evaluation: To demonstrate the usefulness and correctness of our Document Comparison and Topic Modeling Tool, we have devised a multifaceted evaluation plan. The tool's effectiveness will be assessed through its application to diverse real-world document sets, such as academic papers, news articles, and literary works. This will provide insights into its ability to accurately identify common topics and themes, showcasing its practical applicability. Additionally, we will benchmark the tool's performance against existing document comparison and topic modeling tools, evaluating its precision, recall, and overall

accuracy in identifying relevant topics. This comparison will highlight the competitive advantage of our tool. To ensure the robustness and reliability of the software, the implementation will undergo rigorous testing, including unit tests and continuous integration, which will help identify and rectify any issues or bugs. Furthermore, to validate the effectiveness of our topic modeling component, we will compare the topics identified by our tool with summaries generated by ChatGPT. This will provide an additional layer of evaluation, ensuring that our tool produces results that align with human-like understanding and interpretation of text. Lastly, we will collect feedback from potential users, such as researchers, students, and content creators, to evaluate the tool's usability, user interface, and overall satisfaction with the results. This user feedback will be instrumental in making user-centric improvements to the tool. Through this comprehensive evaluation plan, we aim to thoroughly assess the tool's usefulness and correctness, ensuring its value and reliability for various text analysis and information retrieval tasks.

Potential Applications:

- **Author Influence Estimation:** By modeling common topics between documents, the tool can estimate the extent to which the author of one document was influenced by the other. This can be particularly useful in academic research for tracing the flow of ideas and concepts.
- **Authorship Similarity Analysis:** Through semantic analysis, the tool can estimate the likelihood that two documents were written by the same author, or assess how similar the tones of two authors are. This can aid in literary analysis, forensic linguistics, and plagiarism detection.
- **Uncovering Common Topics in Diverse Documents:** The tool can identify common topics between two documents that may not seem relevant at first glance. This can reveal hidden connections and themes, providing insights into interdisciplinary research or content creation.
- **Topic Relevance Ranking:** By setting one document with just a few topic words, the tool can function as a ranking machine, evaluating the relevance of these topics in the second document. This can be useful for targeted content analysis, keyword research, and information retrieval tasks.

Timeline and Task Division (Tentative):

- Core functions now - 4/20
- Front End(Web) 4/21 - 4/25
- Test 4/25 - 4/31
- Prepare for Presentation 4/31 – last day

Our team will collaborate closely throughout the development process, with each member focusing on specific areas of expertise to ensure a well-rounded and efficient workflow. **Bohao** will primarily focus on the construction and optimization of algorithms for document comparison and topic modeling; **Shusen** will be responsible for front-end development, ensuring a user-friendly interface for interacting with the tool; and **Wenjie** will concentrate on data collection and evaluation.

Conclusion: In conclusion, our proposed Document Comparison and Topic Modeling Tool represents a significant advancement in the field of text analysis and information retrieval. By providing a deeper understanding of document content, it has the potential to revolutionize the way we analyze and interpret textual information, ultimately contributing to more informed research and decision-making.