BE PROJECT

COMPREHENSION BOT

Presented by -

Rushi Desai (60004190096)
Saloni Patel (60004190099)
Shourya Kothari (60004190103)

Under guidance of -
Prof. Pankaj Sonawane

INTRODUCTION

- Students are required to refer to various books and resources. This becomes a tedious task to refer to multiple books and read redundant information again and again wasting the student's time.
- Our goal is to create a bot that is given multiple pdfs as input and gives a combined version of them as output, eliminating the redundant information.
- This project will involve use of mutiple technologies such as OCR, multi-document text-summarization, question-answering system etc.
- We plan to use state-of-the art models.

PROBLEM DEFINITION

- For specific coursework, a student must refer to books from several publishers, carefully reading each one to extract only the necessary and relevant information.
- Making a compact version of numerous journals, research papers, and books becomes challenging for research-minded students while incorporating novel concepts.
- Simply combined PDFs contain inessential data and lack images, diagrams, or any visualizations.

We, therefore, suggest a comprehension bot to address these issues and broaden the scope of its application.

SCOPE OF PROJECT

- A comprehensive bot that can read PDF books and respond to questions based on them.
- The bot should also have the ability to read various books, draw links between them, and integrate two (or more) books.
- It should be concise enough to provide all essential information and remove redundancy from extracted data.
- The bot should be able to provide diagrams, charts, and formulas in addition to text.
- Mappings between the uploaded syllabus and the merged PDF copy to help students comprehend their courses.

LITERATURE SURVEY

- Multi-document text summarization involves compressing, extracting, removing redundancy, ranking/selecting sentences, and ordering them, with previous attempts [1][2] relying on statistical tools that have performed poorly.
- Previous extractive summarization works used simple features (key phrases, tf-idf, sentence positions [3][4][5]) to rank sentences and applied compression ratio to select top redundant sentences.
- In MDTS, redundancy removal is crucial, and some papers [6][7] use the MMR technique [8] to select top-ranked sentences and reduce redundancy during summary generation.

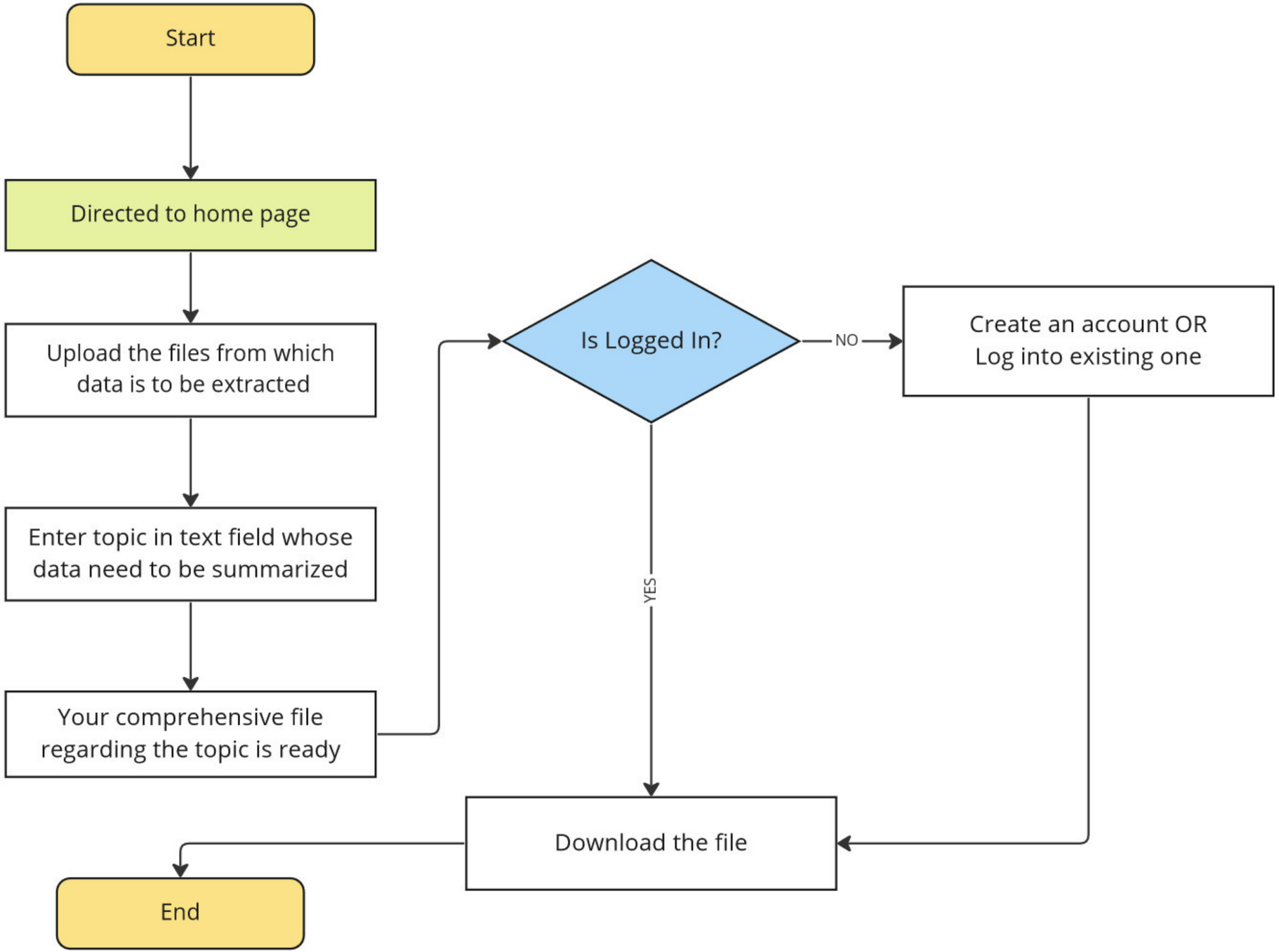
LITERATURE SURVEY

- To prevent redundancy and ensure coverage, clustering-based approaches are used to similar group sentences and choose representative sentences from each group to create a summary [8][9][10], with the quality of clusters depending on the sentence similarity measure used.

- Clustering-based methods are not as accurate as supervised methods that use annotated data since they rely on heuristics and assumptions about similarity.

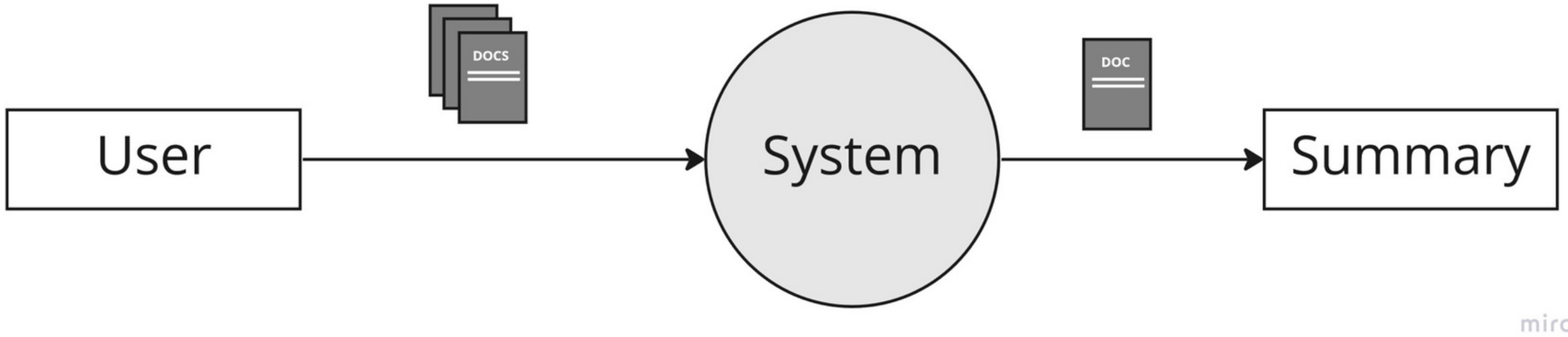
- They may not be effective at capturing the overall coherence and meaning of the source documents, since they only consider sentence-level similarity.

USER FLOW DIAGRAM



FUNCTIONAL MODEL

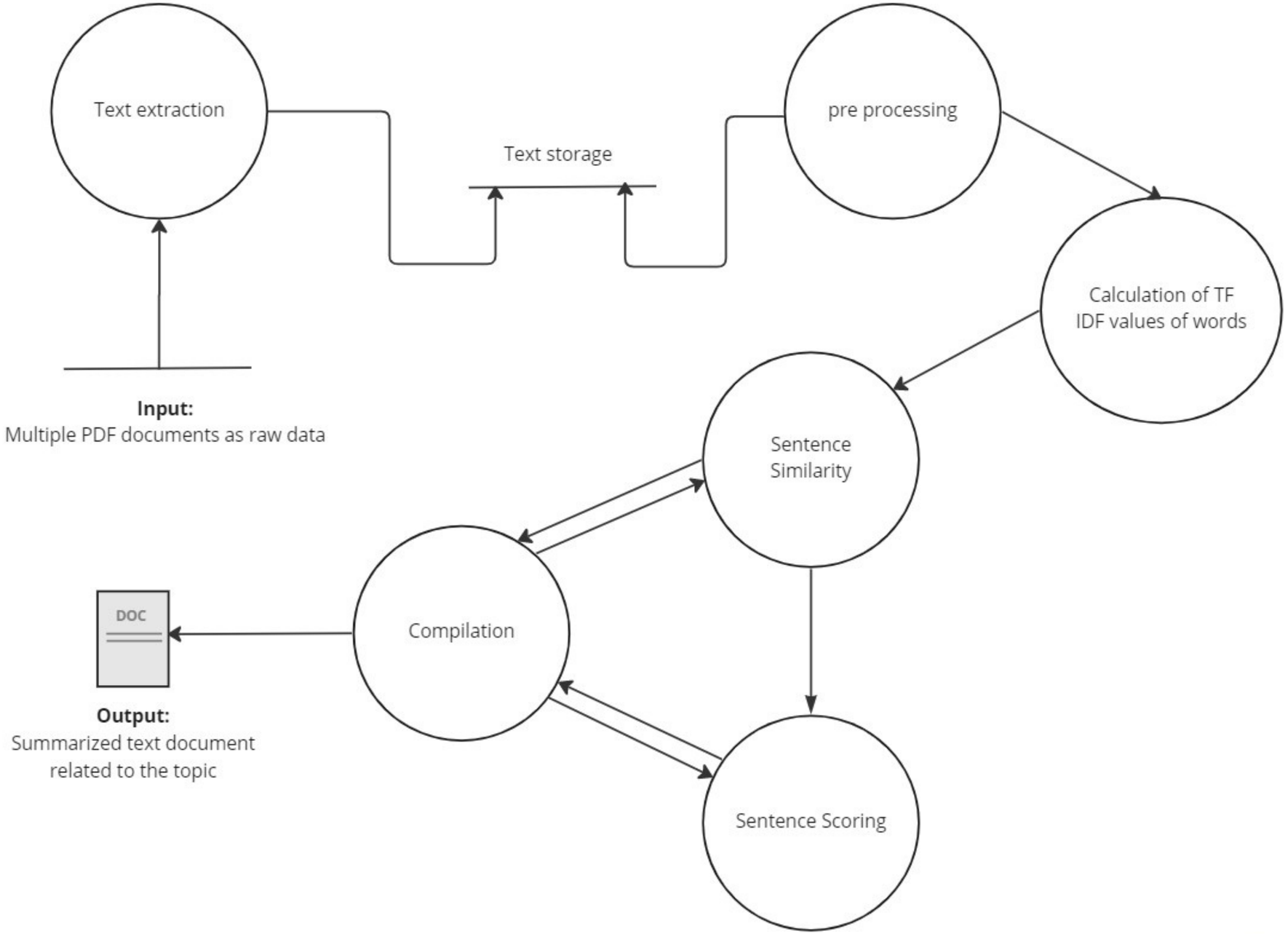
DFD Level 0



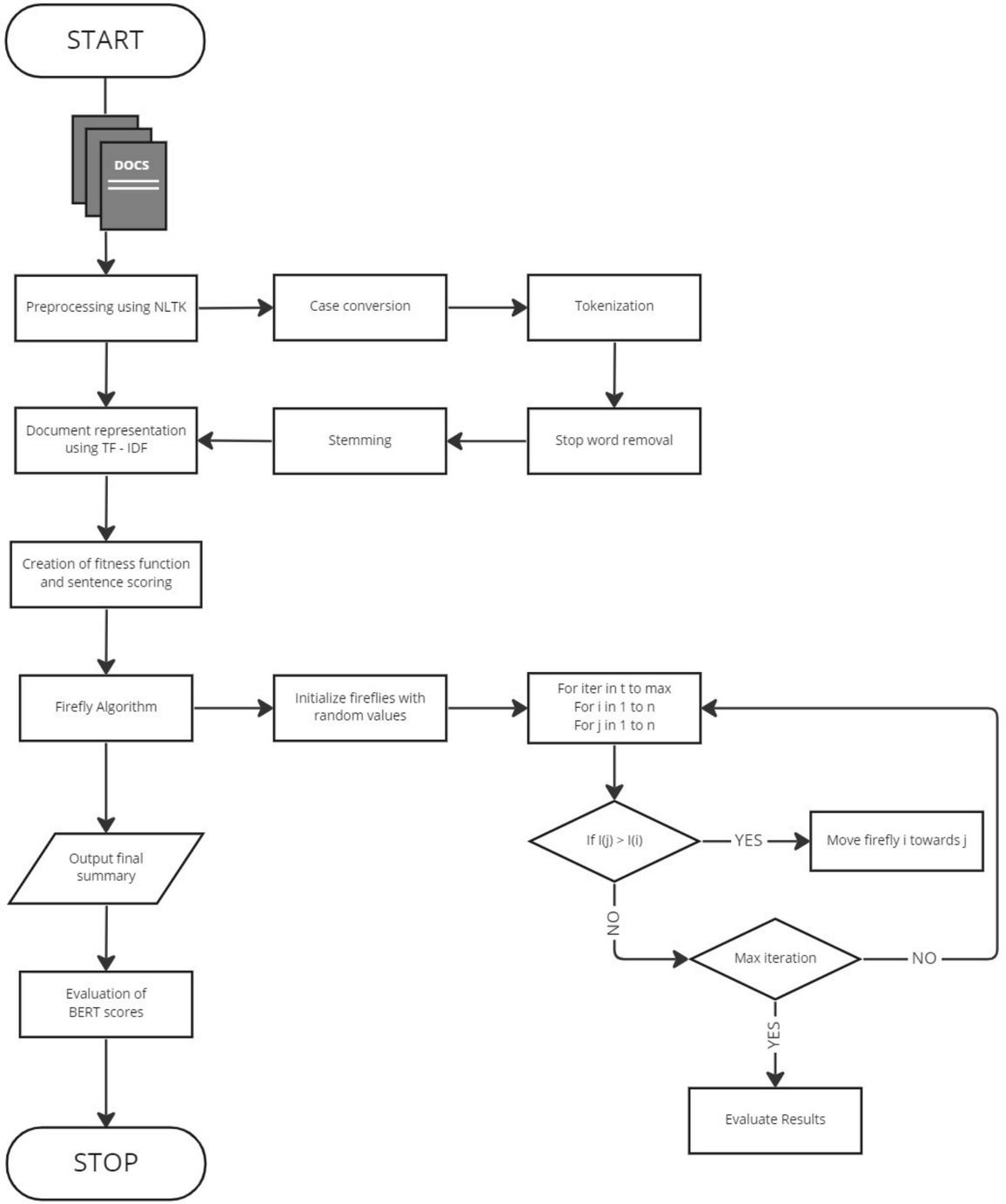
It's designed to be an abstract view, showing the system as a single process with its relationship to external entities. It represents the entire system as a single bubble with input and output data indicated by incoming/outgoing arrows.

FUNCTIONAL MODEL

DFD Level 1



ARCHITECTURAL DIAGRAM



DATASET

- The DUC dataset is a famous collection of papers in document summarization and text-mining studies.
- Both single-document and multi-document summarization tasks, which summarize one or more primary documents into a concise summary, are included in the dataset.
- The DUC dataset has been extensively used in the creation and assessment of summarization algorithms.

EVALUATION METHOD

- The ROUGE error matrix is a method for evaluating the performance of text summarization systems.
- The drawback is that it doesn't account for the semantic similarity between the source and the derived summary.
- BERTScore metric uses pre-trained transformer models such as BERT to compare the semantic similarity between the generated summary and the source documents.
- BERTScore has been shown to correlate strongly with human judgment and outperforms ROUGE metrics in many cases.

REFERENCES

[1] K. Sarkar, K. Saraf and A. Ghosh, "Improving graph based multidocument text summarization using an enhanced sentence similarity measure," 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), 2015, pp. 359-365, doi: 10.1109/ReTIS.2015.7232905.

[2] K. Sarkar, "Automatic Text Summarization Using Internal and External Information," 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT), 2018, pp. 1-4, doi: 10.1109/EAIT.2018.8470412.

[3] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), 2017, pp. 1-6, doi: 10.1109/ICCCSP.2017.7944061.

REFERENCES

[4] O. Tas and F. Kiyani , "A SURVEY AUTOMATIC TEXT SUMMARIZATION", PressAcademia Procedia, vol. 5, no. 1, pp. 205-213, Jun. 2017, doi:10.17261/Pressacademia.2017.591

[5] I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand and P. K. Soni, "Natural Language Processing (NLP) based Text Summarization - A Survey," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1310-1317, doi: 10.1109/ICICT50816.2021.9358703.

[6] Ozsoy, M. G., Alpaslan, F. N., & Cicekli, I. (2011). Text summarization using Latent Semantic Analysis. Journal of Information Science, 37(4), 405–417. https://doi.org/10.1177/0165551511408848

REFERENCES

[7] A. S. Almasoud, S. Ben Haj Hassine, F. N. Al-Wesabi, M. K. Nour, A. Mustafa Hilal et al., "Automated multi-document biomedical text summarization using deep learning model," Computers, Materials & Continua, vol. 71, no.3, pp. 5799–5815, 2022.

[8] 14 - Multi-document Summarization Based on Unsupervised Clustering https://link.springer.com/chapter/10.1007/11880592_46

[9] Sentence Clustering-based Summarization of Multiple Text Documents https://www.researchgate.net/publication/256186750_Sentence_Clustering-based_Summarization_of_Multiple_Text_Documents

[10] An Statistical Tool for Multi-Document Summarization https://www.ijsrp.org/research_paper_may2012/ijsrp-may-2012-23.pdf

THANK YOU