CS 6301 Big Data Security and Privacy

Group Members:
Akshat Gangrade (Axg210048)
Ankit Sahu (AXS210226)
Sarthak Gupta (SXG200139)
Yash Shah (Yxs210015)

Topic:
Secure Content Filtering in Document Summarization

Aim:
This project proposes a novel approach to document summarization that integrates security protocols, ensuring sensitive information is appropriately filtered based on varying clearance levels among readers. This approach is crucial in environments handling classified information, where standard summarization methods may inadvertently compromise confidentiality.

Tools/Libraries:

1. Python (with NLP libraries such as NLTK or spaCy)
2. Secure Document Processing API (Proprietary if applicable)
3. Databricks

Proposed Method:
We plan to innovate beyond traditional extractive and abstractive summarization by introducing a security layer. Our method will leverage advanced NLP techniques for content understanding, combined with a security clearance protocol.

- **Phase 1:** Utilize NLP to parse and understand document content, identifying sensitive information based on predefined criteria.
- **Phase 2:** Apply a clearance check function that dynamically adjusts the summarization process, ensuring each summary is clearance compliant.
- **Phase 3:** Implement a review mechanism, possibly powered by machine learning, to continuously improve the sensitivity and accuracy of the clearance-level filtration process.

Concept:
Our strategy integrates the TextRank algorithm with a unique security clearance check. After splitting the document into sentences and determining sentence similarity, we'll apply clearance level checks. This step ensures that the final summary is both comprehensive and security-compliant, catering to various reader groups without compromising classified information.