Project Progress

Zeyu Liao
zeyu9@illinois.edu
University of Illinois

Liziqiu Yang (leader)
Liziqiu2@illinois.edu
University of Illinois

November 17, 2023

Contents

1 Project Overview

Our project focuses on text analysis and summarization, incorporating tokenization, stemming, TF-IDF computation, and the implementation of Gensim for text summarization. The primary goal is to create a Chrome extension that can extract and process data from web pages.

2 Accomplishments

1. Tokenization and Stemming:
    a) Successfully implemented tokenization and stemming for input data.
    b) These preprocessing steps ensure a standardized format for the text data, enhancing the subsequent analysis.
    c) Top 20 Frequent Words:
        i. Identified and extracted the top 20 frequent words from the tokenized and stemmed data.
        ii. This step provides insights into the most prevalent terms in the text corpus.
    d) TF-IDF Computation :
        i. Calculated TF-IDF scores for the text data.
        ii. TF-IDF helps in understanding the importance of each term in the context of the entire document corpus.
    e) Gensim Implementation:
        i. Integrated the Gensim library into our project for text summarization.
        ii. Utilized the built-in summary function in Gensim, which eliminates the need for explicit TF-IDF computation.

3 Challenges and Solutions

Initial Scope Consideration: Explored the Gensim library and found that it offers a summary function without the need for TF-IDF. This led to a potential scope reduction. Scope Expansion: Considering the reduced workload, we are exploring additional functions to enhance the project's overall capabilities.

4 Upcoming Tasks

1. Chrome Extension Conversion:
    a) Our next major task involves converting the project into a Chrome extension.
    b) This step will enable the extraction of data directly from web pages.

2. Scope Expansion:

   a) As we have identified a reduced workload with the Gensim library, we are actively considering additional features to enrich our project.

   b) Possible enhancements include sentiment analysis, keyword extraction, or advanced summarization techniques.

5 Conclusion

The project has made substantial progress in the implementation of text analysis and summarization techniques. We have successfully overcome challenges and are now poised to transition into the development of a Chrome extension. The potential scope reduction due to the Gensim library's summary function has prompted us to explore additional features, ensuring a more comprehensive and impactful project outcome. Overall, the project is progressing smoothly, and the team is enthusiastic about the upcoming development phases.