DSCI-560 Assignment No. 4 - Part 2

Team Formation

Team name: HealthEdBot

Names & USC ID

Cici Chang (8706354957)

Daoyang Li (7168949829)

Yifan Yang(8386626967)

https://github.com/DSCI560

Minutes of the offline meeting on February 12

Participants: project team members

Conference topic: Forum Analysis & Clustering Algorithms

In the meeting held on February 12th, the team decided that Cici Chang would lead the forum analysis and clustering algorithms section. Given her potential background in data analysis and proficiency with clustering algorithms, Cici is tasked with web scraping, data preprocessing, and implementing clustering algorithms to group similar messages. Her responsibilities include developing web scraping scripts to collect data from forums, cleaning and preparing the data for analysis, selecting and implementing a suitable clustering algorithm for the project, and refining the process to ensure the messages are accurately grouped.

Minutes of the online meeting on February 13

Participants: project team members

Conference topic: Forum Analysis & Clustering Algorithms

During the online meeting on February 13, project team members gathered to further discuss the specifics of the forum analysis and clustering algorithms segment of the HealthEdBot project. This session served as a deep dive into the methodologies and technologies that would be employed by Cici Chang in leading this segment. The team deliberated on several key aspects to enhance the effectiveness and efficiency of the project's initial phase.

Advanced Web Scraping Techniques: The discussion emphasized the need for sophisticated web scraping techniques that can dynamically adapt to different forum structures and layouts. Cici proposed the use of Selenium in conjunction with BeautifulSoup for more complex scraping scenarios that require interaction with web pages, such as navigating through pagination or handling JavaScript-generated content.

Data Preprocessing Enhancements: The team explored advanced data preprocessing techniques to ensure the quality of the data collected. This included the implementation of natural language processing (NLP) methods to clean and normalize the text data, such as removing stopwords, lemmatization, and handling emojis and slang. The goal is to refine the input data for more accurate clustering outcomes.

Clustering Algorithm Selection: A significant portion of the meeting was dedicated to discussing the selection of the most suitable clustering algorithm for the project. Cici presented a comparison between several algorithms, including K-Means, Hierarchical Clustering, and DBSCAN, considering factors such as scalability, efficiency, and the ability to handle noisy data. The team decided to initially experiment with K-Means due to its simplicity and effectiveness for large datasets but agreed to keep options open for algorithm adjustments based on trial results.

Parameter Tuning and Validation: The importance of parameter tuning for the chosen clustering algorithm was highlighted, with a focus on selecting the optimal number of clusters and other hyperparameters. The team discussed using methods like the Elbow Method and the Silhouette Score to evaluate the coherence and separation of clusters. Cici emphasized the need for a validation process involving manual review of clustered messages to ensure the algorithm's effectiveness in grouping similar content.

Collaborative Tools and Communication: Given the complexity of the tasks and the need for continuous collaboration, the team agreed on using collaborative tools such as GitHub for code sharing and version control, and Slack for ongoing communication. Regular updates and peer reviews were scheduled to ensure that everyone is aligned and to facilitate the sharing of insights and challenges encountered during the implementation phase.

Minutes of the offline meeting on February 14
Participants: project team members
Conference topic: Message Content Abstraction
In the meeting on February 15th, Daoyang Li was assigned to handle the message content abstraction component, which is crucial for converting messages into vector values representing their meanings. With a strong understanding of NLP and experience with neural network-based approaches like Doc2Vec, Daoyang is responsible for researching and selecting the most suitable method for message abstraction, implementing the chosen method to convert messages into vectors, integrating this component with the data collected from the forums, and verifying that the vectors accurately represent the content of the messages.

Minutes of the online meeting on February 15
Participants: project team members
Conference topic: Clustering Messages
During the meeting on February 14th, the team allocated the message content abstraction task to Yifan Yang. Yifan's role involves developing an algorithm for clustering messages based on the extracted keywords and the vectors generated during the message content abstraction phase. This task requires expertise in algorithms and natural language processing (NLP). Yifan's duties are to extract keywords from the

text of messages, implement an algorithm for clustering messages based on these keywords and their vector representations, use appropriate libraries or toolkits such as Scikit-Learn or NLTK, and visualize the clusters to ensure the contents within each cluster are similar.

Minutes of the online meeting on February 16

Participants: project team members

Conference topic: Automation

The discussion on February 16th focused on the automation aspect of the project. This component, essential for the system's real-time data processing capability, involves scripting and scheduling tasks to automate data collection, processing, and storage at fixed intervals. Although this requires a collective effort from all team members, one individual might take the lead to ensure smooth integration and automation of the processes. The team agreed on writing scripts for automation, implementing a command-line interface for real-time system interaction, and managing the system's stability while handling potential errors efficiently.

Finally we recorded a video to show our work.

Team name: HealthEdBot
Cici Chang (8706354957)
Daoyang Li (7168949829)
Yifan Yang(8386626967)