DS5110 Final Group Project – Iteration #3

Sharan Giri and Jason Leung

1. **Current and previously completed tasks in the last 2 weeks of our final project as well as who worked on each task.**

Sharan:

1. Dataset Preprocessing 1: Isolating user prompts and bot responses from raw Wildchat files using explode function and aggregating columns of interest back together again
2. Dataset Preprocessing 2: Filtering out all non-English language conversations as well as all toxic conversations and those with redactions containing potentially sensitive personally identifying information
3. Dataset Preprocessing 3: Removed stop words and applying necessary character checks such as return characters in preparation for NLP processing
4. Performed EDA on processed dataset during each preprocessing step to ensure the transformations were carried out 100% correctly
5. Researched and started coding basic Flask application for simple keyword search

Jason

1. Dataset Preprocessing debugging and error fixing
2. Pre-processed dataset EDA and error checking
3. Pre-processed dataset exploration of examples of conversations within Jailbreaking, Coding and Journalism categories, finalizing these categories
4. Researching Journalism and Jailbreaking categories and reading research papers
5. Learning the code and libraries for potential NLP text embedding models, choosing Gensim as our preferred library
6. Learning and developing proficiency with Gensim NLP models and prototyping with our intermediately pre-processed dataset

2. **Revised Timeline of the stages of our final project**

Week 3: Implement Simple Keyword Search module.

Sharan: Finished Wildchat pre-processing steps

Jason: NLP Research, finding text embedding/text summarization and text classification models for possible implementation.

Week 4: Selected the final categories of conversation topics, narrowed down from 5 to 3 with the chosen 3 being: Sharan doing Coding/Programming, Jason doing Jailbreaking/Journalist categories

Jason: NLP prototyping at least in Gensim, prototype one similarity-based text-embedding NLP model

Jason: Research, EDA and coding for Jailbreaking conversation category

Sharan: Research, EDA and coding for Coding/Programming conversation category. Prototype simple Flask interface

Week 5: Start technical report. Specialization and reporting on 1-2 ChatGPT conversation topics and buffer time for previous weeks.

Sharan: Build User Interface and Flask application for simple search based on simple single keyword

Week 6: Working draft of technical report. Integrating the products and tasks worked on in weeks 2 – 5.

Jason: Category #3 Journalism/Plagiarism reproduction of the core of this case study from research paper: https://www.researchgate.net/publication/381579112_Breaking_News_Case_Studies_of_Generative_AI's_Use_in_Journalism

Week 7: Create and practice final presentation. Finish technical report and finalize Github and code base.

Sharan: Contribute to final presentation. Finalize Github and code base.

Jason: Create and finalize final presentation as well as practice presentation dialogue with Sharan. Polish technical report.

Past Weeks:

Week 1: Overall Research and choosing of 3-5 different Chatgpt conversation topic areas: (Initial categories were 1. Jailbreaking 2. Journalism/Plagiarism 3. Coding/Programming

Prompts 4. Midjourney Image AI Generation 5. Education/Students asking questions to ChatGPT acting as a Tutor)

Week 2: Build full scale database management system excluding end user interface

We received authorization for access to full dataset including toxic conversations and we did a large portion of pre-processing. We also began NLP research.