DS5110 Final Group Project – Iteration #2

Sharan Giri and Jason Leung

1. **Project Kickoff questions**

What are the specific goals of this project?

The specific goals for this project include:

1. Effective Keyword search that filters similar chat interactions from the dataset.
2. We will build a full scale database management system from back-end to end user interface.
3. We will do comprehensive overall EDA and investigation of the entire Wildchat corpus within the English language.
4. We will investigate and annotate 3-5 different Chatgpt conversation topic areas and report on similarities and differences between these topics as well as novel details about some topics that the general audience can relate with.
5. We will distill our main findings into a technical report and final presentation tailored to a layperson audience.

How do we define the project scope clearly to avoid scope creep?

The 5 points are manageable core tasks that we will be able to deliver even within the 7-week project timeframe. Each of the 5 bullet points is a modest work product but will allow for extension or greater specialization if we have time in the later stages of the project. We have been careful not to expect grandiose results or unrealistic final products that will lead to scope creep, but we do hope to go in depth into one stage of the project, time permitting (such as NLP) that we can really learn well, develop skills, and specialize in.

What deliverables must be completed at different phases?

The project deliverables include:

1. Final Code with an integrated end-user interface and back-end to search for related ChatGPT interactions based on certain keywords.
2. Final Presentation will distill the main findings tailored to communicate effectively to any type of audience.

3. Technical Report which will include all the individual steps performed in the project from EDA through building database management system.
4. Git Repository which will include Dataset, Code, Final Presentation and Technical Report.

What are the major milestones, and what deadlines should we set?

1 week for each of the goals generally but we want some buffer time at the end of the 7-week timeframe and hopefully we can extend some specialization to be determined in part of that buffer time.

Week 1: Overall, EDA & choosing of 3-5 different Chatgpt conversation topic areas

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

Week 3: Build simple User Interface to search based on keywords. Begin NLP Analysis.

Week 4: Research and integrate OpenAI or another similarity-based text-embedding NLP model

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

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

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

Do the team's capabilities align with these goals? Are there any gaps that need to be addressed early on?

Learn Spark, Flask and NLP analysis. We do not have previous experience here and will need to learn these technologies and integrate to meet set goals. Although we do not foresee Spark and Flask being too difficult.

Do you have a dataset ready to use for the current project?

We already have our full dataset, the 1 million conversation WildChat dataset available here: Primary WildChat research paper as well as the WildVis visualization tool of the

conversations available here we will start our investigation of the dataset WildVis research paper. We have received authorization for use of the full dataset including toxic conversations from the authors of the Wildchat paper and have already downloaded it.

2. Team Discussions

What are the core skills each team member brings to the table?

Sharan – I have experience with coding in Python, and moderate experience with developing UI using Flask and integrating with back end to query data. I value time management and focus on contributing my best to achieve the goals set by my team.

Jason – I have moderate proficiency in Python, was an expert in Matlab and decent proficiency in Java if needed. I have basic general skills in Machine Learning and knowledge of the fundamentals of data science and most major ML models. In terms of soft skills, I would say that I am an adaptable and conscientious team partner, I always have the group goals in mind and am happy to contribute for the benefit of the group. I am patient, detail oriented and analytical and I don't mind helping anyone or being helped or reviewing theory or implementation as needed.

How will each person's expertise contribute to specific tasks?

We have come up with our 7-week project plan with the interests and backgrounds of both of our team members in mind. We will need to develop and practice some new technical skills, especially regarding Spark and NLP, but we are both detail oriented and want to learn in these new areas. We are both flexible in working together if need be on some tasks which end up being more difficult than we have predicted and although we plan to do less than half of the work independently and asynchronously, to increase our group time efficiency, we also specifically want to collaborate together in person in order to learn from each other and because we believe that will lead to the best group output.

What skills are missing that may cause delays or challenges?

At this stage of the project, we do not foresee any huge gaps or missing knowledge or academic theory, but we do need to become at least proficient if not skilled in new areas such as Spark, learning new NLP techniques and especially the implementation side of our project. We might need to build more Big Data skills within Spark or otherwise, that will allow us to handle and efficiently manage our 1 million size Wildchat conversation corpus.

What tools do we have experience with, and what do we need to learn?

Experienced with these Tools: MATLAB, Python (NumPy, Pandas, Scikit Learn), Flask (Intermediate), SQLite.

Tools we need to Learn: Spark, NLP models.

What programming languages and platforms should we select based on our project needs and team experience?

Whenever possible, we will utilize Python in implementation as it is the main language in which we both have proficiency. To handle the large volume of data in our dataset, we will use Spark to query data efficiently. We will be using Pycharm, Jupyter notebook and Google Colab as our coding platform.

3. **Skills & Tools Assessment**

Are there external resources or team members with expertise in the areas where we lack skills?

We plan to ask questions and attend multiple office hours of Professor Nafa for general help but especially if there is an area where we find we lack skills. If the scope of our project can be extended, time permitting, we hope to draw upon the NLP knowledge of Professor Nafa for some last NLP stage applied to the Wildchat dataset.

Which tools, frameworks, and libraries are most suitable for the project's scope?

The main tools and libraries we will be using are Pycharm, Jupyter notebook and Google Colab and within Python we will definitely work with Numpy, Pandas and Scikit-learn as well as a few other yet to be determined libraries. We might need to learn a new specific library or tool within Spark to deal with Big Data throughput.

How can we ensure that each team member is comfortable with the tools selected?

For any tool that is new or more difficult to implement we will work together as we believe that collaborating will allow us each to get more comfortable with any tools and to figure out any issues or confusion.

Have specific tasks been assigned based on individual strengths, and are team members clear on their roles?

Yes, specific tasks have been assigned based on our strengths, and we are clear on our roles. We feel comfortable in taking collective responsibility and staying engaged in every stage of this project. At the end of this stage, we split the learning of tools between the two of us and have started a basic EDA on the dataset. Both of us are flexible and

specifically want to collaborate to achieve the best group output. Jason will take on a note keeping role to type in group meetings, record work progress on tasks and to articulate a group message for written documents. Sharan will handle integrating the code and maintaining the versions and any revisions as we progress through the project.

4. **Initial Setup**

What development environment setup is necessary for this project?

We will be using PyCharm, Jupyter Notebook, and Google Colab as our coding platform for this project.

Have we successfully configured version control (such as Git)? Does everyone have access to the repository?

We have created a private Git Repository and have added us both as collaborators. Currently, our repository houses a third of the project data set.

Have we installed and configured all required software, libraries, and tools?

We have all the software mentioned above installed. We are in the investigation stage of our project and will be finalizing the libraries and tools over the next couple of iterations.

What testing can we run to ensure that the development environment is functioning correctly?

We will be planning to test the integration of our software/tools stack during the last 3 weeks (Integration phase) of our project.

What troubleshooting steps should we take if the setup does not work as expected?

We are facing an issue with uploading the entire dataset in Github, owing to personal account upload limits. We have currently pushed a third of our dataset and our investigating means to push the rest.

5. **Progress Review** (answering all 5 questions together)

Some of the obstacles or blockers that we foresee are learning the new-ish technologies for our project which include Spark, Flask, and NLP analysis tools. We have collaborated well and have done good work together virtually. For the meat of the project, we plan to meet 2 or 3 days a week in person on campus, usually Monday afternoons,

Tuesday afternoons and one of Wednesday, Thursday, or Friday, as well as needed for virtual meetings or quick check-ins. It is hard to exactly estimate milestones and timeline in this planning stage, but we want to have as much flexibility as possible and to complete milestones on the early side, so we have some buffer room before the week of the 26th of November. We are both happy with our progress so far. We narrowed down and decided on our topic 2.5 weeks ago and are confident about our current project status. However, we also think that we have a long way to go to reach our overall project objectives.

6. **Plan Revision**

Based on progress so far, do we need to adjust the project timeline or milestones?

There is a pretty tight timeframe for our project but in this our first stage of the project we do not think we need to adjust the project timeline.

Are any tasks delayed or requiring reassignment due to workload or skill gaps?

There have been no delayed tasks or tasks that at this point in time might need reassignment. We might find the need towards the middle or end of the project, but we have some buffer time built into our weekly plan that can hopefully allow for that.

How can we ensure that all members are clear on the revised plan and their next steps?

We are planning to meet at least 3 times a week and in person as well so we hope that any changes we make to a revised plan we will be able to clearly communicate to each other and agree upon.

What communication strategies can we implement to avoid future delays or misunderstandings?

So far, we have had very good communication. We do plan to check in with our professor multiple times in the 7 weeks of the project timeline and we can use that discussion time as another communication strategy to solidify and clarify anything needed in terms of project plan or technical implementation.

How will we track progress going forward and maintain alignment with the revised plan?

We will regularly refer back to our 7-week project timeline and adjust the plan if our time estimates are off. We should also both be encouraged and free to bring up any issues that come up with our own work tasks and to ask for help from each other or the professor if needed.

7. **Submission for This Iteration**

What specific tasks need to be documented for this iteration's submission?

We have documented all over our planning, team group process and discussion of potential challenges and risk analysis in this, the iteration #2 hand-in, and we have a overall 7-week timeline of tasks to complete the project.

Have we detailed the challenges faced, the solutions implemented, and any adjustments to the plan?

Yes, we have detailed the challenges faced in the previous sections of this document. They include:

1. Uploading Large Data Set to Github.
2. Even though we expect it to be a challenge, we want to pick up skills working with tools such as Spark, NLP and to successfully implement the project goals.

If your data is available online, please provide a link to access it.

Our full dataset is available online but requires release authorization just for the toxic subset of conversations. The nontoxic publicly available data is here: Non-toxic Wildchat dataset . The full 1 million with toxic dataset is here: Wildchat-1M-Full.

Is the PDF using the Overleaf template, and does it reflect the team's actual progress?

For the final technical report, we will use Overleaf, and any template provided. For iteration hand-ins we currently do not plan to use Overleaf.

Has everyone tracked their progress using the Excel file, and are we submitting it along with the PDF?

We will be tracking progress on the large-scale tasks of the project, and we can use a shared Google spreadsheet to map everything out, to take notes and to share and resolve any issues that come up in the workflow of each task.

Does the submission meet all the project requirements, and is it ready for review by stakeholders?

Our project plan should meet all project requirements, and we will try to frame our final presentation in a narrative or story as if we were presenting to some group of stakeholders that have interest in our research. At this point, everything is not ready for review.

Please provide a link to your GitHub repository with the updated files, including the PDF uploaded there as well.

We will be uploading our completed iteration #2 to this Github. We will not make the Github public until the week of final presentations and we should not make the full toxic dataset public even then so we will investigate a way to password protect the toxic dataset on Github.

https://github.com/sharan0276/WILDCHAT