Meeting #1 Minutes – First Client Meeting

Date: 09/07/17

Time/Duration: 9:30am – 10:47am

Participants:

- Seth Forrest
- Paden Rumsey
- Austin Sass
- Dr. Frenzel
- Dr. Hemati

Summary:

Lise was not present due to an error that we made in our initial scheduling. It was not her fault. We recorded the meeting and sent an mp3 file of it to her later. During the meeting we made our introductions. Dr. Frenzel primarily spoke with input from Dr. Hemati. We were brought up to speed on the current state of the project which is little more than a few python scripts. They told us the overall goal of the project was to build an automated web mining system that could generate data from IBM's Watson. From that data we were to generate a "clone" of Watson using machine learning to create a new machine.

Meeting #2 Minutes – Second Team Meeting

Date: 09/11/17

Time/Duration: 9:30am – 10:45am

Participants:

- **Austin Sass**
- **Lise Welch**
- **Paden Rumsey**
- **Seth Forrest**

Summary:

In this meeting we made sure that Lise had listened to the mp3 file, which she had. After this we spent most of the meeting boiling down the requirements from the meeting that we had participated in with the client the week prior. These requirements were written down and placed in a Latex documented by Paden Rumsey with the document being generated by Austin Sass. We then doled out assignments for the next week. Below.

Owner	Task	Start	End	Done ?
Lise/All	Set up Bitbucket repository. Members should join.	9/11	9/18	
Paden/Seth	Type up requirements in design doc. Send to client	9/11	9/18	
Paden/All	Finish typing up contract. Finish amending. Sign	9/11	9/18	
Seth	Reach out to Bruce about being at a meeting	9/11	9/18	
Lise	Transcript of meeting with client.	9/11	9/18	
Austin/All	Look at tools for data/web mining and put in google drive	9/11	9/18	

Meeting #3 Minutes – Third Team Meeting

Date: 09/18/17

Time/Duration: 9:30am – 10:20am

Participants:

- **Austin Sass**
- **Lise Welch**
- **Paden Rumsey**
- **Seth Forrest**

Summary:

Everybody went over some of the contents of their research papers. Seth and Lise read only the abstracts so we didn't go into as much depth. Austin found a good source for text mining in the form of an O'Reilly book and Paden found a list of text mining tools. After that Paden shared some of his concerns about the project. Then we talked about a production schedule for this semester which looks very similar to this

Phase One

- Data mining tool chosen
- Initial production schedule set

Phase Two

- JSON (or whatever format) placed in the database from Watson
- Text parsed, preprocessed as output from online data

Phase Three

- Database has all the info we require
- Text mining is essentially set up.

Owner	Task	Start	End	Done ?
All	Research Papers are read (except Austin's)	9/18	9/25	
All	Preliminary Idea of mining tool	9/18	9/25	
Seth	Reach out to client about question	9/18	9/25	

Meeting #4 Minutes –Fourth Team Meeting

Date: 09/25/17

Time/Duration: 9:30am – 10:45am

Participants:

- **Austin Sass**
- **Lise Welch**
- **Paden Rumsey**
- **Seth Forrest**

Summary:

In this meeting we discussed a couple of things. The first thing we discussed was our results for reading our research papers and trying to find a scraper. Preliminarily we don't have anything concrete although we are getting closer to a solution. Having read all the research papers Paden outlined some of the requirements that we might have for the machine in addition to the recommendations that Dr. Heckendorn gave him. After this we split up the tasks for the snapshot day poster. After those were distributed we assigned specific team roles for the project.

Owner	Task	Start	End	Done ?
Lise	Set up database and put it on bitbucket.	9/25	10/4	
Lise	Poster – Team name/members/sponsors/ Descriptive project title	9/25	10/4	
Seth	Narrow down scrapers to 4 or 5	9/25	10/4	
Seth	Poster – Design goals, Deliverables/Problem Statement	9/25	10/4	
Austin/Paden	Look at research papers focusing on the word "survey."	9/25	10/4	
Austin	Poster – Diagram/Compose list of research in word doc	9/25	10/4	
Paden	Poster – Word doc with Specifications Table/ Schedule for our plan of completion	9/25	10/4	

Meeting #5 Minutes – Fifth Team Meeting

Date: 10/02/17

Time/Duration: 9:30am – 10:30am

Participants:

- **Austin Sass**
- **Lise Welch**
- **Paden Rumsey**
- **Seth Forrest**

Summary:

Seth went over some of the stuff that he researched. He suggested that we used Twitter to find our data. He suggested a tool called Twerpy that is an API to help mine data from Twitter. Next week he will choose out the final scraper. Lise hadn't gotten to the database stuff yet. She will create the database the day after the meeting and start learning some of the commands that MySQL server required. Paden told the group about how IBM does its personality analysis. IBM moved from a tool called Linguistic Inquiry and Word Count (classifier) to a tool from Standford's NLP library called GloVe. GloVe uses data compression to reduce bodies of text to vectors based on co occurrence of words. Austin and Paden will look at GloVe in combination with Gaussian Processes to get a better idea of how to code the machine. For the next meeting with the client we want to nail down the following questions.

How important is a refined scraping tool versus having the data?

We aren't going to be doing word frequency. It's not necessary

Need to know how much money we have for Watson's calls?

What do we want to go in the database?

What's a good meeting time for the future?

Owner	Task	Start	End	Done ?
Seth	Pick final scraping tool	10/02	10/09	
Lise	Look up basic MySQL commands and start learning	10/02	10/09	
Paden/Austin	Look at GloVe code and Gaussian Processes	10/02	10/09	
Paden	Set Agenda for Wednesday	10/02	10/02	

Meeting #6 Minutes – Sixth Team Meeting

Date: 10/09/17

Time/Duration: 9:30am – 10:30am

Participants:

- **Austin Sass**
- **Lise Welch**
- **Paden Rumsey**
- **Seth Forrest**

Summary:

Paden went over last week's activities and evaluated their progress. We also discussed snapshot day and what would be expected from us for that day. Paden then outlined the final goals for this semester. He placed and emphasis on data collection for Seth and Lise while he and Austin worked out how the machine learning algorithm was going to work. He assigned roles for this week.

Owner	Task	Start	End	Done ?
Seth	Update Bruce on team progress and ask about snapshot day	10/09	10/16	
Seth	Look at Kaggle databases and check to see if there's any data that exists. Start working with Tweepy	10/09	10/16	
Austin	Analyze Glove code well enough to explain it to Paden next week	10/09	10/16	
Paden	Get in contact with Statistics Assistance Center to learn more about Gaussian processes. This might include a trip to IRIC	10/09	10/16	
Lise	Look for data on Kaggle. Get in contact with client about where to store database for large amounts of data.	10/09	10/16	
Lise	Revise snapshot day stuff in the google drive to look good. Print out and come early to snapshot day	10/09	10/10	

Meeting #6 Minutes – Sixth Team Meeting

Date: 10/16/17

Time/Duration: 9:45am – 10:20am

Participants:

- **Lise Welch**
- **Paden Rumsey**
- **Seth Forrest**

Summary:

Paden went over the snapshot day. What we could do better and what we really excelled at. We need to increase the portfolio strength. Include the contract for the team, tabs, resources that we are using to base our project on (links and papers). Also we need to make sure the document is up to snuff every step of the way. Paden then went around and asked about the tasks that everyone has. Seth did not check in on Bruce the week before but will do it this week. Informed me that Tweepy only has a limited amount of API calls. 450 every 15 minutes. That's a little over 43,000 a day. We might have to get the four of us to sign up and make a profile. Seth will have pulled a Twitter user by next Monday using Tweepy. Lise has downloaded MySQL and will have indexed an example JSON set from Watson Personality Analysis. Will also have contacted Dr. Frenzel and Hemati about database storage, money for API calls, and a place to run the scraping algorithm. Paden talked about his experience with the Statistics Consulting Center. He was transferred to another person in the department and is awaiting their email. He will meet with Austin this week and talk about GloVe and check to see when they can start programming it. He will also attempt to set up a consult with the SCC this week. Austin was not here.

Owner	Task	Start	End	Done ?
Seth	Update Bruce on team progress. Ask about expectations for the rest of the semester.	10/16	10/23	
Lise	Index example JSON from Watson	10/16	10/23	
Seth	Will pull a twitter user	10/16	10/23	
Lise	Ask Dr. Frenzel/Hemati about data base storage, API costs, and place to run scraping algorithm.	10/16	10/23	
Paden/Austin	Meet together and talk about the GloVe code base. (1 to 2 hours)	10/16	10/23	
Paden	Meet with Statistics Consulting Center	10/16	10/23	

Meeting #8 Minutes – Eighth Team Meeting

Date: 10/23/17

Time/Duration: 9:30am – 10:10am

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**

Summary: Paden first started off by evaluating last week's actions items. With the exception of the meetings that we had tried to schedule we managed to complete all our tasks. Bruce was hard to find for Seth. Lise was unable to meet with Dr. Frenzel or Hemati until tomorrow and the Statistics Assistance center didn't get back to Paden until Friday of last week and didn't respond all last weekend. Seth pulled a twitter user and Lise indexed some JSON from Watson. Seth looked at the optimal methodology from Tweepy and thinks he can send JSON to Lise's database. Paden asked whether or not he could pull the text body from the tweet and store as a separate file. Seth has read he can pull 45,000 users in 15 minutes. Lise finished her task but is now testing optimal format for Watson into a database. Paden then talked about Design review. Before we engage in it we have to get some fundamental questions asked and answered. Each person will be responsible for part of the power point. Seth is responsible for a room and a time. We then talked about the binder and documentation. Seth will include API instructions as well as a summary of his process. Lise will include her organizational structure as well as frequent commands/names of databases. Austin will include the GloVe paper and have code summary/breakdown. Paden will include selections from the Gaussian Process book and a code breakdown/summary.

Owner	Task	Start	End	Done?
Seth	Read methodology for tokenizing/test optimal methodology for tweepy use	10/23	10/30	
Seth	Set up Room time for design review	10/23	11/06	
Lise	Test optimal methodology for indexing data base user	10/23	10/30	
Lise	Ask Dr. Frenzel/Hemati about data base storage, API costs, and place to run scraping algorithm.	10/23	10/30	
Austin	Wikipage – initial startup	10/23	10/30	
Austin	Thoroughly read GloVe code	10/23	10/30	
Austin	Set up meeting with math professor for GloVe paper	10/23	10/30	

Paden	Meet with Statistics Assistance Centwer	10/23	10/30	
Paden	Read Gaussian Process Code	10/23	10/30	

Meeting #9 Minutes – Ninth Team Meeting

Date: 10/30/17

Time/Duration: 9:30am – 10:20am

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**

Summary: The meeting started off with Lise listing all the things we had due next week. This included logbooks and portfolios. Paden will compile the portfolio. Lise also talked about Dr. Hemati and Dr. Frenzel's reponse to our three questions. This generated some controversy and we will set up a meeting with them. Specifically about the adjective filtering that Dr. Hemati talked about. We talked about Seth and the scraper. He wasn't able to complete everything but he's going to use the amount of Re-Tweets to determine who is a bot. We only need 3000 words anyway. We updated what was going in the database. This included the URL (included in JSON object from Tweepy), the name of the author (included in JSON object), Watson Personality Insight's Personality Analysis and our machine's analysis. Austin is going to set up a meeting with Manuel Welhan in the math department to decipher some of this glove stuff. Paden scheduled a meeting with the SCC and read over the Gaussian Process stuff. He said that the file was about 150 lines of code. It should possibly be re-written in C (as it is now in matlab) for speed. The other files included mean, covariance, likelihood and inference functions. We will each generate powerpoint slides for the design review for our respective parts and in the design document due next week as well. This week is going to be focused on handling administrative stuff with some machine learning interspersed. We affirmed out commitments to the November deadlines.

Owner	Task	Start	End	Done?
All	Generate slides for design review powerpoint	10/30	11/06	
All	Generate documentation for each part	10/30	11/06	
Seth	Set up Room time for design review	10/23	11/06	
Lise	Test optimal methodology for indexing data base user	10/23	End of Nov.	
Austin	Take GloVe to Manuel (or set up meeting based on his schedule)	10/23	10/30	
Austin	Update wikipage with stuff from Design Doc that will be placed there.	10/30	11/06	
Austin	Set up meeting with math professor for GloVe paper	10/23	10/30	
Paden	Meet with SCC (Finally)	10/30	10/31	

Paden	Compile Portfolio	10/30	11/06	
Paden/Seth	Meet with Dr. Hemati on Wednesday	10/30	11/01	

Meeting #10 Minutes – Tenth Team Meeting

Date: 10/30/17

Time/Duration: 9:30am – 10:20am

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**

Summary: We started off by going over next week's action items. With the design doc and powerpoint stuff done, we have set up a meeting on Thursday, November 9, 2017. Seth is now able to get a list of users and can run quite a few samples at a time. Austin did not take any of the GloVe stuff to professors yet. But that was more of the professors not responding. He will just go in and talk to them now. Paden gave him some examples of professors: Stefan Teuheano, Robert Ely. Paden and Seth told the team about Dr. Hemati's meeting. About how they have 200,000 amazon reviews that we can use. But we won't be able to run all those through Watson. So everybody is going to sign up for multiple Watson accounts so we can have at least 24,000 Watson samples by Christmas. Austin will run those 200,000 samples through GloVe and Lise will index 24,000 by December 20, 2017 and having 12,000 done before November 20, 2017. Paden will continue to look at Weka and CNN's and Seth will continue to automate the scraper. Final deadlines are

Mid-February for initial results

March 31st for final deliverables

Owner	Task	Start	End	Done?
All	Make Watson Account for Samples	11/06	11/08	
All	Make Presentation for their powerpoint section	10/30	11/09	
Seth	Continue to automate Tweepy	11/06	11/13	
Lise	Get set up on database system	11/06	11/13	
Austin	Take GloVe to professor	11/06	11/13	
Paden	Use tutorial for WEKA	11/06	11/13	
Paden	Study and analyze code for CNN's	11/06	11/13	

Meeting #11 Minutes – Tenth Team Meeting

Date: 11/10/17

Time/Duration: 9:30am – 10:20am

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**

Summary: We went over the powerpoint. Each took turns doing their own respective sections and we gave feedback on what they were saying. Not a diverse meeting. Overall we decided not to make any sweeping changes to any sections. We did assign Lise to talk about the PDF's section of the powerpoint though.

Meeting #12 Minutes – Twelfth Team Meeting

Date: 11/13/17

Time/Duration: 9:30am – 10:20am

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**

Summary: Went over last week's stuff. Some account samples were made by the team. We have 8 in total. We are aiming for twelve. Seth and Lise still need to make more to get up to 3. Seth is continuing his work on the Twitter stuff. Lise tried to get signed up for the database but their server is having issues so that action item was stalled. It will be moved to whenever they get it working again. We all created sections for our powerpoint. Austin took GloVe to a math professor Manuel Welhan who didn't feel comfortable enough with the material to give too much input. Austin has identified several more people who can help him however and will continue to look into that. Paden looked into Weka and figured out some of the features it provides. It has the ability to export code in Java format based on a the model we use. However, it appears as though Gaussian Processes are not among those models. He will attempt to use the CNN samples over Thanksgiving break and get a pipeline running for when the GloVe vectors are trained.

Owner	Task	Start	End	Done ?
All	Make Watson Account for Samples	11/13	Indefinite	
All	Design Review	…	11/14	
Seth	Rewrite code to do multiple calls at a time. Structure output to fit in database	11/13	11/27	
Lise	Get set up on database system	11/06	Indefinite	
Lise	Index as many of 12,000 samples as possible before thanksgiving	11/13	11/27	
Austin	Train GloVe vectors on as many of 200,000 amazon reviews as possible	11/13	11/27	
Paden	Use CNN on sample input from paper	11/13	11/27	
Paden	Identify possible other alternatives for machine learning toolkit.	11/13	11/27	

Meeting #13 Minutes – Twelfth Team Meeting

Date: 11/27/17

Time/Duration: 9:30am – 10:20am

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**

Summary: Review progress over break. Lise has been unsuccessful accessing amazon reviews despite contact with client - follow up with clients and/or grad student. Lise looked into Google Cloud: $300 dollars per account free. No progress update from Seth, but he will continue work on scraper as well as looking into running it on the cloud. Austin was unable to train GloVe vectors without data, but did look at scripts for doing such task. Paden continued study of machine learning toolkits over break. Austin and Paden will have to condense design review slides for snapshot day #2.

Owner	Task	Start	End	Done ?
All	Make Watson Account for Samples	11/13	Indefinite	
All	Snapshot Day #2 Prep.	…	11/1	
All	Portfolio, final edits	11/27	12/8	
Seth	Collect tweets with scraper	11/27	...	
Lise	Update client on database access	11/27	12/4	
Austin	Train GloVe vectors on as many of 200,000 amazon reviews as possible	11/13	11/27	
Austin	Update wiki page	11/1	12/8	
Austin	Search for data sets	11/27	12/4	
Paden	Check Google Cloud support	11/27	12/4	
Paden	Use CNN on sample input from paper	11/13	11/27	
Paden	Identify possible other alternatives for machine learning toolkit.	11/13	11/27	

Meeting #14 Minutes –Fourteenth Team Meeting

Date: 12/04/17

Time/Duration: 9:30am – 10:10am

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**
-

Summary: Snapshot day has been completed and turned out quite well. Seth has set up Tweepy to start collecting multiple profiles. Lise resolved the data crisis that we were having the week before. Apparently the data is in no singular location. We have to go out and find it and parse it ourselves. Austin has identified an amazon database that has 36 million reviews in it. So that's a possibility in addition to the Twitter stuff. We spent a large chunk talking about the Google Cloud Platform. It appears to be extremely viable. We can run multiple instances of our scraping algorithm and store large amounts of data there for awhile for free. Trying to work around the "buckets" (the storage in GCP) is kind of difficult but we will all attempt to find solutions for it. Paden ran the CNN on GCP but ran into some issues. He left something in the issues thread for the creators of the algorithm and will continue to try and fix it himself. He is working through those issues. Our last meeting will be next Tuesday at about 1:30 P.M.

Owner	Task	Start	End	Done?
All	Make Watson Account for Samples	11/13	Indefinite	Ongoing
Lise	Parse amazon reviews	…	12/25	
All	Portfolio, final edits	12/4	12/8	
Seth	Collect tweets with scraper	12/4	Indefinite	Ongoing
Seth	Update client on overall progress	12/4	12/12	
Austin	Train GloVe vectors on as many of 200,000 amazon reviews as possible	11/13	12/25	Ongoing
Austin	Update wiki page	12/4	12/5	
Paden	Attempt workaround for CNN problems and run it	12/4	12/25	

Austin	Request Data for other Amazon reviews	12/4	12/12	
Austin/Paden	Identify possible other alternatives for machine learning toolkit.	11/13	12/25	Ongoing

Semester 2 - Meeting #1 Minutes

Date: 01/16/18

Time/Duration: 9:30am – 10:10am

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**
-

Summary: Thousands of Tweets were collected over the break. No samples indexed through Watson although Amazon 12-15,000 amazon reviews were parsed for entry into Watson. No progress was made on the machine learning bit. There is a part that Paden is stuck on. He believes there might be some versioning problems with the Python software he is working on. We thought maybe running the software on some other computer might do the trick. So Seth is going to attempt to run the algorithm on his machine now that the Twitter stuff is completed. Our weekly meeting with be on Thursday at 11 in JEB. The same room if we can manage it. We are close to being able to generate data. In order to run the CNN the data must first be ran through another personality recognizer based on Linguistic Word Inquiry Count (LIWC). Then we can put the samples through the actual CNN.

Owner	Task	Start	End	Done?
Seth	Try to run the CNN with Python on his own computer	1/16	1/23	Ongoing
Lise	Parse amazon reviews	…	Ongoing	X
Lise	Finish automation script for Watson`	1/16	1/23	
Paden	Run data through Mairesse Personality Recognizer	1/16	1/23	X
Austin	Identify possible other alternatives for machine learning toolkit.	11/13	Ongoing	X Fizzled Out

Semester 2 - Meeting #2 Minutes

Date: 1/25/18

Time/Duration: 11:00 am – 12:00 pm

Location: IRIC

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**

Summary: Seth is still trying to run the CNN on his computer. List is working on getting Amazon reviews through Watson. She is having an issue with CURL on GCP. Paden ran the data through the Mairesse personality analyzer. Austin concluded that there is no suitable alternative to CNN for our skill level. The design review will take place several weeks from now. Running the CNN is a priority so now we will have three people attempt to run the CNN. A new schedule was determined. We will have the CNN running by February 9th. The Watson Script will be completed by February 1st. By March 9th we will have the PDF generator running.

Owner	Task	Start	End	Done?
Seth	Email Bruce and Hemati	2/01	2/08	X
Paden, Austin, Seth	Get the CNN to actually run	1/25	2/08	
Lise	Finish automation script for Watson`	1/16	2/08	X
Paden	Read and understand Python	1/25	2/08	X

Semester 2 - Meeting #3 Minutes

Date: 02/08/18

Time/Duration: 11:00 am – 11:45 pm

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**

Summary: The team update Bruce on all aspects of the parts of the project. The first was the text scraper that we talked about. We have the scraper operating but the data is dirty. We will need to clean it but that's not our priority. Next we talked about the Watson input automation script. Lise made some headway Paden suggested that we work in teams to increase efficiency. Seth is going to help Lise next week in finishing the script. Next we talked about the machine learning bit. Paden got in contact with the author of the original CNN. They didn't use a machine that was much more powerful than any of the team's. So Paden assumed there must be some issue with how we are running the CNN. Paden had a few theories and will follow up on them. Paden and Austin will work together to try and get it running. Niloo will give Paden access to the Watson files that have already been analyzed.

Owner	Task	Start	End	Done?
Seth/Lise	Finish automation script for Watson	2/01	2/08	x
Paden, Austin, Seth	Get the CNN to actually run	1/25	2/08	x
Lise	Finish automation script for Watson`	1/16	2/08	x
Paden	Read and understand Python	1/25	2/08	x
Paden	Send Niloo zip of tweets and essays and get access to Dr. Frenzel's database	2/01	2/08	x

Semester 2 - Meeting #4 Minutes

Date: 02/08/18

Time/Duration: 11:00 am – 11:45 pm

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**

Summary: Updates on the Watson script. CURL is a tool that cannot be used on Google Cloud Platform. Lise will run the script locally and then upload the files. Paden/Austin were able to run the CNN on Austin's GTX 1080i Nvidia graphics card. It appears that most of Theano's applications can only be ran on Nvidia drivers. Now we must change the CNN to score the Watson data as it's not ready for prediction quite yet. We also must change the CNN from a regression problem to a classification one. This might be simple. It might not be. Design review is next week. We will use the other powerpoint as our initial base and then alter that.

Owner	Task	Start	End	Done?
Paden, Austin, Seth	Meet at 1:30 on Friday to go over the code in the CNN	2/08	2/16	X
Austin	Run Paden's changes on the CNN	2/08	2/16	X
Paden	Alter CNN to give scores to Amazon sample (y or n)	2/08	2/16	X
Paden	Get access to lab (not essential now they don't have graphics cards on their machines)	2/08	2/16	
Everyone	Work on their part for Design review	2/08	2/16	X

Semester 2 - Meeting #5 – Design Review

Date: 02/16/18

Time/Duration: 11:30 am – 12:30 am

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**
- **Niloo Hezarajibi**
- **Dr. James Frenzel**
- **Dr. Saied Hemati**

Summary: Went over the current design of the personality analyzer. We first discussed the scraper. We told the client that the information is currently dirty but we are going to work on cleaning it. After that we told them we had a loose collection of files that have yet to be put into a database. While talking about the machine learning stuff they asked about running the CNN on an NVIDIA graphics card and how much those might cost. They informed us that they do not have a graphics card in any of their machines. We also informed them that we might not be able to change the CNN to a regression problem by our own merits. It's possible. After the design review was over Paden assigned everyone next week's roles.

Owner	Task	Start	End	Done?
Paden , Austin	Run classifier over the amazon reviews that we have	2/16	03/01	Ongoing
Paden	Change the CNN to do prediction and run it	2/16	2/23	X
Lise	Continue to run Amazon samples through the Amazon server. Talk with Dr. Hemati about writing a script to calculate PDF's	2/16	2/23	Ongoing
Seth	Continue to collect data and meet with Paden to discuss ways to clean up the data	2/16	2/23	X

Semester 2 - Meeting #6 – Design Review

Date: 02/23/18

Time/Duration: 11:00 am – 11:25 am

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**

Summary: Lise indicated that we have a snapshot day and a portfolio that needs to be turned in as well. Team citizenship forms will be handed in electronically. Last week, Lise managed to run more Amazon reviews through the Watson system and now we have about 7500. Seth will clean the data by removing hashtags and hyperlinks and stripping away emojis. Paden is still working with Austin to get the classification done.

Owner	Task	Start	End	Done?
Paden , Austin	Run classifier over the amazon reviews that we have	2/16	03/01	X
Paden	Fix the CNN's prediction problems	2/23	03/01	X
Lise	Continue to run Amazon samples through the Amazon server. Talk with Dr. Hemati about writing a script to calculate PDF's	2/16	…	Ongoing
Seth	Continue to collect data/clean	2/16	3/09	Ongoing
Seth, Austin	Meet together to understand the CNN	2/23	3/01	X

Semester 2 - Meeting #7 – Design Review

Date: 03/01/18

Time/Duration: 11:00 am – 11:25 am

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**

Summary: Several meetings were performed the prior week and breakthroughs were made. Paden was able to figure out the problem with the classifier with Austin's help. He also met with Seth and Austin and showed them how the code worked so they could help on the Regression problem. Graphics card information was sent by Austin to Dr. Hemati. Austin will be responsible for adding a method to append classification output from CNN to an output file containing the file name and it's text. Paden will figure out why the Mairesse features aren't uniform in size. Preparations for snapshot day need to be done. Citizenship forms need to be completed.

Owner	Task	Start	End	Done?
Austin	Create method to append values to output file for the classifier in preparation for regression.	03/01	03/09	X
Paden	Find out why the Mairesse features aren't uniform.	03/01	03/09	X
Paden	Add description of CNN to portfolio/design doc	03/01	03/06	X
Lise	Continue to run Amazon samples through the Amazon server. Create PDF script.	2/16	…	Ongoing
Lise	Print out slides from Design review to use in Snapshot day	03/01	03/06	X

Seth, Lise, Paden	Print out code for their various sections for the portfolio.	03/01	03/06	X Owner Lise Lise Paden Seth Austin Everyone Paden/Austing
Seth	Continue to collect data/clean	2/16	..	Ongoing

Semester 2 - Meeting #8 – Design Review

Date: 03/19/18

Time/Duration: 1:30 pm – 2 pm

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**

Summary: Seth was able to finish a preliminary data cleaning algorithm. It removes any language that doesn't utilize English characters and removes hyperlinks and references. He now needs to try to filter out English character, non-English languages. Paden and Austin started work on changing the neural network to do regression instead of classification. They are still working on it. Lise talked to Dr. Hemati and he wants a histogram of the different PDF's she will calculate. We told her that going to the Statistical Consulting Center might be beneficial. Aside from this, Austin needs to touch up the wiki and figure out a way to print out the weights of the network to make it easier to do prediction. In addition to this, everyone needs to create a README for their programming sections.

Owner	Task	Start	End	Done?
Lise	Consult SCC to do histogram	03/19	03/26	X
Lise	Make script to input data into a mySQL database	03/01	Mid-April	
Paden	Change the Neural Network to do regression	03/01	Mid-April	X
Seth	Figure out a way to filter out foreign languages in the data	03/19	03/26	
Austin	Touch up wiki	03/19	03/26	X
Everyone	Create a README for their section	03/19	Mid-April	Ongoing
Paden/Austing	Meet Twice this week to work on regression	03/19	03/26	X

Semester 2 - Meeting #9

Date: Thursday, 03/29/18

Time/Duration: 11:00-12:00

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**
- **Dr. Hemati**

Summary: Instead of the SCC Lise consulted Dr. Hemati to get more information about the PDF's. He wants them in a specific histogram style. Seth said that in order to filter out other languages that don't use non-english characters he would have to re-write his cleaner. Paden has rewritten the CNN to do regression. This involved changing the dimensions of the different filters and also changing the objective function that calculated the loss from the network. Dr. Hemati stopped in and said that analyzing a trait like Neuroticism might perform better because the distribution for that data is uniform. Paden said there are still bugs in the CNN and before he can run it he needs to take care of those.

Owner	Task	Start	End	Done?
Lise	Work on the PDF's now that Hemati has been consulted	03/29	Mid-April	Ongoing
Lise	Make script to input data into a mySQL database	03/01	Mid-April	Ongoing
Paden	Take care of the bugs in the neural network and run regression tool	03/29	04/05	X
Seth	Collect more data while thinking of a way around the language barrier issue	03/29	04/05	X
Austin	Run classifier	03/29	04/05	X
Everyone	Create a README for their section	03/19	Mid-April	Ongoing

Semester 2 - Meeting #10

Date: Thursday, 04/05/18

Time/Duration: 11:00-11:30

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise Welch**
- **Dr. Hemati**
- **Niloo**
- **Bruce**

Summary: Seth's scraper is running but there are an unknown number of samples. Paden requested a number of how many we have by next Thursday. Niloo's script was able to filter out non-English languages so she will pass that onto us as a guide to alter our own. Paden was able to print out the weights of the network but was having some issues reloading them back into another model. The regression problem is also working but the data is less than accurate. He gave a number of reasons, which include: lack of data, incorrect loss function, etc. This weekend he will run it on more samples and observe the resulting data. Lise is continuing to work on PDF's. Austin has offered his help and they will meet next week.

Owner	Task	Start	End	Done?
Lise	Work on the PDF's now that Hemati has been consulted	03/29	Mid-April	Ongoing
Lise	Make script to input data into a mySQL database	03/01	Mid-April	Ongoing
Paden	Run the regression CNN on more data	04/05	04/12	X
Seth	Collect more data and implement Niloo's tool in filter.	04/05	04/12	X
Austin/Lise	Help Lise do the histograms for the PDF's	04/05	04/12	X
Everyone	Create a README for their section	03/19	Mid-April	Ongoing

Semester 2 - Meeting #11

Date: Thursday, 04/12/18

Time/Duration: 11:00-11:30

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Niloo**
- **Bruce**

Summary: Seth was able to reimplement the scraper using the files that Niloo gave us and now it filters out non-english characters. Austin will be doing the histogram work this week. Paden ran the regression task on a larger set of the data and the results were promising. The Mean Square Error, the metric by which Watson actually grades it's data, was similar to the error that Watson gets. We got .10-.14 MSE and Watson gets about .12 MSE. This is good news. Paden is cautiously optimistic and wants to run more data to verify the results. This means that the original requirements for the project are almost complete and we can start looking towards delivering the data/project in the future.

Owner	Task	Start	End	Done?
Lise	Works on the histograms for the PDF's	04/12	04/19	Ongoing
Lise	Make script to input data into a mySQL database	04/12	04/19	Ongoing
Paden	Verify the regression results using an entirely new set of data	04/12	04/19	X
Seth	Give an accurate count of the data that we have. Keep collecting data	04/12	04/19	X
Everyone	Create a README for their section	03/19	Mid-April	Ongoing

Semester 2 - Meeting #12

Date: Thursday, 04/19/18

Time/Duration: 11:00-11:30

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise**
- **Niloo**

Summary: Seth gave us a count of how many files that we had 6,000 useable Tweets. But he will keep running the scraper. Lise was finishing up doing the histograms for the PDF's. Paden and Austin continued to work on running the data through and verifying regression. The percentile results seemed to be a little worse on larger sets of data but still maintained quite a bit. The raw results were a lot closer but the uniformity in the data seems like it might be causing this. They are done running data now and they are just going to put all the samples in a single CSV that will be put into a database by Lise. Aside from this, Expo is next week and the poster is going to need to be designed. Seth said he would do this. Paden is going to deliver the Oral Presentation at the expo and is responsible for that. Everyone continues to add documentation to the design document and a final report will need to be generate before the end. Paden and Austin will attend Dr. Hemati and Dr. Frenzel's weekly meeting to inform them of the results of the project.

Owner	Task	Start	End	Done?
Lise	Works on the histograms for the PDF's	04/12	04/19	Ongoing
Lise	Make script to input data into a mySQL database	04/12	04/19	Ongoing
Paden	Verify the regression results using an entirely new set of data	04/12	04/19	X
Seth	Give an accurate count of the data that we have. Keep collecting data	04/12	04/19	X
Everyone	Create a README for their section	03/19	Mid-April	Ongoing

Semester 2 - Meeting #13

Date: Thursday, 04/26/18

Time/Duration: 10:00-10:30 AM

Participants:

- **Austin Sass**
- **Paden Rumsey**
- **Seth Forrest**
- **Lise**

Summary: Essentially we are wrapping up. Here is a list of everything that needs to be done and dumped in Dr. Frenzel's Google Drive:

- README's (Everyone)
- Final Report Final Draft (Paden and Austin)
- PDF Histograms (Lise)
- Portfolio (Paden)
- PDF's (Lise)
- Raw Data (Seth)
- Database (Lise)
- Code (Everyone)
- Copy of the Design Document from Overleaf (Paden)
- Research Papers we found (Paden)

This all has to be compiled and finished by the end of next week. Aside from that we just talked about Expo and there might not be a meeting next week depending on if we finish. Everybody has to get hard copies of their code to Paden before Wednesday so he can compile the portfolio.