Project Descriptions

"CS + X" is a topics course designed by Dr. Yael Gertner, graduate student Kathleen Isenegger, and a team of collaborators for high school students to engage with computer science and its relationship to other areas of discipline at the University of Illinois. Our unit for this summer is called "CS + Linguistics" and was created to show the connections in the two fields through topics such as Python implementation, machine learning, natural language processing, and artificial intelligence. This unit will be taught to three groups of high school students: two sessions of Upward Bound and the other will be with local high school students.

TRIO Upward Bound, through the University of Illinois Minority Student Affairs Division, partners with three local high schools to "provide our students with high quality academic, cultural, and career related activities designed to prepare and equip them to successfully complete high school, enroll in an accredited post-secondary institution and obtain a baccalaureate degree". Read more about the program here.

The course includes a English sentence generator Python application that uses grammar rules to construct sentences in English with a determiner, noun phrase, and verb phrase with extensions to include prepositional phrases and more. Students will utilize parsing trees to plan their own grammar rules and decide terminal cases to ensure the accuracy of their generated phrases through programming. Along the way, students will read and discuss key computer science literature that examines fairness, ethical concerns, equity, and inclusion in the computer science field and community as well as in the creation and implementation of artificial intelligence applications.

The CS173 project is to collect feedback on a course in the University of Illinois undergraduate discrete mathematics course, which due to the nature of the subject is considered to be more conceptual and theoretical compared to other applied computer science courses in the major curriculum. Through the recorded interviews of students, and midterm and final survey responses, Dr. Gertner, graduate student Deb Ghosh, and team are seeking to collect feedback on successful study skills and resources the students utilized to provide guidance to future students. In addition, we hope to analyze the impact of this discrete mathematics course on the undergraduate's sense of belonging in computer science, and we are looking to discover the relationship between a student's study habits, their utilized resources, and their perception of the course, their performance, and their sense of belonging.

Time Worked

I have been working on both of these projects for eight weeks. I plan on working another three and a half weeks minimum to satisfy my requirement for entry to the University of Illinois graduate school; however, I am hopeful that I will be able to continue in my research capacity on these projects during the school year.

I have logged my hours and I have spent approximately 310 hours.

Project Status

The CS + X team facilitated two sessions of Upward Bound that will conclude on Thursday, July 20th and one session with local high school students, which will also be wrapping up by the end of July. Working with these high school newcomers to computer science was incredibly rewarding for me, and I hope the team felt the same way.

As far as my accomplishments go, I am proud to say that I contributed a great deal to the development of the curriculum and lesson plans. When I was brought on to the team, we had a NLP project and a few articles that we wanted the students to read and discuss. From my experience as a music teacher, I know students succeed when they are slowly introduced to new material with scaffolded information, standards-based content, and given time to ask questions and reflect on their learning. I researched computer science education papers and best practices online and found a plethora of resources to comb through and reference during my development.

My curriculum was designed to culminate in the NLP project, but I knew I needed to build student skills along the way. I created Python introductory materials, daily activities that were leveled in their rigor (so more experienced students could be challenged and beginning students could reinforce their knowledge), created English grammar review activities, and additional supplemental work for the project in case any students wanted to explore a new direction.

For my future accomplishments, I assume we will be meeting as a team to discuss the results of the curriculum and any more changes we would like to make for the next group of students or make these resources available to high school computer science teachers; if not, I will suggest we do so. I would like for the team and I to do more research on linguistics education and computer education to see where this intersection lies, so we are able to design more impactful, engaging, and appropriate materials for students.

Although the field is changing, computer science for a long time was an exclusionary major, and I feel that programs such as this are integral for increasing representation in computing. The students come to us with only a minimal background in computing: perhaps they had an introductory experience in elementary or middle school or have attended coding club sessions at their school or have learned on their own. Through our natural language processing curriculum we were able to explore concepts, without getting overwhelmed by detailed mechanics, and have discussions on the impact of NLP and A.I. as well as conversations about human languages and their histories.

The CS173 team is in the middle of our project as we have begun to process and analyze the data to draw statistical conclusions with the survey data, but we are only beginning to analyze the qualitative interview data. I anticipate that this project will continue into the fall and winter, and I am hopeful to remain on the team.

My role in this project started out with data cleaning and validation. There were surveys that were given to a spring cohort of students this past spring semester as well as 17 interviews conducted. Although the interviews were transcribed by a NLP software, there were many mistakes, so I listened to the interviews and hand-corrected the transcripts. This was a long process, but upon reflecting, I am glad that I did this because I felt that I truly engaged with the dataset and have a clearer picture of the verbal responses and their emotional state and intent with their words that a NLP software would not be able to communicate.

A lot of time spent after this initial phase was “code development”. As a team, we needed to identify which keywords we wanted to tag in the interview transcripts to later perform thematic and other qualitative analyses on. We identified major categories of “Strategy”, “Resources”, and “Self” with subcodes such as “Textbook” and “Change in Behavior” and notes to further detail each response. I was fascinated by how the different people on our team interpreted different responses; our perspectives were truly unique, and it took us weeks of meetings to come up with a set of codes and subcodes. We would be proud of our first draft and then we would meet again with a new set of ideas for information tagging, and it felt like we were back to the drawing board.

While I have made contributions with statistical analyses and discovering new cross-sections of data to analyze, my greatest contribution to this project has been the code reconciliation. The PhD student that I am working under asked me to compare our codes to see where we are similar and where we are different and sent an aggregated Excel document. I briefly considered going line by line and marking any differences in an additional column, but once I saw that there were thousands of codes, I thought against it.

I developed a Python program that reads in this aggregated Excel document, parses the data, and stores the data into “Entry” objects, as shown below:

```
@dataclasses.dataclass
class Entry:
    deb_text: str
    angela_text: str
    reference_text: str
    deb_codes: list[Code]
    angela_codes: list[Code]
    similarity: str
    sentiment: Optional[Sentiment] = None
```

I read up on the best way to implement this class in Python, and after many hours of researching I discovered the library dataclasses, so I began to read the documentation and build my project. The data required substantial cleaning and after a lengthy debugging process, I was able to split the data into deb_text (the text the PhD student tagged), angela_text (the text I tagged), reference_text (the interviewee’s response), deb_codes and angela_codes (a list of code objects that represent the aforementioned code, subcode, and note), similarity (a calculated value saved as a string), and a sentiment object, which I will discuss shortly.

Through the power of object-oriented programming, the data is future-proof and neatly packaged. I am able to create methods for these objects to compare identical codes and

dissimilar codes for easy comparison, and this has helped the PhD student and I reconcile our codes so that we are more accurate and reliable moving forward.

My future accomplishments lie in the “sentiment” field of the Entry object. I suggested to explore a sentiment analysis on the interviewee’s responses to see if there were any meaningful conclusions to be drawn by the sentiment score and the keywords tagged. For example, when a student is talking about their sense of belonging in computer science, is it mostly positive, negative, or neutral? This sentiment analysis is a way I have hypothesized to make qualitative information more quantifiable. This is by no means replacing a qualitative analysis, but I am hopeful that there are meaningful conclusions to be discovered.

Challenges and Excitements

For the CS+X project, my biggest challenge was realizing my first draft of the Introduction to Python Google Collab was outside the scope of what was needed for the NLP project. In my efforts to create a beginner-friendly, all-encompassing guide to Python, I lost sight of the necessary skills that the students needed to succeed. I added too many concepts for the length of this assignment, the length of the sessions, and the scope of the project. I realized this once the students were beginning to work through the collab, and they were struggling to not only complete the assignment, but also locate symbols on the keyboard.

I was disappointed in myself because I did not sequence this material appropriately with reference to the curriculum and the short time we had together. For the next session of Upward Bound students, I limited the concepts needed to three: variables, data types, and functions. These concepts, encapsulated at an appropriate amount for beginners in a shortened experiential program, was a flying success. Students were able to grasp this material confidently and identify these concepts in the project.

The most exciting part of the CS+X project is witnessing the growth in the students. High school students are at a pivotal time in their life where they are figuring out who they are, what they like, and future careers they are interested in, so watching students feel comfortable and confident in computer science was the most exciting part for me. Even though we only saw these students for three weeks, I felt that I had a connection to them and each student felt a spark of interest, curiosity, and success. And while they may not ultimately choose a career in computer science, I feel that experiences like this encourage self-esteem and a growth-mindset that they can learn anything, so long as they work diligently.

My main challenge for the CS173 project was the creation of the dataclasses Python program because it was a library that I was unfamiliar with, and I am still learning object-oriented programming. There were a lot of challenges with the data due to its aggregated nature. For example, the PhD student could tag a response and add three codes, while I could have zero, one, two, or up to five codes. Unrolling this data was a messy process, and I tried many paths of regular expression pattern matching and research before I learned of a Python method that could interpret Excel with as_eval. It was a giant break in my data cleaning, and it was the first time I felt success with this program.

Coincidentally, my most exciting moment was seeing the output of my object-oriented code with all the data in the correct place. The output terminal overflowed with my printed objects, and I nearly jumped for joy. While this was a long process, it saved me hours and hours

of sifting through the Excel file line by line making small notes where the PhD student and I tagged differently. It also felt like a “computer science” way of solving the program; meaning, I was able to build a program that helped me achieve my goals with tools that I knew and tools that I had to research and learn on the fly.

Working Environment

The CS+X team meets every week remotely to discuss our progress, the students’ progress, and any new materials we have created. I am also meeting with the students in person and online twice for Upward Bound and once a week with the group of local high school students. The team I work with consists of myself, Professor Gertner, a PhD student, and another graduate student. I am vaguely aware of others who have contributed to this project, but I have not formally met them. With Upward Bound, I am teaching mostly with the other graduate student; however, Professor Gertner and the PhD student have participated in activities as well.

The CS173 team that I meet with once or twice (rarely three times) a week is Professor Gertner, a different PhD student than the one referenced above, and myself. There are several other professors that are on this team, but I have not met with them since the beginning of the summer. I usually meet with Professor Gertner and the PhD student, although I do meet with just one or the other occasionally for progress check-ins and code reconciliations.

I have not attended any formal research talks, but my iCAN cohort meets once a week to discuss papers and our research projects, so I feel that I am connected to their progress.

Closing Thoughts

I am not finding the words to completely express my gratitude for this fellowship and these opportunities to research. At this (more than) halfway point, I feel more capable than I have felt before in computer science. When I first started research, I had a lot of self-doubts as I contemplated how little I’ve spent in the field, with a year under my belt of formal learning through the iCAN program at U of I, and I did not think I would contribute anything meaningful.

As I began to work and engage with my teams and my professors, I felt supported in asking questions and contributing ideas to meetings. I found that I could keep up with the more technical vocabulary and could program useful tools that were once way outside of my comfort zone.

Professor Gertner has been my role model and greatest mentor throughout this process. She affirms my contributions to meetings and encourages me to explore new curriculum development for CS + X and data analysis for CS173. Professor Gertner has lifted me up countless times, and I owe a lot of my newfound self-esteem to her.