This summer, I had the opportunity to complete an internship with McGill’s .txtLab, a laboratory for cultural analytics at McGill. I worked closely with Professors Andrew Piper and Richard Jean So on a project studying semantic and community separation across cultural spheres. I worked with a set of nearly 40,000 Wikipedia articles about films, novels, and television shows to better understand how Wikipedia editors group around certain subjects and how language divides are formed between those subjects. My project used network analysis and natural language processing to study the data set of Wikipedia articles. This project began in November 2018 with a general question in mind: How do editors on Wikipedia argue over language? While this research question has evolved and taken tangents over time, it has largely stayed true to the central premise. This being said, the project is not yet complete: There are many avenues of research that went unexplored, and I am planning on continuing the project into the Fall.

I became fascinated with Wikipedia well before beginning this project, first as a reader and soon after, as an editor. I have now volunteered as an editor on Wikipedia for over two years, a volunteership which I would recommend and encourage to anyone who has interacted with Wikipedia. Through the .txtLab and my ARIA internship, I have had the opportunity to turn my fascination with Wikipedia into a research project. As a computer science student, I have had ample opportunities to learn more about programming paradigms and practices, and this is the first large-scale computer science project where I’ve been able to apply this learning and produce new research about Wikipedia.

For my research, I studied how Wikipedia editors cohere around three different aspects of culture: Genre, medium, and period. Do Wikipedia editors tend to edit articles about works from a single medium, period, or genre? Or do they cross over between them, not grouping around any of these cultural facets? To answer this question, I studied the complete history of three different sets of Wikipedia pages, sourced from lists of American media curated by Wikipedia editors. To that end, I used the Wikipedia API to source the complete metadata and every revision of each page as a separate document. This summer, my work focused on community coherence, that is, how and where editors tend to group on Wikipedia. I created social networks with Wikipedia articles as nodes and the number of editors who edited any two Wikipedia articles as edges and then used Louvain community detection to detect groupings of articles based around the editors who worked on them. I found that editors tend to group very strongly around medium, that is, editors will edit only pages on either films, television, or novels. In addition, while editors tend to group around period for films and television, they do not for novels. Further, editors do not group at all around genre.

I also studied whether there are semantic divisions across each of the three metrics. There were very strong divisions for genre, strong divisions for period in films and television, with weaker divisions in novels, and strong division based on genre. I also close read distinguishing features for each comparison, finding patterns of language depending for each of the metrics. These studies helped clarify the groupings of editors by activity and language.

Over the course of this project, I’ve also learned two new programming languages and many statistical techniques for data analysis. I’ve also learned how to manage large datasets, how to implement practical NLP, how to manage diverse data structures, and how to work with new file types. I also have written thousands of lines of original code, including several practical scripts which can be used for projects outside the scope of my own research. I’ve built a significant repository of scripts on GitHub which allow the work I’ve completed so far to be generalized across any set of Wikipedia pages, not just the pages I studied.

I would like to thank Mr. Harry Samuel, who funded my project this summer and gave me the opportunity to work in a field which fascinates me on a project which has offered me significant opportunity for learning and experience. I completed a large amount of work this summer thanks to this funding, and I am hoping to continue working with the .txtLab this year. This grant has helped me foster connections in professional and academic fields, particularly by allowing me to work full time at the .txtLab. I’m hoping to remain in the field of digital humanities after I graduate, particularly after seeing what it is like to work a DH project full time.