We are continuing prior research done by past student researchers at Calvin College as an extension of research performed in conjunction with CSIRO – Commonwealth Scientific and Industrial Research Organization, Australia’s national research lab.  Our focus is on utilizing Twitter data, Tweets, as a dataset by which we measure the SLO – Social License to Operate – of various mining, gas, and oil companies. SLO is defined as the acceptability of a company’s business operations by its employees, stakeholders, and the general public.  The primary goal of the summer 2019 research project is to investigate and find a methodology by which we can effectively model the topics of all the Tweets in our dataset. Topic modeling is a way of defining abstract “topics” that are prevalent in a corpus of textual documents. It is statistical in nature and is essentially unsupervised machine learning by which we attempt to cluster the Twitter data to find similarities and patterns among groups of words.

To that end, we first utilized standard data science techniques to investigate the nature of our Twitter dataset.  This involves the use of the Python programming language, the Pandas data analysis library, the Matplotlib data visualization library, and other processing and visualization software.  Our discoveries and results are recorded in Jupyter Notebooks – an interactive web-based application that allows researchers to easily share code, equations, visualizations, and text. We also utilize Scikit-Learn, a machine learning software suite, and Gensim, a topic modeling software suite, along with various 3rd party libraries, to implement baseline topic models from which we can begin to investigate how to best extract relevant topics from the Tweet texts.

Our initial data analysis and baseline topic models have been implemented and we are currently attempting to understand the mathematical and statistical constructs behind each topic modeling algorithm in order to better envision what it would entail designing a custom topic modeling algorithm that better suits our needs.  The future end-goal is to create a new topic modeler that is an improvement over CISRO’s current algorithm and can provide the least model perplexity with the maximum topic coherence metrics. By the end of this summer 2019 period, we hope to have begun the initial design towards a new Twitter Tweet topic modeling algorithm.  The research will be continued in the Fall of 2019 by either myself or other student researchers as a senior thesis project.

This experience has been instructional as to what can be expected if I continue onwards to graduate school.  It has been a test of whether I am truly interested in research as this is at minimum a 40-hour work week full time paid position.  So far, it has been challenging yet I am still motivated to keep trekking along; enough so that I put in extra hours voluntarily. Research never ends, unlike course assignments and projects.  Practically, I have learned much about data science, both in analysis and visualization. Learning to digest scholarly articles has probably been the most challenging aspect of research alongside understanding the pseudo-code for algorithms and statistical formulas.  The skills I have obtained and continue to refine should assist me in the future, whether in graduate school or along some other path.