Past efforts towards the creation of Part-of-Speech (POS) have been met with a wide variety of successes. Utilizing proper lexical works, the most successful of these attempts have used tailored output sets parsed by hand. These past efforts, however, have been labor intensive, and the difficulties in deriving such output sets has rendered a huge body of data inaccessible. As the data available for use continues to increase exponentially with the digital era, it is in rendering this data useful that provides a key issue in the drive towards more accurate learners. Herein I use of conversational data as a means by which to create an input-output space, and by mining Twitter, create a set of trainable data that uses simplistic and exclusionary parsing rules to determine the value of words appearing in harder-to-parse contexts using kNN.

Data collection utilized the Twitter REST API, querying for Tweets containing ambiguous words. For each of these terms, a search was then conducted so as to ascertain the most replied-to ten tweets containing these terms on 1-day intervals five days prior to program execution. For each tweet, the author's profile was then searched against in order to find replies that also contained the search term. Each conversation was then parsed into a number of trainable instances. The bit vector for the selected ambiguous word contained in the instance that, when parsed using simplistic exclusionary methods, yielded the lowest entropy, was then selected to serve as the output for each instance derived from the conversation. This project has made use of a bag-of-words style approach, where each instance includes a list of ids that correspond to the tokens occurring within the sentence. Each instance also keeps track of which word was immediately before and after the search terms instantiation.

A k-nearest-neighbor approach was settled upon for use in the experiment, the decision resting upon the relative ease in implementing kNN, its extensibility, as well as the specialized nature of the collected data. With a generic kNN implementation thus established, a series of different distance measures were applied with varying success.

From the onset of this project, it was unclear as to what success this approach might have.