Thomas N. Serrano
Bryan D. Martin
STAT 499
10 March 2020

mRkov: An R package

This quarter, my mentor, Bryan, and I were tasked with creating an R package. The main goal of our project was to create a user-friendly R package that has the potential to get people interested in statistics and its fun applications. To this end, we created mRkov, an R package and Shiny application for text generation using Markov chain Monte carlo methods. We decided to scrape twitter since its abundance of data is readily available and many people love to see analysis done on themselves or important political figures. The shiny app became a large component of this as well. By creating a shiny app, it grants users who do not know how to program the ability to use and experience our package.

Markov chain Monte Carlo methods are a class of algorithms that sample from a probability distribution. Each "link" in a generated chain is decided based upon its previous state. This previous state has a probability distribution of events that could occur after it, which is sampled to decide the next link. This is repeated until we have a desired number of links, or a delimiter condition is met.

Our package allows users to scrape Twitter for tweets (or load in their own textual data), which it then parses and formats for later use, adding metadata about the positions and sentiments of words. Then, using another function from our package, users can sample and generate sentences using a specified body of text. They can choose to prompt the algorithm with words that exist in the body of text, or even use n-gram methods to use larger sequences of words to influence the outcome of sentences. We also made an interactive Shiny app that allows users to play with our tool, as well as learn a bit about the process that makes technology like this work.

Throughout this project, we learned about the software development cycle as well. By deciding what we wanted our package to do and how we wanted to do it, steps one and two are completed. The design phase was completed when we decided how we wanted to go about collecting textual data and cleaning it. We initially settled on the twitteR package, but we

eventually migrated to use Rtweet. Implementation was achieved when we initially put the package together. We had to set up our namespace document and other R package infrastructure before getting the more complex parts put together. Testing was achieved during the production of our shiny app. We would often encounter tokens in bodies of text that would crash the app, forcing us onto the last step of development, maintenance. We are currently bug-fixing and optimizing our package which involves restarting the whole design process all over again, from redesigning how we scrape tweets to adding new ways to generate sentences using n-grams.

Our package is available at the following repository:

https://github.com/serrat839/mRkov