Word Prediction App

DRC

2022-12-04

Summary

For the Data Science Specialization Capstone Project, we were tasked to build an app that would predict the next word when given a phrase by the user. Below is the link to the app.

https://drchachere.shinyapps.io/capstone_project/

For Natural Language Processing (NLP), a corpus is needed to base the predictions from. For this app, a small sample of a corpus containing blog posts, news articles, and tweets were used.

NLP Algorithm

To predict the next word, n-grams were used. For example, if a two word phrase was given, then the algorithm would calculate the respective probabilities for all the three word n-grams beginning with the two word phrase. The matching three word n-grams (3-grams) with the highest probabilities are identified and the last word of these n-grams are the predictions. Feature matrices that counted the frequency of unique words, 2-grams, 3-grams, and 4-grams from the corpus text were loaded and used by the algorithm to calculate the probabilities of the predictions. If no matches could be made from the last three words, then the last two words were used. If no matches could be made from the last two words, then the last word was used.

Computational Demand

This app was developed with the processing constraints of the ShinyApps server in mind. Because of this, a small sample of the original corpus was used. Consequently, not every word from the original corpus is used in creating the n-grams. This makes some phrases (specifically ones that end in words that are not in the sample corpus used) produce more generic predictions based on the overall frequency of unique words.

How To Use The App

To generate predictions, simply type your phrase in the text box on the left side of the screen. Press submit and shortly thereafter the top predictions will appear in the main portion of the window.