Team Project Report - Rough Draft

Feel free to add/change anything where you think it’s needed. I tried to include as many details about the algorithm and the wrapper as I could. There are some sections I didn’t quite know what to put, so if you guys could help me fill those out that’d be great. Make changes, and upload it again!

Blake Cromar

Jordan Tway

Kimberlee Simpkinson

Source of the data:

<https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences>

UCI Machine Learning Repository

For this project we wanted to build a text prediction algorithm. We started searching the web for text databases we could use, and found Amazon, Yelp, and IMDB review datasets, which consisted of average everyday words you might here in a day. Given a word, and a few words preceding it in a sentence, we wanted to build an algorithm that would predict the next word in the sentence. Finally, we wanted to bring this all together in a nice wrapper to be user friendly and easy to read. Now, at the close of this project of those goals, we have completed …

The first dataset we found was a set of random sentences gathered from random Wikipedia sites. After a while of pre-processing and cleaning the data we wanted to look for a better dataset that included words that were used in every day conversations. After searching a little more, we found review data from Amazon, Yelp, and IMDB. We rearranged to data to include only one word per cell and cleaned it up to prepare it for the algorithm. Getting rid of punctuation, excluding extra spaces, and moving the data around to have only one sentence per row were some of the steps we took to tidy the data. Once the data was tidied, we rearranged it one more time to show the current word the user is on at position 0, and the three preceding words in the sentence at positions -1, -2, and -3. Position 1 would then be the target value. We also converted the data to numeric data by assigning each word a number.

There were many things we had to do to clean this dataset up. It was not ‘off-the-shelf’ ready. There were spaces where there shouldn’t be, and a lack of spaces where there should be. There were formatting issues, and ways we had to arrange the data in a custom way to fit the algorithm we wanted to run.

There were a few algorithms we thought about using and that would be best for the problem we were trying to solve. At first, we thought about using apriori or a decision tree or a random forest would be best, but then we turned to a naïve Bayes classifier before we finally decided on a neural network. We thought maybe a decision tree would work if we had attribute labelling the words we nouns and adjectives and such, but with our plain list of words we had we decided to try something else, like an apriori classifier. The only issue there was the order of the words. The apriori algorithm wouldn’t take into account the ordering of the words in a sentence. So, we finally decided to use a neural network, because we knew we could convert our dataset into a numeric dataset.

* 1. Make sure to discuss different things you tried along the way, even if they resulted in dead ends.

One challenge we ran into was when there was a word in the sentence the user was typing, that wasn’t in our database. To solve this, when we ran into that problem, we would assign that word to be an empty space. Another problem we ran into was if we were trying to predict the next word of a one-word sentence. This would mean that there wouldn’t be a set of 3 previous words we could use to help predict the next word. To solve this, we simply used the one word and empty spaces to predict the next word.

Using a neural network algorithm, we were successfully able to predict the next word in a given sentence, and not only that, we were able to build a wrapper that works with this algorithm, making it more user friendly. As the use types their phrases into the textbox, our algorithm will predict the next word and create a suggestion for them.

Our results could be of value to a business or stakeholder because…

Our results constitute something "interesting" because…

* 1. Don't forget to discuss potential limitations or ethical issues.

Even a simple algorithm requires a lot of preparation before it can actually be implemented. The data that was needed and that was prepared for this algorithm was uniquely altered to fit this algorithm and may not work well with many other algorithms. Also, even though we had a significant number of sentences that we were able to train our neural network on, the predictions still weren’t fantastic. They were predictions, and the words went well with the previous word, but didn’t fit the sentence as a whole very well. Creating a more efficient text prediction algorithm, especially one that could be updated frequently with ease, would take a lot more guts than the one we created.

If we were to start this project again, we would use a larger dataset that was more inclusive with the types of words and categories of words. With our review dataset, there were many words that were repeated in the many sentences we had, and so there wasn’t a large of a variety as there would appear to be looking at the length of the dataset alone. We would also consider using different learning algorithms to see if we could find one that was faster to learn, or one that could be updated in a more efficient way.