Machine Learning Project

Contributors:

* Christoph Kramer
* Logan Carpenter

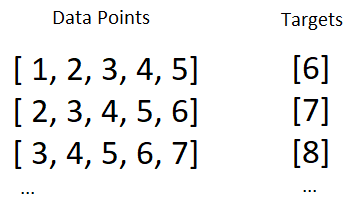
Data Set analyzed

Wikipedia, the simplified pages. Chosen because of the simpler vocabulary and shorter articles. Our hope with this is to have a machine that generates cohesive sentences by using Wikipedia to teach it.

Data Preparation

The first struggle we faced was pulling the article data we wanted from the html code of the website. Thankfully, there was a library for python that solved that exact issue called BeautifulSoup. Once the data we wanted was imported, it was broken up into individual words, using a different library called NLTK. We then made sure to toss any worthless ‘words’, or in this case, punctuation that were being treated as such.

Once that was done, we created 2 instances of our data. 1 set, we set aside, leaving it in its natural order. The other set, we sorted by letter, then deleted any duplicate words, keeping track of how many duplications there were. Those words were then sorted again, this time by the number of times that they were found in the text, with more common words being lower numbers. We used this as the machine’s number/word dictionary, as well as its vocabulary.

Next, taking the original unmodified data se that we put aside, we created a data set from that by taking the first 5 words of the text, and treating that as our inputs, using our dictionary to translate them into numbers. The 6th word was used as our target for training. We then looped through the text, using the next word in the text as the starting point for the next data point. This is represented in the figure below.

Learning from the Data

* Nothing to report yet.

Results

Here is where we show off some sentences that were generated.

Conclusions

IDK man, don’t have the results

Lessons Learned

Language Learning is complicated.