Second Project - Hidden Markov Model

Anna Seib

March 28, 2021

1 Introduction

For this second project, we want to build a Hidden Markov Model of a dataset to perform text prediction and generate new text. With the dataset from Kaggle, we formulate ideas on how machine learning can be used to create word correlations and distributions within the dataset. This Hidden Markov Model was trained on a data set of Shakespeare Plays.

2 Design

I chose this dataset because of how loosely Shakespeare uses his words, so it is more likely to make sense when I generate text from the model. Also, Markov chains are useful with fixed probabilities, since that is how the events are related to each other, and there are fixed probabilities since the dataset is known. To predict the next event in a Markov chain, a simple pair can contain all the information needed, since only the previous event is needed to make the prediction. After the model is trained, the new text is generated using random number generation and probabilities. For every word counted from the text corpus, the next word is stored as well. The first word is chosen randomly to generate new text every time. The previous event (previous word) and next event (next word) are stored as a "pair". Instead of repeating words in the dictionary, a repeated word has the new word appended to the list of words following the word. Every word is sampled randomly from the list of of generated words.

3 How To Run

Simply run:
python3 filename.py

```
anna@a-XPS13:~/College/2021/738/Project_2$ python3 project2_seib_hmm.py
['the', 'Talbot', 'is', 'the', 'messenger."', '"O,', 'are', 'you', 'a', 'foul', 'weather', 'in', 'cou
rse', 'unnatural,"', '"You', 'taught', 'me', 'he', 'my', 'fancy', 'more']
the Talbot is the messenger." "O, are you a foul weather in course unnatural," "You taught me he my f
ancy more
```

Figure 1: Sample Output

3.1 Results

I thought the generated text was surprisingly real, and every once in a while a sentence would generate that I thought made sense and could have been Shakespeare, although most were grammatically incorrect, as the only relation the events have is the previous event, this result makes sense.