

1 Introduction

- **Group members**

Sara Beery, Natalie Bernat, and Eric Zhan

- **Team name**

Voiceless Turtles

- **Division of labor**

We all worked as a team, and helped with the different segments. Eric focused on training the HMM and generating the sonnet, Sara focused on developing the rhyming dictionary and filtering the observation matrix to enforce meter, Natalie focused on NLTK and Old English to Modern English translations

2 Overview

- **Shakespearean Sonnets**

Shakespeare's sonnets follow a precise format. They have 14 lines, which are split into 3 *quatrains*, one of which is the *volta* and each of which has 4 lines, and a single *couplet* of 2 lines. The rhyming scheme is *abab cdcd efef gg*. The volta tends to have a different tone or content than the previous quatrains, and the couplet also has a different tone, as it concludes the sonnet.

Shakespeare's sonnets, like almost all of his writing, are written in *iambic pentameter*. This is a meter which requires all lines to be 10 syllables long, with stress on every other syllable beginning with the second syllable.

- **Hidden Markov Models**

We trained 3 separate HMMs, one for each section of the sonnet (quatrains, volta, and couplet). Each HMM has 30 hidden states, and we trained for 900 iterations.

3 Data Pre-processing

- **Tokenization**

We tokenize the data by letting each word be a token. We remove all punctuation before training, and add them back post-generation of a sonnet based on the contents of each sentence. We also map each token to a number (like an ID) to use during training.

- **Sequences**

We treat each line of a sonnet as an observed sequence (so each sonnet generates 14 sequences). We also consider the sequence in reverse order. This allows us to later seed the last word of a sentence, to enforce rhyming, and generate the rest of the sentence backwards.

- **Models**

The separate sections of the sonnet (the quatrains, volta, and couplet) are trained independently on 3 separate HMMs with 30 hidden states each. This helps distinguish the different "voices" of the different sections within the sonnet.

4 Unsupervised Learning

- **Basic HMM**

We used the HMM implementation from HW5 solutions, which uses the Baum-Welch algorithm. We also used NLTK to extract information about the number of syllables and stress of each word, and we used pickle to save/load our models for comparison later.

- **Number of Hidden states**

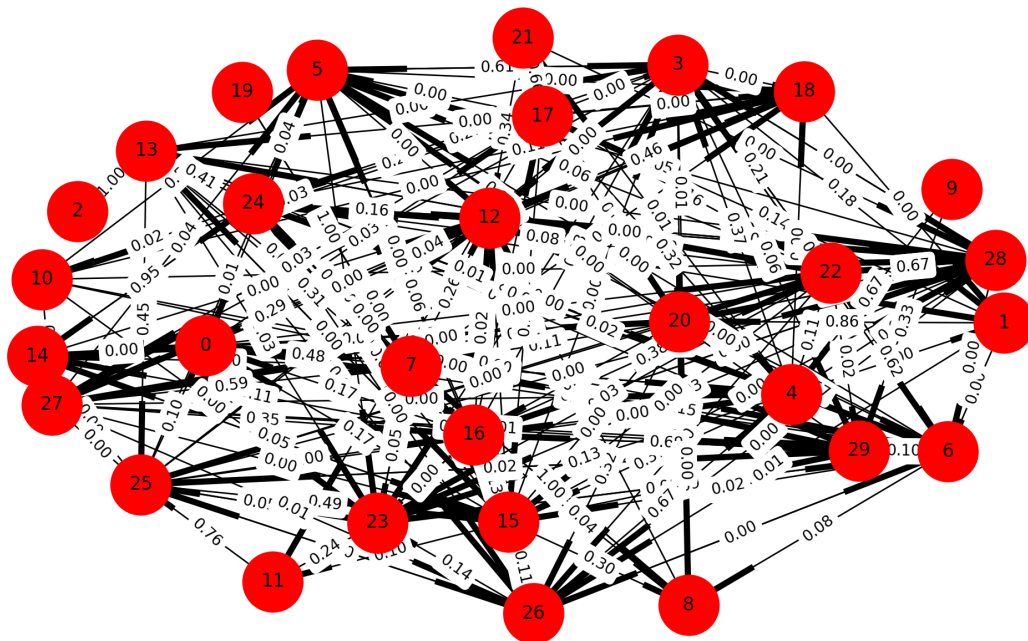
We experimented with a various number of hidden states including 10, 15, 20, and 30. The sonnets generated from the model with 30 hidden states sounded the best, so we ultimately decided to go with that one and train to 900 iterations.

5 Visualization & Interpretation

We trained 3 different HMMS with 30 hidden states each. To keep the report short, we will only discuss the visualization and interpretation of just one of the HMMS, the one trained on the quatrains (the discussion for the other two HMMS would be similar).

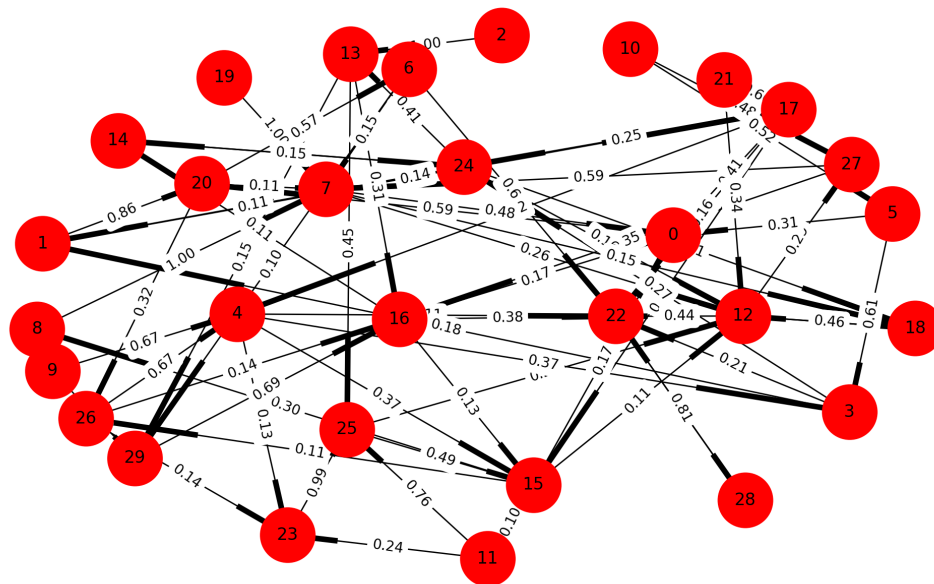
After training the HMM for 900 iterations, we discovered that only 226 of the 900 possible transition probabilities were non-zero (actually, 900 iterations was over-kill because the transition probabilities already converged to 6 degrees of accuracy after 300 iterations). The visualization of the transition probabilities between the 30 states are shown in Figure 1:

Figure 1: Transition Probabilities > 0



Unfortunately, there are still too many edges in this (directed) graph for us to glean anything, so instead we will only display the edges that have transition probabilities greater than 0.1 (there are 74 such edges out of 900) in Figure 2:

Figure 2: Transition Probabilities > 0.1



The graph is much more clear now, and we can immediately see some interesting transitions. For example, we see that hidden state 8 has a 100% probability to transition to hidden state 7. Upon a closer look, we can understand why in Figure 3.

Hidden state 8 has high probability to emit a word that can be both a noun or a verb, and hidden state 7 emits common words that follow after, which can be anything from a noun to adverb phrase to prepositional phrase. For example, states 8 and 7 together can emit phrases like: 'desire to', 'desire your', 'speak of', 'treasure his', and 'time to'. The emissions of three more hidden states are shown in Figure 4.

Figure 3: Hidden States 7 and 8

STATE: 7		STATE: 8	
the	0.256021	thee	0.052957
to	0.127067	treasure	0.034269
his	0.063999	time	0.025035
of	0.055949	dead	0.019821
or	0.030171	desire	0.016035
your	0.029791	rose	0.014687
me	0.026485	frame	0.014687
so	0.023251	any	0.014687
her	0.021367	least	0.014687
would	0.020233	speak	0.014687

Figure 4: Hidden States 2, 15, and 16

STATE: 2		STATE: 15		STATE: 16	
decease	0.018546	of	0.355104	my	0.226495
state	0.018446	with	0.097670	thy	0.098411
same	0.012364	all	0.055226	the	0.096624
sounds	0.012364	upon	0.043698	a	0.071903
end	0.012364	for	0.043615	every	0.042402
confound	0.012364	is	0.028103	this	0.029536
aside	0.012364	so	0.020492	so	0.029090
spent	0.012364	that	0.017142	of	0.025751
doom	0.012364	out	0.013719	it	0.024865
asleep	0.012364	hath	0.012491	their	0.019552

The top 10 words emitted by hidden state 2 all have stress on their final syllables ('decease', 'confound', 'aside', 'asleep', etc.), which suggest that this hidden state might be associated with the end of a sentence (or in our case, the beginning of our sequence). Interestingly, this hidden state had the smallest maximum emission probability out of all the states (0.018546 for 'decease' and 'state'). On the other hand, hidden state 15 had the highest emission probability of the entire matrix: 'of' with a emission probability of 0.355104. This hidden state seems to indicate a preposition because its top 10 emissions also include the words 'with', 'for', 'so', and 'that'. Lastly, another interesting hidden state we found was hidden state 16. This state emits a lot of adjectives, specifically possessive adjectives like 'my', 'thy', and 'their'. All the top 10 emissions of this state usually precede a noun.

6 Poetry Generation

Our process for generating a sonnet is as follows:

- Pre-process the data
- Train HMM for each section of the sonnet (quatrains, volta, couplet)
- Generate each section of the sonnet:
 - Seed the ends of sentences with a rhyming pair
 - Generate emissions (end to beginning of sentence) for each sequence
 - * At each word selection step, enforce meter and syllable count (details explained below)

Our best sonnet is shown on the next page (and shared on Piazza).

Sonnet 155:

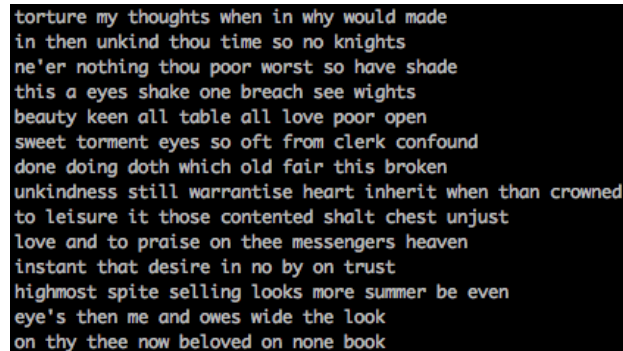
That did my greater love of strangle grave, [1]
It being hate burn seldom choirs decays. [2]
That first do heavy I as lofty have, [3]
Vex me say dian's bid thing limit days. [4]
Shall but have busy when lines present'st write, [5]
Her maladies forsaken are denote. [6]
Watch never can so roses for aright, [7]
When this thou elder for is public note. [8]
Touch shalt find public thy friend judgment's groan, [9]
View all she touches and truth many mine. [10]
Sun earth thoughts whether blot your beauty on, [11]
Before night envy his praise leaving thine. [12]
 Then and give often worth still lesson sin, [13]
 And ever so to angel good begin. [14]

This follows iambic pentameter, has proper rhyming scheme, and even arguably has some continuity of theme! Here's our line-by-line analysis:

- 1: The writer no longer loves someone that they once loved, their love has been "strangled"
- 2: They are now filled with burning hatred that will not decay
- 3: The writer's hatred is heavy, but they cannot rid themselves of it
- 4: "dian's" refers to Diana, the goddess of the moon. The writer is frustrated that the night has a hold on them, the hatred of their past lover is causing insomnia
- 5: They are currently writing this sonnet to pass those long nighttime hours
- 6: They are using this sonnet to list their past lover's terrible qualities
- 7: They now feel that their lover was never beautiful
- 8: And was older then they had believed
- 9: They find that, in recalling her touch, they are repulsed
- 10: And are similarly repulsed by everything that has a connection to her in their mind
- 11: Her beauty has been blotted from their memory (note the interesting change in audience from the public to the lover)
- 12: They are stating that their lover is only praised by nightfall, meaning that they can only be considered beautiful if no one can see them
- 13: The writer is hoping to learn from his past mistakes
- 14: And move on to begin some new love

The irony within this sonnet is that history seems doomed to repeat itself, the writer, having ended a relationship they at one time believed to be good, is now filled with regret. However, they seek only a new "perfect" love, not understanding that perhaps their previous love's failings may have stemmed from unrealistic expectations of their lover.

Figure 5: Our first rhyming sonnet, trained for only 10 iterations.



```
torture my thoughts when in why would made  
in then unkind thou time so no knights  
ne'er nothing thou poor worst so have shade  
this a eyes shake one breach see wights  
beauty keen all table all love poor open  
sweet torment eyes so oft from clerk confound  
done doing doth which old fair this broken  
unkindness still warrantise heart inherit when than crowned  
to leisure it those contented shalt chest unjust  
love and to praise on thee messengers heaven  
instant that desire in no by on trust  
highmost spite selling looks more summer be even  
eye's then me and owes wide the look  
on thy thee now beloved on none book
```

7 Additional Goals

- **Rhyme**

In order to enforce the rhyming scheme, we chose to create lists of rhyming pairs that we could choose from at random in order to seed our HMMs. We would then generate each line of the poem backwards, which ensures that our resulting sonnet follows the desired rhyming scheme. We choose the start state for each sequence to be the one with the highest probability of emitting each particular seed word. Our first rhyming sonnet (trained with only 10 iterations) can be seen in Figure 5. It is starting to look like Shakespeare, but the meter and syllable counts are far from correct, and the sentence structure is far from grammatically correct.

- **Meter & Syllable count**

In order to enforce meter, we employed CMU's NLTK package, which provides a dictionary of information about word pronunciation. From this dictionary, we generated our own dictionary of {word, (syllableCount, emphasisBoolean)}, where emphasisBoolean told whether the word's final syllable was stressed.

We used this information to filter the Observation Matrix as we generated each word in the line. We kept track of the total number of syllables generated so far, as well as the emphasis on the first syllable of the previous word (calculated by the emphasis on its last syllable and the number of syllables in the word, assuming that all words Shakespeare used have alternating emphasis). As each word in a line is generated, we only consider outputs that have appropriate numbers of syllables and emphasis. We do this by setting the temporary output probabilities of all invalid outputs corresponding to the current state to 0 and then normalizing the temporary distribution at each iteration.

Using this logic, we were able to generate our first sonnet that accounted for meter:

```
happy thinking reason december's had  
thus live might crooked cry own roses play'st  
warmed then not ocean self-love that's touches mad  
part thy of guilty my is little sway'st  
i plods is wretched his with muses beard
```

heaven's every without fleeting show
me all hath after sin which very herd
my when i being if thine costly so
your more lodged dial i thine after hide
ashes divining it not building ghost
others policy eyes and verses pride
authority beauty another's costs
do be till learning length self pity hence
turn speechless thy this sweetest thee defence

As you can see, the meter is still inconsistent. We analyzed our algorithm and discovered that most of the errors were coming from inconsistencies in our syllable counts and emphasis information. While generating the pronunciation dictionary, we encountered many words that were unique to Old English/Shakespeare, and were thus not in CMU's dictionary. To deal with these words, we set up a post-training data-cleaning function that made a handful of attempts to find analogous words within the CMU dictionary—e.g. removing 'eth', removing "st", converting 'ou' to 'o', and a few others. By performing this data-cleaning, 52% of the words that could not be found in CMU's dictionary were found (reduced from 434 to 226 unfound words).

Also, we discovered that Shakespeare doesn't always play by his own rules!! We found multiple examples of lines in Shakespeare's sonnets that did not follow iambic pentameter. An example from Sonnet 144:

...
The worser spirit a woman coloured ill.
To win me soon to hell my female evil,
Tempteth my better angel from my side,
And would corrupt my saint to be a devil:
...

Notice that lines 2 and 4 both have 11 syllables, and therefore do not follow iambic pentameter. The last words, evil and devil, do not have emphasis on the final syllable. Another example from Sonnet 87:

...
My bonds in thee are all determinate.
For how do I hold thee but by thy granting,
And for that riches where is my deserving?
The cause of this fair gift in me is wanting,
...

Here we have 11-syllable lines ending in granting and wanting. Since these words are added to our rhyming seeds, and we assume that all the seeds have last-syllable emphasis, this causes issues. In this case, we chose to be consistent with Shakespeare and allow 11-syllable sentences where the first 10 syllables follow iambic pentameter, and the last syllable is unstressed.

8 Conclusion

- **Discoveries**

Our algorithm imposed elements of Shakespearian style on the generated poems by enforcing meter, rhyme, and general sonnet structure. We relied on the HMM to learn grammatical structure and generate sensible content. Since much of the meaning in Shakespeare's work is implicitly embedded in short phrases within each line—thus requiring only short-range correlation between words— and since HMMs in some sense work by generating probabilities of short-range correlations between words, our model was able to generate enough sense to be interpretable by an intelligent, creative human.

- **Challenges**

The largest challenges we faced related to inconsistencies in our training set, and the incompatibility between Shakespearean English and the CMU Dictionary used to get syllable counts and pronunciation information.

Since Shakespeare's sonnets do not always follow the "prescribed" rhyming structure and meter of an English sonnet, they are a challenging dataset to learn from. In some cases, such as Sonnets 99 and 126, the number of lines and rhyme scheme were incorrect. We chose to remove these sonnets from our dataset. In other cases, such as Sonnets 87 and 144, Shakespeare does not consistently follow iambic pentameter and has some 11-syllable lines. In this case, we dealt with the issue by sticking to Shakespeare and allowing 11-syllable lines in our sonnets which mimic the structure we witnessed in his work.

Our other big challenge was translating Shakespearean English to something that could be understood by CMU's dictionary. We analyzed the words seen in Shakespeare's sonnets that were not included in the dictionary, and adapted as many of them as we could with a simple set of rules in order to use the dictionary information as much as possible. There were some words, such as "niggardly" and "churls" which had no simple translation to a modern word. In these cases we did our best to approximate the number of syllables and assumed there was stress on the final syllable. These assumptions frequently caused issues, but were difficult to consistently fix without manually entering the words into the dictionary ourselves. We decided, based on the time constraint of the project, that this was not worth the effort required.

We also found that CMU's dictionary was not always correct, even when looking up modern words.

- **Future Work**

Although our HMMS were able to generate short interpretable phrases, improvements might be made to the overall grammatical structure of the sonnets by providing the HMM with input of both words and part-of-speech tags. In addition, we believe that substantial improvements would be made on our current model by simply providing more training data.

In the future, we would also develop an "Olde English" dictionary providing syllable count and pronunciation information for Shakespearean words no longer used in the current day.

References

- [1] M. Allan and C. K. Williams. Harmonising chorales by probabilistic inference. In *NIPS*, pages 25–32, 2004.
- [2] L. Rabiner and B. Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.