Markov Chain Lab report

Name: Peiyu Shi
Course: CSEN 166

Part 1.

Writer and Dataset:

The two datasets I chose are William Shakespeare's Macbeth and a personal narrative I wrote last quarter. Macbeth consists of approximately 17,000 words, while the personal narrative contains about 1,800 words.
The preprocessing steps include removing punctuation and irrelevant symbols, converting all text to lowercase, and tokenizing the text into words.

Part 2.

Implementation details:

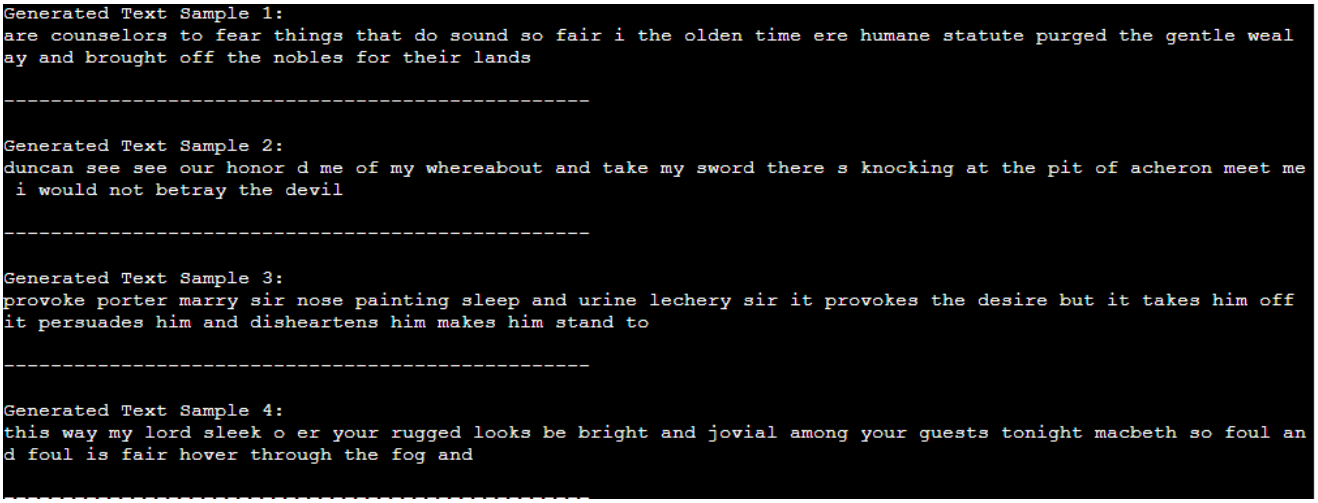
The total Markov chain project I had includes three different functions. One is preprocess_text, transition_model, and generate_text, and at last, my main function.
This function preprocesses the input text. I used Python's re library to replace all special characters with spaces. Then, I used .lower() to convert all letters to lowercase. Finally, the function tokenizes the text into words using .split().
The second function is my transition_model function. This function builds the transition model by calculating the probabilities of word transitions. I used defaultdict from Python's collections library to handle nested dictionaries. The function iterates through the list of words to extract n-grams and their corresponding next words, stopping when there are not enough words left to form a complete n-gram. Each transition is counted in the dictionary. Next, I normalized the transition counts to probabilities. I created a new dictionary to store these probabilities, iterated through each n-gram, calculated the total transitions for each n-gram, and divided each next word's count by the total.
Last is my generated text function. This function generates text from the transition model. It starts with a random n-gram as the initial key, then iteratively selects the next word based on the probabilities. The new word is appended to the result, and the key is updated using the last n words. If no valid next words exist, the generation stops. I then used Python's random.choices to select the next word based on probabilities. The function returns the final generated text as a string.

Part3.

Dataset Size analysis:

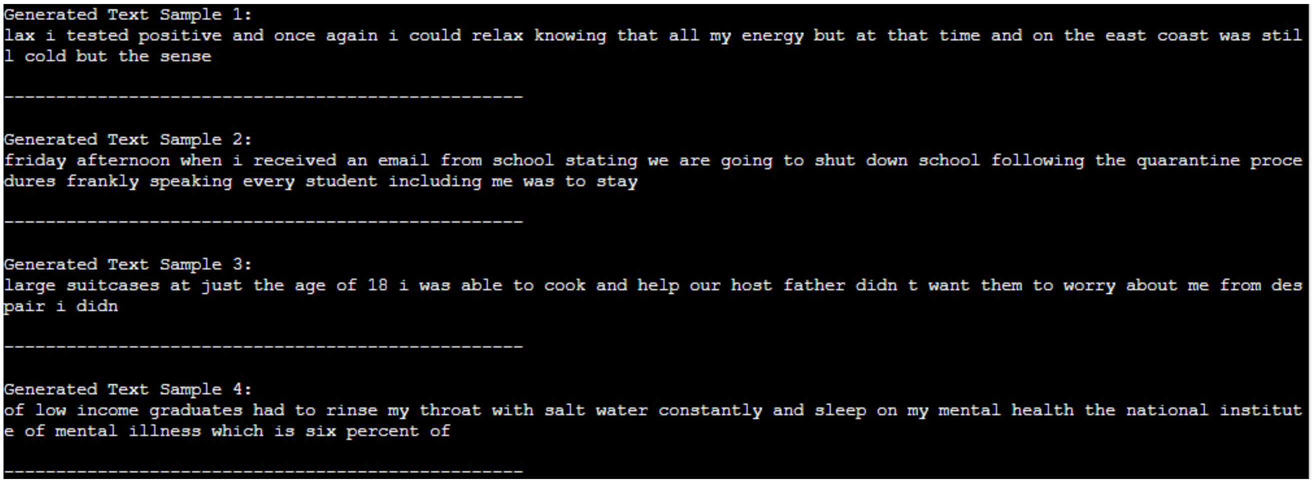
Based on the text generated, my observation is that the texts generated based on length are much harder to understand and text generated from shorter dataset lacks diversity. The text generated from a larger dataset size, Macbeth in this case, generates a more varied text compared to the text generated from a shorter dataset. So basically the limitation with a smaller dataset has fewer unique words and transitions and sometimes it might stop generating if there are no words followed by the current key. A lot of text generated from shorter paragraphs are pretty much generating the original text exactly. On the other hand, the text generated from a larger dataset is less coherent because there are many possibilities of our "next word."

Outputs:



```
Generated Text Sample 1:
are counselors to fear things that do sound so fair i the olden time ere humane statute purged the gentle weal
ay and brought off the nobles for their lands

--------------------------------------------------

Generated Text Sample 2:
duncan see see our honor d me of my whereabout and take my sword there s knocking at the pit of acheron meet me
 i would not betray the devil

--------------------------------------------------

Generated Text Sample 3:
provoke porter marry sir nose painting sleep and urine lechery sir it provokes the desire but it takes him off
it persuades him and disheartens him makes him stand to

--------------------------------------------------

Generated Text Sample 4:
this way my lord sleek o er your rugged looks be bright and jovial among your guests tonight macbeth so foul an
d foul is fair hover through the fog and

--------------------------------------------------
```

(text generated from Macbeth)



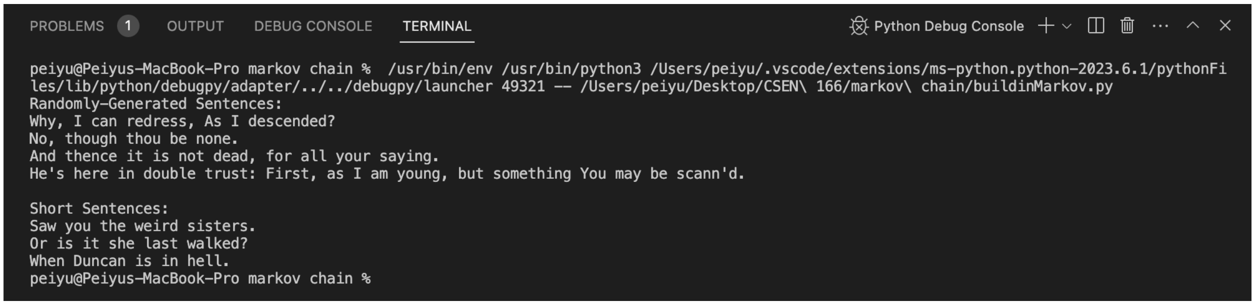
```
Generated Text Sample 1:
lax i tested positive and once again i could relax knowing that all my energy but at that time and on the east coast was stil
l cold but the sense

--------------------------------------------------

Generated Text Sample 2:
friday afternoon when i received an email from school stating we are going to shut down school following the quarantine proce
dures frankly speaking every student including me was to stay

--------------------------------------------------

Generated Text Sample 3:
large suitcases at just the age of 18 i was able to cook and help our host father didn t want them to worry about me from des
pair i didn

--------------------------------------------------

Generated Text Sample 4:
of low income graduates had to rinse my throat with salt water constantly and sleep on my mental health the national institut
e of mental illness which is six percent of

--------------------------------------------------
```

(text generated from my personal narrative)

As you can see, the generated text from Macbeth is somehow hard to understand and even though the sentence structure is mostly correct, the combination of words have made it really hard to understand. Meanwhile the generated text from my personal narrative is much more comprehensive but it lacks diversity. I can recognize many of the sentences in there and not much of a variation of word-choosing.

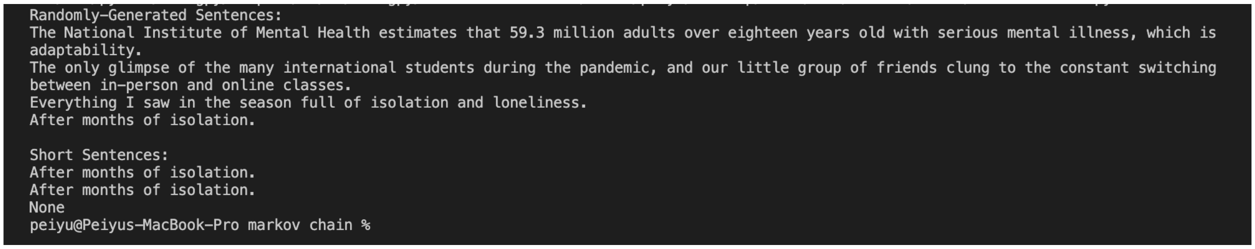
Part4. Python Markovify comparison.

Comparing the generated text from my own Markov Chain and the built-in Markov Chain, there are many differences and similarities. The similarities happen with the text generated from my personal narrative(small dataset), lack of variety and sometimes even no generation could be made. But for the text generated from a large dataset, the output is much easier to understand, but the text generated becomes much shorter.



```
peiyu@Peiyus-MacBook-Pro markov chain %  /usr/bin/env /usr/bin/python3 /Users/peiyu/.vscode/extensions/ms-python.python-2023.6.1/pythonFi
les/lib/python/debugpy/adapter/../../debugpy/launcher 49321 -- /Users/peiyu/Desktop/CSEN\ 166/markov\ chain/buildinMarkov.py
Randomly-Generated Sentences:
Why, I can redress, As I descended?
No, though thou be none.
And thence it is not dead, for all your saying.
He's here in double trust: First, as I am young, but something You may be scann'd.

Short Sentences:
Saw you the weird sisters.
Or is it she last walked?
When Duncan is in hell.
peiyu@Peiyus-MacBook-Pro markov chain %
```

(Text generated using build-in Markovify based on Macbeth)



```
Randomly-Generated Sentences:
The National Institute of Mental Health estimates that 59.3 million adults over eighteen years old with serious mental illness, which is
adaptability.
The only glimpse of the many international students during the pandemic, and our little group of friends clung to the constant switching
between in-person and online classes.
Everything I saw in the season full of isolation and loneliness.
After months of isolation.

Short Sentences:
After months of isolation.
After months of isolation.
None
peiyu@Peiyus-MacBook-Pro markov chain %
```

(Text generated using build-in Markovify based on my personal narrative)