Homework1

Wenyu Yang
Sep 16th

Describe how the choice of training dataset affected your model. What happens if we train a model on one type of data (like music lyrics) and then ask it to work with a different type of data (such as medical reports)? Which datasets do popular language models use, and how might this affect their output? (1 paragraph)
The choice of training dataset will impact a language model's performance. Because the model learns vocabulary and sentences from the dataset. If the dataset is lyrics, then the training yields an output that may be related to the lyrics, which I think may be limiting. If we choose a database of document compositions such as novels, or newspapers, or something like that, the training might produce a more logical result.

How long did this assignment take you? (1 sentence)
I used about three days to complete this assignment.

Whom did you work with, and how? (1 sentence each)
I did this assignment independently.
Based on the fact that the professor's class had already explained the assignment more clearly, I understood the overhaul logic of completing this assignment, but of course before that I brushed up on some of my python syntax knowledge.

Discussing the assignment with others is encouraged, as long as you don't share the code.
Which resources did you use? (1 sentence each)
1. I discussed with classmates some of the requirements of the assignment. For example, should we put an end sign at the end of the document or the end of the sentence?
2. I also reviewed the 5001 class material, mainly for the Python syntax review. (https://northeastern.instructure.com/courses/156795/pages/lesson-10-6-files-as-an-example?module_item_id=9221602)

For each, please list the URL and a brief description of how it was useful.
A few sentences about:
What was the most difficult part of the assignment?
I think the biggest difficulty is that code is very tedious. And one mistake can lead to bugs due to details.

What was the most rewarding part of the assignment?
I learned how language processing models work through a humble model overhaul.
What did you learn doing the assignment?
It sparked my interest in the class. And I reviewed my knowledge of Python syntax.

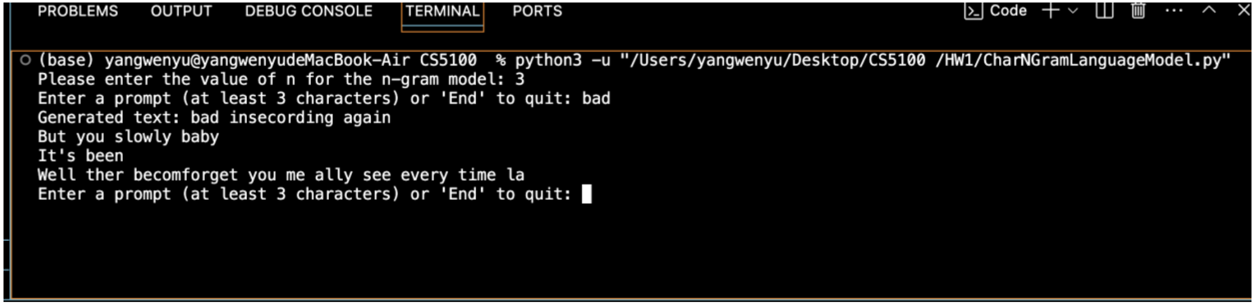
Constructive and actionable suggestions for improving assignments, office hours, and class time are always welcome.

I think I wish the professor could have been a little more clear on the requirements of the assignment. For example, where the end sign should be added, this has stumped me for a long time. I wish the professor could have labeled if for some of the more flexible details, such as... You can choose how you want it to be or something like that.

Output :

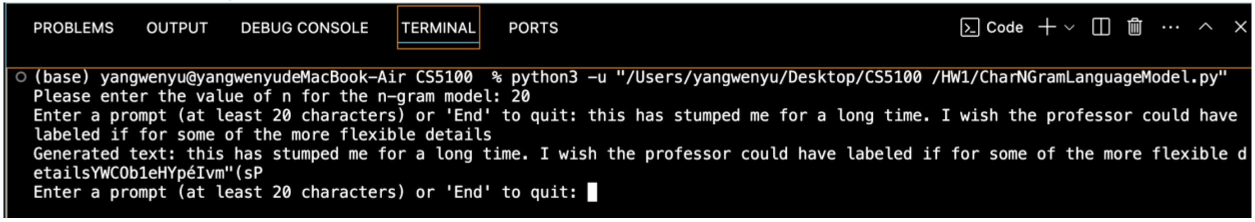
I found that when the value of n is 3, it will be easier to get some correct words or sentences. But when the value of n is 20, the output may not appear in the training data, resulting in the model not being able to find the right historical context to predict the next character. If the data set is large enough, the output may be some logical sentences.

When n =3 (small value):



```
○ (base) yangwenyu@yangwenyudeMacBook-Air CS5100  % python3 -u "/Users/yangwenyu/Desktop/CS5100 /HW1/CharNGramLanguageModel.py"
  Please enter the value of n for the n-gram model: 3
  Enter a prompt (at least 3 characters) or 'End' to quit: bad
  Generated text: bad insecording again
  But you slowly baby
  It's been
  Well ther becomforget you me ally see every time la
  Enter a prompt (at least 3 characters) or 'End' to quit: █
```

When n = 20 (large value)



```
○ (base) yangwenyu@yangwenyudeMacBook-Air CS5100  % python3 -u "/Users/yangwenyu/Desktop/CS5100 /HW1/CharNGramLanguageModel.py"
  Please enter the value of n for the n-gram model: 20
  Enter a prompt (at least 20 characters) or 'End' to quit: this has stumped me for a long time. I wish the professor could have
  labeled if for some of the more flexible details
  Generated text: this has stumped me for a long time. I wish the professor could have labeled if for some of the more flexible d
  etailsYWCOb1eHYpéIvm"(sP
  Enter a prompt (at least 20 characters) or 'End' to quit: █
```