Count-based Autoregressive Language Models

Let's examine this with an example. Consider a text corpus from which we want to predict how likely it is that the sentence "What do I say next?" appears. We would build a giant lookup table like the one shown to the right. The questions we would try to answer would be in the following order:

```
- How likely is it that the word `what` appears?
- How likely is it that the word `do` appears after the word `what`?
- How likely is it that the word `I` appears after the sequence `what do`?
```

... and so on!

The mathematical way of looking at this would be through *probabilities*. In the figure shown to the right, the symbol `P(do | what)` refers to what is known as a *conditional probability*:

```
`P(do | what)` is the probability that the word `do` appears given that the word `what` has already appeared.
```

How can we calculate this? From the lookup table we see four possibilities with varying frequencies of occurrence: "what do" appears 40 times in the text; "what am", 16 times; "what should", 16 times and "what have", 8 times. So we can estimate the probability, `P (do | what)` (to be read out aloud as *P of "do" given "what"*) thus:

$$P(\text{do}|\text{what}) = 40/40+16+16+8 = 40/80 = 0.5 \quad P(\text{do}|\text{what}) = 40+16+16+840 = 8040 = 0.5$$

Now, the probability that the sequence "what do" appears is not just `P(do | what)` but rather:

$$P(\text{what do}) = P(\text{what}) * P(\text{do}|\text{what}) \quad P(\text{what do}) = P(\text{what}) * P(\text{do}|\text{what})$$

, i.e., the probability that the word "what" appears to begin with, multiplied by the probability that it is followed by "do".

Expanding this further, the probability that the sentence "What do I say next?" appears in this corpus of text would be:

$$P(\text{sentence}) = P(\text{what}) * P(\text{do}|\text{what}) * P(\text{I}|\text{what do}) * P(\text{say}|\text{what do I}) * P(\text{next}|\text{what do I say}) \quad P(\text{sentence}) = P(\text{what}) * P(\text{do}|\text{what}) * P(\text{I}|\text{what do}) * P(\text{say}|\text{what do I}) * P(\text{next}|\text{what do I say})$$

Note: There are usually many more possibilities in a typical corpus but we've kept it short and sweet for the sake of this exercise!

Instructions

Can you examine the counts of the next word to see if the probabilities match up? For instance, does the probability of occurrence “monkeys” after “what do” match up with the count tables?

You can use a similar approach as what we did in the exercise. We see four possibilities in our corpus to follow “what do” and they are “I”, “you”, “we” and “monkeys”. Examine the frequency occurrence of “monkeys” after “what do” and divide it by sum of counts of all 4 possibilities to get `P(monkeys | what do)`. Based on this, what would. `P(what do monkeys)` be?