In Class Activity

Large Language Models – Roadmap Activity 5

Creating a Python script to illustrate how large language models (LLMs) like GPT (Generative Pre-trained Transformer) work involves simplifying a ***very*** complex process.

LLMs are based on deep learning in neural networks. They operate on vast datasets, using sophisticated neural network architectures.

What we have in today's activity is a basic script that simulates a very simplified aspect of how a language model might predict the next word in a sentence based on a given context.

This script will not/does not capture the complexity and capabilities of real LLMs but can serve as an educational tool to understand the concept of probability distribution over a vocabulary for generating text.

The script will:

1. Define a simple vocabulary.
2. Simulate a basic probability distribution for the next word prediction based on the current word.
3. Generate a simple text sequence by predicting the next word.

Instructions

1. Students should visit http://tinyurl.com/shaferaicourse and download the files found in the **roadmap5** folder. (There is only one Jupyter notebook file today, whee!!)

2. Review the code, notice that the first cell begins a variable called probabilities. The occurrence of each word results in a probabilistic prediction of the next word.

   The words found in "probabilities" were inspired by the nursery rhyme "Jack and Jill"

   The first cell also breaks down the probabilities dictionary into a simple Python list variable called vocabulary. This is not strictly necessary for the algorithm… but "vocabulary" is handy to have when checking user input.

3. The last cell prompts the user for a start word and generates a sentence using the probable next words.

4. Today's the day we use ChatGPT for the first time! Go get yourself a "free"ChatGPT account if you don't have one already.

   https://chat.openai.com

   (You don't need to get the premium version… ***yet!***)

   CONTINUED

5. Ask ChatGPT to generate some new values for probabilities variable.

   For example: my **prompt** looked like this:

   This python dictionary represents the word probabilities for a children's book. Can you make one just like it using the text from the screen crawl in the original Star Wars movie?

   ```
   probabilities = {
               'jack': {'and': 1.0},
               'and': {'jill': 0.5, 'went': 0.5},
               'jill': {'went': 1.0},
               'went': {'up': 0.5, 'down': 0.5},
               'up': {'the': 1.0},
               'the': {'hill': 1.0},
               'hill': {'.': 1.0},
               'down': {'the': 1.0}
   }
   ```

6. Copy / paste your new variable into your script, replacing the original probabilities variable.
7. Test it.
8. **Find the relevant discussion to post to on canvas.**

   Post your new probabilities variable, and the random sentence your code produced.