CARTE ML Bootcamp Lab 5-2: Word Embeddings - Decoding Transformer Output and Prompt Engineering

This lab is about some of the mechanics of Transformer-based text generation, and the general skill of Prompt Engineering.

1 Decoding

The first part of the lecture described the decoding method - how to choose what the next word is a function of the output probabilities. This notebook, from Huggingface, explores that in more detail using GPT-2: `lab_5_2_Decoding_Huggingface`. Upload the notebook to Colab, and read through it and run it.

2 Viewing Parameters and Probabilities on OpenAI's Models

In this section, we will take a look the inputs and outputs of the GPT3 model known as `text-davinci-003.` Do the following:

1. Go to the OpenAI playground and login as necessary, here: https://platform.openai.com/playground?mode=complete

2. Make sure that the 'Mode' in the upper right hand corner is set as `Complete`, and that the 'Model' is set to be `text-davinci-003.`

3. Type the following sentence into the open box: `List 5 ways to have fun:` and click `Submit` on the lower left of the screen. Do those things sound like fun to you? If not, add some advice that might make it more fun for you, such as like `...ways to have fun that involve eating`. (You'll need to erase the output before you re-submit the revised prompt).

4. What you type to produce the output is called the *prompt* and the subsequent output is called the *completion*. Let's use the same prompt you just used, but trying varying the temperature parameter, situated as a slider on the right hand side, with a default Temperature of 1. Try increasing the temperature above 1, and see if the expected behaviour happens (more diverse, perhaps weird things come out), and below 1, more predictable things.

5. On the lower right hand side, click the box that says `Show probabilities` and select `Full spectrum` and re-run some of your prompts. Notice that the output now has a spectrum of colours. Click on anyword, and you can see the other words that were probable when the selected word was generated.

6. As an interesting aside, click on various words, and you'll also be able to see what portions of a word (or sometimes a whole word) are actually tokenized. During the lectures, we have always referred to words as the inputs and outputs, but the truth is that it is actually a smaller unit called a token, which is often just part of a word. An interesting side note, try the following prompt: "What does idk mean?" Look at what parts of the word `idk` are tokenized in the completion.

7. Play around with the other parameters on the right side - Top P and Frequency Penalty. Perhaps leave the Maximum Length and the Best Of alone, as these can incur heavy token usage if used unwisely, and tokens will cost some money. (text-davinci-003 is the most expensive of this class, at 2 US cents per 1000 tokens (both input and output, so you're not at too much risk here).

8. Click on the following link to bring up many things that you can do with this model: https://platform.openai.com/examples. Try the example "Explain Code" and use something like that to explain code that you used earlier in the week.

9. Try the `Tweet Classifier` and use its example to classify some phrases that you write. Revise this to

3 Prompt Engineering

You may be here because you do have text generation or classification or processing work in your research. This is a chance to use one of these powerful models to do that work for you, using the playground. So, while you're reading the below, please think of how you might use a prompt to do something in your research, whether it is summarization, question answering, conversation generation or anything else.

Read: OpenAI's guide to prompting here: https://platform.openai.com/docs/guides/gpt-best-practices/strategy-write-clear-instructions. There are many good suggestions here, and one to particularly pay attention is the Section titled "Give GPTs time to think".
Here are two more resources to peruse on prompting, perhaps at another time:

1. https://lilianweng.github.io/posts/2023-03-15-prompt-engineering/

2. https://medium.com/mlearning-ai/the-chatgpt-list-of-lists-a-collection-of-1500-useful-mind-blowing-and-st

Do the following:

1. Switch your Mode in the OpenAI playground to 'Chat'. Switch your Model to `gpt-3.5-turbo`. Notice that there are fewer buttons on the right. This is a more powerful, better model to do the things your tried above.

   As discussed in class, recent advances in large pre-trained language models have shown impressive capabilities. These models can do question answering in the form shown above using essentially a zero-shot approach, and they also have some amount of embedded knowledge that allows them to answer a question without context. However, it sometimes takes some careful 'prompt engineering' to be successful. Here are some questions to guide your exploration of the very powerful GPT-3 model:

2. Consider the following problem that arises in automated psychological counseling. We would like take observations/statements of fact that a counselor may perceive about a patient, and turn it into a "more gentle" suggestion that is less direct than simply stating the fact, and would be easier for a patient to take in. For example, the observed fact might be:

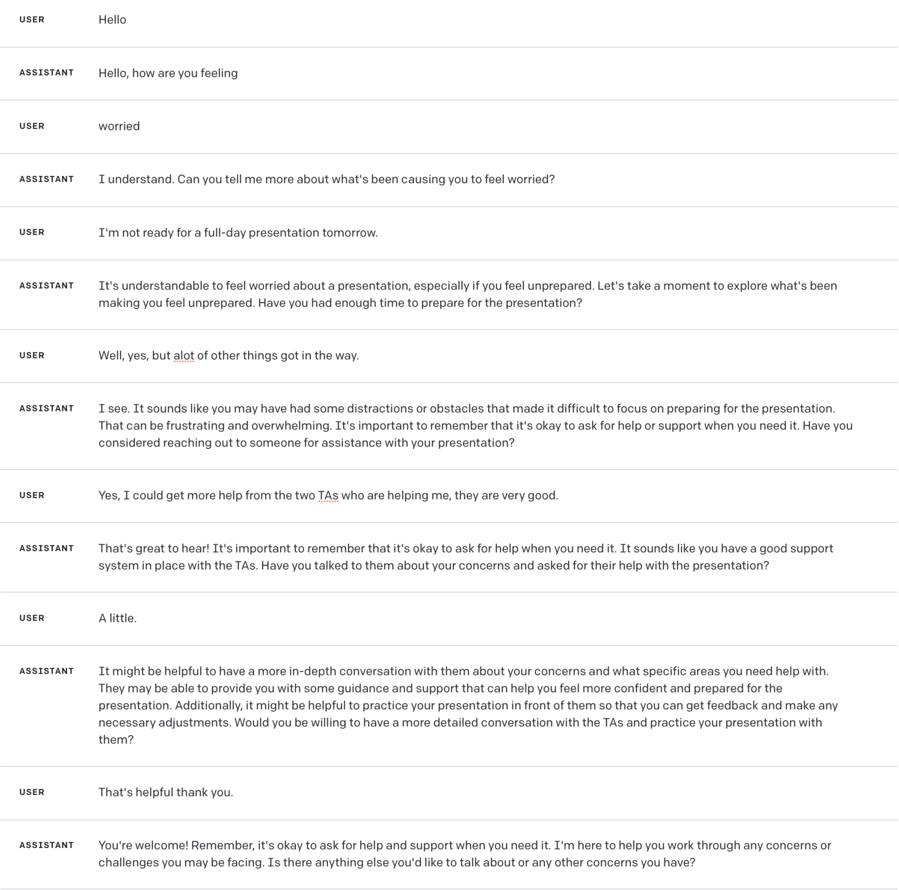
   ```
   You don't like being judged by your family.
   ```

and the softer version might be:

> `You might not like being judged by your family.`

Design a prompt for GPT-3.5-Turbo that takes direct statements of fact and turns them into such "softer" statements. Here is one more example of the original statement: `You're having trouble gett` Try experimenting with different prompts, for this example and one of your own choosing.

3. Try turning GPT-3.5-Turbo by doing the following: set the System Prompt to be "You are a psychotherapist, helping a patient who has been diagnosed with social anxiety disorder." In the User box, type hello. (Make sure that the maximum length parameter is set to be something reasonable like 70 characters or more). Have a conversation, see if it helps. It helped me, see below.



USER	Hello
ASSISTANT	Hello, how are you feeling
USER	worried
ASSISTANT	I understand. Can you tell me more about what's been causing you to feel worried?
USER	I'm not ready for a full-day presentation tomorrow.
ASSISTANT	It's understandable to feel worried about a presentation, especially if you feel unprepared. Let's take a moment to explore what's been making you feel unprepared. Have you had enough time to prepare for the presentation?
USER	Well, yes, but alot of other things got in the way.
ASSISTANT	I see. It sounds like you may have had some distractions or obstacles that made it difficult to focus on preparing for the presentation. That can be frustrating and overwhelming. It's important to remember that it's okay to ask for help or support when you need it. Have you considered reaching out to someone for assistance with your presentation?
USER	Yes, I could get more help from the two TAs who are helping me, they are very good.
ASSISTANT	That's great to hear! It's important to remember that it's okay to ask for help when you need it. It sounds like you have a good support system in place with the TAs. Have you talked to them about your concerns and asked for their help with the presentation?
USER	A little.
ASSISTANT	It might be helpful to have a more in-depth conversation with them about your concerns and what specific areas you need help with. They may be able to provide you with some guidance and support that can help you feel more confident and prepared for the presentation. Additionally, it might be helpful to practice your presentation in front of them so that you can get feedback and make any necessary adjustments. Would you be willing to have a more detailed conversation with the TAs and practice your presentation with them?
USER	That's helpful thank you.
ASSISTANT	You're welcome! Remember, it's okay to ask for help and support when you need it. I'm here to help you work through any concerns or challenges you may be facing. Is there anything else you'd like to talk about or any other concerns you have?

Figure 1: Psychology