In this notebook, we'll explore few-shot learning with GPT-2. While GPT-2 is a less expressive model than GPT-3 (and hence not as a good of a few shot learner), it can fit within the memory and processing constraints of laptops while also being openly available. Can you create a new classification task and design prompts to differentiate between the classes within it?

In [ ]:

```
import torch
from torch.nn import functional as F
```

In [ ]:

```
from transformers import pipeline
```

In [ ]:

```
from transformers import AutoModelForCausalLM, AutoTokenizer
```

In [ ]:

```
tokenizer = AutoTokenizer.from_pretrained('gpt2')
model = AutoModelForCausalLM.from_pretrained('gpt2')
```

In [ ]:

```
def classify_with_prompt(prompt, labels):
    inputs = tokenizer.encode(prompt, return_tensors='pt')
    completion_layer = model(inputs).logits[:, -1, :]
    probabilities = F.softmax(completion_layer, dim=-1)[0]
    pred_idx=torch.argmax(probabilities)
    pred_token = tokenizer.decode(pred_idx.tolist())

    label_ids=[]
    for label in labels:
        token_ids=tokenizer.encode(label)
        # token labels (e.g., Spanish, English) must be 1 token in length
        assert len(token_ids) == 1
        label_ids.append(token_ids[0])

    sorted_args=list(torch.argsort(probabilities[label_ids], descending=True)
    for arg in sorted_args:
            print("%.6f\t%s" % (probabilities[label_ids[arg]], labels[arg]))

    print("\nCompletion with highest probability:\n")
    print(prompt + pred_token)
```

In [ ]:

```
prompt = """X: I love this movie
Y: positive

X: I hate the movie
Y: negative

X: I kind of like the movie
Y: positive

X: This is one of the best movies I've ever seen
Y:"""

classify_with_prompt(prompt, ["positive", "negative"])
```

```
0.000231        positive
0.000102        negative

Completion with highest probability:

X: I love this movie
Y: positive

X: I hate the movie
Y: negative

X: I kind of like the movie
Y: positive

X: This is one of the best movies I've ever seen
Y: positive
```

In [ ]:

```
prompt = """X: Vampires take over the planet during an eclipse
Y: Hor

X: Two friends switch bodies and live each other's lives
Y: Com

X: John turns into a werewolf during a full moon
Y: Hor

X: John is a werewolf who plays basketball
Y: Com

X: A court sentences George to be Jerry's butler
Y: Com

X: A virus outbreak turns everyone into zombies
Y:"""

classify_with_prompt(prompt, ["Hor", "Com"])
```

```
0.000428        Hor
0.000238        Com

Completion with highest probability:

X: Vampires take over the planet during an eclipse
Y: Hor

X: Two friends switch bodies and live each other's lives
Y: Com

X: John turns into a werewolf during a full moon
Y: Hor

X: John is a werewolf who plays basketball
Y: Com

X: A court sentences George to be Jerry's butler
Y: Com

X: A virus outbreak turns everyone into zombies
Y: Hor
```

In [ ]:

```
prompt = """Q: This is a text
A: English

Q: Nel mezzo del cammin' di nostra vita
A: Italian

Q: Je ne sais pas
A:"""

classify_with_prompt(prompt, ["English", "Italian", "French", "Spanish", "Jap
```

```
0.000223        Spanish
0.000058        French
0.000022        English
0.000020        Italian
0.000008        Japanese

Completion with highest probability:

Q: This is a text
A: English

Q: Nel mezzo del cammin' di nostra vita
A: Italian

Q: Je ne sais pas
A: Spanish
```

Q1. Your job is to create a new classification task using prompt design (as in the examples above). You are free to consider binary classification or multiclass classifaction; keep in mind that you have ~1000 tokens to use as a prompt for GPT-2, so be sure to provide enough answered prompts for each class. (Note it is not a requirement that your model performs *well* (we want to assess what is -- and isn't -- learnable) but give it every opportunity to do so. Create 5 test examples to assess whether GPT-2 is able to recognize the class given your fixed prompt. To take the language ID task above, one test example corresponds to one prediction you make for the same set of answered prompts; the following constitutes two test examples for that task:

1.)

```
prompt = """Q: This is a text
A: English

Q: Nel mezzo del cammin' di nostra vita
A: Italian

Q: Je ne sais pas
A:"""
```

2.)

```
prompt = """Q: This is a text
A: English

Q: Nel mezzo del cammin' di nostra vita
A: Italian

Q: Non lo so
A:"""
```

Insights:

- Found that Negative Tweets were more difficult to identify compared with Positive Tweets

Prompt Design:

- Included 10 example prompts before the final prompt/answer of interest
- Included a summary of the task and a title at the top

In [ ]:

```
prompt = """Sentiment Analysis of Tweets

The task is to classify tweets with positive sentiment as “Pos” and tweets wi
Positive tweets often carry a thankful connotation. Negative tweets are often

Prompt: We won!!! love the land!!! #allin #cavs #champions #cleveland #clevel
Answer: Pos

Prompt: no comment!  in #australia  #opkillingbay #seashepherd #helpcovedolph
Answer: Neg

Prompt: It's unbelievable that in the 21st century we'd need something like t
Answer: Neg

Prompt: I am thankful for having a partner. #thankful
Answer: Pos

Prompt: use the power of your mind to #heal your body!! -     #altwaystoheal 
Answer: Pos

Prompt: woohoo!! just over 5 weeks to go!
Answer: Pos

Prompt: yes! received my acceptance letter for my masters so will be back at 
Answer: Pos

Prompt: omg!!! loving this station!!! way to jam out at work!!! while getting
Answer: Pos

Prompt: @user i'm not interested in a #linguistics that doesn't address #race
Answer: Neg

Prompt: #people aren't protesting #trump because a #republican won-they do so
Answer: Neg

Prompt: how the #altright uses amp; insecurity to lure men into #whitesuprema
Answer:"""

classify_with_prompt(prompt, ["Pos", "Neg"])

# Incorrect, "Pos" is identified. However, the true label is "Neg."
# Interestingly, the autocomplete identifies Neg as the most likely next word
```

```
0.009229        Pos
0.006963        Neg

Completion with highest probability:

Sentiment Analysis of Tweets

The task is to classify tweets with positive sentiment as “Pos” and tweets wit
h negative sentiment as “Neg.”
Positive tweets often carry a thankful connotation. Negative tweets are often
about hate groups and other forms of social injustice.

Prompt: We won!!! love the land!!! #allin #cavs #champions #cleveland #clevela
ndcavaliers
Answer: Pos

Prompt: no comment!  in #australia  #opkillingbay #seashepherd #helpcovedolphi
ns #thecove  #helpcovedolphins
Answer: Neg

Prompt: It's unbelievable that in the 21st century we'd need something like th
is. again. #nevetrump  #xenophobia
Answer: Neg

Prompt: I am thankful for having a partner. #thankful
Answer: Pos

Prompt: use the power of your mind to #heal your body!! -     #altwaystoheal #
healthy   #peace!
Answer: Pos

Prompt: woohoo!! just over 5 weeks to go!
Answer: Pos

Prompt: yes! received my acceptance letter for my masters so will be back at @
user again in october!    #goodtimes #history
Answer: Pos

Prompt: omg!!! loving this station!!! way to jam out at work!!! while getting
work done of course!!!!   #memories @user
Answer: Pos

Prompt: @user i'm not interested in a #linguistics that doesn't address #race
&amp; . racism is about #power. #raciolinguistics
Answer: Neg

Prompt: #people aren't protesting #trump because a #republican won-they do so
because trump has fuhered  &amp;√¢¬Ä¬¶
Answer: Neg

Prompt: how the #altright uses amp; insecurity to lure men into #whitesupremac
y
Answer: Neg
```

In [ ]:

```
prompt = """Sentiment Analysis of Tweets

The task is to classify tweets with positive sentiment as “Pos” and tweets wi
Positive tweets often carry a thankful connotation. Negative tweets are often

Prompt: We won!!! love the land!!! #allin #cavs #champions #cleveland #clevel
Answer: Pos

Prompt: no comment!  in #australia  #opkillingbay #seashepherd #helpcovedolph
Answer: Neg

Prompt: It's unbelievable that in the 21st century we'd need something like t
Answer: Neg

Prompt: I am thankful for having a partner. #thankful
Answer: Pos

Prompt: use the power of your mind to #heal your body!! -     #altwaystoheal
Answer: Pos

Prompt: woohoo!! just over 5 weeks to go!
Answer: Pos

Prompt: yes! received my acceptance letter for my masters so will be back at
Answer: Pos

Prompt: omg!!! loving this station!!! way to jam out at work!!! while getting
Answer: Pos

Prompt: @user i'm not interested in a #linguistics that doesn't address #race
Answer: Neg

Prompt: #people aren't protesting #trump because a #republican won-they do so
Answer: Neg

Prompt: if you hold open a door for a woman because she's a woman and not bec
Answer:"""

classify_with_prompt(prompt, ["Pos", "Neg"])

# Incorrect, "Pos" is identified. However, the true label is "Neg".
# The autocomplete identifies Pos as the most likely next word.
```

```
0.015760        Pos
0.003872        Neg

Completion with highest probability:

Sentiment Analysis of Tweets

The task is to classify tweets with positive sentiment as “Pos” and tweets wit
h negative sentiment as “Neg.”
Positive tweets often carry a thankful connotation. Negative tweets are often
about hate groups and other forms of social injustice.

Prompt: We won!!! love the land!!! #allin #cavs #champions #cleveland #clevela
ndcavaliers
Answer: Pos

Prompt: no comment!  in #australia  #opkillingbay #seashepherd #helpcovedolphi
ns #thecove  #helpcovedolphins
Answer: Neg

Prompt: It's unbelievable that in the 21st century we'd need something like th
is. again. #nevetrump  #xenophobia
Answer: Neg

Prompt: I am thankful for having a partner. #thankful
Answer: Pos

Prompt: use the power of your mind to #heal your body!! -     #altwaystoheal #
healthy   #peace!
Answer: Pos

Prompt: woohoo!! just over 5 weeks to go!
Answer: Pos

Prompt: yes! received my acceptance letter for my masters so will be back at @
user again in october!    #goodtimes #history
Answer: Pos

Prompt: omg!!! loving this station!!! way to jam out at work!!! while getting
work done of course!!!!   #memories @user
Answer: Pos

Prompt: @user i'm not interested in a #linguistics that doesn't address #race
&amp; . racism is about #power. #raciolinguistics
Answer: Neg

Prompt: #people aren't protesting #trump because a #republican won-they do so
because trump has fuhered  &amp;√¢¬Ä¬¶
Answer: Neg

Prompt: if you hold open a door for a woman because she's a woman and not beca
use it's a nice thing to do, that's . don't even try to deny it
Answer: Pos
```

In [ ]:

```
prompt = """Sentiment Analysis of Tweets

The task is to classify tweets with positive sentiment as “Pos” and tweets wi
Positive tweets often carry a thankful connotation. Negative tweets are often

Prompt: We won!!! love the land!!! #allin #cavs #champions #cleveland #clevel
Answer: Pos

Prompt: no comment!  in #australia  #opkillingbay #seashepherd #helpcovedolph
Answer: Neg

Prompt: It's unbelievable that in the 21st century we'd need something like t
Answer: Neg

Prompt: I am thankful for having a partner. #thankful
Answer: Pos

Prompt: use the power of your mind to #heal your body!! -     #altwaystoheal
Answer: Pos

Prompt: woohoo!! just over 5 weeks to go!
Answer: Pos

Prompt: yes! received my acceptance letter for my masters so will be back at
Answer: Pos

Prompt: omg!!! loving this station!!! way to jam out at work!!! while getting
Answer: Pos

Prompt: @user i'm not interested in a #linguistics that doesn't address #race
Answer: Neg

Prompt: #people aren't protesting #trump because a #republican won-they do so
Answer: Neg

Prompt: when your having a good weekend and it shows :) #thankful #blessed
Answer:"""

classify_with_prompt(prompt, ["Pos", "Neg"])

# Correct, "Pos" is identified and is the true label.
# The autocomplete identifies Pos as the most likely next word.
```

```
0.018239        Pos
0.002447        Neg

Completion with highest probability:

Sentiment Analysis of Tweets

The task is to classify tweets with positive sentiment as “Pos” and tweets with negative sentiment as “Neg.” Positive tweets often carry a thankful connotation. Negative tweets are often about hate groups and other forms of social injustice.

Prompt: We won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers
Answer: Pos

Prompt: no comment!  in #australia  #opkillingbay #seashepherd #helpcovedolphins #thecove  #helpcovedolphins
Answer: Neg

Prompt: It's unbelievable that in the 21st century we'd need something like this. again. #nevetrump  #xenophobia
Answer: Neg

Prompt: I am thankful for having a partner. #thankful
Answer: Pos

Prompt: use the power of your mind to #heal your body!! -     #altwaystoheal #healthy   #peace!
Answer: Pos

Prompt: woohoo!! just over 5 weeks to go!
Answer: Pos

Prompt: yes! received my acceptance letter for my masters so will be back at @user again in october!    #goodtimes #history
Answer: Pos

Prompt: omg!!! loving this station!!! way to jam out at work!!! while getting work done of course!!!!   #memories @user
Answer: Pos

Prompt: @user i'm not interested in a #linguistics that doesn't address #race &amp; . racism is about #power. #raciolinguistics
Answer: Neg

Prompt: #people aren't protesting #trump because a #republican won-they do so because trump has fuhered  &amp;√¢¬Ä¬¶
Answer: Neg

Prompt: when your having a good weekend and it shows :) #thankful #blessed
Answer: Pos
```

In [ ]:

```
prompt = """Sentiment Analysis of Tweets

The task is to classify tweets with positive sentiment as “Pos” and tweets wi
Positive tweets often carry a thankful connotation. Negative tweets are often

Prompt: We won!!! love the land!!! #allin #cavs #champions #cleveland #clevel
Answer: Pos

Prompt: no comment!  in #australia  #opkillingbay #seashepherd #helpcovedolph
Answer: Neg

Prompt: It's unbelievable that in the 21st century we'd need something like t
Answer: Neg

Prompt: I am thankful for having a partner. #thankful
Answer: Pos

Prompt: use the power of your mind to #heal your body!! -     #altwaystoheal
Answer: Pos

Prompt: woohoo!! just over 5 weeks to go!
Answer: Pos

Prompt: yes! received my acceptance letter for my masters so will be back at
Answer: Pos

Prompt: omg!!! loving this station!!! way to jam out at work!!! while getting
Answer: Pos

Prompt: @user i'm not interested in a #linguistics that doesn't address #race
Answer: Neg

Prompt: #people aren't protesting #trump because a #republican won-they do so
Answer: Neg

Prompt: the happiest baby ive ever known #cute #smiles   #babygirl #beautiful
Answer:"""

classify_with_prompt(prompt, ["Pos", "Neg"])

# Correct, "Pos" is identified and is the true label.
# The autocomplete identifies Pos as the most likely next word.
```

```
0.018036        Pos
0.004628        Neg

Completion with highest probability:

Sentiment Analysis of Tweets

The task is to classify tweets with positive sentiment as “Pos” and tweets wit
h negative sentiment as “Neg.”
Positive tweets often carry a thankful connotation. Negative tweets are often
about hate groups and other forms of social injustice.

Prompt: We won!!! love the land!!! #allin #cavs #champions #cleveland #clevela
ndcavaliers
Answer: Pos

Prompt: no comment!  in #australia  #opkillingbay #seashepherd #helpcovedolphi
ns #thecove  #helpcovedolphins
Answer: Neg

Prompt: It's unbelievable that in the 21st century we'd need something like th
is. again. #nevetrump  #xenophobia
Answer: Neg

Prompt: I am thankful for having a partner. #thankful
Answer: Pos

Prompt: use the power of your mind to #heal your body!! -     #altwaystoheal #
healthy   #peace!
Answer: Pos

Prompt: woohoo!! just over 5 weeks to go!
Answer: Pos

Prompt: yes! received my acceptance letter for my masters so will be back at @
user again in october!    #goodtimes #history
Answer: Pos

Prompt: omg!!! loving this station!!! way to jam out at work!!! while getting
work done of course!!!!   #memories @user
Answer: Pos

Prompt: @user i'm not interested in a #linguistics that doesn't address #race
&amp; . racism is about #power. #raciolinguistics
Answer: Neg

Prompt: #people aren't protesting #trump because a #republican won-they do so
because trump has fuhered  &amp;√¢¬Ä¬¶
Answer: Neg

Prompt: the happiest baby ive ever known #cute #smiles   #babygirl #beautiful
#niece #blessed #xo
Answer: Pos
```

In [ ]:

```
prompt = """Sentiment Analysis of Tweets

The task is to classify tweets with positive sentiment as “Pos” and tweets wi
Positive tweets often carry a thankful connotation. Negative tweets are often

Prompt: We won!!! love the land!!! #allin #cavs #champions #cleveland #clevel
Answer: Pos

Prompt: no comment!  in #australia  #opkillingbay #seashepherd #helpcovedolph
Answer: Neg

Prompt: It's unbelievable that in the 21st century we'd need something like t
Answer: Neg

Prompt: I am thankful for having a partner. #thankful
Answer: Pos

Prompt: use the power of your mind to #heal your body!! -     #altwaystoheal
Answer: Pos

Prompt: woohoo!! just over 5 weeks to go!
Answer: Pos

Prompt: yes! received my acceptance letter for my masters so will be back at
Answer: Pos

Prompt: omg!!! loving this station!!! way to jam out at work!!! while getting
Answer: Pos

Prompt: @user i'm not interested in a #linguistics that doesn't address #race
Answer: Neg

Prompt: #people aren't protesting #trump because a #republican won-they do so
Answer: Neg

Prompt: trump ny co-chair makes racist remarks about michelle obama  #p2 #p21
Answer:"""

classify_with_prompt(prompt, ["Pos", "Neg"])

# Incorrect, "Pos" is identified. However, the true label is "Neg".
# The autocomplete identifies Pos as the most likely next word.
```

```
0.009717        Pos
0.003300        Neg

Completion with highest probability:

Sentiment Analysis of Tweets

The task is to classify tweets with positive sentiment as “Pos” and tweets wit
h negative sentiment as “Neg.”
Positive tweets often carry a thankful connotation. Negative tweets are often
about hate groups and other forms of social injustice.

Prompt: We won!!! love the land!!! #allin #cavs #champions #cleveland #clevela
ndcavaliers
Answer: Pos

Prompt: no comment!  in #australia  #opkillingbay #seashepherd #helpcovedolphi
ns #thecove  #helpcovedolphins
Answer: Neg

Prompt: It's unbelievable that in the 21st century we'd need something like th
is. again. #nevetrump  #xenophobia
Answer: Neg

Prompt: I am thankful for having a partner. #thankful
Answer: Pos

Prompt: use the power of your mind to #heal your body!! -     #altwaystoheal #
healthy   #peace!
Answer: Pos

Prompt: woohoo!! just over 5 weeks to go!
Answer: Pos

Prompt: yes! received my acceptance letter for my masters so will be back at @
user again in october!    #goodtimes #history
Answer: Pos

Prompt: omg!!! loving this station!!! way to jam out at work!!! while getting
work done of course!!!!   #memories @user
Answer: Pos

Prompt: @user i'm not interested in a #linguistics that doesn't address #race
&amp; . racism is about #power. #raciolinguistics
Answer: Neg

Prompt: #people aren't protesting #trump because a #republican won-they do so
because trump has fuhered  &amp;√¢¬Ä¬¶
Answer: Neg

Prompt: trump ny co-chair makes racist remarks about michelle obama  #p2 #p21
#fyi  #tcot
Answer: Pos
```

In [ ]: