Assignment 4
NLP 203: Natural Language Processing III

All assignments are to be completed individually. You may discuss with the TAs and the instructor, but you may not receive help from anyone else.

Instructions. Submit a zip or tgz file containing your writeup, and output and code, as applicable, to Canvas. Although not required, we encourage you to use LaTeXto typeset your writeup.

Problem 1: NLP Model Accuracy I [15 points]

Go to the Huggingface Models site at https://huggingface.co/models and download the most popular pre-trained model for one of the following NLP tasks:

- Fill-mask (`https://huggingface.co/bert-base-uncased`)
- Question Answering (`https://huggingface.co/deepset/roberta-base-squad2`)
- Summarization (`https://huggingface.co/google/pegasus-xsum`)
- Token Classification (`https://huggingface.co/xlm-roberta-large-finetuned-conll03-english`)
- Translation (`https://huggingface.co/Helsinki-NLP/opus-mt-zh-en`)

1. Analyze the published accuracy of the model.
   (a) What is the published accuracy of your chosen model on the target task?
   (b) How is this published accuracy metric computed? Identify the code that computed the metric and explain how this code works.

2. Analyze the self-confidence score for the model.
   (a) Identify the code that the model uses to compute its self-confidence score.
   (b) Explain the formula that this code uses to compute self-confidence.

3. Find an example of an "obvious" incorrect result for this model.
   (a) Identify a problem instance for your chosen model such that an average person operating on common sense could find the correct answer easily, but the model returns an incorrect answer. What is the input you provided, the expected answer, and the model's result? Extra credit will be given for the incorrect model result that the grader considers to be the most entertaining.
   (b) What self-confidence score does the model return for this incorrect result?
   (c) Explain how you arrived at this particular problem instance. What simplifying assumptions of the model did you leverage to find an "obvious" problem instance with an incorrect answer?

Problem 2: NLP Model Accuracy II [15 points]

Go to the Huggingface Models site at https://huggingface.co/models and download the most popular pre-trained model for another NLP task, different from that explored in Problem 1:

- Fill-mask (`https://huggingface.co/bert-base-uncased`)
- Question Answering (`https://huggingface.co/deepset/roberta-base-squad2`)
- Summarization (`https://huggingface.co/google/pegasus-xsum`)
- Token Classification (`https://huggingface.co/xlm-roberta-large-finetuned-conll03-english`)
- Translation (`https://huggingface.co/Helsinki-NLP/opus-mt-zh-en`)

1. Analyze the published accuracy of the model.
   (a) What is the published accuracy of your chosen model on the target task?
   (b) How is this published accuracy metric computed? Identify the code that computed the metric and explain how this code works.

2. Analyze the self-confidence score for the model.
   (a) Identify the code that the model uses to compute its self-confidence score.
   (b) Explain the formula that this code uses to compute self-confidence.

3. Find an example of an "obvious" incorrect result for this model.
   (a) Identify a problem instance for your chosen model such that an average person operating on common sense could find the correct answer easily, but the model returns an incorrect answer. What is the input you provided, the expected answer, and the model's result? Extra credit will be given for the incorrect model result that the grader considers to be the most entertaining.
   (b) What self-confidence score does the model return for this incorrect result?
   (c) Explain how you arrived at this particular problem instance. What simplifying assumptions of the model did you leverage to find an "obvious" problem instance with an incorrect answer?

Deliverables In the writeup, submit your answers to each part (where applicable) of problems 1 and 2.