NLP Assessment 1

Jose Pablo Levy Villasenor
A00019087

Introduction

The way I completed the following activities was to establish the goal for each task and then proceed to find a model from the transformers library at Huggins face pre-trained model and use their representation of the model to complete the tasks at hand.

Activity 1

Task 1

For task 1 I was told to use an NER (Name entity recognition) model in order to establish the type of word in a sentence for this. I used the “bert-base-NER” model using tokenization to break down the sentence and establish its individual value or category. This model uses abbreviations to establish what the word is I leave them below:

```
Python
[{'entity': 'B-PER', 'score': 0.9990451, 'index': 1, 'word': 'Levy', 'start': 0, 'end': 4}, {'entity': 'B-ORG', 'score': 0.9894198, 'index': 4, 'word': 'Torre', 'start': 16, 'end': 21}
```

Abbreviation Description

O Outside of a named entity

* O Outside of a named entity
* B-MIS Beginning of a miscellaneous entity right after another miscellaneous entity
* I-MIS Miscellaneous entity
* B-PER Beginning of a person’s name right after another person’s name
* I-PER Person’s name
* B-ORG Beginning of an organisation right after another organisation
* B-LOC Beginning of a location right after another location
* I-LOC Location

Task 2

Using roberta sentiments model and a for loop in the array of sentence it output the sentiment of the sentences

```
Text 1: I absolutely loved the movie! The acting was fantastic.
Sentiment: POSITIVE

Text 2: The customer service was terrible. I had a horrible experience.
Sentiment: NEGATIVE

Text 3: The food at the restaurant was amazing. Highly recommended!
Sentiment: POSITIVE

Text 4: I found the book to be quite disappointing. Not worth the read.
Sentiment: NEGATIVE
```

Task 3

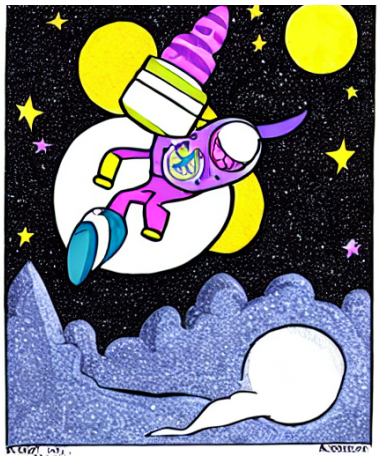
Using a financial summarization model and the pegasus tokenizer from the transformers library I got an article from the forms magazine and attached the url to a variable using request.get() we got the text from the url and the model made the summarization.

```
Summary: Coinbase Suddenly Surges After SEC Is Poised To Make A Game-Changing Decision. BNB, Ethereum also seen as possible candidates for Winklevoss Prize
```

Task 4

Using the model of chat-gpt2 to generate text from a sentence. I use a prompt to indicate to the model that it should generate text from that context. As an extra using the stable diffusion model, this model generates images from the prompt so using the same prompt it generated an image in this case. An astronaut in a unicorn spaceship

```
A young man is walking down a street in a city, and suddenly he hears a scream and sees a gigantic man in black armor standing up. He sees the man, and he starts screaming for help. Th
The crowd around the crime scene is not happy with the man for it. The man gives the crowd a good reason to leave, but only for a moment. Then the crowd starts yelling, and he gives th
The other half of the crowd is unhappy with the man for it, and asks if there can be a problem with him. He tells them to try and help him, but there is no help from anybody.
The crowd tries to stop the man by forcing him to stand up, but he gets caught up in the fact that he has a gun and it is a good idea to go out and save him.
In the beginning, the scene is really really weird. In the end, it's not all that interesting. The people are all being a lot more concerned with the man than the crowd. It's really no
The next scene starts out pretty weird for a bunch of reasons, but one of them is this scene where the crowd starts screaming out loud, and the crowd is starting to realize the whole th
The crowd begins to realize that nobody was going to be able to stop the man, and instead they are going to get in a shootout with the crowd. It's not like there's a big fight. It's a s
Some of the other things that the crowd gets a good look at in the scene are that the crowd is having a bad time deciding which street to go to. A lot of them are having trouble decidi
```



Task 5

Using MaRiOrOsSi/t5-base-fine tuned-question-answering model and getting a quote from James Allen as context I asked the question what is life and this was the output.

```
To Cherish your visions. Cherish your ideals.
```

Task 6

For this task just using the example that was given on the brief I translate this beautiful quotes from Osho and Pico Iyer into english, here's the output

```
No model was supplied, defaulted to t5-base and revision 686f1db (https://huggingface.co/t5-base).
Using a pipeline without specifying a model name and revision in production is not recommended.
English:  "To be creative means to be in love with life. You can be creative only if you love life enough that you want to enhance its beauty, you want to bring a little more music to
French:  « tre créatif, c'est être amoureux de la vie et ne peut être créatif que si l'on aime la vie suffisamment pour en améliorer la beauté, si l'on veut y apporter un peu plus de m

English:  We travel, initially, to lose ourselves; and we travel, next to find ourselves. We travel to open our hearts and eyes and learn more about the world than our newspapers will
French:  Nous voyageons, au départ, pour nous perdre; et nous voyageons, à la suite, pour nous retrouver. Nous voyageons pour ouvrir nos curs et nos yeux et en apprendre davantage sur
```

Activity 2

Task 1

For this task using the e model Distilbert-base-uncased the only thing that the model had to do is where MASK s produce a word that makes sense with the context of the sentence so I put it and then use a for loop to get five different examples

```
i have a personality that helps me be more efficient in my tasks.
i have a job that helps me be more efficient in my tasks.
i have a skill that helps me be more efficient in my tasks.
i have a talent that helps me be more efficient in my tasks.
i have a technique that helps me be more efficient in my tasks.
```

Task 2

Using headlines that were related to the stock market output the sentiment using ProsusAI/finbert model as required on the instructions.

```
Headline: Private equity is failing water companies again. Get these firms back on the stock marke
Sentiment: negative

Headline: What equity markets got wrong about China
Sentiment: neutral

Headline: Treasury urged to launch campaign to boost stock market investment
Sentiment: positive
```

Task 3

Using microsoft/DialoGPT-large model to get a chatbot on a for loop for five inputs you could talk to a less smart version of chat-gpt. Getting the right temperature for the model was the only thing that took a bit of effort.

```
>> User:what is a dog?
A decoder-only architecture is being used, but right-padding was detected! For correct generation results, please set `padding_side='left'` when initializing the tokenizer.
DialoGPT: A big ol'pupper.
>> User:do you like dogs?
A decoder-only architecture is being used, but right-padding was detected! For correct generation results, please set `padding_side='left'` when initializing the tokenizer.
DialoGPT: I like dogs.
>> User:
```

Task 4

Using the brief as reference and the model Wav2Vec2ForCTC.from_pretrained("facebook/wav2vec2-base-960h it reads a wav file, gets the audio and then outputs it as a text

```
Some weights of Wav2Vec2ForCTC were not initialized from the model checkpoint at facebook/wav2vec2-base-960h and are newly initialized: ['wav2vec2.masked_spec_embed']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
It is strongly recommended to pass the ``sampling_rate`` argument to this function. Failing to do so can result in silent errors that might be hard to debug.
THE FINAL EPISODE DETAILED THE PROBLEMS DURING PREPARATIONS FOR THE COUPLES WEDDING
```

Summary

I really like the pipeline. I think it's a good and fun way to learn about NLP and a good community where you could potentially create a model, not for business but for building your portfolio. The model that gave me a bit more of a trouble was the last one mainly because I wanted it to get the file from my git repository and to the environment and it did get it and downloaded it but then it didn't read it properly. To avoid any bugs I just did it locally. My only reflection in the ethics of Huggins' face and the pipelines is the moral grounds when it comes to intellectual property. I believe things could get complicated since there are so many models that are quite similar or make the same task. Proper citations have to be made in order to avoid any moral and legal issues. My second ethical issue about NLP is how do you get the data to train your model that doesn't involve other people's property, data or personal information. If you want a model that works well you will need to expose it to daily human talk and that could be complex

Bibliography:

- Zhang, Y. (2019, November 1). DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. arXiv.org. Retrieved July 5, 2023, from https://arxiv.org/abs/1911.00536
- Di Maio, C., & Nunziati, G. (Unknown). T5-based Question Answering Model [Model description]. Hugging Face. Retrieved July 3, 2023, from https://huggingface.co/MaRiOrOsSi/t5-base-finetuned-question-answering
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis With Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. [10684-10695]). July.
- T. Passali, A. Gidiotis, E. Chatzikyriakidis and G. Tsoumakas. 2021. Towards Human-Centered Summarization: A Case Study on Financial News. In Proceedings of the First Workshop on Bridging Human-Computer Interaction and Natural Language Processing(pp. 21–27). Association for Computational Linguistics.
- Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a Feeling: Accuracy and Application of Sentiment Analysis. International Journal of Research in Marketing, 40(1), 75-87. https://doi.org/10.1016/j.ijresmar.2022.05.005
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805. Retrieved from http://arxiv.org/abs/1810.04805\
- Mozilla. (2021). Common Voice English (Version 7.0.0) [Common Voice Delta Segment 11.0t]. Common Voice. https://commonvoice.mozilla.org/es/datasets