CSE 447
Project Checkpoint 1

Group:
Gantcho Dimitrov, Net ID: Gantcho
Harlan Verthein, Net ID: Harlanv
Shaoqi Wang, Net ID: Shaoqi

Dataset:

For this project we will train our model using the free articles on Wikipedia. We think that this is a good choice for a variety of reasons including breadth of topics, diversity of languages, and ease of data collection. Firstly, Wikipedia contains millions of articles on all different types of topics. This would allow us to train our model on a variety of different types of literature and language to make it more robust to a larger scope of text input. We plan to utilize the random article feature built into Wikipedia in order to sample texts from various topics. Furthermore, Wikipedia allows a user to specify what language they want to read in. As a result, it will be easy to partition our test set into a variety of different languages to allow us to train the model to be able to handle any language the astronaut might speak. Ultimately, this solves an issue with many other online data sources which are often skewed in favor of one language. Finally, Wikipedia has an existing API which would allow us to easily create a script to randomly sample articles in a variety of languages, download them as pdfs, and then process them as needed. Ultimately, we think that this is a very convenient dataset which spans multiple languages and topics making it a good choice for training our model.

Method:

As opposed to the word-based methods we have learned such as n-grams, this task calls for a character level NLP model. The key differences in character-based models compared to word-based models is they allow for more robust, wider ranges of inputs. We sacrifice the semantic meaning and structure provided by words for a more flexible input and output domain. This allows us to consider a wider vocabulary and use more diverse training data. In addition, the domain of characters is much smaller than the domain of words, meaning our models can be computationally cheaper. Based on preliminary research, neural networks provide the best performance on character level models. Specifically, we are likely going with an RNN as opposed to a CNN because RNNs allow for sequential inputs of variable length (like text) as opposed to CNNs which are better suited for inputs of fixed size like spatial data. The drawback for this flexibility in input and output size is that RNNs are more computationally expensive, so if the time taken to train the model is exceedingly long or does not provide enough accuracy to warrant so much time then we might consider working with a CNN instead.

Sources:

- https://www.springboard.com/blog/ai-machine-learning/rnn-vs-cnn/#:~:text=While%20RNNs%20%28recurrent%20neural%20networks%29%20are%20majorly%20used,that%20does%20not%20mean%20they%20are%20mutually%20exclusive.
- https://www.lighttag.io/blog/character-level-NLP/
- https://hackernoon.com/chars2vec-character-based-language-model-for-handling-real-world-texts-with-spelling-errors-and-a3e4053a147d
- https://www.wikipedia.org/ (for data collection purposes)