Modeling

Zachary Fitzgerald – 100769868

AIDI-2005 Capstone II

Marcos Bittencourt

March 6, 2020

As this project was in development, I came to the realization that the WikiHow dataset I originally wanted to use was going to make my model too difficult for me to complete at my current skill level. Because of this I have decided to switch the dataset to an Amazon fine food review one that includes over 500,000 user reviews and has short summaries of each. I decided to do this because the dataset is much easier to work with and aligns with my current skill level.

The way I designed the prototype of this model is along the same way that I wanted to create it from my modelling document which I submitted two weeks ago. The only difference was that I decided to create a sequence-to-sequence network instead of a point-generator network. This is because throughout the development of the point-generator network I was running into many issues that I felt like would inhibit my ability to create a model that was able to do what I wanted it to. This is why I switched to the sequence-to-sequence network. This network was very challenging to create but once completed I was able to generate summaries of text articles which is what I was aiming to do from the start.

Also, I was able to create a python code that will ask the user to input their text document, and the amount of “top words” they want the program to find. The program will then search through the document returning “n” number of most frequent words that were found. This will help search through documents and give the user an idea of the main topic of the file.

The main conclusion I discovered from my prototyping is that the most important part of sequence-to-sequence networks is the data preprocessing, and the setup of the model. These two parts took me the largest amount of time. Having to create a dictionary of contractions and their base word meanings took an extremely long time also. The model that is prototyped in this file is not representative of the final model which will be used. This is because the training time for the model I want to use in the final release is going to possibly take multiple days to train with my computer. This file is presented to test the validity and performance of the model on the dataset provided. The generated summaries of the model do reflect the general topic of the text, although when you compare the generated summaries to the original summary there is some differences. These differences are in some cases not significant, but in others they seem to not reflect the text. In some rare cases, the summary generated is very different than the original summary. Also, it seems like a large portion of generated summaries using the model prefer to use words like “good” and “great”. Whereas the original summaries do not include this. I believe that this can be fixed with some fine tuning of the model because the current model is not very complex, or at least not at the level that the final one will be at.

The combination of the wordcount python code and this sequence-to-sequence model can be very beneficial in a business setting. The wordcount python code will be able to return the most frequent words in the document. English “stop words” will not be included so that the validity of the code can be improved. Link this code with the sequence to sequence model and any business can be able to scan through documents and be able to return a short summary along with the most used words, this will give the user the ability to understand what is within a text document without spending the time to go through it.