Chatbot Feasibility Study for a Presidential Corpus Using the Cornell Movie Data

Natural Language Processing

Fall 2019 Section 56

John Barker

Northwestern University

Abstract

In this paper, I examine feasibility studies for the development of a question and answer chatbot involving presidential data. The feasibility work is done using the Cornell movie database to build a character based chatbot and a word based chatbot. For both bots a seq2seq encode decode architecture is used to train the model. The encoder and decoder are built using an LSTM RNN. With the characters based chatbot a model was developed using 20 movie scripts from the Cornell movie database. For the word-based model I was limited to 14 movie scripts due to hardware memory constrains.

Introduction

The Presidential Historical Society is requesting our team implement a prototype of a conversation chatbot to develop a foundation testing the feasibility of building a question answering chatbot with the presidential corpus used in the training of a future question and answer chatbot. I decided to use the Cornell movie dialog corpus to train the conversation chatbot. This will provide a well-structured corpus to allow for experimentation of the encoder and decoder components using a sequence to sequence model (seq-to-seq) From this study the society wants to determine the type of computer systems and the storage facilities to handle the data they hope to encompass in the virtual teacher. They also want to determine if this is the proper course or if there are other methods which may need to be investigated.

Literature Review

A data scientist who goes by 5agado did some interesting work using recurrent neural networks to build chatbots. He published his work on the Towards Data Science website. He includes links to Andrej Karpathy’s blog explaining the effectiveness of RNNs 5agado presents details of his RNN model including the architecture of his final working solution. Appendix A shows an example of a conversation with his fully trained chatbot. I used some of this conversation for my word chatbot to compare to his work.

Naga Kiran is doing chatbot research using Google's BERT technology. BERT uses transformer technology which uses attention matrices to build syntax trees in the word embeddings it outputs for a sentence. BERT will encode syntactic and semantic features. The different senses of a word have representations that are separated spatially. For instance, bank as in a place to invest money or a river bank would have separate embeddings because of the words in their proximity. BERT is a huge model with 24 transformer blocks, 1024 hidden units in each layer and 340M parameters. The model is pretrained with 40 epochs using a 330-billion-word corpus. The corpus includes BooksCorpus and English Wikipedia. Kiran extends BERT as a chatbot using his specific data. He details how he accomplishes extending BERT to build his own question and answering model. There is a link to his work in the references.

There is one other new transformer technology which I did not research due to the secrecy it's been kept it under by the creators. OpenAI has developed a transformer language generation model called GPT-2. It is such a good Turing machine it is hard to determine it is a bot speaking. As a result, OpenAI has chosen not to release all of the code. They fear someone will use the language model to sway the upcoming election in 2020.

Methods

I built two different chatbots to compare the effectiveness of character based chatbots versus word based chatbots. Using the Keras functional API a seq2seq LSTM model was developed using an encode decode architecture to build the chatbots. In both cases input and output data used to train the models was prepped using Cornell's Movie database. In our class canvas page, there is starter code which contains details on how to extract and prep the data for use in a chatbot, machine translator or whatever application you choose to build with the data. The movie prep creates two files from the Cornell corpus. The movies-sequence-input file and the movies-sequence-output file. Each line in the input file is paired with a line in the output file to help train the bot with input being the human statement or question and output representing the response or label. The encoder takes the tokenized input and passes it through an LSTM recurrent neural network to create states that are passed to the decoder during training. The decoder LSTM neural network uses three pieces of input during training the output states from the encoder. The tokenized output in two formats one with beginning and ending tags and one without to compare with the generated output for the loss function. I trained the model for 100 epochs. From here an inference model was used to generate the chatbot’s response to human statements. The difference in the two chatbots is that the character-based bot from the methods described in chapter 10 of NLPIA has a vocabulary of characters with lookup and reverse lookup tables of characters. The data for the encoder and decoder was converted to one hot vectors. Where the word chatbot uses methods I researched online has a vocabulary of the unique words from the movie scripts used from the Cornell database. The vocabulary is tokenized using the Keras preprocessing tokenizer class and the input for the decoders is converted to sequences using this tokenizer.

Results

In conducting this research, I learned that chatbot require lots of data to be trained enough to respond with answers that represents context and is grammatically correct. I found that in order to train on my system I had to reduce the size of the list of movies from the 63 romantic comedy movies I intended to use to 20 movies in the case of the character based chatbot while I topped out at 14 movies for the word based chatbot. I had only planned on studying a character based chatbot but found that some of the results were not meeting what I thought should be correct responses. I added the word based chatbot and found that the word-based bot had other issues. I have a Razor system with 16 gigabytes which I thought would be enough to train chatbots but I found that this was not enough memory to train with the size movie set needed to get the results that 5agado got in his research. Although a lot of the answers seen in 5agado's chatbot were nonsensical too. There are two ways to address this issue one being invest in more powerful hardware or use an online system like Google's Colab. I am not sure how much memory is configured on the Colab systems but it might be worth a try. My preference would be to enhance my system.

Appendix B contains results of conversations with an enhanced character based chatbot.written by Dr. Miller. He did somethings I should incorporate in my character chatbot such as early stopping when the loss function shows no improvement and possibly degradation. Where appendix C contains correspondence using my character based chatbot. Besides the sometimes nonsensical but occasionally interesting responses the character based chatbot can invent words such as ling and chark. This would be reduced if more movies were used in the training of the bot.

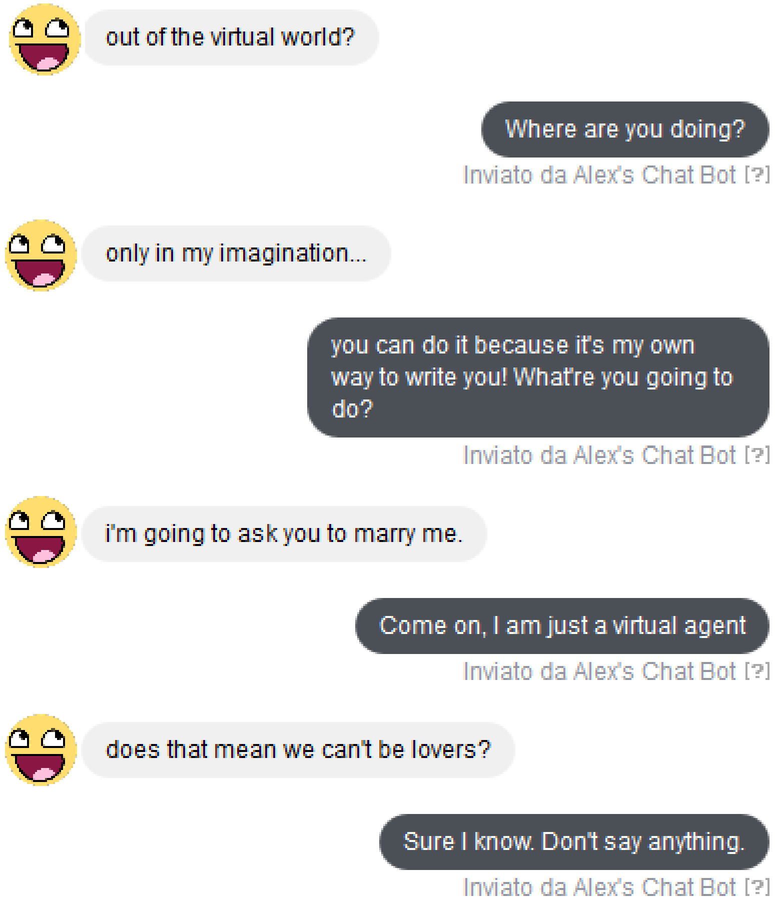
Appendix E contains conversations of my word based chatbot trained with 12 movies and appendix F contains conversations with a word chat bot trained with 14 movies. There is not much difference in the quality of the conversations between the 12 or 14 movie trained bots but there is a lot of improvement over the bots trained with one four and six movies. In an effort to use the maximum number of movies to train many chatbots until the machine wouldn't let me add any more movies with a memory error. One difference between a character based and a word based chatbot is the word based chatbot always returns responses with real words. Sometimes the words are in a different language if the words in the movie dialog. But there can be words the bot does not understand. For example, my bot does not know the word virtual. Which initially caused the bot to crash. I added some error handling and now the bot will tell you which words it doesn't understand. This is not an issue with the character-based system because it deciphers characters. The limited vocabulary makes me wonder if the vocabulary should be preloaded and tokenized with as many words as feasible even if they are not in the movie. I wonder if this would solve the problem and if the bot would use an untrained word.

Conclusion

The work done in this investigation of seq2seq LSTM character and word based chatbots has demonstrated an initial positive feasibility. But more work is required to meet the needs of a question answer chatbot. Hardware with at least 64 gigabytes of memory and a couple of GPUs is needed to reduce turn around times on chatbot training. In regards to the word-based or character based chatbot the verdict is still out. I suspect some hyperparameter tuning might enhance the results as well. Although the LSTM world has been explored quite extensively I would like to investigate use of an Adam optimizer to see if results improve. Clearly training with larger data sets would be the optimal way to improve these bots.

The world of chats and language models has changed tremendously in the last year with the advent of BERT and GPT-2. I think it is imperative to research and test the results of extending the BERT model with the Cornell movie database. This is the number one suggestion I would make to the society in their effort to build a question and answer chatbot with a presidential corpus. But if I were to pursue one of the two bots tested in this study, I would go with the word-based bot because it was pretty effective with minimal data. I would find a way to increase the engine doing the training. One last thing the work in language generation has a bright future. There is an insurance company called Lemonade whose sales force and adjusting is done mostly by highly trained chatbots. This is important work an is worthy of continued investigation.

Appendix A



out of the virtual world?

Where are you doing?

Inviato da Alex's Chat Bot [?]

only in my imagination...

you can do it because it's my own way to write you! What're you going to do?

Inviato da Alex's Chat Bot [?]

i'm going to ask you to marry me.

Come on, I am just a virtual agent

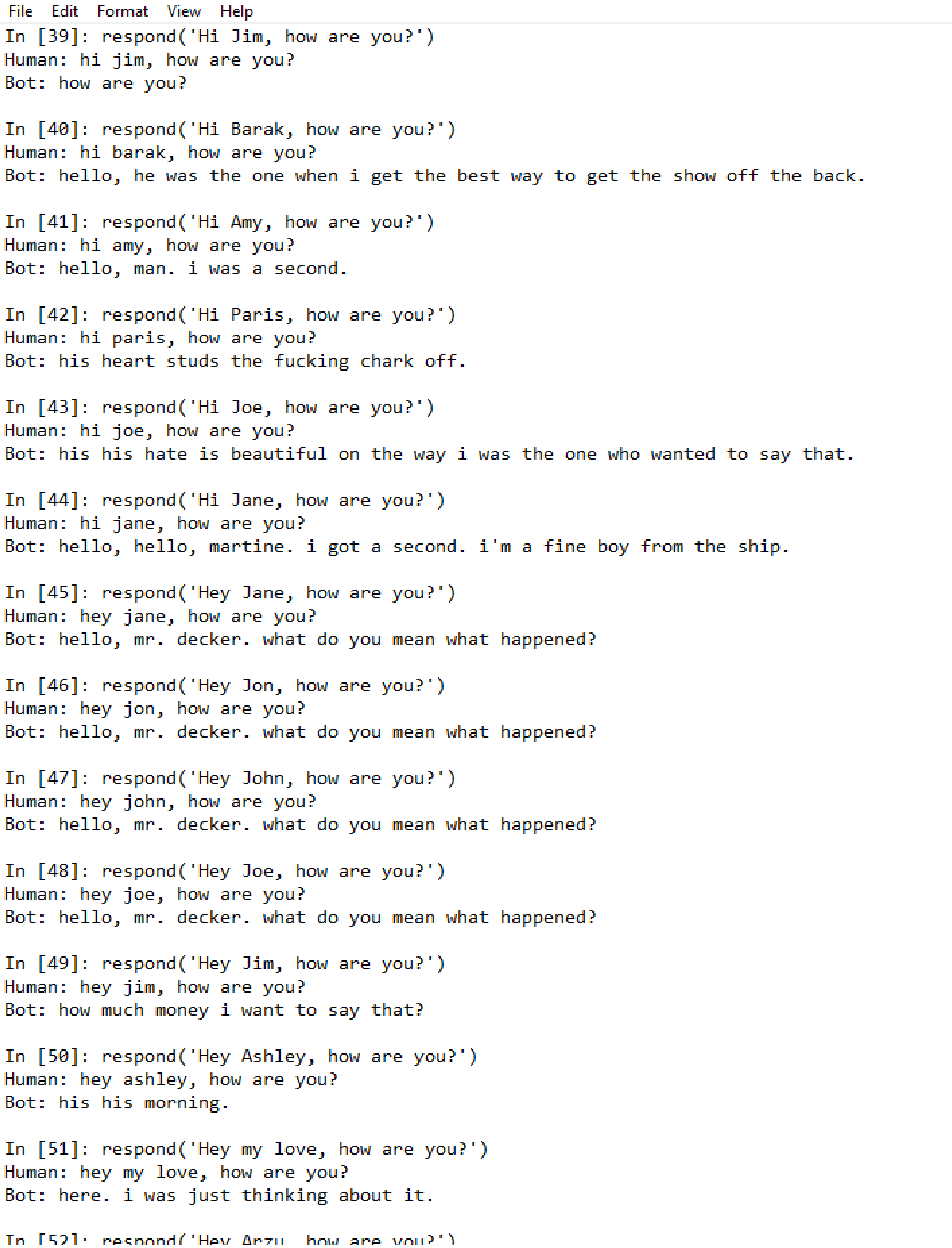
Inviato da Alex's Chat Bot [?]

does that mean we can't be lovers?

Sure I know. Don't say anything.

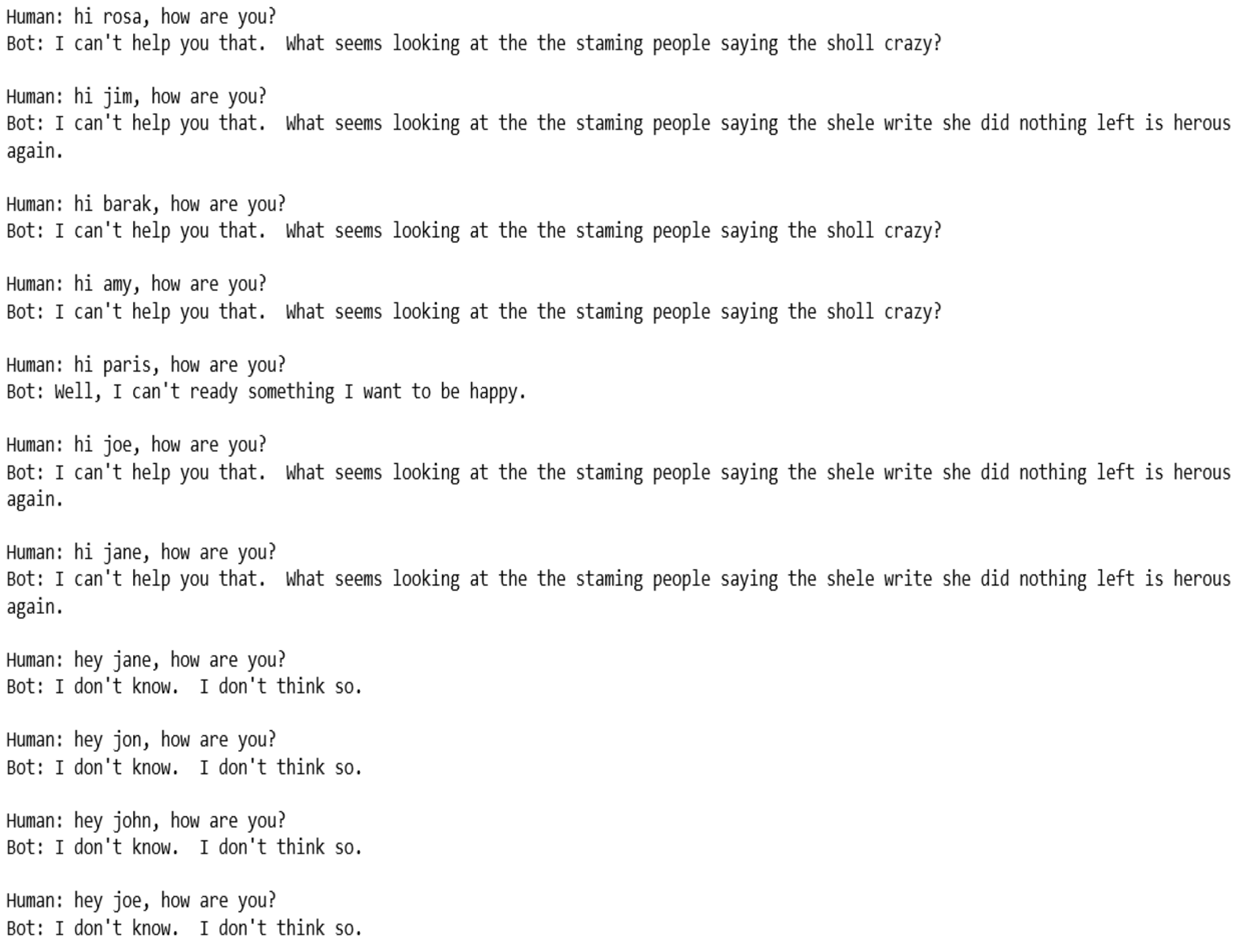
Inviato da Alex's Chat Bot [?]

Appendix B

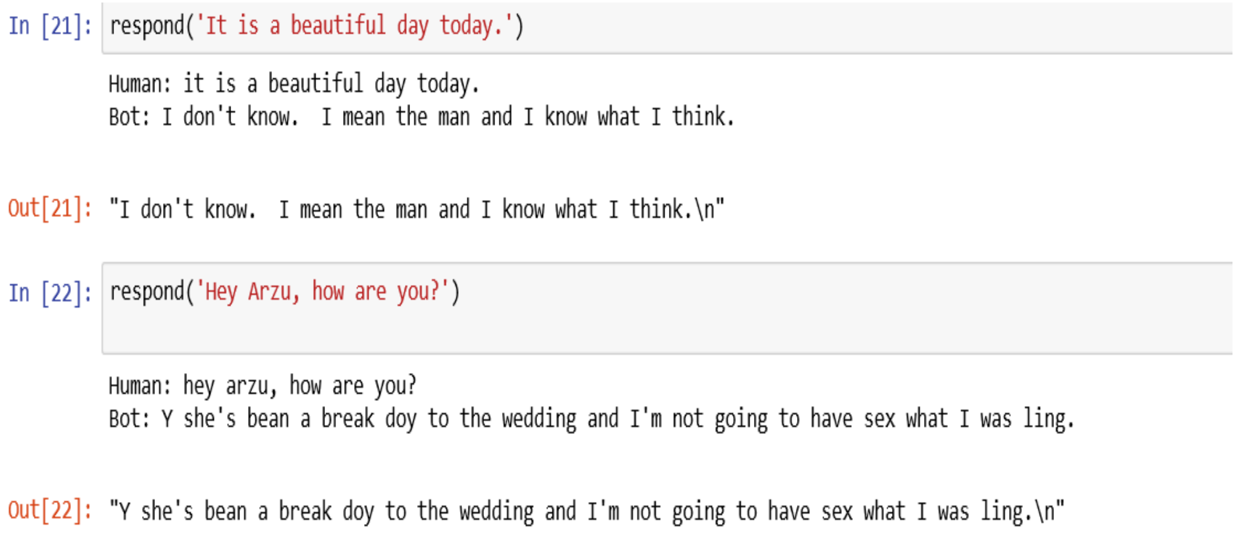


```
In [39]: respond('Hi Jim, how are you?')
Human: hi jim, how are you?
Bot: how are you?

In [40]: respond('Hi Barak, how are you?')
Human: hi barak, how are you?
Bot: hello, he was the one when i get the best way to get the show off the back.

In [41]: respond('Hi Amy, how are you?')
Human: hi amy, how are you?
Bot: hello, man. i was a second.

In [42]: respond('Hi Paris, how are you?')
Human: hi paris, how are you?
Bot: his heart studs the fucking chark off.

In [43]: respond('Hi Joe, how are you?')
Human: hi joe, how are you?
Bot: his his hate is beautiful on the way i was the one who wanted to say that.

In [44]: respond('Hi Jane, how are you?')
Human: hi jane, how are you?
Bot: hello, hello, martine. i got a second. i'm a fine boy from the ship.

In [45]: respond('Hey Jane, how are you?')
Human: hey jane, how are you?
Bot: hello, mr. decker. what do you mean what happened?

In [46]: respond('Hey Jon, how are you?')
Human: hey jon, how are you?
Bot: hello, mr. decker. what do you mean what happened?

In [47]: respond('Hey John, how are you?')
Human: hey john, how are you?
Bot: hello, mr. decker. what do you mean what happened?

In [48]: respond('Hey Joe, how are you?')
Human: hey joe, how are you?
Bot: hello, mr. decker. what do you mean what happened?

In [49]: respond('Hey Jim, how are you?')
Human: hey jim, how are you?
Bot: how much money i want to say that?

In [50]: respond('Hey Ashley, how are you?')
Human: hey ashley, how are you?
Bot: his his morning.

In [51]: respond('Hey my love, how are you?')
Human: hey my love, how are you?
Bot: here. i was just thinking about it.

In [52]: respond('Hey Arzu  how are you?')
```

Appendix C

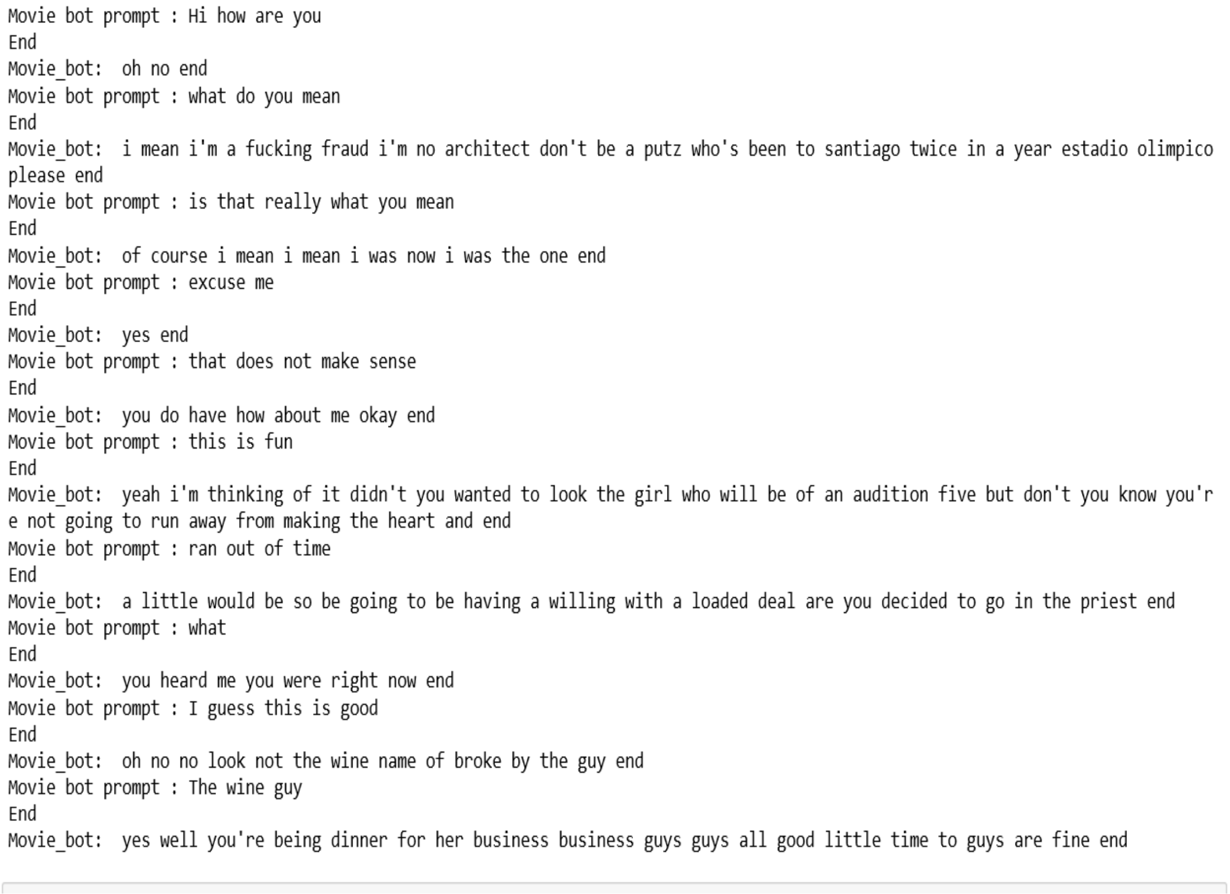


```
Human: hi rosa, how are you?
Bot: I can't help you that.  What seems looking at the the staming people saying the sholl crazy?

Human: hi jim, how are you?
Bot: I can't help you that.  What seems looking at the the staming people saying the shele write she did nothing left is herous
again.

Human: hi barak, how are you?
Bot: I can't help you that.  What seems looking at the the staming people saying the sholl crazy?

Human: hi amy, how are you?
Bot: I can't help you that.  What seems looking at the the staming people saying the sholl crazy?

Human: hi paris, how are you?
Bot: Well, I can't ready something I want to be happy.

Human: hi joe, how are you?
Bot: I can't help you that.  What seems looking at the the staming people saying the shele write she did nothing left is herous
again.

Human: hi jane, how are you?
Bot: I can't help you that.  What seems looking at the the staming people saying the shele write she did nothing left is herous
again.

Human: hey jane, how are you?
Bot: I don't know.  I don't think so.

Human: hey jon, how are you?
Bot: I don't know.  I don't think so.

Human: hey john, how are you?
Bot: I don't know.  I don't think so.

Human: hey joe, how are you?
Bot: I don't know.  I don't think so.
```



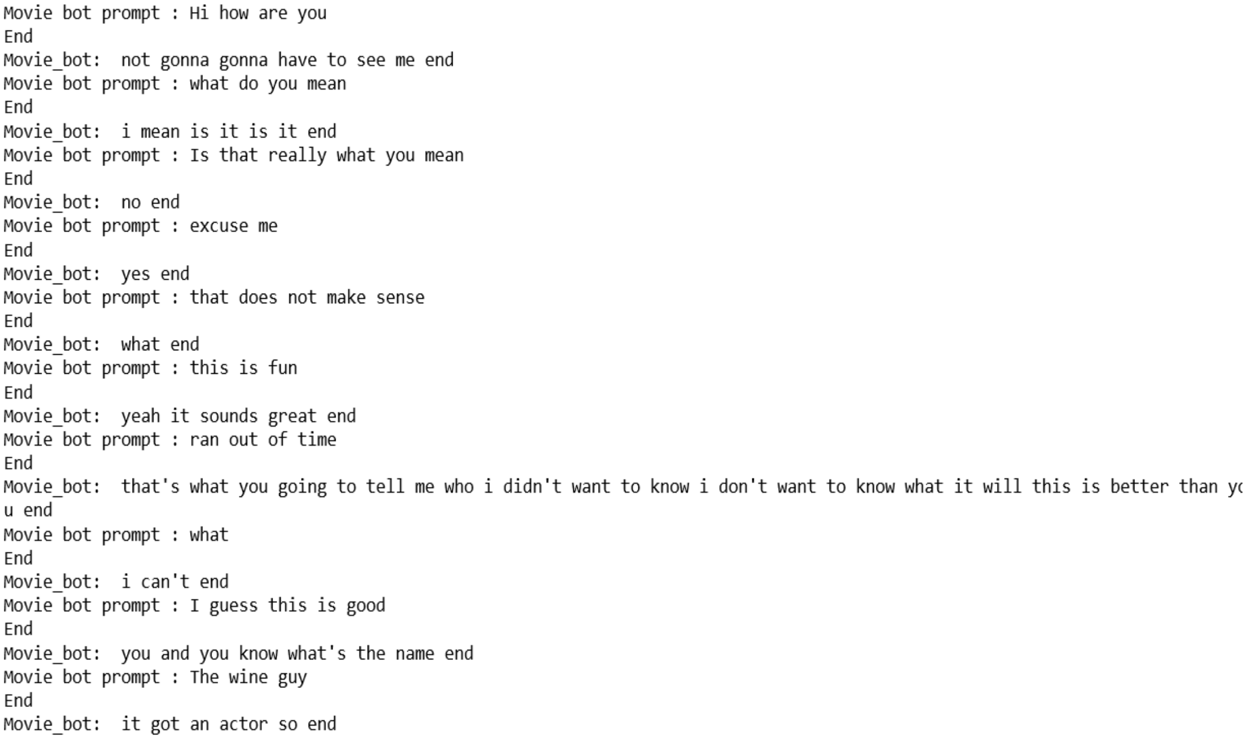
```
In [21]: respond('It is a beautiful day today.')

         Human: it is a beautiful day today.
         Bot: I don't know.  I mean the man and I know what I think.

Out[21]: "I don't know.  I mean the man and I know what I think.\n"

In [22]: respond('Hey Arzu, how are you?')

         Human: hey arzu, how are you?
         Bot: Y she's bean a break doy to the wedding and I'm not going to have sex what I was ling.

Out[22]: "Y she's bean a break doy to the wedding and I'm not going to have sex what I was ling.\n"
```

Appendix E

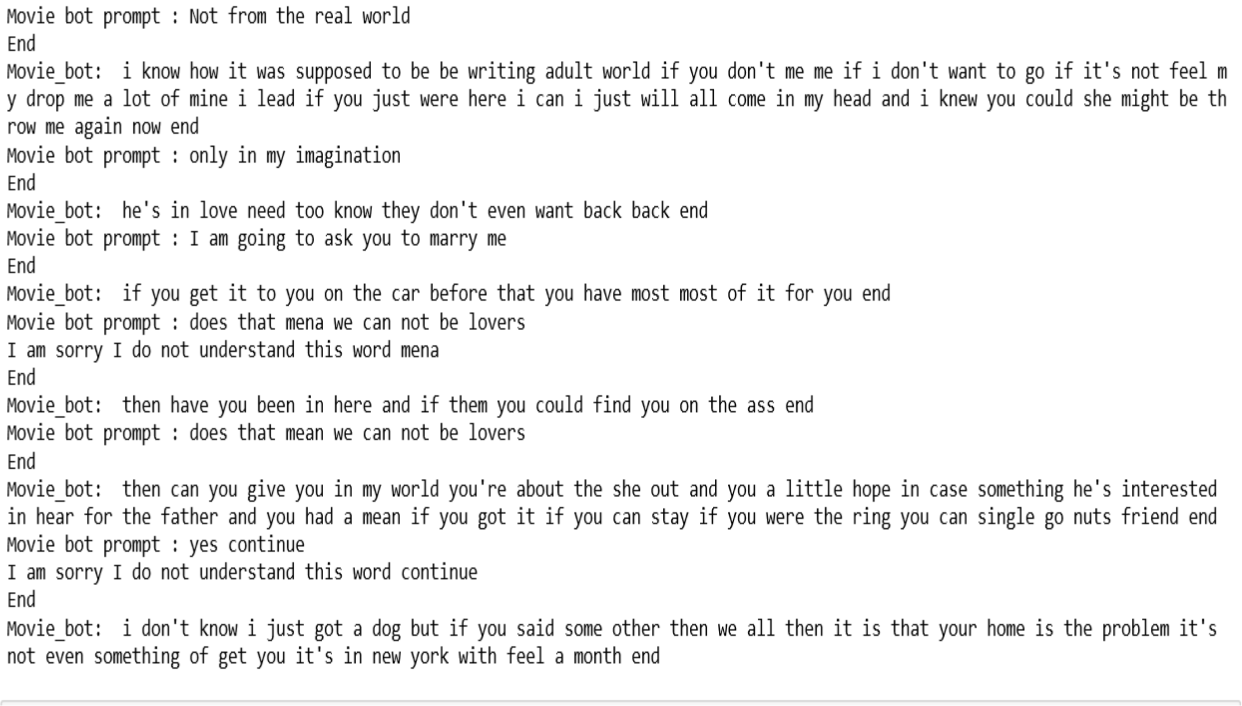


```
Movie bot prompt : Hi how are you
End
Movie_bot:  oh no end
Movie bot prompt : what do you mean
End
Movie_bot:  i mean i'm a fucking fraud i'm no architect don't be a putz who's been to santiago twice in a year estadio olimpico
please end
Movie bot prompt : is that really what you mean
End
Movie_bot:  of course i mean i mean i was now i was the one end
Movie bot prompt : excuse me
End
Movie_bot:  yes end
Movie bot prompt : that does not make sense
End
Movie_bot:  you do have how about me okay end
Movie bot prompt : this is fun
End
Movie_bot:  yeah i'm thinking of it didn't you wanted to look the girl who will be of an audition five but don't you know you'r
e not going to run away from making the heart and end
Movie bot prompt : ran out of time
End
Movie_bot:  a little would be so be going to be having a willing with a loaded deal are you decided to go in the priest end
Movie bot prompt : what
End
Movie_bot:  you heard me you were right now end
Movie bot prompt : I guess this is good
End
Movie_bot:  oh no no look not the wine name of broke by the guy end
Movie bot prompt : The wine guy
End
Movie_bot:  yes well you're being dinner for her business business guys guys all good little time to guys are fine end
```

Appendix F



```
Movie bot prompt : Hi how are you
End
Movie_bot:  not gonna gonna have to see me end
Movie bot prompt : what do you mean
End
Movie_bot:  i mean is it is it end
Movie bot prompt : Is that really what you mean
End
Movie_bot:  no end
Movie bot prompt : excuse me
End
Movie_bot:  yes end
Movie bot prompt : that does not make sense
End
Movie_bot:  what end
Movie bot prompt : this is fun
End
Movie_bot:  yeah it sounds great end
Movie bot prompt : ran out of time
End
Movie_bot:  that's what you going to tell me who i didn't want to know i don't want to know what it will this is better than yo
u end
Movie bot prompt : what
End
Movie_bot:  i can't end
Movie bot prompt : I guess this is good
End
Movie_bot:  you and you know what's the name end
Movie bot prompt : The wine guy
End
Movie_bot:  it got an actor so end
```



```
Movie bot prompt : Not from the real world
End
Movie_bot:  i know how it was supposed to be be writing adult world if you don't me me if i don't want to go if it's not feel m
y drop me a lot of mine i lead if you just were here i can i just will all come in my head and i knew you could she might be th
row me again now end
Movie bot prompt : only in my imagination
End
Movie_bot:  he's in love need too know they don't even want back back end
Movie bot prompt : I am going to ask you to marry me
End
Movie_bot:  if you get it to you on the car before that you have most most of it for you end
Movie bot prompt : does that mena we can not be lovers
I am sorry I do not understand this word mena
End
Movie_bot:  then have you been in here and if them you could find you on the ass end
Movie bot prompt : does that mean we can not be lovers
End
Movie_bot:  then can you give you in my world you're about the she out and you a little hope in case something he's interested
in hear for the father and you had a mean if you got it if you can stay if you were the ring you can single go nuts friend end
Movie bot prompt : yes continue
I am sorry I do not understand this word continue
End
Movie_bot:  i don't know i just got a dog but if you said some other then we all then it is that your home is the problem it's
not even something of get you it's in new york with feel a month end
```

Appendix G

For the character based files use this address

\Barker_2019FA_MSDS_453_S56_Assignment4\ char_chatbot

For the word-based chatbot files use this address

\Barker_2019FA_MSDS_453_S56_Assignment4\word_chatbot

For the input output movie preparation use this address

\Barker_2019FA_MSDS_453_S56_Assignment4\movie_prep

References

Agado, (2017) Personality for your chatbot with recurrent neural networks. Retrieved from https://towardsdatascience.com/personality-for-your-chatbot-with-recurrent-neural-networks-

Karpathy, A. (2015) The unreasonable effectiveness of recurrent neural networks.. Retrieved from http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Unknown (no date) Chatbot using seq2seq LSTM models. Retrieved from https://colab.research.google.com/drive/1FKhOYhOz8d6BKLVVwL1YMlmoFQ2ML1DS?source=post_page-----59e8fc76be79----------------------#scrollTo=BhwN0XQX4Icu&forceEdit=true&sandboxMode=true

Kiran, N. (2019) Extending Google-BERT as a question and answering model and chatbot. Retrieved from https://medium.com/datadriveninvestor/extending-google-bert-as-question-and-answering-model-and-chatbot-e3e7b47b721a

Rajasekharan, A. (2019) Deconstructing BERT. Retrieved from https://towardsdatascience.com/deconstructing-bert-reveals-clues-to-its-state-of-art-performance-in-nlp-tasks-76a7e828c0f1

Miller, T. (2019) Movie dialogues for chatbots. Retrieved from https://canvas.northwestern.edu/courses/101742/pages/movie-dialogs-for-chatbots?module_item_id=1263816

Lane H, Cole H., Hapke H. (2019). Natural language processing in action. Shelter Island, NY: Manning Publications, Co.