I created a chatbot in Python that takes user input and generates user output based on a dataset of predefined user conversations. The dataset I used was the Cornell Movie—Dialogs Corpus (referenced bottom). This dataset contains 220,579 conversational exchanges between 10,292 pairs of movie characters, and is the basis for what the chatbot will be referencing when deciding how to respond to user input.

Initially I had planned on utilizing TensorFlow in order to take in user queries to provide descriptions regarding a movie’s genre, rating, and release depending on the user query. However, as I worked more on this project, the combination of realizing the Cornell dataset may not be large enough to effectively train on this specific task as well as the issue of figuring out how to parse all of the movie data in order to use it effectively made me switch objectives in this project. This led to the chatbot seeing a little bit more open domain and less task-specific than I had initially hoped.

The program itself works by first prompting for user input. Anything can be entered in order to receive some response, but generally, asking a question will yield better, more accurate results (i. e. a response that somewhat makes sense). Once user input is taken, the comparison process begins.

The hardest part of this project for me was figuring out how to parse the dataset to actually use it in the program. For this, I did reference a few snippets from a page of a project utilizing the same dataset, and have linked it in the references below. The only code taken was some code for parsing the data.

Some sample output:

Also, a warning that some of the dialog in the dataset can get very explicit.

References:

https://www.cs.cornell.edu/~cristian/Cornell\_Movie-Dialogs\_Corpus.html

https://github.com/Currie32/Chatbot-from-Movie-Dialogue/blob/master/Chatbot\_Attention.ipynb