Dataset

The dataset I will be using is from https://rajpurkar.github.io/SQuAD-explorer/. This data set is in the form of a json file that contains dictionaries with questions and associated answers. This allows the model to recognize various formulations for the same question and seek the appropriate answer. I chose this dataset since the information is easy to access and incorporate into the model. There is also a training dataset over which the model can be trained and evaluated.

Methodology

i. Data Preprocessing
The dataset is comprised of dictionaries with the following information: sample questions and their associated answers, given an article on Wikipedia. I will have to extract the question-answer pairs from the dictionaries and use those to train and test my model.

ii. Machine learning model
First, I will need to train my model to recognize if a given question has an answer by recognizing if it is an existing question. Then, if it is a valid question, the answer will be returned from the question answer pair. The model will then be tested with the testing dataset. To recognize valid questions, I will need to identify common formulations, nouns, verbs, and other characteristics of the English language. I will need to use a classification method that will map the given question to either an existing question or N/A for invalid questions. Then, I will need to find questions that contain similar words to existing questions without having an answer or without having the same answer as the existing question. Finally, for a valid question, I will train the model to return the answer based on the formulation used in the question.

iii. Final conceptualization
Once the model is finalized, it will be integrated into a simple webpage. The user will be able to enter a question into a chatbox and the answer will be returned by the chatbot.