FAQ system

Siddhartha Satpathi

May 23, 2022

Abstract

We would like to build an FAQ system that generates frequently asked questions given a document. The purpose of this project is to allow users to read through a document quickly by going through the FAQs from the document. This would allow the user to understand the contents of the document quickly. After this step, the user can ask specific questions to the document depending on the the user's needs through a question-answering module which would allow the user to get what they want from the document.

1 GPT-3

GPT-3 is a large scale language model that is available to fine-tune for under specific tasks through a user friendly API. In GPT-3 I learnt the following tasks:

1. Few-shot learning: In this approach, we give the model a prompt consisting of few examples and expect it to learn and generate text completion given the examples. This does not change the weights of the model. (https://beta.openai.com/docs/guides/completion)

2. Fine-tuning: In this approach, the GPT-3 model is given a few hundred prompts and text completions as training data which changes the weights of the model. We can thus input a new prompt and the fine-tuned model can provide us with answers. (https://beta.openai.com/docs/guides/fine-tuning)

2 FAQ system using GPT-3

In order for us to build an FAQ system using pre-trained models like GPT-3, there are two approaches we can take -

2.1 Build a fine-tuned model that generates FAQs as the text completion given a document

In this approach we could break the document into several paragraphs. We could input few paragraphs as GPT-3 prompt and expect the GPT-3 module to provide 2-3 questions with answers as output. We can then combine the outputs from these paragraphs to make FAQ for the entire document. The training input for fine-tuning GPT-3 would consist of the prompt as few paragraphs and the 2-3 question-answers as completion. In the GPT-3 documentation for complex generation tasks like that of FAQs it states that about 500 training data points are a good start (see here).

Why I did not take this approach? It felt that in this approach the model needs to generate both questions and answers from a given text. To me this seemed like a difficult task for the model and I thought we might need a lot of training examples for this approach. I tried fine-tuning and it did not provide satisfactory results with about 30 training examples.

2.2 Treat Questions generation and Answers generation as two separate tasts

In this approach we could break the document into several paragraphs. Each paragraph is fed as a prompt to a fine-tuned GPT-3 module and it spits out few important questions from the paragraph.

We can then use a question answering module to generate answers for the above questions. We could combine the results from different paragraphs to make the full set of FAQ.

2.2.1 Question generation module

For this module, I was provided with 14 documents (links provided below in Appendix) to generate the training data. I could only generate training data from two documents. The training data has paragraphs followed by 3-4 questions and answers from the paragraph. There are some short questions answers, but mostly I focused on questions whose answers were long since they closely resemble FAQs. The details of the training data is given in the Appendix.

There are the following styles for question generation.

1. Zero-shot learning - In this approach I planned to ask the pre-trained 'davinci-instruct-beta-v2' model to generate questions. A sample code for that is given in 'Zero-shot-GPT3.ipynb'.

2. Few-shot learning - In this approach, I planned to provide the model 'davinci-instruct-beta-v2' few examples and then generate the questions from it.

3. Fine-tuning - In this approach, I train the 'davinci-instruct-beta-v2' model with prompts as paragraphs and the questions as completions. The dataset can be put in a csv file with prompts and completions as columns. Following that, one can use the tools provided in the GPT-3 website to clean and prepare the dataset for fine-tuning. (see here) The total amount I spent doing experiments was less than $50. The GPT-3 website mentions about 500 examples for generation tasks to work well. Nevertheless, the fine-tuned model with 20-30 examples generated good questions. An example is provided in the Appendix below. It is possible that we can see that some questions are repetitions. One can remove questions which are similar using the following approach - use GPT-3 embeddings to calculate similarity between two questions and remove questions whose similarity score is above a threshold.

2.2.2 Question answering module

There are two approaches for this -

1. Pre-trained QA system - GPT-3 has a question answering endpoint which takes few examples, and given a paragraph, and a question, it generates answers from the paragraph. This is well documented in the following link:

2. Fine-tune ones own QA system - This fits better for our purpose because we do not want short answers as is usually provided by the answers endpoint of GPT-3. The process of fine-tuning GPT-3 for QA is detailed in a tutorial provided by GPT-3 in the Github project - (olympics 1-3 covers the question answering modules).

3 APPENDIX

The data for the above project is provided in https://github.com/spikuflexday/FAQ

3.1 Training document links

Document 1 Document 2 Document 3 Document 4 Document 5 Document 6 Document 7 Document 8 Document 9 Document 10 Document 11 Document 12 Document 13 Document 14

3.2 Data collection

Document one has 23 paragraphs and document two has 15. I have completely gone through document one but partially gone through document two. I copy the paragraphs from the document and paste them in a txt file. Then I write the questions and answers. Each section is separated by /// signs. The beginning or end of paragraphs has more than ten / slashes, the end of questions or answers if they are followed by another question have more than two but less than ten / slashes. Example -

$//$
$Paragraph1$
$//////////////////////$
$Q1$
$///$
$ANS1$
$///$
$Q2$
$///$
$ANS2$
$//$
$Paragraph2$
$//////////////////////$
$Q2$
$///$
$ANS2$
$/////////////////////////////////////$

I wrote a program to read these from the txt file and saved them in a dictionary data type in python. The dictionary has the paragraphs as keys and questions answers as a list of tuples: '$Paragraph1' : [(Q1, ANS1), (Q2, ANS2), (Q3, ANS3)],' Paragraph2' : \ldots$

3.3 Fine-tuned model question generation example

3.4 Prompt -

B. Selecting a Mode of Transportation Use of Special Fares Non-contract carriers sometimes offer restricted or unrestricted coach fares to the general public, which are lower than the government contract fares. In such cases, the lower fare(s) may be used in accordance with procedures described below. They should not be used simply to avoid use of the contract carrier. However, when a non-contract carrier offers a commercial fare lower than its contract fare, the lower fare should be obtained provided the traveler can meet the requirements of the lower fare. One or more of the following conditions, which must be certified on the travel authorization in advance of the trip, must apply if a carrier other than the contract carrier is used for travel within a contract route: • Seating space or the scheduled flight is not available in time to accomplish the purpose of travel, or use of contract service would require the traveler to incur unnecessary overnight lodging costs that would increase the total cost of the trip; or • The contract carrier's flight schedule is inconsistent with those who are required to travel during normal working hours; or • A non-contract carrier offers a lower fare available to the general public, the use of which will result in a lower trip cost to the Government. This determination should be based on a cost comparison to include the combined cost of transportation, lodging, meals, and related expenses; or Note: This exception DOES NOT APPLY if the contract carrier offers a Comparable fare and has seats available at that fare, or if the lower fare offered by a non-contract carrier is restricted to Government and military travelers on official business and may only be purchased with a CBA contractor issued charge card. Rail service is available and such service is cost effective and consistent with mission requirements. Use of Super- Saver Fares A super-saver fare is generally defined as a non-refundable, discounted, airline ticket that must be purchased within 24 hours of making the reservation and that requires a Saturday night stay at the destination. The special fare is usually offered only for a limited time. Depending on the market served, a requirement may exist to purchase the ticket at least 14 days prior to departure. The use of a super-saver fare requires that individual cost comparisons be computed between contract and non-contract carriers. When doing individual cost comparisons, we may assume that the cancellation penalty will not be invoked each time a super-saver fare is used, as the decision on how to compute penalty fees associated with the cancellation of non- contract carriers could be based on overall ED experience. In our judgment, we believe it is reasonable to assume that no more than one out of every two super-saver trips will be modified or canceled. Accordingly, extended trips using non-contract carriers may be approved if the penalty does not exceed 150 percent of the overall trip savings. In order for a super-saver fare to be used, the following must hold true: • the super-saver fare is the lower cost fare; • the fare is available to the general public, and is not a "me too" fare; the "me too" fare is a companion type fare. For example: the airlines are offering a special promotion where one traveler purchases a ticket for a certain price and the second traveler purchases a ticket for a penny; • use of the fare will lower the overall trip cost; and • the penalty the non-contract carrier may impose will not exceed 150 percent of the expected savings. It is very important that cancellation of super-saver fares be kept to a minimum if savings are to be realized. The EXO must approve requests of this nature and maintain records for review by OCFO. These records must show the cost comparison and reflect penalty fees incurred. Because of various legal considerations, the traveler must initiate requests for extended stay to take advantage of super-saver fares. Under these circumstances, it is assumed that no overtime pay or other compensation will be paid to the employee, except for additional per diem when travel dates are extended to achieve overall trip savings.

3.5 Completion

What are the conditions for using special fares?
Which carrier should be used for travel within a contract route?
What are the conditions for using super saver fares?