The Scot Bot: Question and Answer AI using Limited Keyword Approach

Wyatt Mayor

Supervised by Dr. Logan Mayfield and Dr. Robert Utterback

Monmouth College®

Department of Mathematics, statistics, and Computer Science

Abstract

A recent explosion in AI has brought the discipline to the forefront of computer science. This technology is used to solve an amplitude of problems such as generation, data analysis, and much more! The capabilities of generating and data analysis were utilized in this project to create a question-and-answer AI bot called “The Scot Bot”. The Scot Bot uses a combination of previous solutions and a novel solution to produce an AI that can answer questions based on a small dataset of syllabi. Throughout this poster, the multiple parts that coherently work together will be broken down and explained in detail.

The Scot Bot

If you have ever taken a college course… you know how important syllabi are to your success in that course. There is often a lot of information about the course that is necessary to know. One problem that always seems to resurface is disorganization. This usually results in lost syllabi. So, a simple solution is to put the syllabi all in one spot. However, there is still a inconvenience factor with syllabi often ranging from 1-10 pages worth of information. The Scot Bot’s purpose is to diminish syllabi overload and

Figure 1: Scot the Bot

alleviate the inconvenience of searching through pages of information. The Scot Bot is a bot that will answer questions that students have about syllabi. This bot uses AI technology, specifically a variant of ChatGPT, to answer questions from a dataset of syllabi. The Scot Bot utilizes a combination of Zero-Shot Learning and a Limited Keyword method to limit the dataset and provide concise and accurate responses. A feat of The Scot Bot is its generality. It can be implemented with ease for other universities and colleges.

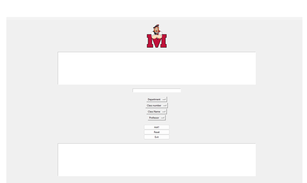

Figure 2: The Scot Bot UI

Data Problem

A common limitation in machine learning is obtaining, cleaning, and preparing large amounts of data that are required for high accuracy. Models that are specifically tasked with open-domain question-and-answer, usually average training datasets with around 90,000 data points (text snippets). When working with syllabi there are only so many syllabi that are created for the specific school. The Scot Bot alleviates the data problem using a hybrid approach of zero-shot learning and a limited keyword method.

Zero-Shot Learning

Zero-shot learning is a machine learning approach that uses previously trained models to achieve problems it hasn’t seen before. Traditionally, models require large amounts of labeled data which is hard to obtain. Zero-shot uses semantics and knowledge transfer that is not given during the training process. Zero-shot learning is a machine learning approach that uses previously trained models to achieve problems it hasn’t seen before. Traditionally, models require large amounts of labeled data which is hard to obtain. Zero-shot uses semantics and knowledge transfer that is not given during the training process.

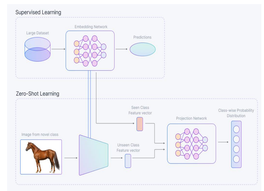

Figure 4: Zero-Shot learning diagram

Large Language Model

Large Language Models are machine learning models that are trained on terabytes of text that come from Wikipedia and other internet sources. During the training of an LLM, a function removes words from a paragraph and the model is tasked with filling in the blanks. This type of training results in impressive generation abilities. The Scot bot utilizes this generation ability to create responses to the user’s questions.

Figure 3: Large Language Model diagram

Limited Keyword Method

The inability to fine-tune a model to my specific dataset due to size and the limitations of the max input size for Large Language Models, a method was required to narrow the information in the syllabi. The limited keyword approach is a method I created that utilizes the assumption that a user’s question will provide keywords to rank passages that contain the answer. This method splits a question into each a list of words and then removes common English words such as “we, our, and, the, etc.”. The keywords produced are then paired with common endings which are used to expand the keyword list. Finally, this method ranks passages based on the number of present keywords. This method helps limit the scope of data that is being filtered and sent through the machine learning model. Through user testing, this approach results in the highest accuracy responses to the user.

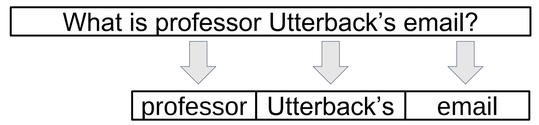

Figure 5: Limited Keyword diagram

Conclusion

At the beginning of this project, three main goals were set for this project which were generality, implementing machine learning with a limited dataset, and answering questions to users about syllabi with relatively high accuracy. The current state of the software achieves the second and third goals by utilizing a simplistic approach of keyword extraction and a Zero-shot Learning approach. This combination allows for a small dataset while also producing accurate responses. The first goal was also achieved by making the project modular. The modular aspect of this project allows a user to only provide a syllabi folder and a logo to convert to another university or college. With the accomplishment of the three main goals, The Scot Bot was a success!

References

“Introduction to Large Language Models.” *Cohere AI*, Cohere, https://docs.cohere.ai/docs/introduction-to-large-language-models.

Kundu, Rohit. “What Is Zero Shot Learning in Image Classification? [Examples].” *What Is Zero Shot Learning in Image Classification? [Examples]*, V7, 2 Mar. 2023, https://www.v7labs.com/blog/zero-shot-learning-guide.