QuizBot: An Automated Retrieval-Based Conversation System for Machine Learning Mastery

Introduction

Several nontraditional education programs have emerged to increase the supply of machine learning talent by providing less expensive options for people interested in developing machine learning and data science skills, particularly in comparison to traditional computer science degree programs. However, due to the short duration of most nontraditional training programs, thorough mastery of certain topics (e.g., systems design, data structures and algorithms, etc.) will need to occur outside of the classroom. Therefore, an educational retrieval-based chatbot will be developed in order to facilitate the mastery of key topics needed to pass technical interviews and successfully begin a career in data science.

Methods

A dataset consisting of 168 machine learning and data science questions and answers was created. Questions were categorized as general data science, data structures and algorithms, and SQL and database design. Word clouds were generated to visualize frequently used words using the standard word cloud function in the python word cloud library. The following 8 word clouds were created: (1) all questions; (2) all answers; (3) general data science questions; (4) data structures and algorithms questions; (5) SQL and database design questions; (6) general data science answers; (7) data structures and algorithms answers; (8) SQL and database design answers.

In order to develop the minimal viable product, the question and answer dataset was simplified and converted to a text document consisting of the 60 general data science answers from the original dataset. The text in the chatbot document was tokenized using the sent_tokenize function from the python Natural Language Toolkit (NLTK) library. The resulting sentences were subsequently tokenized using the word_tokenize function from the python NLTK library. Following tokenization, all tokens were lemmatized using the NLTK WordNet lemmatizer. The WordNet lemmatizer algorithmically searches through the WordNet database to identify the morphological base form of a word. This will allow users to utilize any version of a base word and get more accurate and relevant results. Lastly, the text was normalized (i.e., converted to lower case and punctuation was removed) prior to TF-IDF vectorization and training.

The TfidfVectorizer method from the scikit-learn library was used to convert the corpus and user's answer into feature vectors. It was initialized with a custom tokenizer and the stop words parameter was set to English in order to exclude words that have very low semantic value or information content. The Term Frequency–Inverse Document Frequency (TF-IDF) vectorizer was fitted on the sentence-tokenized version of the text and IDF values were evaluated utilizing the fit method. In addition, features are mapped to IDF values and the transform method returned a sparse 2D TF-IDF feature document matrix.

The cosine similarity function from the sci-kit learn library was used to find the similarities between the user's answer and the corpus. The cosine similarity function measures the cosine of the angle between two vectors [i.e., the angle can be between 0° to 90°, cos 0° = 1 (vectors are identical) and cos 90° = 0 (vectors are dissimilar)]. The argsort function is used to sort the feature vectors by their respective cosine similarity values (i.e., from 0 to 1) and negative array slicing is then used to find the text from the

corpus that match the user answer (i.e., the most related feature vectors will be at the end of the sorted array since they will have highest cosine similarity values; the sentence with the second highest cosine similarity value will be selected since the user's answer will have the highest cosine similarity value of 1). Of note, the similarity comparisons are flattened so that it returns a float (i.e., the cosine similarity value) and not a 2D array. If the cosine similarity value was 0, then a matching vector does not exist in the corpus and the chatbot will indicate a lack of understanding in its reply to the user. If the cosine similarity is not equal to 0, the chatbot will return the matching sentence (i.e., the sentence at index -2). Finally, a loop is created which represents a study session with the chatbot. If the user's answer is not "bye", then it's appended to the sentence corpus, tokenized, vectorized, weighted by TF-IDF, and compared to the corpus using the cosine similarity function. The matched sentence is returned as the bot's answer. The loop will continue executing until the user enters the word "bye".

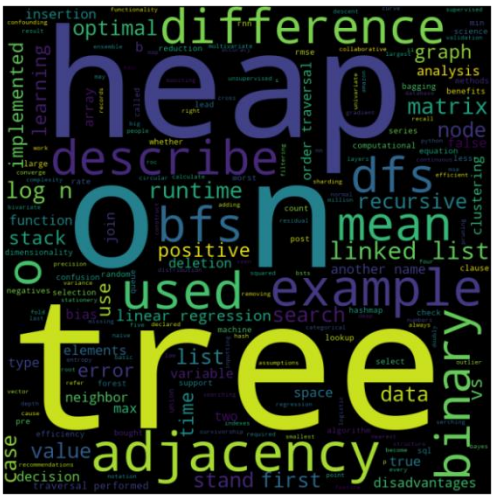
Results

Word Frequency Analysis

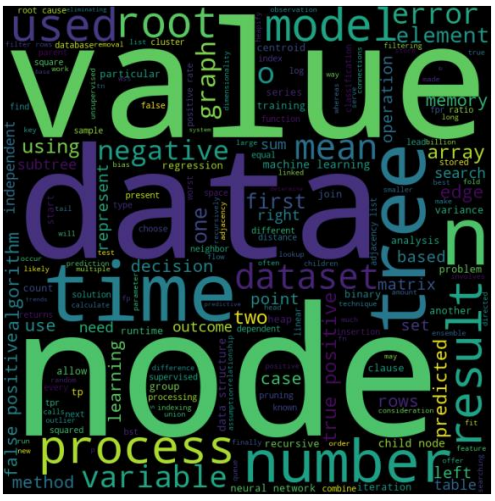
A visual representation of the frequency of words found in the questions and answers of the original dataset was used to determine the most frequently occurring concepts covered in the original dataset (**Figure 1**). Based on this analysis, questions regarding tree-based data structures (e.g., heap) and big O notation tend to occur more frequently. This is also reflected in the answers with words such as "node", "tree", "root", and "time" occurring more frequently than other pertinent terms or concepts.

Figure 1. Word Cloud Based on (A) Questions and (B) Answers in the ML Education QA Dataset

(A) Questions



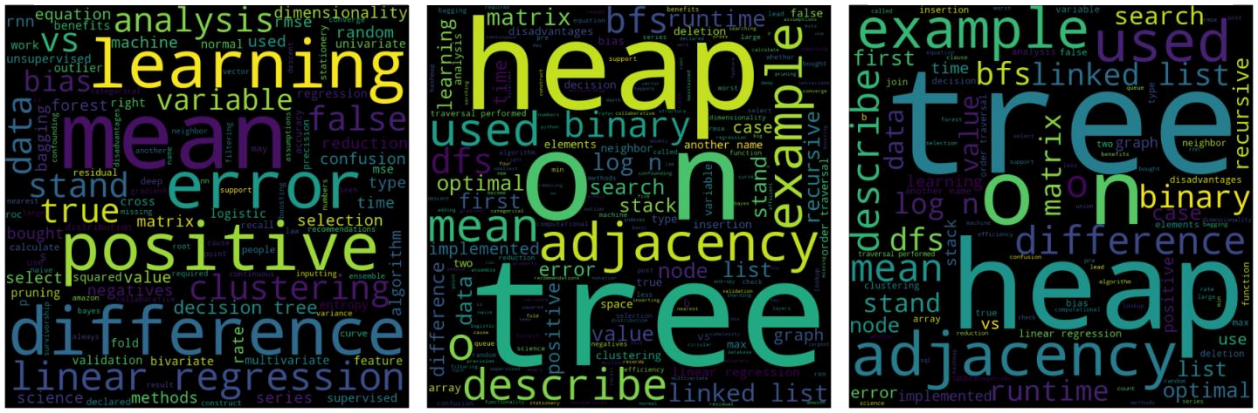
(B) Answers



A word cloud visualization of the questions by category was also generated (**Figure 2**). Based on the visualization, questions in the general data science category pertaining to supervised and unsupervised learning algorithms (i.e., linear regression, logistic regression, decision trees, and clustering), error estimates for linear regression models, and basic concepts (e.g., bias, confusion matrices, and dimensionality reduction) occurred more frequently. Of all questions, 60% of questions fall in the data structures and algorithms category. Therefore, most of the concepts in this category (e.g., tree-based data structures and big O notation) are predominantly represented in the word cloud in Figure 1A. Lastly, only 6% of questions fell into the SQL and database design category. As a result, the word cloud that was generated for questions and answers is not representative of the concepts actually covered in this category.

Figure 2. Word Cloud Based on (A) Data Science, (B) Data Structures, and (C) SQL Questions

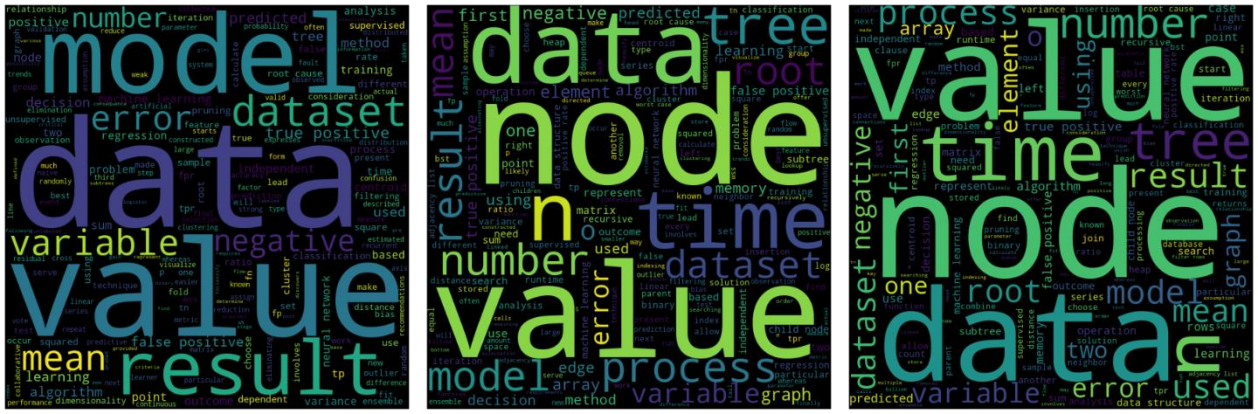
(A) General Data Science **(B) Data Structures & Algorithms** **(C) SQL & Database Design**



The concepts present in the questions are also represented in the answers (**Figure 3**). Of note, in the general data science category, concepts such as “true positive” and “false positive” indicate explanations of how models are assessed in the questions pertaining to confusion matrices.

Figure 3. Word Cloud Based on (A) Data Science, (B) Data Structures, and (C) SQL Answers

(A) General Data Science **(B) Data Structures & Algorithms** **(C) SQL & Database Design**



QuizBot Study Session

Of the 168 questions and answers in the original dataset, the corpus only included text from the 60 general data science answers. The study session begins by welcoming the user and displaying instructions for ending the session (**Figure 4**). The user enters "start" and the chatbot responds with a data science question. In this case, the first question is "What is logistic regression?". The user enters what they think the correct answer is and the chatbot returns the correct answer to the question (i.e., the most relevant response based on its cosine similarity). The user can then check to see if their answer was correct and assess how close they were to the correct answer. The user then enters "next" and the chatbot responds with another data science question. In this case, the question is "What is one way to select K for K-means?". The user enters their answer and the chatbot returns the correct answer. The user enters "next" again and third question is "What is collaborative filtering?". The user seems confused and enters a wrong answer (i.e., a truncated form of the definition of content filtering). However, the answer was close enough to the actual definition (i.e., most likely due to the term "user") and the chatbot was able to retrieve and respond with the correct answer. This allows the user to see that they were incorrect.

The user enters "next" again and the fourth questions is "What is entropy?". The user has a background in chemistry and seems to get confused and enters the chemistry-based definition instead of the data science definition. Once again the chatbot is able to retrieve and respond with the correct answer, likely due to the inclusion of the word "entropy" in the user's answer. The user enters "next" and we get the "What is collaborative filtering?" again in the same session so the user enters "next" again to move on. The sixth question is "What is one way to select K for K-means?" and the user decides to be funny and enter "Eenie Meenie Miney Mo." In this case, the correct answer refers to the elbow method. Since the answer is completely outside of the realm of what could be correct and clearly none of these words will be found in the corpus, the cosine similarity between the user's answer and text in the corpus is 0 and the chatbot responds with "I am sorry! I don't understand you!".

Figure 4: Example of General Data Science Study Session

```
QuizBot: Welcome! If you want to end the study session, just type bye!
start
QuizBot: What is logistic regression?
Logistic regression is a statistical model used to find the relationships that exist between a dependent binary variable and
one or more independent variables.
QuizBot: Logistic regression is a statistical model used to find the relationships that exist between a dependent binary var
iable and one or more independent variables.
next
QuizBot: What is one way to select K for K-means?
The most popular method for selecting k for the k-means algorithm is using the elbow method.
QuizBot: The most popular method for selecting k for the k-means algorithm is using the elbow method.
next
QuizBot: What is collaborative filtering?
Based on user activity
QuizBot: Collaborative filtering is a form of content filtering that uses similarities between different users to make recom
mendations.
next
QuizBot: What is entropy?
Entropy is a chemical term, measure of disorder.
QuizBot: Entropy is a measure of the level of uncertainty or impurity that's present in a dataset.
next
QuizBot: What is collaborative filtering?
next
QuizBot: What is one way to select K for K-means?
Eenie Meenie Miney Mo
QuizBot: I am sorry! I don't understand you!
bye
QuizBot: Bye! See you later!
```

Conclusion

In conclusion, an educational chatbot was created to assist in the mastery of machine learning concepts. In the future, the original dataset consisting of all 3 categories of questions and answers (i.e., general data science, data structures and algorithms, and SQL and database design) will be used to develop the chatbot. In addition, frequently occurring words (e.g., data) will be removed from the dataset prior to conducting the word cloud frequency analysis in order to get a more accurate idea of the frequency of the most important concepts covered. Lastly, the chatbot will be converted to a voice chatbot in the future and deployed using flask.