

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

Background

- ▶ 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

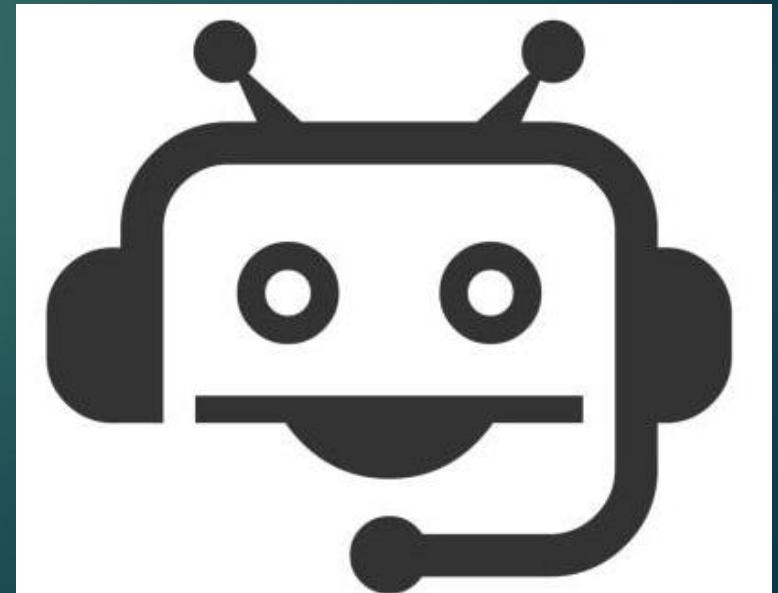
Problem Identification

- ▶ The task of quickly developing proficiency in the topics needed to successfully enter the data science and machine learning field can be daunting. As such, an educational chatbot was created to facilitate the mastery of the most important data science and machine learning concepts typically covered in technical interviews



“Are you ready
to study?”
“What is collaborative
filtering?”

QuizBot



Word Cloud Methodology

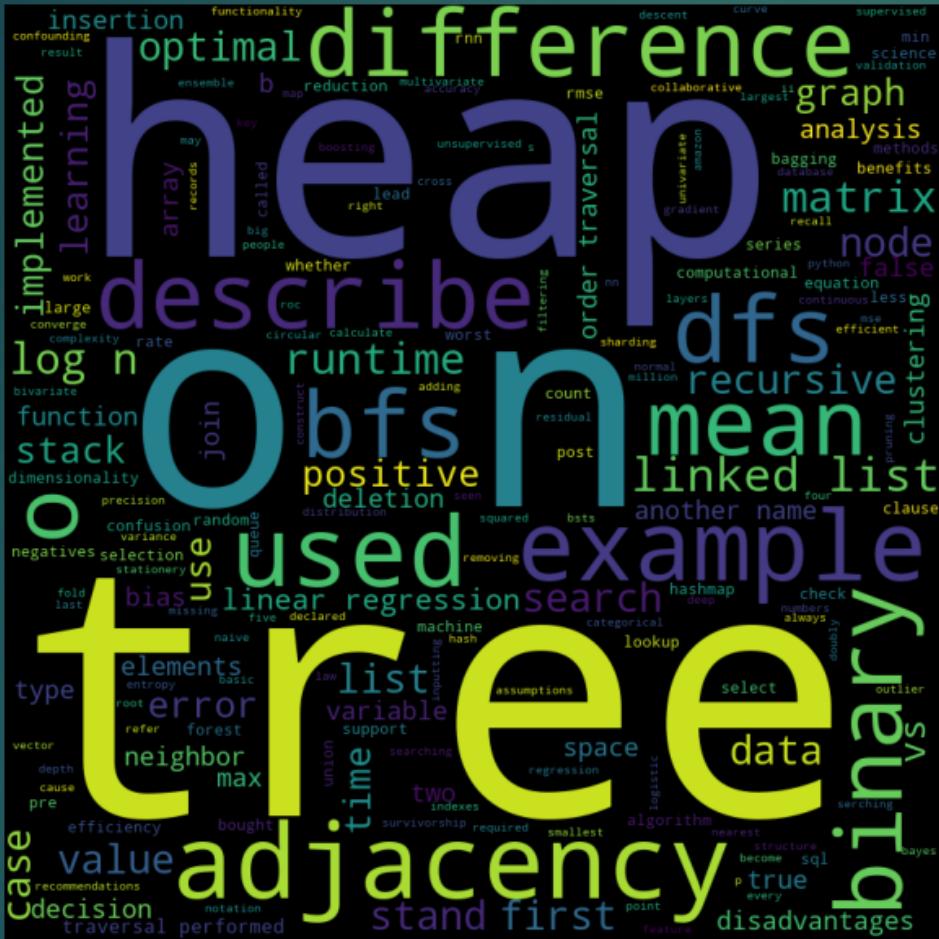
- ▶ A dataset consisting of 168 machine learning and data science questions and answers was created
- ▶ Questions were categorized as general data science, data structures and algorithms, and SQL and database design
- ▶ The following 8 word clouds were created:
 - ▶ All questions
 - ▶ All answers
 - ▶ General data science questions
 - ▶ Data structures and algorithms questions
 - ▶ SQL and database design questions
 - ▶ General data science answers
 - ▶ Data structures and algorithms answers
 - ▶ SQL and database design answers

Chatbot Study Session Methodology

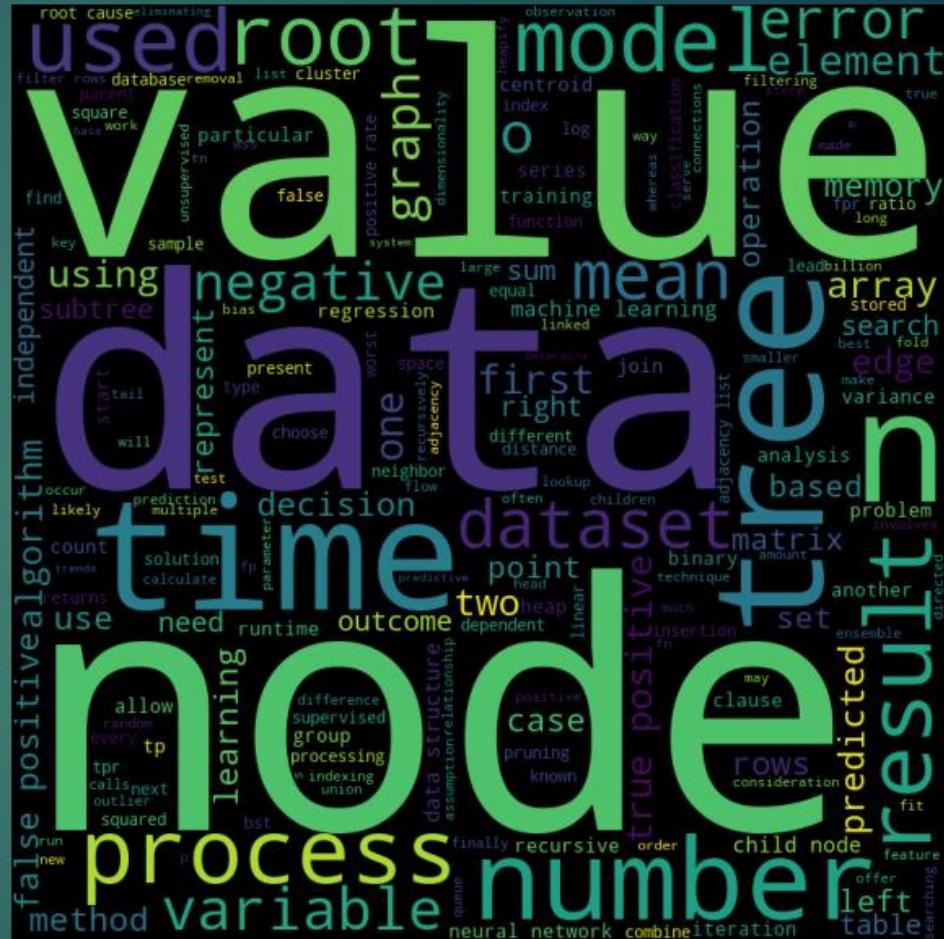
- ▶ In order to develop the minimal viable product, the question and answer dataset was simplified and converted to a text document consisting of the 60 general data science answers from the original dataset
- ▶ The corpus and the user's answers were preprocessed (i.e., tokenized, vectorized, weighted by TF-IDF) and compared using the cosine similarity function
- ▶ Ultimately, the chatbot returns the correct answer [i.e., matched text (text with second highest cosine similarity value)] is returned as the bot's answer
- ▶ The study session continues until the user enters the word “bye”

Word Frequency Analysis: All Questions & Answers

All Questions



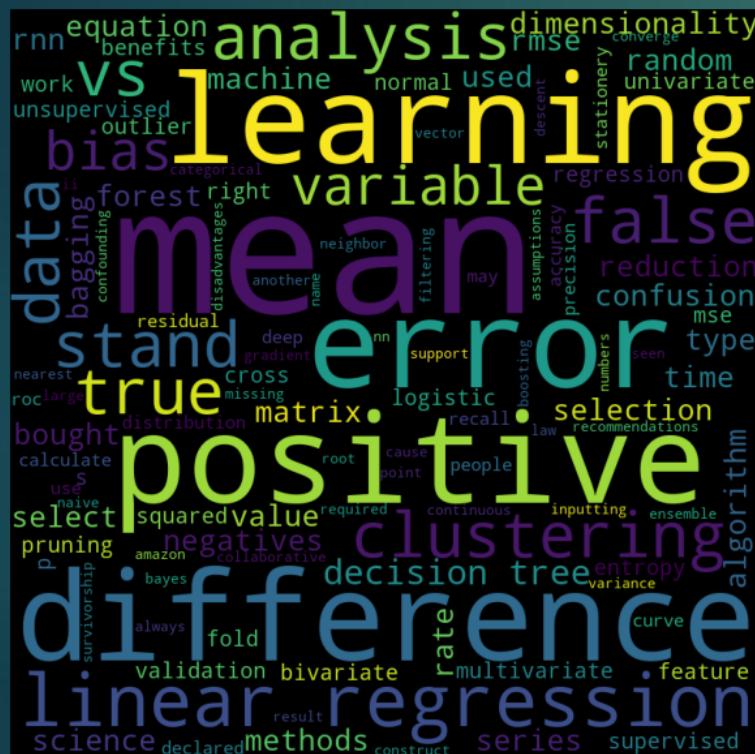
All Answers



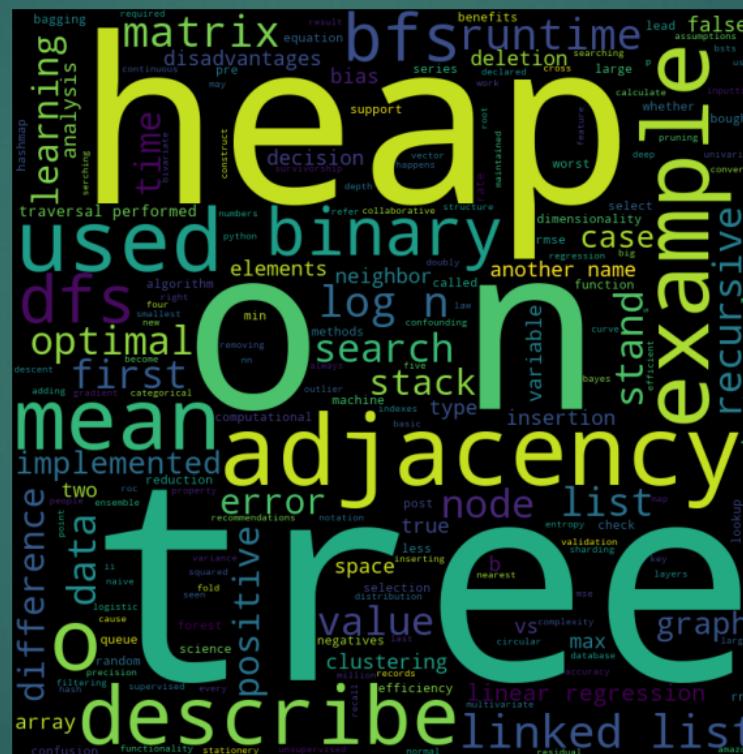
Based on this analysis, questions regarding tree-based data structures (e.g., heap) and big O notation tend to occur more frequently; this is also reflected in the answers with words such as “node”, “tree”, “root”, and “time” occurring more frequently than other pertinent terms or concepts

Word Frequency Analysis: Questions By Category

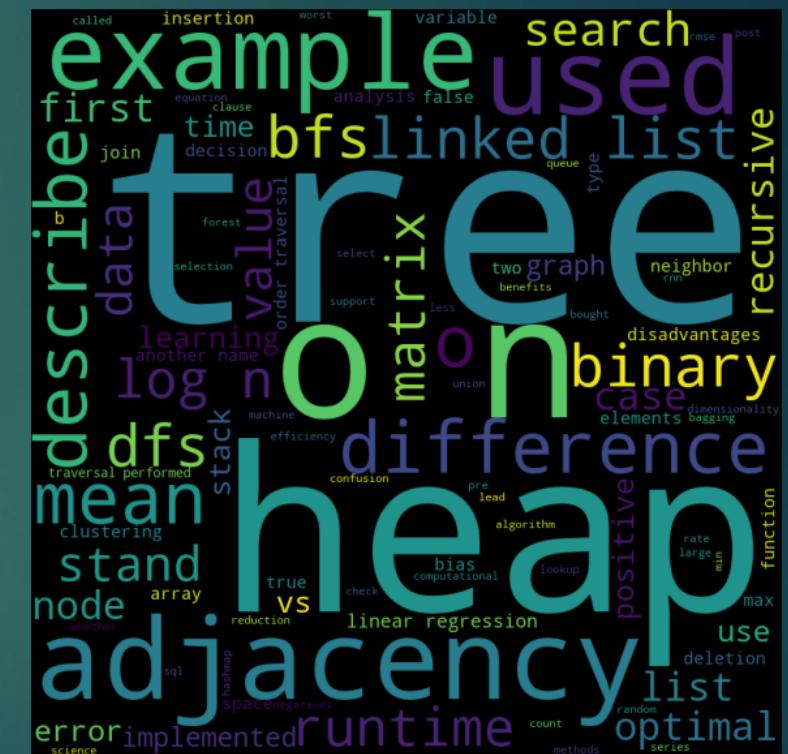
General Data Science Questions



Data Structures & Algorithms Questions



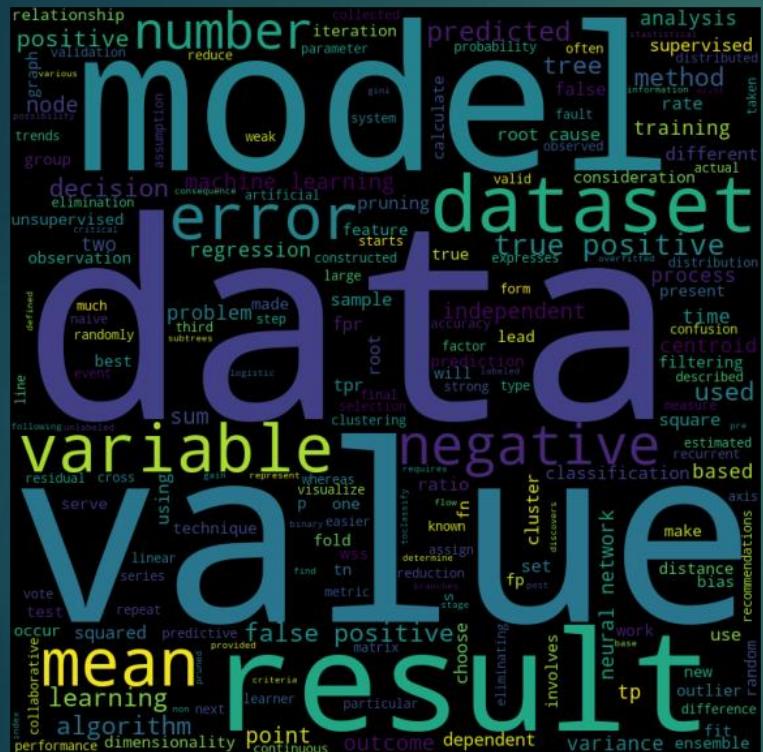
SQL & Database Design Questions



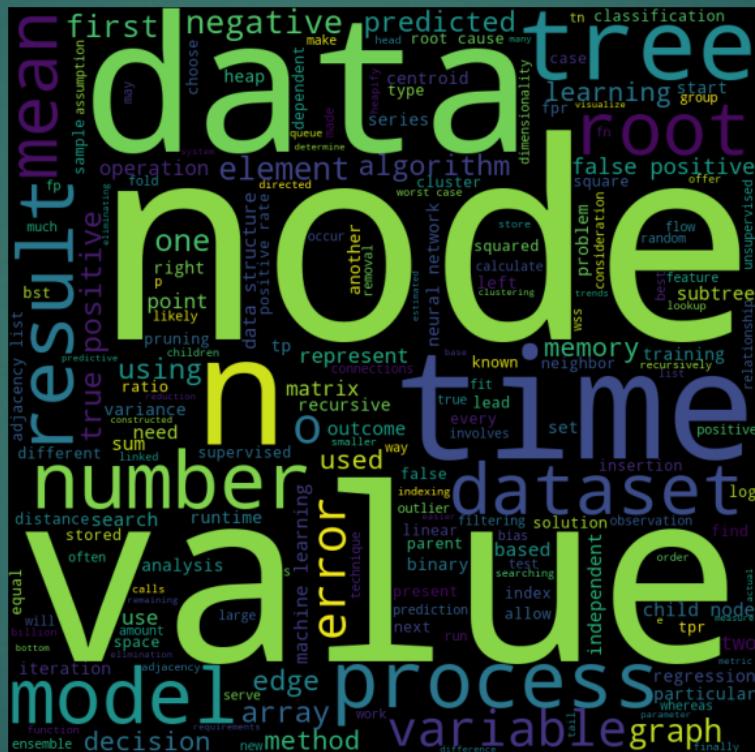
Frequently Occurring Concepts: General Data Science - supervised and unsupervised learning algorithms, error estimates, and basic concepts (e.g., bias, confusion matrices, and dimensionality reduction); DSA - tree-based data structures and big O notation; SQL - NR

Word Frequency Analysis: Answers By Category

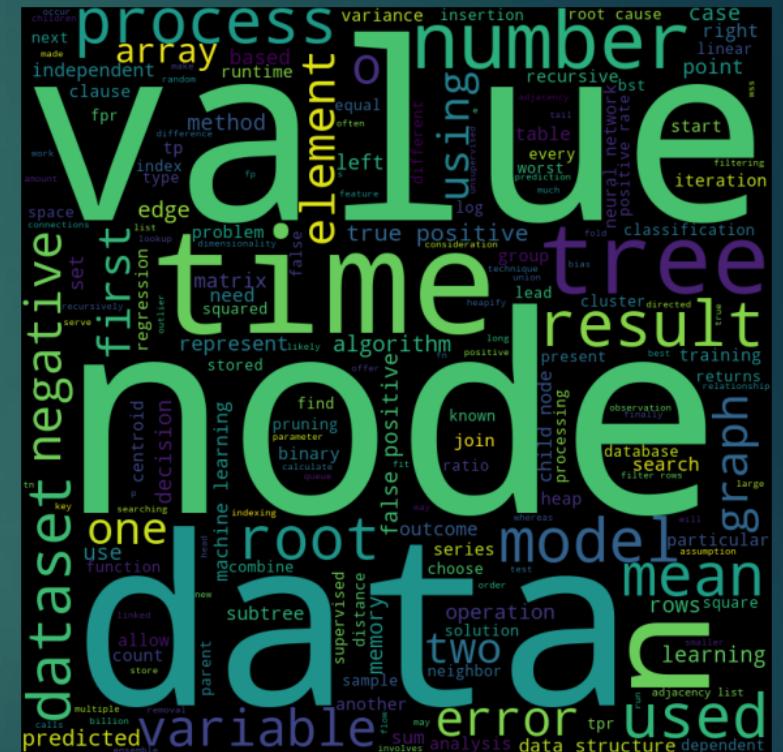
General Data Science Answers



Data Structures & Algorithms Answers



SQL & Database Design Answers



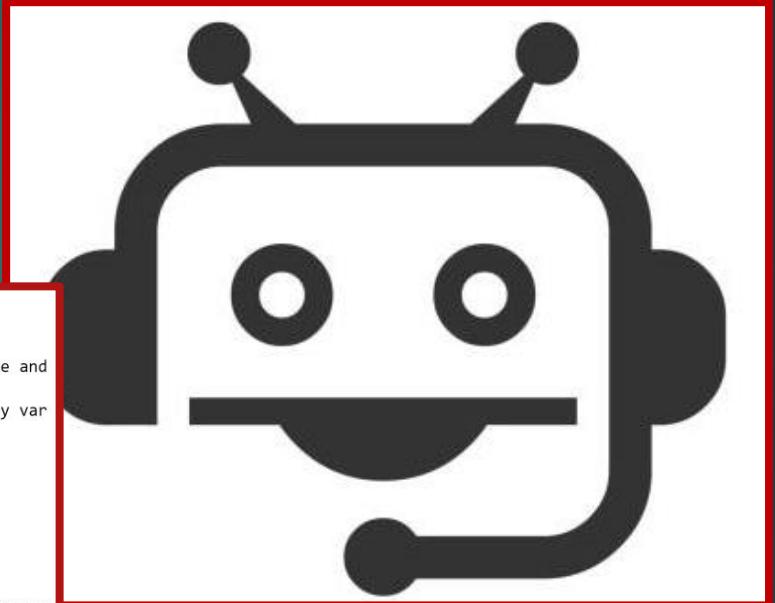
The concepts present in the questions are also represented in the answers; concepts such as “true positive” and “false positive” indicate explanations of how models are assessed in the questions pertaining to confusion matrices in the general data science category

QuizBot Study Session Example



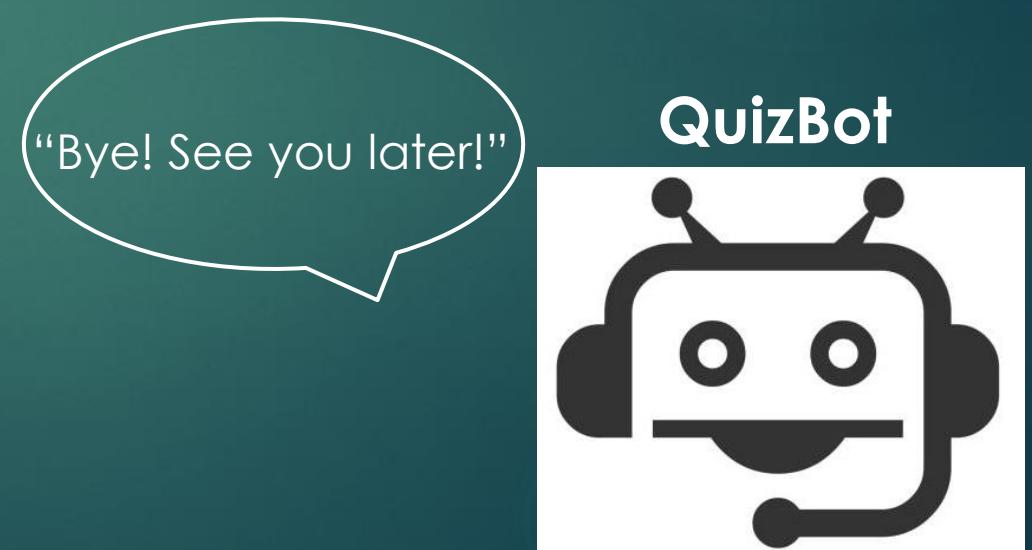
```
QuizBot: Welcome! If you want to end the study session, just type bye!
start
QuizBot: What is logistic regression?
Logistic regression is a statistical model used to find the relationships that exist between a dependent binary variable and one or more independent variables.
QuizBot: Logistic regression is a statistical model used to find the relationships that exist between a dependent binary variable 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 recommendations.
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!
```

QuizBot



Conclusion

- ▶ An educational chatbot was created to assist in the mastery of data science and 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
 - ▶ 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
 - ▶ The chatbot will be converted to a voice chatbot in the future and deployed using flask





Thank You!