Project Details

Contents

- Problem Statement
- Objectives
- Database Schema
- Dataset Information
- Flow Diagram for Objective 1
- Algorithm Selection
- Flow Diagram for Objective 2
- Technology Stack
- Conclusion

Problem Statement

Design and develop a QnA web platform with Natural Language Processing Capabilities for Automated Categorization and Enhanced User Experience.

Objectives

- Enable user registration and account creation for personalized profiles.
- Implement question and doubt posting and sharing functionalities.

Objectives

- Utilize Natural Language Processing (NLP) to analyze and identify frequently asked keywords.
- Develop NLP and Machine Learning algorithms for automatic question categorization.

Database Schema

User:

- username: CharField (unique=True)
- password: CharField
- email: EmailField

Database Schema

Question:

- author: ForeignKey to User
- title: CharField (max_length=200)
- description: TextField
- created_at: DateTimeField
- updated_at: DateTimeField
- category_id: IntegerField

Database Schema

Answer:

- answer_text: TextField
- for_question: ForeignKey to Question
- given_by: ForeignKey to User
- given_at: DateField
- updated_at: DateTimeField

Dataset Information

Total Samples: 120,000 rows

Classes:

World, Sports, Business, Sci/Tech

Columns:

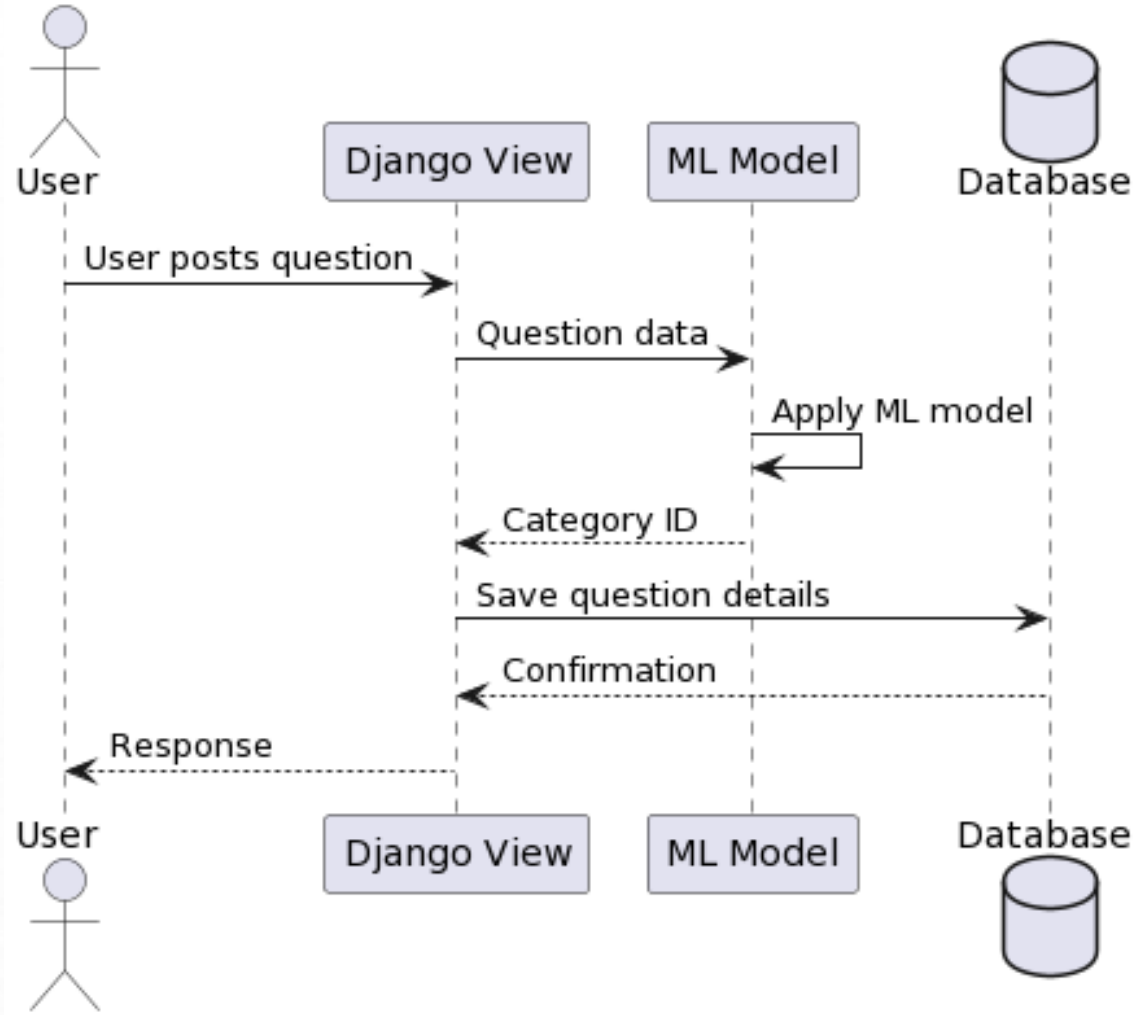
Class Index: Consists of class IDs ranging from 1 to 4.

Title: Contains the title of the news articles.

Description: Contains the description of the news articles.

Click to get more information about dataset : link

Category Prediction Flow


User
Django View
ML Model
Database
User posts question
Question data
Apply ML model
Category ID
Save question details
Confirmation
Response
User
Django View
ML Model
Database

Algorithm Selection

Tested Algorithms (Notebook Link):

Support Vector Classifier (SVC)

Random Forest Classifier

Accuracy Results:

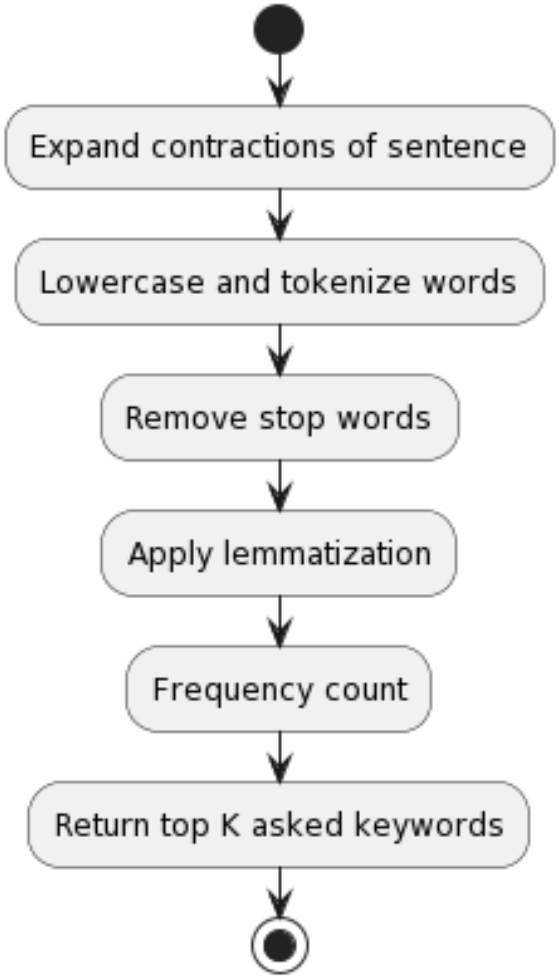
SVC: 91%

Random Forest Classifier: 79%

Chosen Algorithm:

SVC (Support Vector Classifier) for its higher accuracy of 91%.

Keyword Extraction Flow


Expand contractions of sentence
Lowercase and tokenize words
Remove stop words
Apply lemmatization
Frequency count
Return top K asked keywords

Technology Stack

Conclusion

Project aims to create a user-friendly Q&A website, integrating NLP for automatic categorization and keyword analysis. I anticipate choosing the Support Vector Classifier (SVC) over Random Forest, given its higher accuracy (91% vs. 79%). This decision highlights the importance of algorithm selection. Overall, project showcases the effective use of NLP and ML to enhance user experience and knowledge management.


thank you