Assumptions Made:

- **Data Preprocessing**: It is assumed that the initial dataset containing questions required cleaning and preprocessing. Initially, there were no null values, but after preprocessing, 35 null values had to be removed. Additionally, special characters were treated separately to eliminate the characters without removing the attached words. For example, to retain the word "NCF" while removing the question mark from "What is NCF?". Furthermore, white spaces (empty characters) were removed separately to prevent the preprocessing model from considering empty spaces as significant characters.
- **Data Supervision**: There are two potential approaches for handling the data: manual supervision or treating it as unsupervised.
- **Categorization**: The goal is to find keywords in the questions and then group them into categories.
- **Visualization**: Word clouds have been utilized for data visualization, which aids in identifying the most frequently occurring words in the dataset.

Data Cleaning:

The dataset required thorough cleaning and preprocessing. Initially, it was free from null values; however, during the preprocessing phase, 35 null values were identified and subsequently removed. Special characters were handled separately to preserve words while eliminating unwanted characters, ensuring that, for instance, "NCF" would not lose its context when removing the question mark from "What is NCF?". Additionally, white spaces (empty characters) were removed independently to prevent them from being treated as meaningful characters by the preprocessing model.

Exploratory Data Analysis (EDA):

To gain insight into the dataset, a word cloud was generated from the processed question column. This visualization technique

highlights words with the highest frequencies, offering valuable direction for the categorization process.

Models:

Three models were employed, including two unsupervised models and one supervised model.

- Unsupervised Models:
    - LDA2vec with Gensim and pyLDAvis: This model utilizes LDA2vec, which is a combination of Latent Dirichlet Allocation (LDA) and word embeddings, and it was evaluated using pyLDAvis.
    - LDA with TF-IDF: Latent Dirichlet Allocation was applied using Term Frequency-Inverse Document Frequency (TF-IDF) representation.
- Supervised Model:
    - c. Multinomial Naïve Bayes: A supervised model that classifies questions into categories.

For the unsupervised models, coherence scores were calculated. Coherence measures the semantic similarity between high-scoring words within a topic, providing an indicator of topic quality.

LDA2vec Coherence Score: 0.6342565171493334

LDA Coherence Score: 0.6613569332680389

- Using LDA I generated 5 topics. The best coherence score was achieved when number of topics were 5.
- After going through the key words present in each topic and assessing pyLDAvis visualization, I assigned them names, as follows:
    - LDA2vec:
        - Child Health Concerns
        - Learning Concerns
        - Gov. Schemes Info.
        - Skill Development
        - General knowledge

- LDA with TF-IDF
    - Early Childhood Concerns
    - Health Concerns
    - Gov. Schemes Info.
    - Learning and Skill Development
    - General Knowledge

  The difference in nomenclature is due to difference in key words generated by each model.

I've attached python notebook, where you can see the pyLDAvis interactive visualization for the models.