Task-3:

The task requires using the dataset provided in this link - https://www.kaggle.com/rajanand/rajyasabha. This dataset contains information regarding the questions asked in the Rajya Sabha. The task is to **predict the ministry**, provided with all the other details. You are expected to provide different approaches that would be feasible for this problem. We are expecting you to justify your choice for the model, hyper-parameters, etc. It is strongly encouraged to support your arguments through exploratory analysis of the data.

A few points to note:

1. The ministry in question appears in the initial part of the answer for every question. But, this part of the answer will be stripped off, during testing of the idea.
2. Please do not suggest finding the ministry using the name of the Minister/Ministers in charge.

<< Add your solutions here >>

As thlis is a multiclass classification problem(predicting the ministry), a problem of this type can be solved in various methods like neural networks(single or multi-layer perceptron), Decision trees,k-nearest neighbors, Naive Bayes , SVM etc. But I will take the neural network approach here as the dataset is large(and neural networks perform well in large datasets).

The main part of this process will be pre-processing the data than building the neural network.First we have to know number of inputs tokens into the neural network. So we combine all questions and answers into a single text file,remove stop words , stem the words , remove all special characters using the nltk library and common string

library. The size of this resulting words will be the size of the input layer. The input layer is one-hot encoded, ie given a sentence, the presence of a word is represented by 1 in the input layer and 0 if it is not. The size of the neural network will be determined by trial and error (Too less hidden layers may lead to underfitting whereas too many layers may lead to overfitting of the data) and number of neurons in the output layer will be equal to the number of ministries. Activation functions like sigmoid or tanh maybe used and softmax maybe used for the last output layer.The neuron having the highest probabilty will determine the ministry (ie if the fifth neuron represnts the finance ministry and has the highest probability ,then the answer will be finance ministry).
Training of the neural network maybe done using Adam Optimizer,or by using mini batch gradient descent.Tensorflow and keras provide easy implementation in doing so.