Features:

1) Length of the question

2) Lexical Feature

3) Semantic feature

For Lexical features, first make a vocabulary of the entire dataset, and find the 500 most frequent 1-gram. Then among those 500 most frequent 1-gram, find which are present for each training instance ‘training[i]’.

Perform POS tagging, and perform similar types of operations.

So, the dataset, after appending the features ***might*** look like-

Class	Question	Length	Lexical	Syntactic
HUM	Who was the inventor of silly putty ?	7	Who, was, inventor	Pronoun, Verb
DESC	What is the history of skateboarding ?	6	What, is the, history	Pronoun, Noun, Verb

“Who, was, inventor, What, is, the “ are among the Most-frequent-500 1-gram words.

“Pronoun, Verb, Noun”

- ***this can be a way to create the data, for building the Decision Tree model***

Now, let us say, we are currently at a decision node, where we have to separate the data on the basis of its values into different child nodes.

We can use some discretization methods or some other methods discussed in class when we are using the Length feature to segregate the data.

Next we have to do it for the case of Lexical feature or Syntactic feature. As multiple words could be present, so we can use the Naïve Bayes type of calculation? Meaning, we first compute the probability of each word in 1-gram method, conditional

probability in case of 2-gram or 3-gram methods. We will then multiply them accordingly. Following that, we will use the probability score to perform splitting.

This understanding can be followed to work on Assignment-4. You all can approach the problem in this manner.