Q2

To get a training data, I took get a large chunk of text (a novel), divided it in sentences and found out which sentences have a proper noun or numeric data. Based on this, I categorically divided the sentences and obtained the training dataset.

To find which sentences have proper noun and numeric data, I used pos\_tag function of nltk.

The algorithm –

data={'Sentence':'Category'}

for i in doc.split('\n'): [\\i](file:///\\i) contains the paragraphs

  for j in sent\_tokenize(i): [\\j](file:///\\j) contains the sentences

    j=j.strip()

    sent\_arr=pos\_tag(j.split()) [\\sent\_array](file:///\\sent_array) contains words with tags

    if(len(sent\_arr)>3):

      data[j]=0

      for k in range(len(sent\_arr)):

        if(sent\_arr[k][1]=='NNP' or sent\_arr[k][1]=='CD'):

          data[j]=1

          break