EXPOSIS DATA LABS

01

Question Type Identification

NLP Assignment

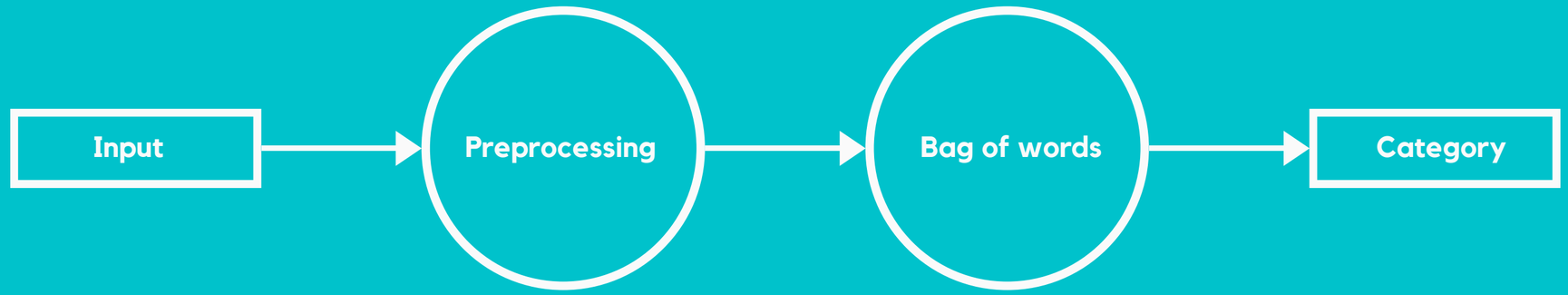
Submitted By- Hritik Kumar

02

Objective

Our aim was to identify the question type whether the question is of type and the four categories to handle for this assignment are: Who, What, When, Affirmation(yes/no).

03



Flow Diagram

Preprocessing

04

Remove Punctuation

One way of doing this is by looping through the Series with list comprehension and keeping everything that is not in string.punctuation, a list of all punctuation we imported at the beginning with import string.

Remove Apostrophe

In writing, an apostrophe is used to indicate the place of missing letters.

Remove Numbers

Since we are dealing with text, so the number might not add much information to text processing. So, numbers can be removed from the text. We can use regular expressions (regex) to get rid of numbers or isdigit() also can be used.

Remove Stopwords

We usually want to remove these because they have low predictive power. There are occasions when you may want to keep them though. Such as, if your corpus is very small and removing stop words would decrease the total number of words by a large percent.

Stemming

Stemming is the process of reducing inflected/derived words to their word stem, base, or root form. The stem need not be identical to the original word.

Bag of words

05

- The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). A bag-of-words model, or BoW for short, is a way of extracting features from the text for use in modeling.
- A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:
  - A vocabulary of known words.
  - A measure of the presence of known words.

Classification

06

Since we have a bag of words with normalized frequencies, we can classify the questions. For classification, we have to perform data preprocessing on that question which we have done earlier. After that, we will match the respective word from the question and measure the score so that we can get to know that this question falls in which category.

Conclusion

After performing all the operations we have tested it for the same dataset and we are getting an accuracy of **79.63%** which is quite good. We can improve it further by using more methods used for text classification.

Thank you!