Subject:Natural Language processing

Attay Rasool

February 2023

1 Summary

When Urdu script is written in digital form, white space is not used for word boundary alone but also serves as a sub-word boundary marker as discussed in subsequent sections. Due to this absence of a clear word boundary marker, Urdu exhibits complex segmentation issues for natural language processing as well as information retrieval. In this paper, we present a system to solve this problem of word tokenization. First we import our data using ”import csv” , then our required file urdu-corpus.txt print in the form of ”utf-8”. Then we tokenize our all the text .Tokenize mean we have to separate all words and then print. Then we apply segment Algorithm on our tokenize data .First it will check end part of the sentence by using comparison statement. when our code find end word of the sentence then it will print (-) underscore at the end of sentence and remove all the spaces.