AI-4001 Fundamentals of Natural Language Processing

Due Date: Thursday, March 14st by 11:55pm on Google Classroom. Assignments are to be done individually. No late assignments will be accepted.

Sentence Segmentation in Urdu

In this assignment, the goal is to implement and practice some basic text processing techniques in NLP for the Urdu language. The Urdu language is written in a different style as compared to English, making segmentation a challenging task. There can be several reasons but Space Insertion Problem and Space Omission Problems are the major ones. In this assignment, your task is to perform Urdu Sentence Segmentation. This assignment is designed to be completed from scratch. You are free to use basic libraries if you are comfortable doing so and you can improve existing libraries like urduhack (https://urduhack.com/), but the functions available in these libraries do not use perform up to the mark.

You are provided with the starter file (a1.ipynb) which contains some initial code that is written in python and will help you load the dataset and shows how the available function in UrduHack performs. A trial Urdu corpus is provided as urdu-corpus.txt.

You must also write a function to evaluate the performance of your segmentation technique. This requires you to utilize your problem solving skills!

Sentence segmentation is the process of determining longer processing units consisting of more than one word. This task involves identifying sentence boundaries between words in different sentences. Here is an example of text combination of two sentences:

پشاور زلمی کے اوپننگ بیٹر صائم ایوب نے کہا کہ لیگ کرکٹ ہوتی انٹرٹینگ ہے پی ایس ایل میں ہمیشہ مزہ آتا ہے پشاور زلمی کے ماحول کو انجوائے کر رہا ہوں

You have to develop a technique that will perform segmentation of sentences and remove extra white spaces in sentences for example by passing the above statement, your model should generate output:

پشاور زلمی کے اوپننگ بیٹر صائم ایوب نے کہا کہ لیگ کرکٹ ہوتی انٹرٹینگ ہے۔ پی ایس ایل میں ہمیشہ مزہ آتا ہے۔ پشاور زلمی کے ماحول کو انجوائے کر رہا ہوں

You need to make several decisions in implementing your sentence segmentation techniqiue and its evaluation:

1. How can you detect patterns from the text?
2. Can you identify end-words of a sentence in Urdu?
3. How can you evaluate your approach

   Note: Simply counting the number of segments may not be enough!

Write your name and e-mail id in a comment line in on top of each source file. You are required to submit a single zip file containing an archive of your documentation and ipython notebook on Google Classroom. You should name your notebook as i21-XXXX.ipynb where i21-XXXX represents your student id. You can document your algorithm/technique within the notebook using text cells.