Fundamentals of Natural Language Processing

Text Segmentation in Urdu

1 Introduction

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 and word 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 (*a*1*.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!

2 Background

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 a text combination of two sentences:

ے چاری عوام چونکہ ہمیشہ سے دھوکہ کھانے کی عادی رہی ہے اس لئے “تبدیلی سرکار” کی چکنی چپڑی باتوں میں آگئی اور اپنے بہتر مستقبل کے لئے نئی

حکومت کو اقتدار کے ایوانوں تک پہنچا دیا

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:

بے چاری عوام چونکہ ہمیشہ سے دھوکہ کھانے کی عادی رہی ہے۔ اس لئے ”تبدیلی سرکار“ کی چکنی چپڑی باتوں میں آگئی اور اپنے بہتر

مستقبل کے لئے نئی حکومت کو اقتدار کے ایوانوں تک پہنچا دیا۔

The word segmentation problem in Urdu refers to the process of dividing the written text into individual words. In Urdu, the written language does not include spaces between words, making it difficult to identify where one-word ends and the next begins.

For example, consider the following sentence in Urdu:

نوہشخبدوخجآ

This sentence, without proper word segmentation, is difficult to understand. However, with proper word segmentation, the sentence can be separated into individual words:

نوه شخب دوخ جآ

This makes the sentence much easier to read and understand.

3 Challenges in Implementing a Model

You need to make several decisions in implementing your sentence segmentation technique 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!