**Name Muammad Shaheer**

**Assignment 1 NLP**

**Roll no p20-0480**

Documentation for the code:

**Urdu Sentence Segmentation**

**Introduction**

This code implements and demonstrates basic text processing techniques in Natural Language Processing (NLP) for the Urdu language. It focuses on sentence segmentation, which is the process of identifying and separating sentences in a text. The code is designed to work in the Google Colab environment and requires the installation of the UrduHack library.

**Setup**

Before running the code, make sure to mount your Google Drive using the `drive.mount()` function provided by Google Colab. This allows the code to access files stored in your Google Drive.

**python**

from google.colab import drive

drive.mount('/content/drive')

Next, install the UrduHack library using pip to handle Urdu text processing:

**python**

!pip install urduhack[tf]

**Usage**

**Reading Input Files**

The code reads input text files from your Google Drive. Make sure to provide the correct file paths.

**Sentence Segmentation**

The code uses the UrduHack library to perform sentence segmentation. It tokenizes the input text into sentences using the `sentence\_tokenizer()` function.

**python**

import urduhack

from urduhack.tokenization import sentence\_tokenizer

sentences = sentence\_tokenizer(text)

**Evaluation**

The code includes a function `evaluate\_segmentation()` to evaluate the performance of the sentence segmentation technique. It calculates precision, recall, and F1-score based on reference and segmented sentences.

**python**

precision, recall, f1\_score = evaluate\_segmentation(reference\_sentences, segmented\_sentences)

**Tokenization into Sentences**

Additionally, the code provides a custom function `tokenize\_into\_sentences()` for tokenizing text into sentences. This function handles Urdu punctuation marks and line breaks.

**python**

sentences = tokenize\_into\_sentences(text)

**Conclusion**

This code serves as a basic demonstration of Urdu sentence segmentation using the UrduHack library.