Name: Hifza Majeed

Roll No: 19P-1652

Section: BS(SE) 8A

Assignment No 1Tokenization and segmentation

**Introduction**

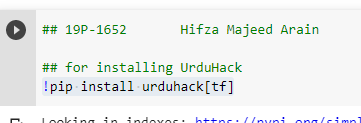
The language I worked with for this project was Urdu, and I separated words and sentences  from passages.

**Steps to solve the Problem**

**Step: 1**

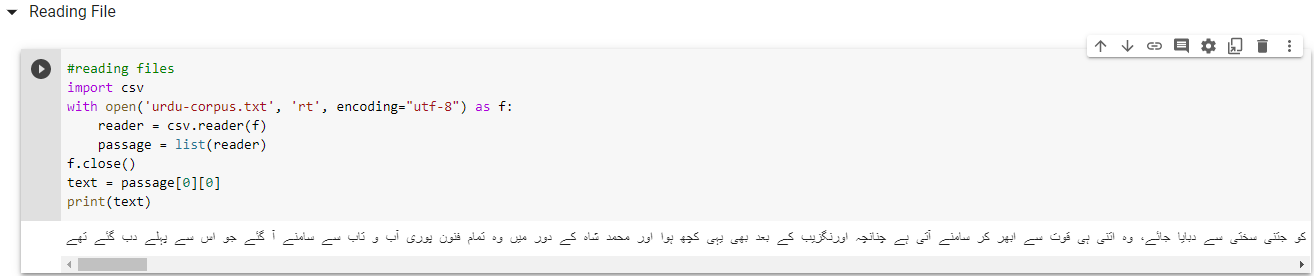
**To install the Urdu hack and import the Urdu hack.**

Install the urdu hack through !pip install urduhack[tf].



**Step: 2**

**Reading file:**



**Step: 3**

**Split full string into list of char**

We splits the paragraph into the words.

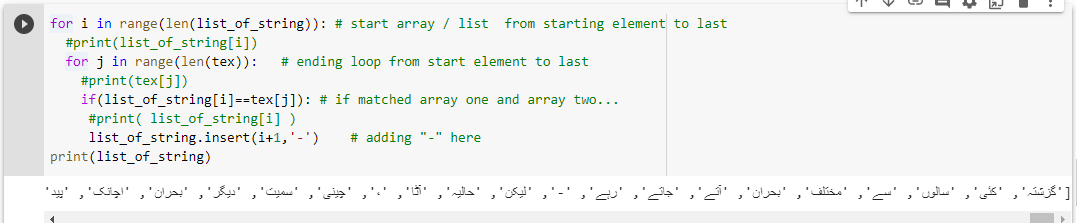
Graphical user interface, text, application

Description automatically generated

**Step: 4**

**Adding the ‘–‘ at the end of sentences**

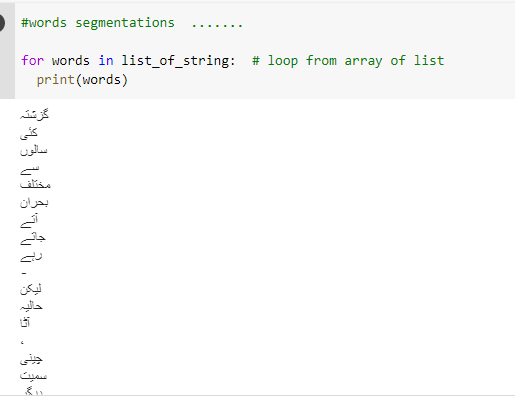
we are comparing the paragraph with end sentences word list than add the ‘–‘ at end of sentences .



**Step: 5**

**Words tokenization**

in list of string having all the word just we apply the loop on the list of string and print the words



**Step: 6**

**Sentences tokenization**

Graphical user interface, text, application, email, Teams

Description automatically generated