Urdu Words and Sentence Segmentation

Noman Siddique

February 7, 2023

Abstract

This model is created to separate different words into a group of words and sentences from a piece of writing, This model is built from scratch which will be effective in urdu writing.

1 Introduction

Urduhack is the library that does the work of NLP on urdu library for which it is built while keeping in mind the ideas of general urdu writing like stop words, notations, etc. This model is a step to build a system that can work like the urduhack, like words and sentence segmentation.

2 TASKS

2.1 Installation

```
[ ]  !pip install 'urduhack[tf]'

     Requirement already satisfied: h5py>=2.9.0 in /usr/local/lib/python3.8/dist-packages (from t
     Requirement already satisfied: google-pasta>=0.1.1 in /usr/local/lib/python3.8/dist-packages
     Requirement already satisfied: keras-preprocessing>=1.1.1 in /usr/local/lib/python3.8/dist-p
     Requirement already satisfied: typing-extensions>=3.6.6 in /usr/local/lib/python3.8/dist-pac
     Requirement already satisfied: grpcio<2.0,>=1.24.3 in /usr/local/lib/python3.8/dist-packages
     Requirement already satisfied: flatbuffers<2,>=1.12 in /usr/local/lib/python3.8/dist-package
     Requirement already satisfied: keras<2.10.0,>=2.9.0rc0 in /usr/local/lib/python3.8/dist-pack
     Requirement already satisfied: tensorflow-estimator<2.10.0,>=2.9.0rc0 in /usr/local/lib/pyth
     Requirement already satisfied: absl-py>=1.0.0 in /usr/local/lib/python3.8/dist-packages (fro
     Requirement already satisfied: protobuf<3.20,>=3.9.2 in /usr/local/lib/python3.8/dist-packag
     Requirement already satisfied: wrapt>=1.11.0 in /usr/local/lib/python3.8/dist-packages (from
     Requirement already satisfied: setuptools in /usr/local/lib/python3.8/dist-packages (from te
     Requirement already satisfied: tensorboard<2.10,>=2.9 in /usr/local/lib/python3.8/dist-packa
     Requirement already satisfied: termcolor>=1.1.0 in /usr/local/lib/python3.8/dist-packages (f
     Requirement already satisfied: numpy>=1.20 in /usr/local/lib/python3.8/dist-packages (from t
     Requirement already satisfied: astunparse>=1.6.0 in /usr/local/lib/python3.8/dist-packages (
     Requirement already satisfied: six>=1.12.0 in /usr/local/lib/python3.8/dist-packages (from t
```

To use the urduhack library you first need to install the urduhack library.

2.2 Words Segmentation

```
def find_matches(text):
    matches = []
    i = 0
    while i < len(text):
        start = i
        while i < len(text) and text[i].isalnum():
            i += 1
        if start < i:
            matches.append(text[start:i])
        while i < len(text) and not text[i].isalnum():
            i += 1
    return matches

urdu_text = text
matches = find_matches(urdu_text)
print(matches)
```

```
['کھودنا', 'گہری', 'ذرا', 'قبر', 'کی', 'حقائق', 'اچھاہے', 'سب', 'تھے', 'گئے', 'دب', 'پہلے', 'سے', 'اس', 'جو', 'گئے', 'آ', 'سامنے', 'سے', 'تاب', 'و', 'آب',
```

This code snippet is doing the job of words Segmentation, split function will take data and a loop will work from start to end in order to look for gaps, and alphanumeric character i-e issalnum() is true the code will scan until the non-alphanumeric character is encountered.

2.3 Sentence Segmentation

The split sentence Function will take the data from the source and check for gaps, question marks, and full stops to separate the sentence from the rest of the writing.

```
[ ] def split_sentences(text):
        sentences = []
        start = 0
        i = 0
        while i < len(text):
            if text[i] in ('.', '?', '!'):
                sentences.append(text[start:i+1].strip())
                start = i + 1
            i += 1
        sentences.append(text[start:].strip())
        return sentences

    urdu_text = text
    sentences = split_sentences(urdu_text)
    print(sentences)
```

```
ے رہے لیکن حالیہ آٹا ، چینی سمیت دیگر بحران اچانک پید ا ہوئے اور ان پر جے آئی ٹی تشکیل دے دیں گئیں تاکہ عوام ؛
```