CS 4063 - Natural Language Processing

Due Date: Friday, Oct 15th by 11:55pm.

Assignments are to be done individually. No late assignments will be accepted.
Submissions that do not comply with the specifications given in this document will not be marked and a zero grade will be assigned.
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 i18-XXXX.ipynb where i18-XXXX represents your student id.

Sentence Segmentation in Urdu

1 Introduction

In this assignment, your goal is to implement and practice some basic concepts of NLP in the Urdu language. The Urdu language is written in different styles as compared to the English language. Urdu Word Segmentation is a challenging task. There can be several reasons but Space Insertion Problem and Space Omission Problems are the major ones. So, in this assignment, our task is to do 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**.
You are provided with the starter file which contains some initial code that is written in python and will help you to load your dataset. The dataset is already split into training and test sets and provided as **Urdu_train.txt** and **Urdu_test.txt**.You can test the performance of your code by using **acc** function.

2 Background

Sentence segmentation is the process of determining the longer processing units consisting of one or more words. This task involves identifying sentence boundaries between words in different sentences. For this assignment, you are given an Urdu corpus. Here is an example of text combination of two sentences:

بے چاری عوام چونکہ ھمیشہ سے دھوکہ کھانے کی عادی رھی ھے 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 train a model 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 ھمیشہ سے دھوکہ کھانے کی عادی رھی ھے۔ 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۔

3 Challenges in Implementing a Model

You need to make several decisions in implementing segmentation Model:

1. How will you extract patterns from the text?
2. How will you identify Urdu End words of a sentence?

Honor Policy

This assignment is a learning opportunity that will be evaluated based on your ability to think in a group setting, work through a problem in a logical manner and write a research report on your own. You may however discuss verbally or via email the assignment with your classmates or the course instructor, but you are to write the actual report for this assignment without copying or plagiarizing the work of others. You may use the Internet to do your research, but the written work should be your own. **Plagiarized reports or code will get a zero**. If in doubt, ask the course instructor.