AI4001/CS4063 - Fundamentals of Natural Language Processing

Due Date: Sunday, February 27th by 11:55pm on Google Classroom.

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 the first text block on top of each python notebook. You are required to submit a python notebook, and a single zip file containing any data files that you have used to run your code, on Google Classroom. You should name your notebook as i19-XXXX.ipynb where i19-XXXX represents your student id. You should document your algorithm/technique within the notebook using text cells.

Transliteration: Roman Urdu - Urdu

1 Introduction

Roman Urdu is the name used for the Urdu language written with the Latin script, also known as the Roman script. The Urdu language is written in a different style as compared to English, making typing a challenging task, and hence the popularity of Roman Urdu.

In this assignment, the goal is to implement and practice some basic text manipulation techniques for the Urdu language. The system of Romanisation used most often by native speakers differs from the formal systems presented in most English language sources. It contains no diacritics or special characters, usually just the 26 letters of the core English alphabet.

In this assignment, your task is to perform transliteration from Roman Urdu to Urdu ($roman2urdu$) and Urdu to Roman Urdu ($urdu2roman$). 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 cannot use online API's such as GoogleTranslate and iJunoon so that your program can also execute offline.

There are several problems with existing Roman Urdu schemes. Either they are not reversible to Urdu script or they don't allow pronouncing the Urdu words properly. Another shortcoming is that a lot of Roman Urdu schemes confuse the Urdu letter *Choti He* which has the sound of voiceless glottal fricative with *Do Chasham He* which is used as a digraph for aspirated consonants in Urdu script. The digraphs *Sh* for letter *Shin* and *Zh* for letter *Zhe* also cause problems as they could be interpreted as the letter *Sin* and *Choti He* or letter *Ze* and *Choti He* respectively. Most Roman Urdu schemes also do not take much consideration of Urdu orthography and the spelling system.

2 Background

Challenges in transliteration involve determining longer textual processing units consisting of more than one character. This task involves identifying grapheme boundaries within different words. Following is an example of Roman Urdu to Urdu transliteration:

```
Maafi say baarh kay abb raha kiya hay?
```

معافی سے بڑھ کے اب رہا کیا ہے؟

You have to develop technique that will perform transliteration that is readable.

3 Implementation Challenges

You need to make several decisions in implementing your tranliteration techniqiue:

1. How can you detect character-level patterns in text?
2. How simple and elegant is your solution?
3. How can you evaluate your approach?

4 Submission and Marking Criteria

Your submission must contain two methods $roman2urdu()$ and $urdu2roman()$.
Your submission will be marked based on the readability of its output, generalizability, computational complexity, code readability and elegance as well as the novelty of your solution. You are not expected to develop machine learning models for this assignment.

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.