INFORMATION RETRIEVAL

Assignment 2: Urdu Spell Corrector

Instructions:

- **This is an individual assignment.**
- **You are not allowed to use NLTK, scikit-learn or any other machine learning or language modeling toolkit.**
- Make a report in PDF format, clearly mentioning question/ part number against your answers.
- Plagiarism of any sort is not acceptable and will be severely punished.
- Late submission will not be accepted.
- **Submit your report, and code as one compressed file (.zip, .rar) on Google Classroom. The name of file should be your roll number i.e. <Roll#>.zip**
- Deadline to submit this assignment is: **Tuesday, April 21, 2020, 11:59 pm.**

Problem:

In this assignment, you'll implement a very import text pre-processing step "Spell Correction". By the end of this assignment you will have your very own "Urdu Spell Checking" software. State-of-the-art spell checker are based on Noisy Channel Model and are represented by below equation:

$$\hat{w} = \operatorname*{argmax}_{w \in V} P(x|w)\,P(w)$$

In this assignment you will implement a simpler version of Noisy Channel model i.e. you will assign same channel probability to each candidate word. You are given following files:

1. jang.txt (950 sentences from Jang Urdu newspaper)
2. jang_errors.txt (50 sentences from Jang Urdu newspaper with error words)
3. jang_nonerrors.txt (Same 50 sentences as jang_errors.txt but without error words)
4. wordlist.txt (Dictionary of Urdu words)

Steps:

1. Train a bigram language model on Jang corpus (jang.txt).
2. For all the error words in jang_errors.txt, find the candidate words that are one and two edit distance away from the error word. Use dictionary (wordlist.txt) to reduce your search space i.e. remove invalid candidates.

Hint: Calculating candidate words is much simpler problem than Minimum Edit Distance.

3. For all the error words, rank the candidate words on the basis of prior probability obtained from the language model.

(REFER TO THE SOLVED EXAMPLE IN LECTURE ON SPELL CORRECTION TO BETTER UNDERSTAND THESE STEPS)

4. Write a report, clearly mentioning error word, top 10 candidate words with probability in a tabular format (separate table for each error word). Is the actual right word (written in jang_nonerrors.txt) is in the list of top 10 candidates (Yes or No)? If yes, highlight/ bold it in the table of candidates.