CS 4063 - Natural Language Processing

Due Date: Friday, October 25th by 10:00am.

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 top of your notebook in a Text cell. You are required to submit a single zip file containing the corpus that you have used to train your language model, and your ipython notebook on Google Classroom. You should name your notebook as i19-XXXX.ipynb where i19-XXXX represents your student id.

Story Generation in Roman Urdu

1 Introduction

After the first few assignments, you have developed a fairly good understanding of the structure of Urdu sentences. In this assignment, you will use *n*-gram language modeling to generate a story using the **spaCy** library for develping the text processing pipeline. For the purpose of this assignment, the generated story should consist of three paragraphs, each containing five to ten sentences. Below is an example of a manually generated paragraph in Urdu.

رات کا اندھیرا چھا چکا تھا اور تمام دنیا نیند کی آغوش میں جا چکی تھی۔ اچانک دروازے پر دستک ہوئی اور میں چونک گیا۔ کون ہو سکتا
تھا؟ دل میں ہزاروں خیال گزرے، مگر ہمت کر کے دروازہ کھولا۔ سامنے ایک اجنبی کھڑا تھا، چہرے پر خوف اور آنکھوں میں کچھ ایسا تھا
جیسے اسے مدد کی ضرورت ہو۔

The task is to print three such paragraphs with an empty line between each. The generational model can be trained on the provided Story Corpus, which contains classic Urdu short stories. You will train unigram, bigram and trigram models using this corpus. These models will be used to generate the story.

2 Assignment Task

The task is to generate a story using different models. You will generate the story sentence by sentence until all paragraphs have been generated. The story generation problem can be solved using the following algorithm:

---

1. Load the Story Corpus
2. Tokenize the corpus in order to split it into a list of words
3. Generate *n*-gram models (unigram, bigram, etc.)
4. For each of the paragraph
    - For each sentence
        - Generate a random number in the range [5...19]
        - Select the first word randomly from a list of starting words
        - for each word from 2 to N Use the *n*-gram model to select the most probable next word
        - **[bonus]** If not the first sentence, try to maintain narrative consistency with the previous sentence
        - Print generated sentence
    - Print empty line after paragraph

2.1 Implementation Challenges

Among the challenges of solving this assignment will be the selection of subsequent words once we have chosen the first word of the sentence. But, to predict the next word, what we want to compute is: what is the most probable next word out of all of the possible next words? In other words, find the set of words that occur most frequently after the already selected word and pick the next word from that set. We can use a Conditional Frequency Distribution (CFD) to figure that out! A CFD tells us: given a condition, what is the likelihood of each possible outcome. [bonus] Maintaining a coherent narrative across sentences can also be a challenge.

2.2 Standard *n*-gram Models

We can develop our model using the Conditional Frequency Distribution method. First develop a unigram model, then the bigram model. Select the first word of each sentence randomly from starting words in the vocabulary and then use the bigram model to generate the next word until the sentence is complete. Generate the next sentences similarly. Follow the same steps for the trigram model and compare the results of the *n*-gram models.

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.