Natural Language Processing - Assignment 02

Deadline 1: 24/09/2021 11:59:59 PM **Maximum Marks: 50**

Deadline 2: 06/10/2021 11:59:59 PM (Without penalty; however, you will lose the chance for earning the overall bonus across assignments at the end of the course. Beyond this 25% penalty will be applicable.).

Instructions:

- The assignment is to be attempted in a group of max 2. You have to clearly mention the contribution of each member in README.pdf.
- Language allowed: Python
- You are allowed to use libraries such as NLTK, Panda, etc., for data preprocessing.
- For Plagiarism, institute policy will be followed.
- You need to submit README.pdf, Code files (it should include both .py files and .ipynb files), and Output.pdf.
- Mention methodology, preprocessing steps, all parameters, and assumptions in README.pdf.
- Mention your predicted answers along with the question for the development set in Output.pdf.
- You are advised to prepare a well-documented code file.
- Submit code, readme, and output files in ZIP format with the following name: A2_<roll_no>_<roll_no>.zip
- Use classroom discussion for any doubt.

Task:

Next Word Prediction: Given a sequence of words in a sentence, predict the most appropriate VERB among a set of possible verbs.

Dataset: Download here

Train set: A list of sentences.

Test/Development set: (Test set will be released during evaluation/demo)

- **Question** - **A sequence of words** containing a masked token marked as XXXX.

`` They are very kind old ladies in their way , " XXXXX the king ; `` and were nice to me when I was a boy . ""

- **Options** - A **list** of **possible words as** replacements for the masked token
  ['christening', 'existed', 'hear', 'knows', 'read', 'remarked', 'said', 'sitting', 'talking', 'wearing']

- **Answer** - The most appropriate word for replacement. (For evaluation)
  ['said']

Training of Language Model

- Extract sentences from the train set to learn the bi-gram model. **[10]**
- You should employ **Laplace (add-1)** and **add-k** (k is arbitrary, you can choose any value) smoothing techniques for learning the language model. Show their effect on the performance. **[10+10]**

Note:

- You have to train at least three models -- without smoothing, with Laplace and with add-k.)
- You have to compute the co-occurrence matrix and utilize it for the probability computation in each case.
- In case you're facing difficulties in training LM due to the limited computational resources, you can reduce the vocabulary size based on frequency. (If you do so, please be aware that it might introduce unknown tokens at the inference time as well. You have to handle it accordingly.)

Testing/Next word prediction

Given 'Question' till the masked token, find out the most probable word from the 'Options' list.

Evaluation

- You have to compute accuracy values. For each question, if the predicted word is not the correct answer, it should be treated as misclassification and vice-versa.
- The development set (a set of questions) should be used for tuning and reporting the performance. **[5]**
- During demonstration, a custom test set will be used for evaluating the accuracy of the language model. **[15]**

Bonus

Show you innovation by considering the future context while predicting the masked token as well. A partial solution would not be considered. Also, the solution should be justifiable. **[5]**