CS7.401: Introduction to NLP **Spring 2021**

Assignment 2

Deadline : 04-04-2021, 23:55 Hrs

Instructor: Dr. Manish Shrivastava TA: Guru Ravi Shanker, Roopal, Tanvi, Prashant

1 General Instructions

1. The assignment can be implemented in Python.
2. Ensure that the submitted assignment is your original work. Please do not copy any part from any source including your friends, seniors, and/or the internet. If any such attempt is caught then serious actions including an F grade in the course is possible.
3. A single .zip file needs to be uploaded to the Moodle Course Portal.
4. Your grade will depend on the correctness of answers and output. In addition, due consideration will be given to the clarity and details of your answers and the legibility and structure of your code.

2 Problem Statement

In the previous assignment we have seen how to develop statistical LMs based on corpora. In this assignment we will develop Neural LMs as discussed in class, and compare the performance between the statistical and neural LMs.

1. Create Neural Language Models trained on Brown Corpus. You are expected to clean the data to remove any special characcters and use it for training the Neural LM. You can use libraries for tokenisation. and use either PyTorch or TensorFlow as your deep learning frameworks.

   You can download the corpora from this link.

   Sample 10,000 sentences as validation set and 20,000 sentences as test set.

2. Calculate the perplexity scores
   - for each sentence of train, validation and test corpus.
   - average perplexity score/corpus/LM on the train, validation and test corpus.
3. Compare the perplexity scores of the following models and your analysis of results, along with any necessary visualisations in a PDF report:
   - Neural LM trained on the Brown corpus shared here.

- Statistical LMs, as created by you in Assignment 1, trained on Brown corpus shared here.

Compare and analyze the behaviour of the different LMs and put your analysis and visualisation in a report.

3 Submission Format

Zip the following into one file and submit in the Moodle course portal. Filename should be RollNum_Assignment1.zip, ex 2021xxxxxx_Assignment1:

1. Source Code along with README on instructions to execute the code.
2. Report containing the perplexity scores of all the LMs and your analysis of the results, along with any visualisations in a PDF.
    (a) for each LM submit the text file with perplexity scores in the following format Format : Sentence TAB perplexity-score, at the end , average score

    (b) Naming must be: roll_number-LM1-train-perplexity.txt, roll_number-LM1-test-perplexity.txt, etc
3. Readme File :on how to execute the code, how to get the preplexity of a sentence. Any other information.

4 Grading

1. Evaluation will be individual and will be based on your viva, report, submitted code review.
2. In the slot you are expected to walk us through your code, explain your experiments, and report.