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

Assignment 1

Deadline : 27-01-2021, 23:55 Hrs

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

1 General Instructions

1. The assignment can be implemented in Python.
2. No standard libraries for calculating n-grams, LMs or tokenization should be used.
3. 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.
4. A single .zip file needs to be uploaded to the Moodle Course Portal.
5. 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

You have been given two corpus : one from Health domain, and one from Technical domain. Your task is to design Language Models for both of these corpora using smoothing.

1. Create language models with following parameters
   (a) On health domain corpus
      i. LM 1 : tokenization + 4-gram LM + Kneyser Ney smoothing
      ii. LM 2 : tokenization + 4-gram LM + Witten Bell smoothing
   (b) On technical domain corpus
      i. LM 3 : tokenization + 4-gram LM + Kneyser Ney smoothing
      ii. LM 4 : tokenization + 4-gram LM + Witten Bell smoothing
2. For each of these corpora, create a test set by randomly choosing 1000 sentences. This set will not be used for training LM.
   (a) Calculate perplexity score for each sentence of health domain corpus and technical domain corpus for each of the above models and also get average perplexity score/corpus/LM on the train corpus.
   (b) Report the perplexity scores for all the sentences in the test set. Report the perplexity scores on the test sentences as well, in the same manner as above.

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

3 Training corpus

Please use the following links to download the text corpora for training the models:

1. Technical Domain Corpus
2. Health Domain Corpus

4 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
   (a) language_model.py: Runs the language model given the following:

```
$ python3 language_model.py <n value> <smoothing type> <path_corpus>
```

   where smoothing type can be k for Kneyser Ney or w for Witten Bell. On running the file, the expected output is a prompt, which asks for a sentence and provides the probability of that sentence using the two smoothing mechanisms. Therefore an example would be:

```
$ python3 language_model.py k ./corpus.txt
input sentence: I am a man.
0.899742021
```

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.

5 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.