CS565: Intelligent Systems and Interfaces

Assignment 2

Topics: N-Gram Language Models, Smoothing, Vector Semantics

Marks: 60 **Due Date: November 5, 2020**

Instructions

- Each student will require to submit the assignment individually. No group submission will be allowed.
- Each student will submit a report. The report will include brief write up of the analysis performed and its results along with the relevant plots.
- You are expected to write the code in Google Colab Notebooks and include the notebook link in your report. The code must reproduce the results discussed in your report. Add appropriate comments and instructions in the code wherever required.
- If there is any query related to assignments, please post/message via Piazza only.

Dataset

Use the same English Wikipedia dataset given for the first assignment. You can download it from [English] in case you have deleted it.

Prepare the training, development and test set as follows: After sentence segmentation and word tokenization, randomly shuffle sentences and split them into two parts of 90% (part-1) and 10% (part-2) of all sentences. Part-2 constitutes the independent test set and will remain fixed. Split the Part-1 to create training and development sets.

[**Remark**: You are free to choose your sentence segmentation and tokenization method. However, your code and report must have the entire details along with the code for splitting data into three subsets.]

N-gram language model [30 marks]

1. Implement a tri-gram language model using Discounting and Interpolation smoothing methods.
2. Use Part-1 to split the data into two parts of 90% (training data) and 10% (development/validation) set. Do this five times to prepare five different pairs of training and development sets.
3. Report model's performances in terms of the perplexity and Likelihood on a held-out (development/validation) sets. Do you obtain the same set of parameters?
4. Report model's performance on independent test set.

5. Summarize results in the report including discussions on which smoothing methods had a least variance and what happens if only Laplace smoothing is used.

Vector Semantics: GloVE implementation [30 marks]

1. Implement GloVe embedding method using the given corpus.
2. In your implementation, you can use either Gradient Descent or AdaGrad as an optimization method.
3. Compare embeddings obtained using your implementation and any pre-trained GloVe embeddings of same dimensions.
4. For your comparison, you can use any standard word-similarity benchmark datasets. Do mention the same in your report.
5. In your report, please write down the necessary expression of derivatives required in optimization algorithm (gradient descent or AdaGrad).