CS 772: Assignment 1

Problem Statement

- In this assignment, you will have to implement the backpropagation algorithm from scratch. After implementing backpropagation ab-initio, train CBow and Skip-gram with backpropagation. (This link might help as a quick refresher for Skip-gram and CBoW.)
- The task is to compare the performance of CBoW and Skip-gram embeddings on the word analogy task.
    - Analogy task: Given an analogy, find a word by correctly determining its relationship with another word. For example,
      man:woman :: king:_____
      (man is to woman, what king is to ____)
      The blank should be filled with “queen”.
- **Input**: An analogy pair with one blank. For e.g.,
    - Delhi:India :: Paris:_____
- **Output**: The correct word to satisfy the analogy given in the input.
- You will have to report on the validation data:
    - P, R, F1-scores
    - Compare the performance of CBoW and Skip-gram models.
    - Perform detailed error analysis

Dataset & other details

- Gutenberg corpus: This is the dataset you will use for training your CboW and Skip-gram models. You will have to augment this dataset as described in the point below.
- Analogy dataset:
    - This dataset contains analogy pairs.
    - For every word in each pair, you need to get k sentences which contain these words from resources like Wikipedia, Concordancer, Wordnet etc. (What is k? A design choice! You can choose a value for k.)
    - Add these sentences to the Gutenberg corpus. You will use the Gutenberg dataset + the extracted sentences for training.

  [Datasets will be shared with you soon]
- Parameters to play with: Embedding dimension, k, no. of iterations, learning rate.
- Validation data: We will provide validation data containing analogy pairs. You will use this data to test and compare the performance of your CBoW and Skip-gram models.
- Test data: A hidden test set will be used to rank the evaluation groups. This will be done using Kaggle In-class. More details regarding this will be shared later.

Submission instructions

- The assignment is to be submitted in groups (Same group for every assignment and project)

- You are supposed to implement the code on your own. Do not copy code from the Internet. Your code will be checked for plagiarism.

Deadline

- The first evaluation will be around the second week of February. The exact dates will be announced soon.