Assignment #4: Word Embedding Network
Due: Friday, April 30th, 2021 at 11:59 PM (EST)

Description

In this assignment you will practice how to create a Word Embedding Network in PyTorch. First, you will finish some functions to parse the data, build the corpus and construct the skip pair. Then, you will construct a word embedding network by follow the specific requirements and architectures. Finally, you will train the network and visualize the result. The goals of this homework are:

- To implement and understand Word Embedding Networks.

Instructions

In this assignment, you need to fill the block of code in the python notebook file. The descriptions of all the functions you need to implement are as follow:

- **Setup** (10 points): Set some of the parameters for your model. In particular, set SUID equal to your SUID number as a random seed. You can also select between the different source texts.
- **build_dictionary** (10 points): Extract the word from the input. Build a non-duplicate word dictionary
- **one_hot_encoding** (10 points): Every word is represented as a tensor containing 1 at its position in the vocabulary
- **build_word_index_mapping** (10 points): Given a word, the function should return the index of this word via a dictionary. Given an index, the function should retrieve the word.
- **build_skip_pair** (10 points): Build the word pairs with given window size.
- **Net** (10 points): Define all the layers you will use in the embedding network. Define the network layer connectivity.
- **Optimizer and Criterion** (10 points): Implement your optimizers and two different loss models.
- **Learning** (10 Points): Follow the instructions in the notebook to learn the embeddings of the words.
- **nearest_indices** (10 points): Given a list of distances from a given embedding, return the indices of the closest embeddings. The closest will always be the original word.

The final ten (10) points will be for a report. In the report please answer the following questions.

- What are the closest five (5) words to 'she' and 'queen' in the two different source texts?
- Does the embedding from the first or second layer work better for your model on the first text? Use the plotting images at the bottom to justify your answer.

Notes:

- The notebook has comments that will walk you through the implementation. Furthermore, they have explanations in each block of code that you have to fill in.
- The number of points available for each block of code is in the comment with the instructions.
- Comment your code.
- Do not call the print function in your final submission.
- *** Do NOT edit any of the code outside of the TODO blocks. ***

Submission

Your submission ZIP archive will contain one (1) python notebook named: '**Assignment_4.ipynb**', and one (1) PDF report named '**report.pdf**'. (Do not change the names of the python files!)

- Zip file named via the following convention:
  - <SU-EMAIL>_<FIRST-Name>_AS4.zip
  - Ex. dprider_Daniel_AS4.zip
- Upload the zip file to blackboard before 11:59PM (EST Time) 04/30/2021