Homework#3: Word Embedding

Name:________________________________ Student ID:___________________________

Q1: Use your own tokenizer (aka word segmentation model) to define word boundaries and split the given text file into words. Capture the screenshot of your code segment that loads the word segmentation model and uses the model to segment the text files.

Q2: “UNK” is often used to represent an unknown word (a word which do not exist in your dictionary/training set). You can also represent a rare word with this token as well. How do you define a rare word in your program? Explain in your own words and capture the screenshot of your code segment that is a part of this process.

Q3: The negative samples are sampled from sampling_table. Look through Keras source code to find out how they sample negative samples. Discuss the sampling technique taught in class and compare it to the Keras source code.

Q4: In your own words, discuss why Sigmoid is chosen as the activation function in the skip-gram model.

Q5: Visualize the model using TSNE (scikit-kearn) and Tensorboard Projector include the image(s) of your visualization here and discuss what you observe. **Include the link to your TensorBoard Projector here.**

Q6: Use the word embeddings from the skip-gram model as pre-trained weights in a classification model. Compare the result the with the same classification model that does not use the pre-trained weights. (it's okay if the result is not what you expected, if that is the case tell us what do you expect to see.)