**Assignment 3 documentation**

**Word2Vec tasks**

Used the pretrained glove-twitter-100 model. Pre-processed the data(Removed mentions, stop words, punctuations, numbers, and emojis). Stored offensive and non-offensive tweets in different variables.

**compare\_texts\_word2vec implementation:**

Made use of python Counter to get the common words from both the files and stored them in different variables. Used similarity method to find the cosine similarity between the words present at a particular index in both the files. For example, the words present at index 0 in the common words list of both the files are compared and the similarity is found. For each of the words present in the common words list of each of the files, we find the k most similar words using the most\_similar method. The top k similar words of each of the common words list are then stored in two lists. These two lists are then used to get the unique and intersection words.

Graphical user interface, text

Description automatically generated

**MLP tasks:**

Used the pretrained glove-twitter-25 model. Pre-processed the data. Trained the MLP model with 1, 2 and 3 layers. The accuracy obtained is 0.27 when the number of layers are set to 2 and 3. But the accuracy is 0.72 when there is only 1 layer. This means that the accuracy decreases as the number of layers increases.

A picture containing text

Description automatically generated

Extra credit:

Trained a new word2vec model with the words present in the training set

