**Word2Vec Embeddings**

**HW5**

**Harsha Vardhan, Khurdula.**

**CS59000 Natural Language Processing**

**Professor: Jon, Rusert.**

**Introduction**

Word2Vec is one of the many popular word embeddings used, in order to create a numerical vector representation of words over a distribution. This was first published in 2013. Word2Vec uses a neural network model. This model learns word associations from large corpus of textual data. Post training the model will be capable of understanding meaningful relations within text, words or suggest additions.

The current task at hand is to work with Word2Vec embeddings, and understand them, while implementing two methods that would determine similarity between texts, and documents respectively.

**Data Preparation**

The dataset that was suggested to be used in this task is same as the previously used Toxic set. Which consists of train set and test set with a corresponding label set. However within the method definition we don’t seem to be passing features, rather simple text as input.

In order to make this possible we shall create two sets of text, i.e. TOXIC\_subset, and NOT\_TOXIC\_subset, and it is also necessary that we should not create these within Word2Vec task. So I shall perform this prior to having developed the two required methods.

This would be achieved by data\_handler notebook that would create the subsets that are asked to be tested on.