**Project's Title:** Skip Gram implementation from scratch

**Motivation:** This is our Assignment for CS772: Deep Learning for NLP. The aim is to study and implement the skip gram algorithm from scratch in python.

Cluster of “animal words”: dog","tiger","cat","elephant","horse","bear","cow","monkey","deer","snake","lion"

Cluster of “bird words”: sparrow","parrot","dove","eagle","woodpecker","crane","owl","duck","cuckoo","penguin"

Create word clusters as mentioned above. Run a concordance for obtaining the neighboring words of these words. Train a skip-gram model with these words. Collect the word representations. Ensure that “animal” words are close to other “animal” words and so are “bird” words; Inter-cluster distance should be large compared to intra-cluster distance; use cosine similarity and other cluster-quality measures.

**Team Members:**

1. Tejas Gunaghe - 213194002
2. Tanishq Awasthi - 213190003
3. Vaibhav Singh Panwar – 213190004

**GitHub repository:** <https://github.com/TejasGunaghe/Skip-Gram-implementation-from-scratch->

**Build Status:** Improving model based on suggestions from TAs

**Programming Language**: Python

**Libraries/API used:**

Text

Description automatically generated

**Input File:** Corpus.txt – attached with the code files. File can be downloaded from below icon.



**Model Tuning Parameters:** Note: Tuning can be done inside the skipgram class in code

1. **Learning Rate:** Adaptive starting from 0.001
2. **Neuron considered in hidden layer:** 30
3. **Max epochs:** 1000
4. **Tolerance for stopping condition:** 0.001

**Running the skip gram model: (Details are provided in section Code Details)**

Step1: Import/Upload the Corpus.txt file attached with code files in the python notebook.

Step2: Perform all steps (code lines). Run all lines of the code

Step3: Get the output in “predict context” section of the code.

Step4: Get cosine similarity & inter and intra cluster distances in respective sections

**Code Details:**

1. **Corpus.txt: 1724 sentences**

The word corpus is taken from Lextutorusing’ the Wikipedia corpus. Collected 25-30 sentences for each of the animal and bird words from the concordancer and the Input corpus is in the form of txt file named “cluster.txt”. This corpus is used to train the model. This file includes 1724 sentences for all words.

1. **Data Preprocessing and encoding:**

The corpus sentences are converted to corpus set by performing following actions:

1. Removing Punctuations
2. Remove spaces
3. Convert the corpus to lower case.
4. One hot encoding for input as well as context (output array)
5. Context window size considered: -2 words and +2 words, can be changed inside the skigram class in code
6. **Skipgram Object (model):**

The skipgram object includes following set of function/parameters listed below:

* 1. **Activation function:** SoftMax

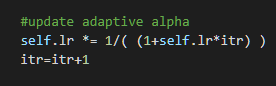
Text

Description automatically generated

* 1. **Loss Function:** Negative log likelihood loss function



* 1. **Learning Rate**: Adaptive starting from 0.001



* 1. **Initializing weights:** weights are initialized with random numbers from -0.8 to 0.8

Text

Description automatically generated

* 1. **Train Function:** This function includes the neural network codes like feedforward and backpropagation algorithm.

Text

Description automatically generated

1. **Predicting context words:**

Text

Description automatically generated

Following set of code can be used to get the predicted context words:

Graphical user interface, application

Description automatically generated

1. **Cosine -Similarity functions**:

Text

Description automatically generated

1. **Inter and Intra cluster distances and DB index:**

Text

Description automatically generated

**Credits:**

1. IIT Bombay course material for CS772: DL for NLP
2. https://www.geeksforgeeks.org