**NLP LAB 1 EVALUATION**

**Team Members:**

Raksha P Rao – 1PI13IS081

M.S.Saicharan – 1PI13IS112

Swetha Banagiri – 1PI13IS114

V.P. Ramachandran – 1PI13IS119

**Objective**: Design a system that follows the principles of Language models.

**Goal**: To determine similarity between two words in the given dataset and evaluate the same by comparing the scores obtained by our method and gensim’s word2vec

Steps involved:

1. File was cleaned by removing hashtags, RTs, links and special characters followed by converting all characters to lowercase.
2. A vocabulary was created out of the same.
3. Extraction of triples – a count was maintained for the same.
4. Random pairs were picked and we found the similarity based on the words present on the left and right and a score was assigned to each of them.
5. We used word2vec from gensim.models to test the similarity of the words in Step 4.

The experiment produced the following tangible results:

|  |  |  |
| --- | --- | --- |
| **Word\_pairs** | **Word2Vec(observed)** | **Word2vec(Gensim)** |
| { modi, namo} | 0.120752515677 | 0.201586963325 |
| { leader,pm} | 0.154905335628 | 0.0170196905491 |
| { is, was} | 0.147383935151 | 0.12309618876 |
| {red, blue} | 1.0 | 0.128824182122 |
| { serviceman, modi} | 1.0 | 0.17664773754 |
| { modi ,pm} | 0.147383935151 | 0.167948699689 |
| { lalit, narendra} | 0.0300263574995 | 0.0385659282736 |
| { narendra, leader} | 0.0533964402373 | 0.123011968701 |
| { dubai, government} | 0.17671273155 | 0.0741578431156 |
| { software, cricket} | 1.0 | 0.105087330884 |

Notes and Observations:

1. When the observed Word2Vec value = 1, it may or may not exist in the corpus or the two words are not semantically related.
2. The computed error values for our observation is

Average = (sum of all observed values)/10 = **0.38**

Average = (sum of all gensim values)/10 = **0.132**

**Average Error = (average of observed values - average of gensim values) = 0.38 – 0.132 = 0.248**