**Assignment Report**

Name: Prajwala Mugajji Shambulingappa

**Aim:** Working with word2vec embeddings to understand them.

**Observation Criteria and Analysis:**

1. Packages Used: nltk, re, pandas, NumPy, nltk - 'stopwords', nltk -'punkt', collections – ‘Counter’ and genism – ‘downloader’
2. Dataset: Dataset used for is provided by Kaggle which has a large number of Wikipedia comments that have been labelled by toxic behavior by human reader. Also, used ‘War of Worlds, ‘On Liberty’ and ‘King Arthur’ from Gutenberg.org.
3. Creating dataset: Before preprocessing, I made stwo subsets toxic and non-toxic dataset of the Wikipedia comments using the labels.
4. Preprocessing: As part of preprocessing, I have created tokens and normalized these tokens by lowercasing, removing whitespaces, splitting, and removing empty strings. After which, I have removed the stopwords using nltk package, as the task stated that I need to find k most common non-stop words. If not done, I noticed that stop words such as ‘this’ and ‘and’ were part of the most common words.
5. Pretrained model used: word2vec-google-news-300
6. Compare\_texts\_w2v method: Initially, I took the count of every non-stop word in the preprocessed toxic and non-toxic corpus. After which, I took the top ‘k’ words using the count from the previous step and then I calculated the similarity score between the top words for both the corpus. Initially, I was not using the try catch block but that kept throwing error as there were no embeddings for some words such as ‘faggot’.
7. doc\_overview\_w2v method: Initially, I took the count of every non-stop word in the preprocessed ‘War of Worlds, ‘On Liberty’ and ‘King Arthur’ dataset. After which, I took the top ‘k’ words using the count from the previous step. After which, I found the ‘n’ similar words to every top ‘k’ words.

**Summary:**

1. Method to take two dataset files and compare them at the word level by leveraging word2vec and another method to get an overview by summarizing the top k words and their nearest neighbors for a given document were implemented successfully.
2. Top words in training dataset when k = 5:

|  |  |  |  |
| --- | --- | --- | --- |
| K = 5 | | | |
| Toxic | | Non-Toxic | |
| Word | Frequency | Word | Frequency |
| ‘fuck’ | 8613 | ‘article’ | 54043 |
| ‘shit’ | 3583 | ‘page’ | 43499 |
| ‘dont’ | 3566 | ‘wikipedia’ | 32297 |
| ‘like’ | 3477 | ‘talk’ | 30030 |
| ‘nigger’ | 3289 | ‘please’ | 28700 |

1. Top words in training dataset when k = 10:

|  |  |  |  |
| --- | --- | --- | --- |
| K = 10 | | | |
| Toxic | | Non-Toxic | |
| Word | Frequency | Word | Frequency |
| ‘fuck’ | 8613 | ‘article’ | 54043 |
| ‘shit’ | 3583 | ‘page’ | 43499 |
| ‘dont’ | 3566 | ‘wikipedia’ | 32297 |
| ‘like’ | 3477 | ‘talk’ | 30030 |
| ‘nigger’ | 3289 | ‘please’ | 28700 |
| ‘wikipedia’ | 3260 | ‘would’ | 28203 |
| ‘fucking’ | 3192 | ‘one’ | 26579 |
| ‘suck’ | 3034 | ‘like’ | 24228 |
| ‘go’ | 2833 | ‘dont’ | 22536 |
| ‘hate’ | 2614 | ‘see’ | 20575 |

1. Top words in training dataset when k = 20:

|  |  |  |  |
| --- | --- | --- | --- |
| K = 20 | | | |
| Toxic | | Non-Toxic | |
| Word | Frequency | Word | Frequency |
| ‘fuck’ | 8613 | ‘article’ | 54043 |
| ‘shit’ | 3583 | ‘page’ | 43499 |
| ‘dont’ | 3566 | ‘wikipedia’ | 32297 |
| ‘like’ | 3477 | ‘talk’ | 30030 |
| ‘nigger’ | 3289 | ‘please’ | 28700 |
| ‘wikipedia’ | 3260 | ‘would’ | 28203 |
| ‘fucking’ | 3192 | ‘one’ | 26579 |
| ‘suck’ | 3034 | ‘like’ | 24228 |
| ‘go’ | 2833 | ‘dont’ | 22536 |
| ‘hate’ | 2614 | ‘see’ | 20575 |
| ‘ass’ | 2596 | ‘also’ | 19815 |
| ‘u’ | 2583 | ‘think’ | 18807 |
| ‘get’ | 2283 | ‘im’ | 17416 |
| ‘gay’ | 2223 | ‘articles’ | 17016 |
| ‘know’ | 2181 | ‘know’ | 16807 |
| ‘page’ | 2112 | ‘edit’ | 16676 |
| ‘die’ | 2076 | ‘people’ | 15890 |
| ‘im’ | 2052 | ‘use’ | 15880 |
| ‘fat’ | 1967 | ‘may’ | 15251 |
| ‘faggot’ | 1954 | ‘time’ | 14345 |

1. Top words in test dataset when k = 5:

|  |  |  |  |
| --- | --- | --- | --- |
| K = 5 | | | |
| Toxic | | Non-Toxic | |
| Word | Frequency | Word | Frequency |
| ‘fuck’ | 5534 | ‘article’ | 21103 |
| ‘nigger’ | 2248 | ‘page’ | 13905 |
| ‘fucking’ | 1764 | ‘would’ | 10930 |
| ‘bitch’ | 1590 | ‘wikipedia’ | 10365 |
| ‘dicks’ | 1440 | ‘one’ | 10105 |

1. Top words in test dataset when k = 10:

|  |  |  |  |
| --- | --- | --- | --- |
| K = 10 | | | |
| Toxic | | Non-Toxic | |
| Word | Frequency | Word | Frequency |
| ‘fuck’ | 5534 | ‘article’ | 21103 |
| ‘nigger’ | 2248 | ‘page’ | 13905 |
| ‘fucking’ | 1764 | ‘would’ | 10930 |
| ‘bitch’ | 1590 | ‘wikipedia’ | 10365 |
| ‘dicks’ | 1440 | ‘one’ | 10105 |
| ‘faggot’ | 1350 | ‘please’ | 9431 |
| ‘gay’ | 1293 | ‘like’ | 9316 |
| ‘shit’ | 1242 | ‘dont’ | 8494 |
| ‘die’ | 1241 | ‘think’ | 7667 |
| ‘youfuck’ | 1165 | ‘see’ | 7664 |

1. Top words in test dataset when k = 20:

|  |  |  |  |
| --- | --- | --- | --- |
| K = 20 | | | |
| Toxic | | Non-Toxic | |
| Word | Frequency | Word | Frequency |
| ‘fuck’ | 5534 | ‘article’ | 21103 |
| ‘nigger’ | 2248 | ‘page’ | 13905 |
| ‘fucking’ | 1764 | ‘would’ | 10930 |
| ‘bitch’ | 1590 | ‘wikipedia’ | 10365 |
| ‘dicks’ | 1440 | ‘one’ | 10105 |
| ‘faggot’ | 1350 | ‘please’ | 9431 |
| ‘gay’ | 1293 | ‘like’ | 9316 |
| ‘shit’ | 1242 | ‘dont’ | 8494 |
| ‘die’ | 1241 | ‘think’ | 7667 |
| ‘youfuck’ | 1165 | ‘see’ | 7664 |
| ‘like’ | 1121 | ‘also’ | 7194 |
| ‘ass’ | 1070 | ‘im’ | 7071 |
| ‘hate’ | 1069 | ‘know’ | 6141 |
| ‘niggers’ | 1062 | ‘people’ | 6067 |
| ‘boob’ | 1005 | ‘articles’ | 5885 |
| ‘poop’ | 994 | ‘use’ | 5693 |
| ‘u’ | 944 | ‘may’ | 5456 |
| ‘suck’ | 910 | ‘talk’ | 5435 |
| ‘stupid’ | 905 | ‘edit’ | 5374 |
| ‘youi’ | 888 | ‘time’ | 5182 |

1. Similarity scores:

|  |  |  |  |
| --- | --- | --- | --- |
| K values | 5 | 10 | 20 |
| Train data Similarity Scores | 0.181 | 0.236 | 0.194 |
| Test data Similarity Score | 0.123 | 0.177 | 0.166 |

1. War of words dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| K = 5, N= 5 | | | |
| Top Word | Frequency | Similar words | Respective Frequency |
| ‘one’ | 195 | ‘only’, ‘two’, ‘three’, ‘five’, ‘four’ | 0.598, 0.576, 0.568, 0.5628, 0.5627 |
| ‘upon’ | 173 | ‘on’, ‘Upon’, ‘Performance\_Ratios\_Based’, ‘on the’, ‘whereupon’ | 0.506, 0.505, 0.391, 0.380, 0.374 |
| ‘martians’ | 167 | ‘Martians’, ‘Venusians’, ‘humanoid\_aliens’, ‘Gungans’, ‘outerspace’ | 0.647, 0.631, 0.617, 0.610, 0.6080 |
| ‘said’ | 166 | ‘says’, ‘explained’, ‘siad’, ‘noted’, stressed’ | 0.776, 0.693, 0.685, 0.657, 0.655 |
| ‘people’ | 161 | ‘peole’, ‘people’, ‘individuals’, ‘folks’, ‘peple’ | 0.605, 0.590, 0.5827, 0.579, 0.578 |

1. On Liberty dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| K = 5, N= 5 | | | |
| Top Word | Frequency | Similar words | Respective Frequency |
| ‘one’ | 242 | ‘only’, ‘two’, ‘three’, ‘five’, ‘four’ | 0.598, 0.576, 0.568, 0.5628, 0.5627 |
| ‘may’ | 219 | ‘might’, ‘could’, ‘should’, ‘can’, ‘will’ | 0.829, 0.726, 0.680, 0.676, 0.627 |
| ‘opinion’ | 153 | ‘opinions’, ‘opinon’, ‘opnion’, ‘Opinions’, ‘opinons’ | 0.716, 0.633, 0.561, 0.5496, 0.5492 |
| ‘would’ | 138 | ‘could’, ‘might’, ‘will’, ‘should’, ‘did’ | 0.832, 0.730, 0.674, 0.671, 0.653 |
| ‘others’ | 115 | ‘those’, ‘people’, ‘other’, ‘them’, ‘Others’ | 0.564, 0.551, 0.533, 0.5078, 0.5077 |

1. King Arthur dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| K = 5, N= 10 | | | |
| Top Word | Frequency | Similar words | Respective Frequency |
| ‘sir’ | 2029 | ‘ma'am’, ‘m'am’, ‘Yes\_Ma'am’, ‘ma'm’, ’maam’, ‘Ma'am’, ’Yes\_sir’, ‘M'am’, ‘KARSNIA\_OK’, ‘Wham\_bam\_thank’ | 0.7133, 0.675, 0.673, 0.666, 0.661, 0.651, 0.620, 0.619, 0.609, 0.608 |
| ‘king’ | 947 | ‘kings’, ‘queen’, ‘monarch’, ‘crown\_prince’, ‘prince’, ‘sultan’, ‘ruler’, ‘princes’, ‘Prince\_Paras’, ‘throne’ | 0.713, 0.651, 0.641, 0.620, 0.615, 0.586, 0.579, 0.564, 0.543, 0.542 |
| ‘said’ | 901 | ‘says’, ‘explained’, ‘siad’, ‘noted’, ‘stressed’, ‘told’, ‘added’, ‘adding’, ‘remarked’, ‘acknowledged’ | 0.776, 0.693, 0.685, 0.657, 0.655, 0.646, 0.644, 0.638, 0.627, 0.601 |
| ‘knight’ | 679 | ‘knights’, ‘shinning\_armor’, ‘gallant\_knigh’, ‘nobleman’, ‘medieval\_knight’, ‘knight\_errant’, ‘Roman\_centurion’, ‘prince’, ‘steed’, ‘chivalrous\_knight’ | 0.732, 0.664, 0.603, 0.596, 0.595, 0.571,0.5397, 0.5390, 0.536, 0.532 |
| ‘thou’ | 475 | ‘Thou’, ‘wherefore\_art’ ‘thou\_Romeo’. ‘thy’, ’Thou\_art’, ‘knowest’, ‘thou\_attitude’, ‘Holier’, ‘wilt\_thou’, ‘thee’ | 0.659, 0.621, 0.620, 0.604, 0.587, 0.575, 0.564, 0.559, 0.537, 0.536 |