**Report for Word2Vec**

First I have created separate list for offensive and non offensive labels from original data frame.

I have created word2vec embeddings for both two list separately. On the basis of learned vocab of both I have calculated frequency of each word in their respective vocab and retrieved the top 10 words of each file.

I have created two document matrix for word embedding for each file for top 10 words.

Each matrix dimension is having 10 by 100 size. As each vector is of 100 for the embedding. I have calculated mean of each vector in embedding matrix resulting in two list of each consists of 10 mean points.

Each list is of dimension 1 by 10. I have applied cosine similarity coming out to be as follows:

**My list1 is offensive and list2 is not offensive for the following results:**

List 1 ['liberals', 'gun', 'like', 'control', 'people', 'antifa', 'shit', 'maga', 'conservatives', 'trump']

List 2 ['liberals', 'antifa', 'gun', 'control', 'maga', 'conservatives', 'like', 'people', 'trump', 'one']

**cosvalue = compare\_texts\_word2vec(file\_one1, file\_one2, 10)**

**cosvalue**

Cosine Value:

-0.9919267062728845

Cosine similarity is coming out to be negative in my case and majority of words common between twolist.

It can be concluded that majority of words are common between two list but along with oth

For second question we have created two similiarity matrices in which for each ten word, 10 similar words are over there for the respective file.

Each matrix is of 10 by 10 size.

Similar words for each respective file for each top 10 words in the vocab

[['conservatives',

'trump',

'like',

'women',

'maga',

'way',

'left',

'could',

'antifa',

'man'],

['laws',

'need',

'country',

'nothing',

'would',

'people',

'want',

'go',

'also',

'us'],

['antifa',

'conservatives',

'get',

'going',

'liberals',

'time',

'people',

'know',

'president',

'us'],

['laws',

'law',

'also',

'violence',

'right',

'liberals',

'want',

'nothing',

'country',

'stop'],

['like',

'know',

'trump',

'time',

'nothing',

'country',

'us',

'go',

'get',

'shit'],

['like',

'time',

'conservatives',

'get',

'country',

'still',

'going',

'liberals',

'never',

'could'],

['like',

'conservatives',

'trump',

'maga',

'antifa',

'left',

'people',

'never',

'going',

'liberals'],

['conservatives',

'liberals',

'us',

'get',

'trump',

'women',

'shit',

'like',

'could',

'going'],

['trump',

'liberals',

'like',

'maga',

'going',

'antifa',

'women',

'us',

'left',

'could'],

['conservatives',

'liberals',

'women',

'never',

'people',

'like',

'maga',

'us',

'get',

'one']]

List of 10 by 10 for second document

[['way',

'may',

'conservatives',

'always',

'also',

'see',

'still',

'women',

'big',

'really'],

['laws',

'strict',

'sensible',

'guns',

'legislation',

'chicago',

'control',

'nra',

'policies',

'california'],

['way',

'going',

'still',

'know',

'trying',

'man',

'liberals',

'much',

'since',

'antifa'],

['laws',

'strict',

'sensible',

'guns',

'nra',

'sense',

'violence',

'chicago',

'safety',

'california'],

['new',

'could',

'country',

'time',

'many',

'would',

'really',

'keep',

'know',

'part'],

['look',

'say',

'conservative',

'still',

'today',

'even',

'party',

'someone',

'like',

'conservatives'],

['many',

'got',

'kavanaugh',

'fact',

'yet',

'anyone',

'trying',

'money',

'see',

'point'],

['kag',

'trump',

'president',

'qanon',

'walkaway',

'please',

'follow',

'thank',

'great',

'let'],

['liberals',

'conservative',

'point',

'someone',

'world',

'say',

'made',

'always',

'party',

'going'],

['president',

'please',

'america',

'twitter',

'supporters',

'great',

'got',

'obama',

'kag',

'help']]

uniquelist1 = (flat1.difference(overlapping))

Unique list in one file:

{'right', 'law', 'get', 'nothing', 'us', 'need', 'maga', 'go', 'one', 'never', 'shit', 'stop', 'left', 'people', 'want'}

uniquelist2 = (flat2.difference(overlapping))

{'strict', 'supporters', 'follow', 'kavanaugh', 'guns', 'always', 'fact', 'help', 'anyone', 'someone', 'america', 'say', 'yet', 'party', 'keep', 'new', 'policies', 'money', 'thank', 'nra', 'twitter', 'since', 'qanon', 'trying', 'point', 'really', 'made', 'part', 'california', 'kag', 'great', 'see', 'conservative', 'please', 'legislation', 'sense', 'look', 'got', 'chicago', 'control', 'even', 'may', 'walkaway', 'many', 'big', 'world', 'obama', 'sensible', 'much', 'today', 'safety', 'let'}

Overlapping

{'also', 'antifa', 'conservatives', 'could', 'country', 'going', 'know', 'laws', 'liberals', 'like', 'man', 'president', 'still', 'time', 'trump', 'violence', 'way', 'women', 'would'}

There are 19 overlapping words between the two files. Therefore, it signifies two files are have something in common that is depicted by cosine similarity of the two files.

**cosvalue = compare\_texts\_word2vec(file\_one1, file\_one2, 5)**

**cosvalue**

['liberals', 'gun', 'like', 'control', 'people']

['liberals', 'antifa', 'gun', 'control', 'maga']

0.9883611747038017

Doc 1 similar words:

[['like', 'conservatives', 'women', 'would', 'country', 'going', 'people', 'one', 'get', 'still'], ['country', 'laws', 'go', 'need', 'like', 'away', 'people', 'maga', 'want', 'good'], ['liberals', 'one', 'people', 'antifa', 'still', 'conservatives', 'democrats', 'could', 'left', 'get'], ['laws', 'every', 'want', 'support', 'like', 'great', 'us', 'liberals', 'people', 'many'], ['like', 'democrats', 'liberals', 'antifa', 'get', 'also', 'say', 'let', 'believe', 'see']]

Doc 2 similar words

[['get', 'liberal', 'party', 'going', 'got', 'think', 'conservatives', 'still', 'say', 'used'], ['laws', 'strict', 'chicago', 'guns', 'sensible', 'common', 'control', 'nra', 'california', 'rights'], ['still', 'one', 'people', 'also', 'nothing', 'even', 'could', 'would', 'really', 'liberals'], ['laws', 'strict', 'gun', 'guns', 'sensible', 'chicago', 'nra', 'safety', 'support', 'california'], ['also', 'democrats', 'new', 'see', 'work', 'government', 'big', 'liberals', 'nothing', 'saying']]

Overlapping:

'also', 'conservatives', 'could', 'democrats', 'get', 'going', 'laws', 'liberals', 'one', 'people', 'say', 'see', 'still', 'support', 'would'}

Unique in list1:

{'women', 'want', 'go', 'maga', 'antifa', 'away', 'believe', 'left', 'need', 'good', 'country', 'let', 'many', 'us', 'every', 'like', 'great'}

Unique in list2:

{'new', 'chicago', 'even', 'rights', 'really', 'guns', 'got', 'big', 'government', 'sensible', 'saying', 'used', 'strict', 'party', 'work', 'nra', 'liberal', 'control', 'gun', 'california', 'safety', 'think', 'common', 'nothing'}

**cosvalue = compare\_texts\_word2vec(file\_one1, file\_one2, 20)**

**cosvalue**

List1 ['liberals', 'gun', 'like', 'control', 'people', 'antifa', 'shit', 'maga', 'conservatives', 'trump', 'get', 'know', 'one', 'think', 'go', 'fuck', 'would', 'ass', 'good', 'right']

List2 ['liberals', 'antifa', 'gun', 'control', 'maga', 'conservatives', 'like', 'people', 'trump', 'one', 'get', 'know', 'think', 'would', 'right', 'good', 'us', 'time', 'love', 'want']

0.9685318076084142

Doc1:

[['like', 'conservatives', 'women', 'would', 'country', 'going', 'people', 'one', 'get', 'still'], ['country', 'laws', 'go', 'need', 'like', 'away', 'people', 'maga', 'want', 'good'], ['liberals', 'one', 'people', 'antifa', 'still', 'conservatives', 'democrats', 'could', 'left', 'get'], ['laws', 'every', 'want', 'support', 'like', 'great', 'us', 'liberals', 'people', 'many'], ['like', 'democrats', 'liberals', 'antifa', 'get', 'also', 'say', 'let', 'believe', 'see'], ['like', 'get', 'one', 'think', 'women', 'country', 'conservatives', 'liberal', 'antifa', 'people'], ['country', 'like', 'maga', 'go', 'laws', 'need', 'make', 'away', 'good', 'people'], ['liberals', 'one', 'antifa', 'get', 'people', 'still', 'left', 'could', 'conservatives', 'said'], ['laws', 'people', 'want', 'stupid', 'stop', 'every', 'like', 'great', 'nothing', 'support'], ['like', 'democrats', 'liberals', 'get', 'antifa', 'also', 'think', 'one', 'see', 'president'], ['like', 'liberals', 'think', 'could', 'let', 'get', 'know', 'never', 'people', 'conservatives'], ['democrats', 'liberals', 'conservatives', 'one', 'like', 'get', 'antifa', 'back', 'think', 'country'], ['trump', 'said', 'liberals', 'nothing', 'like', 'president', 'much', 'conservatives', 'going', 'get'], ['us', 'liberals', 'one', 'know', 'like', 'could', 'get', 'democrats', 'still', 'antifa'], ['maga', 'liberals', 'think', 'say', 'anyone', 'conservatives', 'antifa', 'good', 'still', 'country'], ['like', 'liberals', 'think', 'left', 'one', 'still', 'antifa', 'got', 'conservatives', 'people'], ['conservatives', 'antifa', 'like', 'man', 'liberals', 'get', 'left', 'back', 'one', 'still'], ['like', 'liberals', 'still', 'conservatives', 'think', 'get', 'us', 'said', 'democrats', 'country'], ['liberals', 'get', 'antifa', 'one', 'trump', 'people', 'stop', 'good', 'like', 'still'], ['like', 'left', 'liberals', 'good', 'women', 'say', 'get', 'maga', 'bad', 'one'], ['like', 'country', 'get', 'left', 'also', 'conservatives', 'good', 'back', 'said', 'antifa'], ['could', 'liberals', 'like', 'antifa', 'want', 'get', 'never', 'one', 'nothing', 'trump'], ['still', 'like', 'conservatives', 'liberals', 'said', 'country', 'see', 'know', 'antifa', 'one'], ['country', 'think', 'trump', 'like', 'liberals', 'say', 'one', 'go', 'get', 'president'], ['left', 'liberals', 'think', 'get', 'even', 'time', 'country', 'conservatives', 'antifa', 'trump']]

Doc2:

[['get', 'liberal', 'party', 'going', 'got', 'think', 'conservatives', 'still', 'say', 'used'], ['laws', 'strict', 'chicago', 'guns', 'sensible', 'common', 'control', 'nra', 'california', 'rights'], ['still', 'one', 'people', 'also', 'nothing', 'even', 'could', 'would', 'really', 'liberals'], ['laws', 'strict', 'gun', 'guns', 'sensible', 'chicago', 'nra', 'safety', 'support', 'california'], ['also', 'democrats', 'new', 'see', 'work', 'government', 'big', 'liberals', 'nothing', 'saying'], ['liberal', 'party', 'conservatives', 'going', 'got', 'think', 'used', 'still', 'trying', 'would'], ['laws', 'strict', 'sensible', 'guns', 'nra', 'chicago', 'control', 'shooting', 'california', 'support'], ['one', 'still', 'people', 'also', 'liberals', 'show', 'nothing', 'got', 'would', 'use'], ['laws', 'strict', 'sensible', 'chicago', 'gun', 'guns', 'nra', 'strictest', 'safety', 'support'], ['also', 'democrats', 'new', 'see', 'one', 'still', 'would', 'think', 'believe', 'everyone'], ['party', 'never', 'liberals', 'right', 'also', 'conservatives', 'think', 'always', 'let', 'said'], ['use', 'democrats', 'thought', 'big', 'thing', 'america', 'make', 'liberal', 'someone', 'time'], ['trump', 'kag', 'walkaway', 'qanon', 'president', 'follow', 'thank', 'twitter', 'tcot', 'kavanaugh'], ['party', 'today', 'liberals', 'still', 'right', 'us', 'got', 'liberal', 'time', 'also'], ['president', 'maga', 'thank', 'twitter', 'america', 'kavanaugh', 'usa', 'best', 'kag', 'anything'], ['another', 'got', 'still', 'also', 'one', 'every', 'think', 'long', 'saying', 'even'], ['still', 'man', 'long', 'always', 'well', 'also', 'since', 'never', 'woman', 'made'], ['still', 'also', 'show', 'even', 'another', 'could', 'got', 'get', 'going', 'think'], ['always', 'still', 'got', 'get', 'good', 'one', 'liberals', 'trying', 'also', 'someone'], ['another', 'would', 'still', 'going', 'everyone', 'say', 'get', 'bad', 'yet', 'good'], ['soros', 'first', 'also', 'show', 'business', 'since', 'problem', 'strong', 'saying', 'making'], ['never', 'could', 'also', 'even', 'someone', 'still', 'long', 'nothing', 'see', 'liberals'], ['thing', 'twitter', 'home', 'today', 'still', 'see', 'come', 'since', 'children', 'big'], ['think', 'everyone', 'say', 'even', 'still', 'get', 'trying', 'go', 'one', 'would'], ['many', 'left', 'party', 'first', 'time', 'even', 'conservatives', 'get', 'think', 'true']]

Overlapping:

{'also', 'bad', 'believe', 'conservatives', 'could', 'democrats', 'even', 'every', 'get', 'go', 'going', 'good', 'got', 'laws', 'left', 'let', 'liberal', 'liberals', 'maga', 'make', 'man', 'many', 'never', 'nothing', 'one', 'people', 'president', 'said', 'say', 'see', 'still', 'support', 'think', 'time', 'trump', 'us', 'would'}

Unique in one list1:

{'want', 'anyone', 'antifa', 'away', 'stupid', 'need', 'much', 'know', 'country', 'back', 'stop', 'women', 'like', 'great'}

Unique in second list:

{'new', 'chicago', 'business', 'come', 'walkaway', 'always', 'thought', 'children', 'true', 'rights', 'making', 'qanon', 'yet', 'anything', 'really', 'guns', 'show', 'follow', 'big', 'government', 'someone', 'america', 'right', 'sensible', 'woman', 'saying', 'kag', 'used', 'strict', 'since', 'first', 'strong', 'party', 'kavanaugh', 'another', 'work', 'nra', 'shooting', 'home', 'made', 'long', 'strictest', 'gun', 'control', 'california', 'everyone', 'thank', 'usa', 'thing', 'safety', 'tcot', 'use', 'twitter', 'well', 'best', 'today', 'trying', 'problem', 'soros', 'common'}

**Observations:**

**The top occurrence words are common among both files most of the words are common except few. The commonality of overlapping words show that many words are common among two files and unique words are less within each file. Therefore there exist some similiarity among documents in terms of content but it might be the case that few occurrence of negative words which is not having enough frequency and unique in document making the doc to get negative label.**

**Similarity index value is astonishing to me as I was expecting it to be zero but I think context of the documents is same but occurrence of few negative words making it negative.**

**Related context words are common between the two. The offensive list also has some positive words.**