**Observations Criteria:**

Note:

1. Cosine similarity angle between the two files has been calculated using the cosine similarity formula
2. **preprocessed\_data** was used for calculating the cosine similarity. To understand to what extent the data pre-processing can help in improving the cosine similarity.
3. While doing the data pre-processing, stop words have been removed and the second level of stop word removal is implemented during k’s most common non-stop word detection. This improved the cosine similarity a little bit to some extent.

* **When K = 5**

**The cosine similarity between the two files is**: 0.870228 (radians)

**File one Unique Word's List:**

['ganda', 'Ang\_dami', 'nong', 'sayang', 'kayo\_sa', 'ang\_Pinoy', 'Pwede', 'naniniwala\_sa', 'yong\_mga', 'bagay\_sa']

**File two Unique Word's List:**

['really', 'weird', 'crazy', 'kind', 'maybe', 'loooove', 'anymore', 'Kinda\_reminds', 'definitely', 'kinda\_fishy', 'peole', 'poeple', 'individuals', 'folks', 'peple', 'peo\_ple', 'peope', 'citizens', 'Americans', 'People']

**Overlapped words between both the files are:**

{'rifle', 'Tea\_Party\_adherents', 'Democrats', 'Tea\_Partiers', 'controlled', 'grip', 'liberal\_elitists', 'handgun', 'ideologues', 'contol', 'statists', 'controls', '9mm\_handgun', 'Controlling', '9\_mm\_handgun', 'Control', 'liberal', 'pistol', 'con\_trol', 'controllable', '.##\_caliber\_handgun', 'Republicans', 'conrol', 'firearms', 'controlling', 'handguns', 'conservatives', 'progressives', 'firearm', 'guns'}

* **When K = 10**

**The cosine similarity between the two files is**: 0.487700 (radians)

**File one Unique Word's List:**

['trumps', 'trumping', 'supersede', 'trumped', 'supercede', 'prevail', 'outweigh', 'trample', 'overshadow', 'dictate', 'tell', 'think', 'do', 'understand', 'guess', 'tomorrow\_Camardelle', 'knew', "don'tknow", 'say', 'Nobody\_cares']

**File two Unique Word's List:**

['sh\_\*\_t', 'crap', 'fucking', 'sh\_t', 'fuckin', 'sh\_\*\*', 'fuck', 'f\_\*\_cking', "y'know", 'shyt', 'amps', 'guitar\_amp', 'amplifier', 'preamp', 'tube\_amp', 'Fender\_Bassman', 'Fender\_Twin', 'watt\_amp', 'subwoofer', 'amplifiers']

**Overlapped words between both the files are:**

{'rifle', 'Tea\_Party\_adherents', 'poeple', 'Democrats', 'Americans', 'Tea\_Partiers', 'sayang', 'grip', 'controlled', 'Kinda\_reminds', 'weird', 'liberal\_elitists', 'handgun', 'ideologues', 'contol', 'statists', 'ang\_Pinoy', 'kinda\_fishy', 'loooove', 'really', 'controls', 'kind', '9mm\_handgun', 'liberals', 'Tea\_Party\_activists', 'crazy', '9\_mm\_handgun', 'Control', 'Controlling', 'peple', 'bagay\_sa', 'nong', 'conservatism', 'yong\_mga', 'liberal', 'naniniwala\_sa', 'pistol', 'con\_trol', 'definitely', 'controllable', 'conservative', '.##\_caliber\_handgun', 'kayo\_sa', 'evangelicals', 'peo\_ple', 'Ang\_dami', 'maybe', 'peope', 'Republicans', 'ganda', 'conrol', 'firearms', 'anymore', 'controlling', 'folks', 'handguns', 'conservatives', 'progressives', 'Pwede', 'firearm', 'People', 'individuals', 'peole', 'guns', 'citizens'}

* **When K = 20**

**The cosine similarity between the two files is:** 0.397355 (radians)

**File one Unique Word's List:**

['loved', 'adore', 'loves', 'passion', 'hate', 'loving', 'Ilove', 'affection', 'undying\_love', 'absolutely\_adore', 'needed', 'needs', 'must', 'needing', 'should', 'needto', 'necessary', 'ways', 'how', 'thing', 'things', 'path', 'everything', 'someway', 'it', 'manner']

**File two Unique Word's List:**

['sh\_\*\_t', 'crap', 'fucking', 'sh\_t', 'fuckin', 'sh\_\*\*', 'fuck', 'f\_\*\_cking', "y'know", 'shyt', 'f\_\*\_ck', 'f\_\*\*\_k', 'shit', 'fucked', 'f\_\*\*\*', 'f\_ck', 'f\_---', 'Fuck', 'butt', 'azz', 'asses', 'dick', 'hiney', 'dude', 'd\_\*\_ck', 'nads', '\*\*\_es']

**Overlapped words between both the files are:**

{'excellent', 'poeple', 'Americans', 'weeks', 'suppose', 'grip', 'trumps', 'months', 'ideologues', 'nice', 'need', 'day', 'controls', 'really', 'thought', '9\_mm\_handgun', 'bagay\_sa', 'nong', 'understand', 'liberal', 'naniniwala\_sa', 'now', 'con\_trol', 'controllable', 'bad', 'tell', 'overshadow', 'peope', 'can', 'gotta', 'ganda', 'anymore', 'goingto', 'coming', 'fielder\_Jeromy\_Burnitz', 'progressives', 'amps', 'solid', 'peole', 'guns', 'Tea\_Party\_adherents', 'Democrats', 'controlled', 'prefer', 'liberal\_elitists', 'handgun', 'trumping', 'kinda\_fishy', 'Right', 'kind', 'Nobody\_cares', 'periods', 'Controlling', 'Control', 'crazy', '##.Help\_us', 'fielder\_Ambiorix\_Concepcion', 'do', 'peple', 'going', 'definitely', 'kayo\_sa', 'evangelicals', 'peo\_ple', 'know', 'guess', 'Ang\_dami', 'maybe', 'hours', 'guitar\_amp', 'conrol', 'firearms', 'wanted', 'trumped', 'firearm', 'citizens', 'wrong', 'rifle', 'Fender\_Twin', 'supersede', 'sayang', 'dictate', 'Goodwill\_Catanese', 'weird', 'contol', 'ang\_Pinoy', 'think', 'subwoofer', 'liberals', 'Tea\_Party\_activists', 'say', 'fielder\_Joe\_Borchard', 'go', 'knew', 'conservatism', "don'tknow", 'better', 'great', 'times', '.##\_caliber\_handgun', 'outweigh', 'probably', 'fantastic', 'tube\_amp', 'trample', 'lousy', "'ll", 'prevail', 'controlling', 'folks', 'wanting', 'amplifier', 'conservatives', 'Pwede', 'wants', 'People', 'supercede', 'ROGER\_FEDERER\_Yeah', 'Tea\_Partiers', 'gonna', 'Kinda\_reminds', 'decent', 'statists', 'Fender\_Bassman', 'loooove', '9mm\_handgun', 'when', 'NOTE\_Xactly\_Incent', 'believe', 'preamp', 'yong\_mga', 'pistol', 'conservative', 'want', 'left', 'moment', 'period', 'tomorrow\_Camardelle', 'days', 'wanna', 'obviously', 'watt\_amp', 'wantto', 'amplifiers', 'Republicans', 'terrific', 'handguns', 'individuals'}

**SUMMARY OF ANALYSIS OF RESULTS:**

1. Overall, it seems that the trained word2vec model has captured the training dataset well. With the increase in the number of the k most common non-stop words the cosine similarity angle is dropping.
2. Smaller the angle between two file vectors higher would be the similarity between the two documents.
3. When k is 5, the trained Word2Vec model predicts a higher cosine angle of 0.87 which means the documents are not that similar
4. When the number of most common words gets doubled(k=10), the cosine similarity angle magically dropped to half 0.48, which indicates that the word2vec model performs well with the given set of information.
5. With k=20, the cosine similarity angle further dropped to 0.39, depicting more similarity between the two documents as a larger number of words are taken into consideration.
6. By observing the unique word lists for each of the k most common words, 5, 10, and 20, it can be assumed that most of the words in that list is a not vocabulary words, with special characters in them, which can be removed by using the advanced pre-processing techniques.
7. However, the overlapped list of words is getting bigger with the increase in the k most common words list. This shows the more overlapping words the more the similar two documents could be. This also happened by leveraging the word2vec to find the 10 most similar words for each k-word.
8. In the overlapping word list too, which is created using word2vec, includes words that are not vocabulary words still they are in the list, this is noticeable factor.