Documents:

Document 1: The Ball and the Cross by G. K. Chesterton

Data definition and collection:

The book was directly retrieved from the gutenberg nltk library and the raw() function was used for the Data Extraction.

```
emmatext = nltk.corpus.gutenberg.raw(file8)
```

The next step was to convert the text into tokens. Since the book was directly taken from the library the preprocessing of the data was already done before tokenization.

Preprocessing of the Data:

1. Conversion to the lowercase was the first step in the process of preprocessing the data because words such as “DAY” and “day” would hold the same meaning but they are in different cases.
2. Removal of the nonalphabetic characters is the next step as they don't contribute much to the analysis. However, there are some more words that hold no value to the analysis, such as filler words.
3. Such stop words were removed from the data by filtering them with a unique list of stop words and a list of pre-defined nltk stop words.

Document 2: Wildfire related tweets

Data definition and collection:
Tweepy library from python was used for data collection.

```
# Collect tweets
tweets = tw.Cursor(api.search,
              q=search_words,
              lang="en",
              since=date_since).items(5)

# Iterate and print tweets
for tweet in tweets:
    print(tweet.text)
```

```
[ ] ltweets[:10]

['Chemical aging of particles from #wildfires can lead to cloud formation and intense storm development. @CAPS_CMU… https://t.co/44oiWDpWTG',
 'RT @LancashireFRS: Wildfires are not only a problem in summer 🔥🚨🚒\n\n#ClimateChange means we are seeing more #wildfires in spring than ever b…',
 'RT @CLEANBOSS2: Effects of #global #climatechange include frequent #wildfires #droughts intensity of #tropical #storms.Global #sealevel ris…',
 'RT @CLEANBOSS2: Effects of #global #climatechange include frequent #wildfires #droughts intensity of #tropical #storms.Global #sealevel ris…',
 'RT @CLEANBOSS2: Effects of #global #climatechange include frequent #wildfires #droughts intensity of #tropical #storms.Global #sealevel ris…',
 'Uncovering patterns in California's blazing wildfires\n#SantaAnaWinds \n#Wildfires\n#Mudslides\n#California\n#LosAngeles… https://t.co/RWdsI8dznH',
 'RT @LancashireFRS: Wildfires are not only a problem in summer 🔥🚨🚒\n\n#ClimateChange means we are seeing more #wildfires in spring than ever b…',
 'RT @LancashireFRS: Wildfires are not only a problem in summer 🔥🚨🚒\n\n#ClimateChange means we are seeing more #wildfires in spring than ever b…',
 'Wildfires are not only a problem in summer 🔥🚨🚒\n\n#ClimateChange means we are seeing more #wildfires in spring than e… https://t.co/jRwcO9RY7I',
 'RT @sahoo_biswaj: Save #Mayurbhanj #wildfires #Odisha https://t.co/A4WjivDIDI']
```

Tweepy Library

1. Even though tweepy library is a good tool for the collection of tweets, but it has its limitations.
2. Twitter API has a lot of restrictions such as only 3200 tweets can be collected

Data Preprocessing

1. We use nltk tweet tokenizer for tweet tokenization
2. We complete the preprocessing by following the identical steps of removing non alphanumeric words and converting the words to lowercase.

Results:

Document 1: The Ball and the Cross by G. K. Chesterton

T**he unprepossessed top 50 words ;**

Note:The **‘,’** (comma) is a most frequent word which is natural since the text is a part of literature it would consist of punctuations.

```
(',', 0.053732909478102525)
('the', 0.051027231669804426)
('.', 0.04112013003713877)
('and', 0.027396273777558305)
('of', 0.026274909210621072)
('a', 0.023219447959425 12)
("'", 0.019351254590907 69)
('``', 0.016563274796045 39)
('to', 0.016254642346429638)
('in', 0.014433710893696697)
('he', 0.013641554273016264)
('that', 0.012005802290052776)
('it', 0.011686882092116 5)
('you', 0.010966739709679742)
('i', 0.010267172823884037)
('was', 0.009557318189767806)
```

```
('his', 0.009464728454883 08)
('with', 0.008384514881227946)
('is', 0.0074791930290217384)
('as', 0.007119121837803 36)
('said', 0.006707611904982357)
(';', 0.00655329568017448)
('but', 0.005750851311173523)
('had', 0.005719988066211948)
('turnbull', 0.005596535086365647)
('not', 0.005329053630031995)
('at', 0.004763227472403115)
('they', 0.004680925485838914)
('on', 0.004650062240877339)
('for', 0.004639774492556814)
('macian', 0.004372293036223162)
('be', 0.004197401314774235)
('?', 0.004187113566453 7104)
('or', 0.004104811579889 5096)
('all', 0.003981358600043209)
('have', 0.003960783103402158)
('--', 0.003899056613479 0077)
('this', 0.003806466878594282)
('him', 0.003487546680658 0046)
('man', 0.003456683435696429)
('we', 0.003456683435696429)
('if', 0.003456683435696429)
('are', 0.003425820190734 8537)
('like', 0.003353805952491 1784)
('which', 0.003343518204170 6532)
('were', 0.003261216217606 4523)
('an', 0.003250928469285 9275)
('one', 0.003250928469285 9275)
('up', 0.003024598006234 3753)
('by', 0.003004022509593 3253)
```

Non Alphanumeric filtered for top 50 words:

```
('the', 0.060229991135505 34)
('and', 0.032337190805211836)
('of', 0.031013588177435 61)
('a', 0.027407074595329747)
('to', 0.019186166531068232)
('in', 0.017036830153853627)
('he', 0.016101808114048 4)
('that', 0.014171048317567485)
('it', 0.013794610872970577)
('you', 0.012944590836784 01)
('i', 0.012118857087345631)
('was', 0.011280980194533156)
```

```
('his', 0.011171691904166312)
('with', 0.00989666184988646)
('is', 0.008828065232966205)
('as', 0.008403055214872923)
('said', 0.007917329479909169)
('but', 0.006788017146118444)
('had', 0.006751587715996162)
('turnbull', 0.006605869995507037)
('not', 0.006290148267780598)
('at', 0.005622275382205438)
('they', 0.0055251302352126875)
('on', 0.005488700805090406)
('for', 0.005476557661716312)
('macian', 0.005160835933989872)
('be', 0.004954402496630278)
('or', 0.004845114206263434)
('all', 0.004699396485774308)
('have', 0.004675110199026l2)
('this', 0.004492963048414712)
('him', 0.004116525603817804)
('man', 0.004080096173695522 4)
('we', 0.004080096173695522 4)
('if', 0.004080096173695522 4)
('are', 0.004043666743573242)
('like', 0.003958664739954584)
('which', 0.0039465215965804904)
('were', 0.00384937644958774)
('an', 0.0038372333062136467)
('one', 0.0038372333062136467)
('up', 0.0035700841519835827)
('by', 0.003545797865235395)
("'s", 0.003460795861616738)
('do', 0.003424366431494457)
('there', 0.0034400080144746269)
('what', 0.003242219280883049)
('out', 0.003011499556775267)
('from', 0.002987213270027079)
('then', 0.002890068123034329)
```

Top 50 word with Stop words filtered:

```
('said', 0.007917329479909169)
('turnbull', 0.006605869995507037)
('macian', 0.005160835933989872)
('man', 0.004080096173695522 4)
('like', 0.003958664739954584)
('one', 0.0038372333062136467)
("n't", 0.0021614795205886998)
```

```
('two', 0.0021493363772146060)
('us', 0.0019671892266031987)
('quite', 0.0017486126458695099)
('even', 0.0017000400723731345)
('evan', 0.0017000400723731345)
('know', 0.0016878969289990406)
('god', 0.0016514674988767593)
('face', 0.0016271812121285715)
('little', 0.0015664654952581025)
('say', 0.001542179208509915)
('seemed', 0.0014693203482653520)
('see', 0.0014450340615171644)
('really', 0.0014408604631394883)
('thing', 0.0013721752012726015)
('world', 0.0013600320578985076)
('something', 0.0013600320578985076)
('still', 0.0012750300542798510)
('looked', 0.0012628869109057570)
('come', 0.0012628869109057570)
('made', 0.0012507437675316633)
('men', 0.0012386006241575695)
('voice', 0.0012386006241575695)
('long', 0.0012143143374093819)
('well', 0.0011900280506611940)
('cried', 0.0011778849072871004)
('may', 0.0011778849072871004)
('upon', 0.0011535986205389126)
('went', 0.0011535986205389126)
('back', 0.0011414554771648190)
('almost', 0.0011414554771648190)
('think', 0.0011414554771648190)
('old', 0.0011171691904166314)
('saw', 0.0011171691904166314)
('head', 0.0011050260470425375)
('came', 0.0010928829036684436)
('time', 0.0010564534735461620)
('eyes', 0.0010564534735461620)
('garden', 0.0010443103301720684)
('first', 0.0010443103301720684)
('mr.', 0.0010321671867979745)
('last', 0.0009957377566756930)
('mean', 0.0009957377566756930)
('sort', 0.0009835946133015994)
```

Observation:The top 50 words where stop words are filtered out have the most amount of words.

Bigram Analysis:

Not Processed and Raw frequency:

```
(('.', '``'), 0.007746674485355391)
((',', "''"), 0.007016244354598109)
((',', 'and'), 0.0063681162104050285)
(('of', 'the'), 0.0063269652171229285)
(('.', "''"), 0.005647973827968273)
(('in', 'the'), 0.004135674824851085)
(("''", 'said'), 0.0039813586000043209)
(("''", '``'), 0.0038064668678594282)
(('?', "''"), 0.0030040225095933253)
(('.', 'the'), 0.0029320082713496496)
((',', '``'), 0.002859994033105974)
(('``', 'i'), 0.0025102105902081212)
(('.', 'he'), 0.0024073331070028703)
(('with', 'a'), 0.002294167875477094)
(("''", 'he'), 0.002108988405707643)
(('and', 'the'), 0.0019958231741818667)
(('turnbull', ','), 0.0019958231741818667)
(('it', 'was'), 0.0019649599292202915)
(('in', 'a'), 0.0019546721808997663)
(('it', 'is'), 0.0018415069493739906)
(('.', 'but'), 0.0018209314527329404)
(('to', 'the'), 0.0018209314527329404)
((',', 'the'), 0.0017180539695276895)
((',', 'but'), 0.0016871907245661142)
((',', 'with'), 0.0016357519829634888)
(('of', 'a'), 0.0016254642346429638)
(('he', 'said'), 0.0015328744997582379)
(('macian', ','), 0.0015225867514377129)
(('he', 'had'), 0.0015020112547966626)
(('he', 'was'), 0.0014917235064761376)
(('.', 'it'), 0.0014505725131940372)
(('with', 'the'), 0.0014197092682324620)
(('!', "''"), 0.0013991337715914117)
(('.', 'i'), 0.0013682705266298365)
(('at', 'the'), 0.0013682705266298365)
(('on', 'the'), 0.0012962562883861608)
(('the', 'other'), 0.0012859685400656358)
(('to', 'be'), 0.0012859685400656358)
((',', 'as'), 0.0012756807917451108)
(('i', 'am'), 0.0012756807917451108)
(('like', 'a'), 0.0012551052951040606)
(('said', 'turnbull'), 0.0012345297984630104)
(('said', 'the'), 0.0012242420501424854)
((';', 'and'), 0.0012139543018219602)
((',', 'i'), 0.0011419400635782847)
(('all', 'the'), 0.0011007890702961843)
```

```
((',', 'in'), 0.0010905013219756593)
((':', '``'), 0.0010390625803730338)
(('said', ','), 0.0010287748320525086)
(('of', 'his'), 0.0010184870837319836)
```

Not filtered PMI:

```
(("'ampstead", "'eath"), 16.568713220318084)
(('10', '1991'), 16.568713220318084)
(('1000', '1997'), 16.568713220318084)
(('10000', '2004'), 16.568713220318084)
(('113', '1739'), 16.568713220318084)
(('1500', '1998'), 16.568713220318084)
(('1739', 'university'), 16.568713220318084)
(('1971', 'july'), 16.568713220318084)
(('1997', 'august'), 16.568713220318084)
(('1998', 'october'), 16.568713220318084)
(('2003', 'november*'), 16.568713220318084)
(('2004', 'january*'), 16.568713220318084)
(('9000', '2003'), 16.568713220318084)
(('_altiora', 'peto_'), 16.568713220318084)
(('_idee', 'fixe_'), 16.568713220318084)
(('_natura', 'naturans_'), 16.568713220318084)
(('_vae', 'victis'), 16.568713220318084)
(('_virgin', 'martyr_'), 16.568713220318084)
(('absent', 'minded'), 16.568713220318084)
(('abstractedly', 'boring'), 16.568713220318084)
(('accommodated', 'indoors'), 16.568713220318084)
(('adult', 'reformatory'), 16.568713220318084)
(('adverse', 'theories'), 16.568713220318084)
(('albert', 'memorial'), 16.568713220318084)
(('angler', 'stares'), 16.568713220318084)
(('applaud', 'tyrannicide'), 16.568713220318084)
(('applicable', 'taxes'), 16.568713220318084)
(('architectural', 'arrangement'), 16.568713220318084)
(('assyrian', 'bas-reliefs'), 16.568713220318084)
(('astonishing', 'transparency'), 16.568713220318084)
(('august', '1500'), 16.568713220318084)
(('barren', 'steeps'), 16.568713220318084)
(('bearing', 'witness'), 16.568713220318084)
(('beetle', 'strayed'), 16.568713220318084)
(('ben', 'crowder'), 16.568713220318084)
(('bile', 'beans'), 16.568713220318084)
(('black-and-white', 'beads'), 16.568713220318084)
(('black-haired', 'skull'), 16.568713220318084)
(('blackcoated', 'bustle'), 16.568713220318084)
(('bllcr10.zip', 'corrected'), 16.568713220318084)
(('bllcr11.txt', 'versions'), 16.568713220318084)
```

```
(('boorish', 'moralities'), 16.568713220318084)
(('bourgeois', 'speaks'), 16.568713220318084)
(('bulging', 'bravery'), 16.568713220318084)
(('californian', 'tourist'), 16.568713220318084)
(('captured', 'coach'), 16.568713220318084)
(('cardinal', 'manning'), 16.568713220318084)
(('carthage', '_was_'), 16.568713220318084)
(('chartered', 'lunacies'), 16.568713220318084)
(('constables', 'attended'), 16.568713220318084)
```

Observations:

Since i have not read the book, according to what i can notice that in non processed frequency diagrams,doesn't provide much meaning as most of them are filler words such as is ,the etc.But on the other hand if we consider PMI scoring mechanism it give us some relevant information for example i have highlighted some dates that are mentioned in the book.

Bigram Analysis non Alphanumeric characters filtered;

Bigram Analysis:Non Alphanumeric raw frequency

```
(('of', 'the'), 0.0063269652171229285)
(('in', 'the'), 0.0041356748248510 85)
(('with', 'a'), 0.0022941678754770 94)
(('and', 'the'), 0.0019958231741818667)
(('it', 'was'), 0.0019649599292202915)
(('in', 'a'), 0.0019546721808997663)
(('it', 'is'), 0.0018415069493739906)
(('to', 'the'), 0.0018209314527329404)
(('of', 'a'), 0.0016254642346429638)
(('he', 'said'), 0.0015328744997582379)
(('he', 'had'), 0.0015020112547966626)
(('he', 'was'), 0.0014917235064761376)
(('with', 'the'), 0.0014197092682324 62)
(('at', 'the'), 0.0013682705266298365)
(('on', 'the'), 0.0012962562883861608)
(('the', 'other'), 0.0012859685400656358)
(('to', 'be'), 0.0012859685400656358)
(('i', 'am'), 0.0012756807917451108)
(('like', 'a'), 0.0012551052951040606)
(('said', 'turnbull'), 0.0012345297984630104)
(('said', 'the'), 0.0012242420501424854)
(('all', 'the'), 0.0011007890702961843)
(('of', 'his'), 0.0010184870837319836)
(('as', 'if'), 0.0009876238387704084)
(('and', 'he'), 0.0009773360904498831)
(('is', 'a'), 0.0009361850971677829)
(('in', 'his'), 0.0009053218522062077)
(('was', 'a'), 0.0009053218522062077)
```

```
(('a', 'man'), 0.0008847463555651575)
(('into', 'the'), 0.0008847463555651575)
(('out', 'of'), 0.0008847463555651575)
(('you', 'are'), 0.0008847463555651575)
(('that', 'the'), 0.0008744586072446323)
(('said', 'macian'), 0.0008538831106035822)
(('by', 'the'), 0.0008435953622830571)
(('do', "n't"), 0.0008435953622830571)
(('if', 'you'), 0.0008435953622830571)
(('i', 'have'), 0.0007715811240393815)
(('from', 'the'), 0.0007407178790778062)
(('is', 'the'), 0.0007407178790778062)
(('the', 'same'), 0.0007407178790778062)
(('do', 'you'), 0.0007304301307572812)
(('for', 'the'), 0.0007304301307572812)
(('the', 'world'), 0.0007304301307572812)
(('sort', 'of'), 0.0007201423824367561)
(('that', 'he'), 0.0007201423824367561)
(('and', 'i'), 0.000709854634116231)
(('one', 'of'), 0.000709854634116231)
(('the', 'man'), 0.0006995668857957059)
(('they', 'had'), 0.0006995668857957059)
```

Bigram Analysis:Non Alphanumeric pmi:

```
(("'ampstead", "'eath"), 16.568713220318084)
(('_altiora', 'peto_'), 16.568713220318084)
(('_idee', 'fixe_'), 16.568713220318084)
(('_natura', 'naturans_'), 16.568713220318084)
(('_vae', 'victis'), 16.568713220318084)
(('_virgin', 'martyr_'), 16.568713220318084)
(('absent', 'minded'), 16.568713220318084)
(('abstractedly', 'boring'), 16.568713220318084)
(('accommodated', 'indoors'), 16.568713220318084)
(('adult', 'reformatory'), 16.568713220318084)
(('adverse', 'theories'), 16.568713220318084)
(('albert', 'memorial'), 16.568713220318084)
(('angler', 'stares'), 16.568713220318084)
(('applaud', 'tyrannicide'), 16.568713220318084)
(('applicable', 'taxes'), 16.568713220318084)
(('architectural', 'arrangement'), 16.568713220318084)
(('assyrian', 'bas-reliefs'), 16.568713220318084)
(('astonishing', 'transparency'), 16.568713220318084)
(('barren', 'steeps'), 16.568713220318084)
(('bearing', 'witness'), 16.568713220318084)
(('beetle', 'strayed'), 16.568713220318084)
(('ben', 'crowder'), 16.568713220318084)
(('bile', 'beans'), 16.568713220318084)
(('black-and-white', 'beads'), 16.568713220318084)
(('black-haired', 'skull'), 16.568713220318084)
```

```
(('blackcoated', 'bustle'), 16.568713220318084)
(('bllcr10.zip', 'corrected'), 16.568713220318084)
(('bllcr11.txt', 'versions'), 16.568713220318084)
(('boorish', 'moralities'), 16.568713220318084)
(('bourgeois', 'speaks'), 16.568713220318084)
(('bulging', 'bravery'), 16.568713220318084)
(('californian', 'tourist'), 16.568713220318084)
(('captured', 'coach'), 16.568713220318084)
(('cardinal', 'manning'), 16.568713220318084)
(('carthage', '_was_'), 16.568713220318084)
(('chartered', 'lunacies'), 16.568713220318084)
(('constables', 'attended'), 16.568713220318084)
(('continual', 'tournament'), 16.568713220318084)
(('contributing', 'scanning'), 16.568713220318084)
(('conversational', 'substitute'), 16.568713220318084)
(('corrupt', 'data'), 16.568713220318084)
(('cotton', 'wool'), 16.568713220318084)
(('cyclopean', 'sea-beast'), 16.568713220318084)
(('daggers', 'glinted'), 16.568713220318084)
(('deafening', 'density'), 16.568713220318084)
(('deathly', 'germs'), 16.568713220318084)
(('deepest', 'rabbit-holes'), 16.568713220318084)
(('degradedly', 'diving'), 16.568713220318084)
(("devil's", 'magnifying'), 16.568713220318084)
(('devilish', 'piston'), 16.568713220318084)
```

Observation:
Bigram analysis was not able to find many names using the raw frequency but it consisted of other pairs which were of relevance to the stories in the book.

Bigram Analysis Stop words filtered;

Bigram Analysis:Stop words raw frequency

```
(('said', 'turnbull'), 0.0012345297984630104)
(('said', 'macian'), 0.0008538831106035822)
(('ca', "n't"), 0.0002983447012952275 4)
(('said', 'evan'), 0.0002880569529747 0243)
(('let', 'us'), 0.0002469059596926021)
(('project', 'gutenberg'), 0.00022633046305155192)
(('asked', 'turnbull'), 0.0002160427147310268)
(('mr.', 'turnbull'), 0.0002160427147310268)
(('like', 'one'), 0.00019546721808997664)
(('evan', 'macian'), 0.00018517946976945155)
(('mr.', 'macian'), 0.00018517946976945155)
(('old', 'man'), 0.00018517946976945155)
(('cried', 'turnbull'), 0.00017489172144892647)
(('first', 'time'), 0.00017489172144892647)
```

```
(('young', 'man'), 0.00017489172144892647)
(('flying', 'ship'), 0.00016460397312840138)
(('james', 'turnbull'), 0.0001543162248078763)
(('mr.', 'wilkinson'), 0.0001543162248078763)
(('_the', 'atheist_'), 0.00014402847648735121)
(('little', 'man'), 0.00014402847648735121)
(('cumberland', 'vane'), 0.00012345297984630104)
(('dr.', 'quayle'), 0.00012345297984630104)
(('answered', 'turnbull'), 0.00011316523152577596)
(('blue', 'eyes'), 0.00011316523152577596)
(('looked', 'like'), 0.00011316523152577596)
(('modern', 'world'), 0.00011316523152577596)
(('turnbull', 'said'), 0.00011316523152577596)
(('answered', 'macian'), 0.00010287748320525086)
(('asked', 'macian'), 0.00010287748320525086)
(('ludgate', 'hill'), 0.00010287748320525086)
(('swung', 'round'), 0.00010287748320525086)
(('cried', 'macian'), 9.258973488472578e-05)
(('every', 'man'), 9.258973488472578e-05)
(('fleet', 'street'), 9.258973488472578e-05)
(('macian', 'said'), 9.258973488472578e-05)
(('repeated', 'turnbull'), 9.258973488472578e-05)
(('count', 'gregory'), 8.230198656420069e-05)
(('macian', 'looked'), 8.230198656420069e-05)
(("n't", 'know'), 8.230198656420069e-05)
(("n't", 'want'), 8.230198656420069e-05)
(('replied', 'turnbull'), 8.230198656420069e-05)
(('small', 'print'), 8.230198656420069e-05)
(('ten', 'minutes'), 8.230198656420069e-05)
(('turnbull', 'looked'), 8.230198656420069e-05)
(('two', 'men'), 8.230198656420069e-05)
(('cried', 'evan'), 7.201423824367561e-05)
(('madeleine', 'durand'), 7.201423824367561e-05)
(('old', 'gentleman'), 7.201423824367561e-05)
(('professor', 'lucifer'), 7.201423824367561e-05)
(('project', 'gutenberg-tm'), 7.201423824367561e-05)
```

Bigram Analysis:Stopped words pmi frequency:

```
(("'ampstead", "'eath"), 16.568713220318084)
(('10', '1991'), 16.568713220318084)
(('1000', '1997'), 16.568713220318084)
(('10000', '2004'), 16.568713220318084)
(('113', '1739'), 16.568713220318084)
(('1500', '1998'), 16.568713220318084)
(('1739', 'university'), 16.568713220318084)
(('1971', 'july'), 16.568713220318084)
(('1997', 'august'), 16.568713220318084)
(('1998', 'october'), 16.568713220318084)
(('2003', 'november*'), 16.568713220318084)
```

```
(('2004', 'january*'), 16.568713220318084)
(('9000', '2003'), 16.568713220318084)
(('_altiora', 'peto_'), 16.568713220318084)
(('_idee', 'fixe_'), 16.568713220318084)
(('_natura', 'naturans_'), 16.568713220318084)
(('_vae', 'victis'), 16.568713220318084)
(('_virgin', 'martyr_'), 16.568713220318084)
(('absent', 'minded'), 16.568713220318084)
(('abstractedly', 'boring'), 16.568713220318084)
(('accommodated', 'indoors'), 16.568713220318084)
(('adult', 'reformatory'), 16.568713220318084)
(('adverse', 'theories'), 16.568713220318084)
(('albert', 'memorial'), 16.568713220318084)
(('angler', 'stares'), 16.568713220318084)
(('applaud', 'tyrannicide'), 16.568713220318084)
(('applicable', 'taxes'), 16.568713220318084)
(('architectural', 'arrangement'), 16.568713220318084)
(('assyrian', 'bas-reliefs'), 16.568713220318084)
(('astonishing', 'transparency'), 16.568713220318084)
(('august', '1500'), 16.568713220318084)
(('barren', 'steeps'), 16.568713220318084)
(('bearing', 'witness'), 16.568713220318084)
(('beetle', 'strayed'), 16.568713220318084)
(('ben', 'crowder'), 16.568713220318084)
(('bile', 'beans'), 16.568713220318084)
(('black-and-white', 'beads'), 16.568713220318084)
(('black-haired', 'skull'), 16.568713220318084)
(('blackcoated', 'bustle'), 16.568713220318084)
(('bllcr10.zip', 'corrected'), 16.568713220318084)
(('bllcr11.txt', 'versions'), 16.568713220318084)
(('boorish', 'moralities'), 16.568713220318084)
(('bourgeois', 'speaks'), 16.568713220318084)
(('bulging', 'bravery'), 16.568713220318084)
(('californian', 'tourist'), 16.568713220318084)
(('captured', 'coach'), 16.568713220318084)
(('cardinal', 'manning'), 16.568713220318084)
(('carthage', '_was_'), 16.568713220318084)
(('chartered', 'lunacies'), 16.568713220318084)
(('constables', 'attended'), 16.568713220318084)
```

Observation:
So as to obtain the proper relevance in the outcome,one has to experiment versatile filters.Since there are always different documents for analysis we tend to get different outcomes.

Trigram Analysis non Alphanumeric characters filtered;

Trigram Analysis:Non Alphanumeric raw frequency

```
(('a', 'sort', 'of'), 0.0006071571687046909)
(('i', 'do', "n't"), 0.0004007237313450 96)
(('out', 'of', 'the'), 0.00037643744459690836)
(('that', 'it', 'was'), 0.0003642943012228145)
(('of', 'the', 'world'), 0.0003400080144746269)
(('one', 'of', 'the'), 0.0003400080144746269)
(('it', 'is', 'a'), 0.00030357858435234547)
(('it', 'was', 'a'), 0.00030357858435234547)
(('a', 'kind', 'of'), 0.0002914354409782516)
(('he', 'did', 'not'), 0.00027929229760415784)
(('with', 'a', 'sort'), 0.00027929229760415784)
(('as', 'if', 'he'), 0.0002671491542300 64)
(('it', 'is', 'not'), 0.0002671491542300 64)
(('i', 'do', 'not'), 0.00025500601085597015)
(('do', 'you', 'mean'), 0.00024286286748187636)
(('what', 'do', 'you'), 0.00024286286748187636)
(('seemed', 'to', 'be'), 0.00023071972410778255)
(('the', 'end', 'of'), 0.00023071972410778255)
(('he', 'said', 'with'), 0.00021857658073368873)
(('said', 'the', 'other'), 0.00021857658073368873)
(('there', 'was', 'a'), 0.00021857658073368873)
(('for', 'a', 'moment'), 0.00020643343735959492)
(('for', 'the', 'first'), 0.00020643343735959492)
(('if', 'he', 'were'), 0.00020643343735959492)
(('man', 'in', 'the'), 0.00020643343735959492)
(('the', 'edge', 'of'), 0.00020643343735959492)
(('the', 'first', 'time'), 0.00020643343735959492)
(('the', 'top', 'of'), 0.00020643343735959492)
(('he', 'had', 'been'), 0.00019429029398550108)
(('he', 'said', 'i'), 0.00019429029398550108)
(('he', 'was', 'a'), 0.00019429029398550108)
(('he', 'would', 'have'), 0.00019429029398550108)
(('said', 'turnbull', 'with'), 0.00019429029398550108)
(('do', "n't", 'you'), 0.00018214715061140726)
(('if', 'he', 'had'), 0.00018214715061140726)
(('it', 'was', 'not'), 0.00018214715061140726)
(('man', 'with', 'the'), 0.00018214715061140726)
(('that', 'they', 'were'), 0.00018214715061140726)
(('and', 'all', 'the'), 0.00017000400723731344)
(('for', 'an', 'instant'), 0.00017000400723731344)
(('the', 'little', 'man'), 0.00017000400723731344)
(('there', 'is', 'no'), 0.00017000400723731344)
(('to', 'his', 'feet'), 0.00017000400723731344)
(('said', 'in', 'a'), 0.00015786086386321963)
(('then', 'he', 'said'), 0.00015786086386321963)
(('top', 'of', 'the'), 0.00015786086386321963)
(('turnbull', 'with', 'a'), 0.00015786086386321963)
(('at', 'the', 'same'), 0.00014571771204891258)
(('edge', 'of', 'the'), 0.00014571771204891258)
(('end', 'of', 'the'), 0.00014571771204891258)
```

Trigram Analysis:Non Alphanumeric PMI

```
(('literary', 'archive', 'foundation'), 26.904109594756882)
(('gutenberg', 'literary', 'archive'), 25.614602977561898)
(('small', 'print', 'statement'), 23.199565478283056)
(('project', 'gutenberg', 'literary'), 22.807248055504296)
(('mr.', 'cumberland', 'vane'), 20.227238347754195)
(('st.', 'paul', "'s"), 19.860322798093517)
(('this', 'small', 'print'), 18.2531465184879)
(('the', 'eighteenth', 'century'), 17.38287269164611)
(('half', 'an', 'hour'), 17.126397658247363)
(('of', '_the', 'atheist_'), 16.99253098379732)
(('does', 'not', 'exist'), 16.86520753621971)
(('mr.', 'wilkinson', "'s"), 16.772859956843185)
(('at', 'any', 'rate'), 16.708204293661577)
(('editor', 'of', '_the'), 16.510379288659937)
(('among', 'other', 'things'), 15.993661179735174)
(('i', "'m", 'afraid'), 15.899520641119661)
(('i', "'m", 'sure'), 15.459608473201790)
(('a', 'few', 'moments'), 15.426049252808664)
(('mr.', 'james', 'turnbull'), 15.236143900976089)
(('a', 'few', 'yards'), 14.904096549613303)
(('on', 'each', 'side'), 14.762002537454332)
(('could', 'be', 'seen'), 14.65341640463824)
(('a', 'great', 'deal'), 14.59280697517229)
(('i', 'am', 'sure'), 14.306362212906105)
(('do', "n't", 'believe'), 14.043712313555162)
(('ca', "n't", 'be'), 13.974785423742247)
(('do', "n't", 'want'), 13.89396519405048)
(('the', 'project', 'gutenberg'), 13.79415805606385)
(('might', 'have', 'been'), 13.740486123117385)
(('for', 'an', 'instant'), 13.673162305573968)
(('i', 'ca', "n't"), 13.532311666931431)
(('the', 'flying', 'ship'), 13.338478572287660)
(('may', 'i', 'ask'), 13.330653503032984)
(('after', 'a', 'pause'), 13.24547700716684)
(('my', 'name', 'is'), 13.171142673474876)
(('across', 'the', 'lawn'), 13.15790632664584)
(('you', 'ca', "n't"), 13.074645870124805)
(('want', 'to', 'fight'), 12.938682688543665)
(('do', "n't", 'think'), 12.811051556764884)
(('mr.', 'evan', 'macian'), 12.710932207699983)
(('i', 'am', 'going'), 12.649249926429114)
(('would', 'have', 'been'), 12.557894709838905)
(('the', 'whole', 'universe'), 12.537382640701736)
(('you', 'mean', 'asked'), 12.524635122087144)
(('do', "n't", 'know'), 12.509733741552807)
(('we', 'shall', 'be'), 12.478777114750159)
(('looking', 'at', 'him'), 12.472987831967547)
(('you', 'will', 'find'), 12.450242629805757)
(('do', 'you', 'mean'), 12.425830112031662)
(('they', 'could', 'see'), 12.349494097805191)
```

Document 2: Wildfire related tweets

Top 50 words before preprocessing:

```
('…', 0.03743377102050219)
(':', 0.029371112646855563)
('the', 0.028449665975581663)
('rt', 0.024072794287030637)
('#wildfires', 0.02038700760193504)
('.', 0.01716194425247639)
('to', 0.013706519235199263)
(',', 0.013591338401290025)
('a', 0.013360976733471551)
('in', 0.013015434231743837)
('of', 0.011863625892651462)
('we', 0.011633264224832988)
('wildfires', 0.011172540889196038)
('are', 0.0101359133840129)
('that', 0.009790370882285187)
('and', 0.009560009214466713)
('-', 0.009444828380557475)
('for', 0.009099285878829763)
('(', 0.008523381709283575)
('is', 0.008062658373646624)
('wildfire', 0.007601935038009675)
(')', 0.007256392536281963)
('2020', 0.006910850034554251)
('california', 0.006795669200645013)
('have', 0.006795669200645013)
('grey', 0.0065653075328265375)
('change', 0.0064501266989173)
('cross', 0.0064501266989173)
('studios', 0.006219765031098825)
('more', 0.005874222529371113)
('fire', 0.005874222529371113)
('by', 0.005759041695461875)
('forests', 0.005643860861552638)
('#surrealism', 0.005643860861552638)
('global', 0.005298318359824925)
('from', 0.004952775858097213)
('new', 0.0047224141902787375)
('fires', 0.004492052522460262)
('wild', 0.00414651002073255)
('’', 0.0040313291868233 12)
('#heatwave', 0.004031329186823312)
('can', 0.003916148352914075)
('&', 0.003916148352914075)
('on', 0.0038009675190048375)
('—', 0.0038009675190048375)
```

```
('has', 0.0038009675190048375)
('it', 0.0036857866850956)
('been', 0.0035706058511863624)
('as', 0.0034554250172771253)
('future', 0.0033402441833678877)
```

Non alphanumeric characters filtered from the top 50 words:

```
('the', 0.034891933888967365)
('rt', 0.029523944059895467)
('#wildfires', 0.025003531572255967)
('to', 0.01681028393840938)
('a', 0.016386495267693178)
('in', 0.015962706596976975)
('of', 0.014550077694589632)
('we', 0.014267551914112163)
('wildfires', 0.013702500353157225)
('are', 0.012431134341008617)
('that', 0.012007345670292414)
('and', 0.011724819889814945)
('for', 0.01115976832886001)
('is', 0.0098884023167114)
('wildfire', 0.009323350755756462)
('california', 0.008334510524085324)
('have', 0.008334510524085324)
('grey', 0.008051984743607854)
('change', 0.00791072185336912)
('cross', 0.00791072185336912)
('studios', 0.007628196072891651)
('more', 0.007204407402175449)
('fire', 0.007204407402175449)
('by', 0.007063144511936714)
('forests', 0.00692188162169798)
('#surrealism', 0.00692188162169798)
('global', 0.0064980929509817777)
('from', 0.006074304280265575)
('new', 0.005791778499788106)
('fires', 0.005509252719310637)
('wild', 0.0050854640485944345)
('#heatwave', 0.0049442011583557)
('can', 0.004802938268116966)
('on', 0.004661675377878231)
('has', 0.004661675377878231)
('it', 0.004520412487639497)
('been', 0.004379149597400763)
('as', 0.004237886707162028)
('future', 0.004096623816923294)
('s', 0.00395536092668456)
('important', 0.00395536092668456)
```

```
('study', 0.00395536092668456)
('intensity', 0.0038140980364458257)
('if', 0.0038140980364458257)
('plan', 0.0036728351462070913)
('i', 0.0036728351462070913)
('so', 0.0036728351462070913)
('want', 0.0035315572255968357)
('oregon', 0.0035315572255968357)
('march', 0.0033903309365729623)
```

Stop words filtered filtered from the top 50 words:

```
('rt', 0.042600896860986545)
('#wildfires', 0.036078271504280474)
('wildfires', 0.019771708112515288)
('wildfire', 0.013452914798206279)
('california', 0.012026090501426825)
('grey', 0.011618426416632695)
('change', 0.01141459437423563)
('cross', 0.01141459437423563)
('studios', 0.0110069302894415)
('fire', 0.010395434162250305)
('forests', 0.009987770077456177)
('#surrealism', 0.009987770077456177)
('global', 0.009376273950264982)
('new', 0.008357113738279657)
('fires', 0.007949449653485529)
('wild', 0.007337953526294333)
('#heatwave', 0.007134121483897269)
('future', 0.0059111292295148795)
('important', 0.005707297187117815)
('study', 0.005707297187117815)
('intensity', 0.0055053465144720 75)
('plan', 0.005299633102323685)
('want', 0.005095801059926621)
('oregon', 0.005095801059926621)
('march', 0.004891969017529555)
('absolutely', 0.004891969017529555)
('#onev1', 0.004891969017529555)
('@sclinton60', 0.004688136975132491)
('#climat', 0.004688136975132491)
('#california', 0.004484304932735426)
('service', 0.004280472890338361)
('season', 0.004280472890338361)
('#surrealart', 0.004280472890338361)
('february', 0.0040766408479412965)
('activity', 0.0040766408479412965)
('#copernicusatmosphere', 0.0040766408479412965)
('monitoring', 0.0040766408479412965)
```

```
('@m_parrington', 0.0038728088055442317)
('assimilation', 0.0038728088055442317)
('sys', 0.0038728088055442317)
('whilst', 0.0036689767631471666)
('us', 0.003465144720750102)
('year', 0.003465144720750102)
('almost', 0.003465144720750102)
('become', 0.003465144720750102)
('shown', 0.003261312678353037)
('naturally', 0.003261312678353037)
('occuring', 0.003261312678353037)
('vital', 0.003261312678353037)
('ecosytems', 0.003261312678353037)
```

Bigram analysis:

Not filtered raw frequency:

```
(('…', 'rt'), 0.01520387007601935)
(('(', 'the'), 0.0066804883667357755)
(('the', '2020'), 0.0066804883667357755)
(('grey', 'cross'), 0.006450126698917 3)
(('cross', 'studios'), 0.006219765031098825)
(('studios', '#surrealism'), 0.005643860861552638)
(('global', 'fire'), 0.0046072333563695)
(('2020', 'california'), 0.004146510020732 55)
(('#heatwave', '#wildfires'), 0.004031329186823312)
(('california', 'wildfires'), 0.003916148352914075)
(('wild', 'fires'), 0.003916148352914075)
(('the', 'future'), 0.0033402441833678877)
(('have', 'been'), 0.003225063349458 65)
(('.', 'it'), 0.0028795208477309375)
((':', 'we'), 0.0028795208477309375)
(('#surrealism', '…'), 0.0027643400138217)
(('#wildfires', '#onev1'), 0.0027643400138217)
(('a', 'plan'), 0.0027643400138217)
(('absolutely', 'change'), 0.0027643400138217)
(('can', 'absolutely'), 0.0027643400138217)
(('change', '.'), 0.0027643400138217)
(('change', 'the'), 0.0027643400138217)
(('future', 'if'), 0.0027643400138217)
(('have', 'a'), 0.0027643400138217)
(('if', 'we'), 0.0027643400138217)
(('important', 'that'), 0.0027643400138217)
(('is', 'so'), 0.0027643400138217)
(('it', 'is'), 0.0027643400138217)
(('plan', '🙌'), 0.0027643400138217)
(('so', 'important'), 0.0027643400138217)
(('that', 'we'), 0.0027643400138217)
```

```
(('want', 'change'), 0.0027643400138217)
(('we', 'can'), 0.0027643400138217)
(('we', 'have'), 0.0027643400138217)
(('we', 'want'), 0.0027643400138217)
(('🙌', '#wildfires'), 0.0027643400138217)
(('#climat', '…'), 0.0026491591799124624)
(('#onev1', '#climat'), 0.0026491591799124624)
(('@sclinton60', ':'), 0.0026491591799124624)
(('a', 'new'), 0.0026491591799124624)
(('rt', '@sclinton60'), 0.0026491591799124624)
(('2020', 'oregon'), 0.0025533978346003225)
(('oregon', 'wildfires'), 0.0025533978346003225)
(('#surrealism', '#surrealart'), 0.0024187975120939877)
(('new', 'study'), 0.0024187975120939877)
(('#copernicusatmosphere', 'monitoring'), 0.00230361667818475)
(('&', 'intensity'), 0.00230361667818475)
(('2021', 'global'), 0.00230361667818475)
((':', 'a'), 0.00230361667818475)
(('activity', '&'), 0.00230361667818475)
```

Not filtered pmi:

```
(('#ab', 'https://t.co/jesmcpkmcd'), 13.083811707252881)
(('#energynews', '#news'), 13.083811707252881)
(('#firefighting', 'toolbox'), 13.083811707252881)
(('#gis', '#spatial'), 13.083811707252881)
(('#homes', '#houses'), 13.083811707252881)
(('#hydrology', '#wildfire'), 13.083811707252881)
(('#infrastructure', '#homes'), 13.083811707252881)
(('#instaislam', '#fgrf'), 13.083811707252881)
(('#morningcommute', 'https://t.co/foodlsy4kh'), 13.083811707252881)
(('#nature', '#naturephotography'), 13.083811707252881)
(('#neon', '#neoncolor'), 13.083811707252881)
(('#neoncolor', '#neons'), 13.083811707252881)
(('#neoncolorsart', '#neon'), 13.083811707252881)
(('#neons', '#inthesha'), 13.083811707252881)
(('#pantanal', 'cont'), 13.083811707252881)
(('#pge', '#energynews'), 13.083811707252881)
(('#photography', '#oakridgeoregon'), 13.083811707252881)
(('#pollution', '#instaislam'), 13.083811707252881)
(('#santarosa', '@santarosafire'), 13.083811707252881)
(('#smoky', '#morningcommute'), 13.083811707252881)
(('#sonomacounty', '#recycle'), 13.083811707252881)
(("#southamerica's", '#pantanal'), 13.083811707252881)
(('#tornados', '#heatwaves'), 13.083811707252881)
(('#tree', '#trees'), 13.083811707252881)
(('#trees', '#air'), 13.083811707252881)
(('+', '#nuclear'), 13.083811707252881)
(('..', 'however'), 13.083811707252881)
(('100', 'million'), 13.083811707252881)
(('3lv', 'https://t.co/upjd3ml4eh'), 13.083811707252881)
```

```
(('461', 'acres'), 13.083811707252881)
((';', 'boise'), 13.083811707252881)
(('@buzzsolutions1', 'aims'), 13.083811707252881)
(('@cafiresafe', '#highway168firesafecouncil'), 13.083811707252881)
(('@cityofsantarosa', '@countyofsonoma'), 13.083811707252881)
(('@countyofsonoma', '#sonomacounty'), 13.083811707252881)
(('@nickcowern', '@ricbollen1'), 13.083811707252881)
(('@orgovkatebrown', 'throug'), 13.083811707252881)
(('@poppepk', 'promises'), 13.083811707252881)
(('@ricbollen1', '@rovanzon'), 13.083811707252881)
(('@santarosafire', '@cityofsantarosa'), 13.083811707252881)
(('@sce', '@cafiresafe'), 13.083811707252881)
(('ana', 'winds'), 13.083811707252881)
(('another', 'asset'), 13.083811707252881)
(('any', 'per'), 13.083811707252881)
(('arizona', 'mapped'), 13.083811707252881)
(('art', 'youtube'), 13.083811707252881)
(('audit', 'finds'), 13.083811707252881)
(('auroras', 'bursts'), 13.083811707252881)
(('az', 'https://t.co/qxn0dm06lx'), 13.083811707252881)
(('bananas', 'idea'), 13.083811707252881)
```

Bigram Analysis non Alphanumeric characters filtered;

Bigram Analysis:Non Alphanumeric raw frequency

```
(('grey', 'cross'), 0.00791072185336912)
(('cross', 'studios'), 0.0076281960728916 51)
(('studios', '#surrealism'), 0.00692188162169798)
(('global', 'fire'), 0.005650515609549371)
(('the', 'california'), 0.0050854640485944345)
(('#heatwave', '#wildfires'), 0.0049442011583557)
(('california', 'wildfires'), 0.004802938268116966)
(('wild', 'fires'), 0.004802938268116966)
(('the', 'future'), 0.004096623816923294)
(('have', 'been'), 0.00395536092668456)
(('#wildfires', '#onev1'), 0.003390309365729623)
(('a', 'plan'), 0.003390309365729623)
(('absolutely', 'change'), 0.003390309365729623)
(('can', 'absolutely'), 0.003390309365729623)
(('change', 'it'), 0.003390309365729623)
(('change', 'the'), 0.003390309365729623)
(('future', 'if'), 0.003390309365729623)
(('have', 'a'), 0.003390309365729623)
(('if', 'we'), 0.003390309365729623)
(('important', 'that'), 0.003390309365729623)
(('is', 'so'), 0.003390309365729623)
(('it', 'is'), 0.003390309365729623)
(('plan', '#wildfires'), 0.003390309365729623)
```

```
(('so', 'important'), 0.003390309365729623)
(('that', 'we'), 0.003390309365729623)
(('want', 'change'), 0.003390309365729623)
(('we', 'can'), 0.003390309365729623)
(('we', 'have'), 0.003390309365729623)
(('we', 'want'), 0.003390309365729623)
(('#onev1', '#climat'), 0.0032490464754908886)
(('@sclinton60', 'we'), 0.0032490464754908886)
(('a', 'new'), 0.0032490464754908886)
(('rt', '@sclinton60'), 0.0032490464754908886)
(('the', 'oregon'), 0.0032490464754908886)
(('oregon', 'wildfires'), 0.0031077835852521543)
(('#surrealism', '#surrealart'), 0.00296652069501342)
(('new', 'study'), 0.00296652069501342)
(('#copernicusatmosphere', 'monitoring'), 0.0028252578047746855)
(('activity', 'intensity'), 0.0028252578047746855)
(('february', 'global'), 0.0028252578047746855)
(('fire', 'activity'), 0.0028252578047746855)
(('forests', 'are'), 0.0028252578047746855)
(('from', '#copernicusatmosphere'), 0.0028252578047746855)
(('intensity', 'from'), 0.0028252578047746855)
(('monitoring', 'service'), 0.0028252578047746855)
(('service', 'global'), 0.0028252578047746855)
(('@m_parrington', 'february'), 0.0026839949145359516)
(('assimilation', 'sys'), 0.0026839949145359516)
(('fire', 'assimilation'), 0.0026839949145359516)
(('in', 'the'), 0.0026839949145359516)
```

Bigram Analysis:Non Alphanumeric pmi:

```
(('#ab', 'https://t.co/jesmcpkmcd'), 12.789329860092433)
(('#chernobyl', 'https://t.co/vkwkdaezqx'), 12.789329860092433)
(('#energynews', '#news'), 12.789329860092433)
(('#fgrf', 'https://t.co/xevb4q9dcj'), 12.789329860092433)
(('#firefighting', 'toolbox'), 12.789329860092433)
(('#forestm', 'renewable'), 12.789329860092433)
(('#gis', '#spatial'), 12.789329860092433)
(('#homes', '#houses'), 12.789329860092433)
(('#houses', 'https://t.co/s8302h99la'), 12.789329860092433)
(('#hydrology', '#wildfire'), 12.789329860092433)
(('#infrastructure', '#homes'), 12.789329860092433)
(('#instaislam', '#fgrf'), 12.789329860092433)
(('#losangeles', 'https://t.co/rwdsi8dznh'), 12.789329860092433)
(('#morningcommute', 'https://t.co/foodlsy4kh'), 12.789329860092433)
(('#nature', '#naturephotography'), 12.789329860092433)
(('#neon', '#neoncolor'), 12.789329860092433)
(('#neoncolor', '#neons'), 12.789329860092433)
(('#neoncolorsart', '#neon'), 12.789329860092433)
(('#neons', '#inthesha'), 12.789329860092433)
(('#news', 'https://t.co/42ep2mt7p0'), 12.789329860092433)
(('#pantanal', 'cont'), 12.789329860092433)
```

```
(('#pge', '#energynews'), 12.789329860092433)
(('#photography', '#oakridgeoregon'), 12.789329860092433)
(('#pollution', '#instaislam'), 12.789329860092433)
(('#recycle', 'https://t.co/mumzrlks5d'), 12.789329860092433)
(('#reforestation', 'https://t.co/m162tlit7f'), 12.789329860092433)
(('#santarosa', '@santarosafire'), 12.789329860092433)
(('#similipal', 'https://t.co/jzwrgjgvp4'), 12.789329860092433)
(('#smoky', '#morningcommute'), 12.789329860092433)
(('#sonomacounty', '#recycle'), 12.789329860092433)
(("#southamerica's", '#pantanal'), 12.789329860092433)
(('#tornados', '#heatwaves'), 12.789329860092433)
(('#tree', '#trees'), 12.789329860092433)
(('#trees', '#air'), 12.789329860092433)
(('#wildfire', 'https://t.co/ugnnyoaeu9'), 12.789329860092433)
(('2nd', 'colorado'), 12.789329860092433)
(('3lv', 'https://t.co/upjd3ml4eh'), 12.789329860092433)
(('@avantaventures', '@buzzsolutions1'), 12.789329860092433)
(('@buzzsolutions1', 'aims'), 12.789329860092433)
(('@cafiresafe', '#highway168firesafecouncil'), 12.789329860092433)
(('@caps_cmu', 'https://t.co/44oiwdpwtg'), 12.789329860092433)
(('@cdcenvironment', '@nist'), 12.789329860092433)
(('@cityofsantarosa', '@countyofsonoma'), 12.789329860092433)
(('@countyofsonoma', '#sonomacounty'), 12.789329860092433)
(('@dollyslibrary', 'she'), 12.789329860092433)
(('@eparegion9', '@epa'), 12.789329860092433)
(('@ms_took', 'swirls'), 12.789329860092433)
(('@natgeotravel', 'https://t.co/gyhavhotie'), 12.789329860092433)
(('@nickcowern', '@ricbollen1'), 12.789329860092433)
(('@niosh', '@cdcenvironment'), 12.789329860092433)
```

Bigram Analysis Stop words filtered;

Bigram Analysis:Stop words raw frequency

```
(('grey', 'cross'), 0.01141459437423563)
(('cross', 'studios'), 0.0110069302894415)
(('studios', '#surrealism'), 0.009987770077456177)
(('global', 'fire'), 0.008153281695882593)
(('#heatwave', '#wildfires'), 0.007134121483897269)
(('california', 'wildfires'), 0.006930289441500204)
(('wild', 'fires'), 0.006930289441500204)
(('#wildfires', '#onev1'), 0.004891969017529555)
(('absolutely', 'change'), 0.004891969017529555)
(('change', 'future'), 0.004891969017529555)
(('change', 'important'), 0.004891969017529555)
(('future', 'want'), 0.004891969017529555)
(('important', 'plan'), 0.004891969017529555)
(('plan', '#wildfires'), 0.004891969017529555)
(('want', 'change'), 0.004891969017529555)
(('#onev1', '#climat'), 0.004688136975132491)
(('@sclinton60', 'absolutely'), 0.004688136975132491)
```

```
(('rt', '@sclinton60'), 0.004688136975132491)
(('oregon', 'wildfires'), 0.004484304932735426)
(('#surrealism', '#surrealart'), 0.004280472890338361)
(('new', 'study'), 0.004280472890338361)
(('#copernicusatmosphere', 'monitoring'), 0.0040766408479412965)
(('activity', 'intensity'), 0.0040766408479412965)
(('february', 'global'), 0.0040766408479412965)
(('fire', 'activity'), 0.0040766408479412965)
(('intensity', '#copernicusatmosphere'), 0.0040766408479412965)
(('monitoring', 'service'), 0.0040766408479412965)
(('service', 'global'), 0.0040766408479412965)
(('@m_parrington', 'february'), 0.0038728088055442317)
(('assimilation', 'sys'), 0.0038728088055442317)
(('fire', 'assimilation'), 0.0038728088055442317)
(('rt', '@m_parrington'), 0.0038728088055442317)
(('#climat', 'rt'), 0.0034651447207501O2)
(('fires', 'new'), 0.0032613126783530 37)
(('fires', 'whilst'), 0.0032613126783530 37)
(('naturally', 'occuring'), 0.0032613126783530 37)
(('occuring', 'vital'), 0.0032613126783530 37)
(('shown', 'wild'), 0.0032613126783530 37)
(('study', 'shown'), 0.0032613126783530 37)
(('vital', 'ecosytems'), 0.0032613126783530 37)
(('whilst', 'naturally'), 0.0032613126783530 37)
(('#dyk', '#oldgrowth'), 0.0030574806359559724)
(('#oldgrowth', 'forests'), 0.0030574806359559724)
(('#wildfires', 'become'), 0.0030574806359559724)
(('@helpingrhinos', 'wild'), 0.0030574806359559724)
(('ecosytems', 'spr'), 0.0030574806359559724)
(('forests', '#wildfires'), 0.0030574806359559724)
(('forests', 'resilient'), 0.0030574806359559724)
(('resilient', 'worsening'), 0.0030574806359559724)
(('rt', '@helpingrhinos'), 0.0030574806359559724)
```

Bigram Analysis:Stop words pmi:

```
(('#ab', 'https://t.co/jesmcpkmcd'), 12.260331518557603)
(('#america', 'despite'), 12.260331518557603)
(('#biodiversity', 'lives'), 12.260331518557603)
(('#chernobyl', 'https://t.co/vkwkdaezgx'), 12.260331518557603)
(('#energynews', '#news'), 12.260331518557603)
(('#fgrf', 'https://t.co/xevb4g9dcj'), 12.260331518557603)
(('#firefighting', 'toolbox'), 12.260331518557603)
(('#forestm', 'renewable'), 12.260331518557603)
(('#gis', '#spatial'), 12.260331518557603)
(('#homes', '#houses'), 12.260331518557603)
(('#houses', 'https://t.co/s8302h991a'), 12.260331518557603)
(('#hydrology', '#wildfire'), 12.260331518557603)
(('#infrastructure', '#homes'), 12.260331518557603)
(('#instaislam', '#fgrf'), 12.260331518557603)
```

```
(('#losangeles', 'https://t.co/rwdsi8dznh'), 12.260331518557603)
(('#morningcommute', 'https://t.co/foodlsy4kh'), 12.260331518557603)
(('#nature', '#naturephotography'), 12.260331518557603)
(('#neon', '#neoncolor'), 12.260331518557603)
(('#neoncolor', '#neons'), 12.260331518557603)
(('#neoncolorsart', '#neon'), 12.260331518557603)
(('#neons', '#inthesha'), 12.260331518557603)
(('#news', 'https://t.co/42ep2mt7p0'), 12.260331518557603)
(('#pantanal', 'cont'), 12.260331518557603)
(('#pge', '#energynews'), 12.260331518557603)
(('#photography', '#oakridgeoregon'), 12.260331518557603)
(('#pollution', '#instaislam'), 12.260331518557603)
(('#recycle', 'https://t.co/mumzrlks5d'), 12.260331518557603)
(('#reforestation', 'https://t.co/ml62tlit7f'), 12.260331518557603)
(('#santarosa', '@santarosafire'), 12.260331518557603)
(('#similipal', 'https://t.co/jzwrgjgvp4'), 12.260331518557603)
(('#smoky', '#morningcommute'), 12.260331518557603)
(('#sonomacounty', '#recycle'), 12.260331518557603)
(("#southamerica's", '#pantanal'), 12.260331518557603)
(('#tornados', '#heatwaves'), 12.260331518557603)
(('#tree', '#trees'), 12.260331518557603)
(('#trees', '#air'), 12.260331518557603)
(('#us', 'ravaged'), 12.260331518557603)
(('#utilities', "they're"), 12.260331518557603)
(('#wildfire', 'https://t.co/ugnnyoaeu9'), 12.260331518557603)
(('2nd', 'colorado'), 12.260331518557603)
(('3lv', 'https://t.co/upjd3ml4eh'), 12.260331518557603)
(('@avantaventures', '@buzzsolutions1'), 12.260331518557603)
(('@buzzsolutions1', 'aims'), 12.260331518557603)
(('@cafiresafe', '#highway168firesafecouncil'), 12.260331518557603)
(('@caps_cmu', 'https://t.co/44oiwdpwtq'), 12.260331518557603)
(('@cdcenvironment', '@nist'), 12.260331518557603)
(('@cityofsantarosa', '@countyofsonoma'), 12.260331518557603)
(('@countyofsonoma', '#sonomacounty'), 12.260331518557603)
(('@dollyslibrary', 'gave'), 12.260331518557603)
(('@epa', 'partnering'), 12.260331518557603)
```

Notes and Observation:

1. The Frequency filter along with the PMI scoring mechanism usually works but more experimentation with other filters is much recommended to check for the proper usage of the filters
2. More information is provided by the bigram analysis for the stop words and Non alphanumeric characters

Trigram Analysis non Alphanumeric characters filtered;

Trigram Analysis:Non Alphanumeric raw frequency

```
(('grey', 'cross', 'studios'), 0.005897219882055603)
(('cross', 'studios', '#surrealism'), 0.0053355798932884025)
(('the', 'california', 'wildfires'), 0.004773939904521202)
(('#copernicusatmosphere', 'monitoring', 'service'),
0.002808199943836001)
(('activity', 'intensity', 'from'), 0.002808199943836001)
(('february', 'global', 'fire'), 0.002808199943836001)
(('fire', 'activity', 'intensity'), 0.002808199943836001)
(('from', '#copernicusatmosphere', 'monitoring'), 0.002808199943836001)
(('global', 'fire', 'activity'), 0.002808199943836001)
(('intensity', 'from', '#copernicusatmosphere'), 0.002808199943836001)
(('monitoring', 'service', 'global'), 0.002808199943836001)
(('service', 'global', 'fire'), 0.002808199943836001)
(('@m_parrington', 'february', 'global'), 0.0026677899466442012)
(('fire', 'assimilation', 'sys'), 0.0026677899466442012)
(('global', 'fire', 'assimilation'), 0.0026677899466442012)
(('rt', '@m_parrington', 'february'), 0.0026677899466442012)
(('a', 'new', 'study'), 0.002386969952260601)
(('studios', '#surrealism', '#surrealart'), 0.002386969952260601)
(('and', 'vital', 'for'), 0.002246559955068801)
(('fires', 'a', 'new'), 0.002246559955068801)
(('fires', 'whilst', 'naturally'), 0.002246559955068801)
(('for', 'some', 'ecosytems'), 0.002246559955068801)
(('has', 'shown', 'that'), 0.002246559955068801)
(('naturally', 'occuring', 'and'), 0.002246559955068801)
(('new', 'study', 'has'), 0.002246559955068801)
(('occuring', 'and', 'vital'), 0.002246559955068801)
(('shown', 'that', 'wild'), 0.002246559955068801)
(('some', 'ecosytems', 'have'), 0.002246559955068801)
(('study', 'has', 'shown'), 0.002246559955068801)
(('that', 'wild', 'fires'), 0.002246559955068801)
(('vital', 'for', 'some'), 0.002246559955068801)
(('whilst', 'naturally', 'occuring'), 0.002246559955068801)
(('wild', 'fires', 'a'), 0.002246559955068801)
(('wild', 'fires', 'whilst'), 0.002246559955068801)
(('@helpingrhinos', 'wild', 'fires'), 0.0021061499578770007)
(('ecosytems', 'have', 'been'), 0.0021061499578770007)
(('have', 'been', 'spr'), 0.0021061499578770007)
(('rt', '@helpingrhinos', 'wild'), 0.0021061499578770007)
(('#wildfires', '#onev1', '#climat'), 0.001965739960685201)
(('@sclinton60', 'we', 'can'), 0.001965739960685201)
(('a', 'plan', '#wildfires'), 0.001965739960685201)
(('absolutely', 'change', 'the'), 0.001965739960685201)
(('assimilation', 'sys', 'rt'), 0.001965739960685201)
(('can', 'absolutely', 'change'), 0.001965739960685201)
(('change', 'it', 'is'), 0.001965739960685201)
(('change', 'the', 'future'), 0.001965739960685201)
(('future', 'if', 'we'), 0.001965739960685201)
(('have', 'a', 'plan'), 0.001965739960685201)
(('if', 'we', 'want'), 0.001965739960685201)
(('important', 'that', 'we'), 0.001965739960685201)
```

Trigram Analysis:Non Alphanumeric PMI

```
(('@sacbee_news', 'professor', 'scott'), 20.9522772509514)
(('itself', 'l', 'wolfe'), 20.9522772509514)
(('life', 'itself', 'l'), 20.9522772509514)
(('professor', 'scott', 'stephens'), 20.9522772509514)
(('scott', 'stephens', 'discusses'), 20.9522772509514)
(('stephens', 'discusses', '#forest'), 20.689242845117608)
(('book', '#flamesofextinction', 'hits'), 20.426208439283815)
(('having', 'days', 'where'), 20.203816017947368)
(('careless', 'debris', 'burners'), 20.011170940004973)
(('discusses', '#forest', 'management'), 20.01117094000497)
(('@john_pickrell', 'hot', 'off'), 19.788778518668526)
(('#cleanerindoorairchallenge', 'informational', 'webinar'),
19.426208439283812)
(('10am', 'pst', '7pm'), 19.426208439283812)
(('pst', '7pm', 'cet'), 19.426208439283812)
(('#global', '#climatechange', 'include'), 19.274205345838766)
(('will', 'disappear', 'quietly'), 19.274205345838766)
(('trouble', 'after', 'historic'), 19.10428034439645)
(('@zachkunstk', 'what', 'inspired'), 19.01117094000497)
(('nc', 'each', 'year'), 18.923708098754634)
(('#climatechange', 'include', 'frequent'), 18.788778518668522)
(('year', 'having', 'days'), 18.701315677418187)
(('who', 'find', 'beauty'), 18.689242845117608)
(('debris', 'burners', 'your'), 18.63265931675124)
(('run', '@beingelenala', 'i'), 18.551739321367677)
(("here's", 'how', "we're"), 18.551739321367673)
(('investigate', 'west', '@invw'), 18.551739321367673)
(('@espm_berkeley', '@cal_fire', '#treemortality'), 18.466850423781157)
(('@ucanr', '@espm_berkeley', '@cal_fire'), 18.466850423781157)
(('#climatecrisis', 'will', 'disappear'), 18.426208439283812)
(('#treemortality', 'day', 'workshop'), 18.367314750230243)
(('@cal_fire', '#treemortality', 'day'), 18.367314750230243)
(('day', 'workshop', 'registration'), 18.367314750230243)
(('air', 'quality', 'r'), 18.31073122186388)
(('report', '#colorado', 'forests'), 18.22981122648031)
(('last', 'year', 'having'), 18.186742504588427)
(('7pm', 'cet', 'i'), 18.13670182208883)
(('@beingelenala', 'i', 'was'), 18.13670182208883)
(('what', 'inspired', 'you'), 18.13670182208883)
(('those', 'who', 'hope'), 18.10428034439645)
(('communities', 'most', 'likely'), 18.033891016505056)
(('nature', 'will', 'find'), 18.01117094000497)
(('will', 'find', 'themselves'), 18.01117094000497)
(('summer', '#climatechange', 'means'), 17.984698728643778)
(('cet', 'i', 'm'), 17.966776820646515)
(('our', '#cleanerindoorairchallenge', 'informational'),
17.895693722585037)
(('at', 'am', 'pt'), 17.788778518668522)
(('#storms', 'global', '#sealevel'), 17.719616494161126)
```

```
(('#tropical', '#storms', 'global'), 17.719616494161126)
(('global', '#sealevel', 'ris'), 17.719616494161126)
(('how', "we're", 'working'), 17.703742414812723)
```

Final Analysis

The first document which I have taken is The Ball and the Cross by G. K. Chesterton which is a compilation of stories.

The second document which I have taken is a compilation of all the tweets on the same topic in this case **"wildfire".**

The main difference between both document is the authors ,the first document has a single author so the words usage, writing style remain same throughout the text so n gram analysis makes more sense but for the second document the tweets are regarding the same topic but not by the same author it requires more preprocessing like removing words with the similar meaning to make sense of the n gram analysis.Since I haven't tried out the last part I cannot verify the hypothesis but according to the experimentation done by me I think that should be the next step in preprocessing of the tweets.