Harshit Joshi

IST 664 HW1

1) Selection of documents to be compared: The Man Who Was Thursday by G. K. Chesterton (1908) & Julius Caesar by William Shakespeare which was written in 1599 .

   Both of these texts were obtained from Gutenberg corpus in NLTK library. I choosed these two documents because of their:

   - **huge differences in creation in history (more than 3 centuries)**
   - **different genre : Narrative mystery fiction (*The Man Who Was Thursday by G. K. Chesterton*) & Dramatic Play (*Julius Caesar by William Shakespeare*)**

2) **Question for comparison:** Explain how *The Man Who Was Thursday by G. K. Chesterton* & *Julius Caesar by William Shakespeare* **differ or are similar** in terms of word frequency, Bigrams, Trigrams, or mutual information between words in both these documents.
   - **More importantly explain how the vocabulary in these documents differ because of their differences in creation in history an genres.**

3) **Task 1 Data Collection** : texts were obtained from Gutenberg corpus in NLTK library. With nltk.corpus.gutenberg.raw(file number) I was able to obtain these text in raw format.
   I then explored these texts to make sure they are correct with list indexing to look at first 100 words.

4) **Task 2:**
   a. **Tokenization:** for tokenization I used .word_tokenize function which is defined in NLTK library in Python. I used this because it considers whitespace, punctuation to separate words as tokens. Tokenization is done in order for ease to perform NLP tasks.
   b. Lowercase: using .lower() function, I converted these tokens to lowercase so that words with capital and words with all lower case can be categorized properly. For eg. Thursday and thursday will both be regarded as Thursday. This will make it easier to find frequency distributions of words.
   c. Removing words which are non-alphabetic I used Regular Expression library re in python.
      - I then compiled the filter that will filter out words which are all characters using character filter function cff().
      - Using list comprehension I parsed through documents and filtered out non alphabetic words
      - For the text The Man Who Was Thursday I filtered out around 10,000 non-alphabetic words using this function. And 5000 for Julius Caesar
   d. **Stopwords:** For removing stopwords I used predefined stopword list in english in NLTK. By removing stopwords from tokens of documents, I further filtered the texts and removed around 35000 Stopwords which will have no significance for our analysis.
   e. **Additional step for stopwords**: I made an addition to initial dictionary:
      - I noticed that words like don't were tokenized as do and n't which were not available in the initial stopwords list.
      - Another problem was that words tokenized with apostrophe were not included in this list. I also added this to initial list of stopwords.

f. **Stemming and Lemmetization:** I opted for Stemming as it bound **more** similar words than lemmatization for my documents. I opted for Porter stemmer as it is less stricter in filtering stems from words.

g. **For Top 50 words with normalized frequency I used version of text without stemming ie. till stopwords. To illustrate, the following are top 50 words between both text.**

- **Additional step: I used list comprehension to make the list a horizontal rather than long vertical list of 50 elements. I joined a blank space in between tuples.**

50 Top words with normalized frequency:

- **For** *The Man Who Was Thursday by G. K. Chesterton*:

```
In [209]: Thurs_normalized[:50]

          #Most common words in chesterton-thursday.txt
          horizontal_list = ' '.join(str(item) for item in Thurs_normalized[:50])
          print(horizontal_list)
```

```
('syme', 1.8641618497109829) ('said', 1.8316473988439306) ('man', 0.9898843930635839) ('like', 0.9320809248554914) ('one', 0.71
53179190751445) ('professor', 0.5346820809248555) ('face', 0.4299132947976878) ('gregory', 0.40462427745664736) ('bull', 0.3901
734104046243) ('see', 0.35765895953757226) ('sunday', 0.3504335260115607) ('men', 0.32875722543352603) ('seemed', 0.32875722543
352603) ('even', 0.32514450867052025) ('old', 0.32153179190751446) ('little', 0.3070809248554913) ('great', 0.3070809248554913)
('quite', 0.29985549132947975) ('say', 0.2890173410404624) ('know', 0.2890173410404624) ('us', 0.2854046242774566) ('looked',
0.2781791907514451) ('dr.', 0.2781791907514451) ('long', 0.2709537572254335) ('asked', 0.2709537572254335) ('came', 0.267341040
46242774) ('two', 0.26734104046242774) ('still', 0.26734104046242774) ('eyes', 0.26734104046242774) ('upon', 0.2637283236994219
5) ('first', 0.26372832369942195) ('think', 0.2565028901734104) ('almost', 0.2565028901734104) ('well', 0.2565028901734104) ('p
resident', 0.2565028901734104) ('cried', 0.25289017341040465) ('black', 0.24927745664739884) ('never', 0.24927745664739884) ('t
ime', 0.24927745664739884) ('went', 0.24566473988439308) ('marquis', 0.24566473988439308) ('thing', 0.2420520231213873) ('secre
tary', 0.2420520231213873) ('last', 0.23482658959537572) ('saw', 0.23121387283236997) ('really', 0.22760115606936418) ('good',
0.2239884393063584) ('something', 0.2239884393063584) ('voice', 0.2203757225433526) ('made', 0.2203757225433526)
```

- **For** *Julius Caesar*:

```
In [210]: caesar_normalized[:50]

          #Most common words in Julius Caesar by William Shakespeare
          horizontal_list2 = ' '.join(str(item) for item in caesar_normalized[:50])
          print(horizontal_list2)
```

```
('caesar', 1.752605848168988) ('brutus', 1.4851028502905637) ('bru', 1.411308919841343) ('haue', 1.3559634720044278) ('shall',
1.153030163269071) ('thou', 1.0607877502075453) ('cassi', 0.9869938197583248) ('cassius', 0.7840605110229684) ('antony', 0.6918
180979614427) ('come', 0.6825938566552902) ('let', 0.6549211327368324) ('good', 0.6456968914306799) ('know', 0.627248408818374
7) ('enter', 0.5903514435937643) ('men', 0.5903514435937643) ('heere', 0.5442302370630016) ('vs', 0.5350059957568489) ('thy',
0.5165575131445438) ('man', 0.5073332718383913) ('thee', 0.5073332718383913) ('vpon', 0.44276358269532334) ('ant', 0.4427635826
9532334) ('well', 0.44276358269532334) ('lord', 0.40586661747071306) ('day', 0.39664237616456044) ('yet', 0.3874181348584079)
('go', 0.37819389355225536) ('selfe', 0.3597454109399502) ('caes', 0.3597454109399502) ('noble', 0.3597454109399502) ('like',
0.3597454109399502) ('rome', 0.35052116963379765) ('heare', 0.35052116963379765) ('caesars', 0.35052116963379765) ('cask', 0.35
052116963379765) ('night', 0.35052116963379765) ('say', 0.3412969283276451) ('may', 0.3412969283276451) ('see', 0.3412969283276
451) ('brut', 0.3412969283276451) ('tell', 0.3412969283276451) ('speake', 0.32284844571533994) ('stand', 0.32284844571533994)
('giue', 0.32284844571533994) ('hath', 0.32284844571533994) ('loue', 0.31362420440918737) ('one', 0.31362420440918737) ('cask
a', 0.30439996310303474) ('vp', 0.2951757217968822) ('doth', 0.27672723918457703)
```

h. **Top 50 Biagrams:**

- **For** *The Man Who Was Thursday by G. K. Chesterton*:

```
(("'", 'l'), 10.100490946662747)
(('caius', 'ligarius'), 10.075616278023718)
(('metellus', 'cymber'), 10.075616278023718)
(('wee', "'"), 10.007381542271265)
(('mine', 'owne'), 9.027117760332528)
(('any', 'thing'), 8.94316598200007)
(('fell', 'downe'), 8.837456540828954)
(('mark', 'antony'), 8.288319008307367)
(('was', 'ambitious'), 7.879891807149351)
(('at', 'philippi'), 7.866940735441776)
(('marke', 'antony'), 7.658268618057672)
(('good', 'morrow'), 7.587879290166276)
(('most', 'noble'), 7.531295761799909)
(('be', 'satisfied'), 7.424380557883394)
(('what', 'trade'), 7.349791241462713)
(('too', 'much'), 7.181109406871032)
(('thou', 'hast'), 7.141132931160094)
(('honourable', 'men'), 7.138626075549519)
(("did'st", 'thou'), 7.100490946662747)
(('lou', "'d"), 7.0315958654516795)
(('mou', "'d"), 7.0315958654516795)
(('haue', 'seene'), 7.009343058604552)
(('our', 'selues'), 6.9473905678617935)
(('haue', 'beene'), 6.938953730713154)
(('caius', 'cassius'), 6.836150343328329)
(('my', 'lord'), 6.779190246888943)
(('offer', "'d"), 6.768561459617887)
(('enter', 'lucius'), 6.73096810663627)
(('euery', 'man'), 6.620300767858652)
(('let', 'vs'), 6.616324788087507)
(('thou', 'art'), 6.594138280637958)
(('haue', 'heard'), 6.576383651328445)
(('your', 'selues'), 6.530415264341457)
(('he', 'loues'), 6.478375447524126)
(('come', 'downe'), 6.476000081754952)
(('no', 'more'), 6.456862868998687)
(('his', 'body'), 6.436347357744644)
(("'t", 'is'), 6.408200630827961)
(('durst', 'not'), 6.3627414920910095)
(('art', 'thou'), 6.304631663442974)
(('good', 'night'), 6.2468423723312085)
(('ides', 'of'), 6.1564473526367625)
(('tell', 'them'), 6.116918988450124)
(('bring', 'me'), 6.077158442832124)
(('pardon', 'me'), 6.077158442832124)
(('my', 'selfe'), 5.953219646663992)
(('haue', 'done'), 5.9008186018263835)
(('shall', 'finde'), 5.892733871694697)
(('there', "'s"), 5.869165400556291)
(('put', 'it'), 5.802051204697753)
```

- **For Julius Caesar:**

```
(('let', 'vs'), 0.0006336382717516138)
(('mark', 'antony'), 0.0005148310957981862)
(('marke', 'antony'), 0.0004752270381371034)
(('thou', 'art'), 0.00043562631182923446)
(('art', 'thou'), 0.00035642152786028277)
(('enter', 'brutus'), 0.00035642152786028277)
(('noble', 'brutus'), 0.00035642152786028277)
(('thou', 'hast'), 0.00035642152786028277)
(('caesar', 'caes'), 0.0003168191358758069)
(('good', 'morrow'), 0.0003168191358758069)
(('good', 'night'), 0.0003168191358758069)
(('haue', 'done'), 0.0003168191358758069)
(('lord', 'bru'), 0.0003168191358758069)
(('antony', 'ant'), 0.0002772167438913 31)
(('enter', 'lucius'), 0.0002772167438913 31)
(('come', 'downe'), 0.00023761435190685517)
(('euery', 'man'), 0.00023761435190685517)
(('haue', 'seene'), 0.00023761435190685517)
(('caesar', 'shall'), 0.00019801195992237932)
(('caius', 'cassius'), 0.00019801195992237932)
(('caius', 'ligarius'), 0.00019801195992237932)
(('decius', 'brutus'), 0.00019801195992237932)
(("did'st", 'thou'), 0.00019801195992237932)
(('enter', 'antony'), 0.00019801195992237932)
(('fell', 'downe'), 0.00019801195992237932)
(('great', 'caesar'), 0.00019801195992237932)
(('haue', 'beene'), 0.00019801195992237932)
(('haue', 'heard'), 0.00019801195992237932)
(('honourable', 'men'), 0.00019801195992237932)
(('metellus', 'cymber'), 0.00019801195992237932)
(('mine', 'owne'), 0.00019801195992237932)
(('shall', 'finde'), 0.00019801195992237932)
(('shall', 'haue'), 0.00019801195992237932)
(('caesar', 'doth'), 0.00015840956793790345)
(('caesar', 'hath'), 0.00015840956793790345)
(('euery', 'one'), 0.00015840956793790345)
(('hee', 'put'), 0.00015840956793790345)
(('heere', 'comes'), 0.00015840956793790345)
(('honourable', 'man'), 0.00015840956793790345)
(('l', 'heare'), 0.00015840956793790345)
(('messala', 'messa'), 0.00015840956793790345)
(('mine', 'eyes'), 0.00015840956793790345)
(('noble', 'antony'), 0.00015840956793790345)
(('noble', 'caesar'), 0.00015840956793790345)
(('tell', 'thee'), 0.00015840956793790345)
(('thou', 'shalt'), 0.00015840956793790345)
(('wilt', 'thou'), 0.00015840956793790345)
(('ye', 'gods'), 0.00015840956793790345)
(('yong', 'octauius'), 0.00015840956793790345)
(('brother', 'cassius'), 0.00011880717595342758)
```

g. Triagrams:

- **For** *The Man Who Was Thursday by G. K. Chesterton:*

```
('said', 'dr.', 'bull') 0.0003457565585697204
('professor', 'de', 'worms') 0.0002449108956535519
('de', 'st.', 'eustache') 5.76260930949534e-05
('central', 'anarchist', 'council') 4.321956982121505e-05
('marquis', 'de', 'st.') 4.321956982121505e-05
('said', 'syme', 'impatiently') 4.321956982121505e-05
('said', 'syme', 'seriously') 4.321956982121505e-05
('ask', 'one', 'man') 2.88130465474767e-05
('british', 'police', 'force') 2.88130465474767e-05
('de', 'saint', 'eustache') 2.88130465474767e-05
('dr.', 'bull', 'smiled') 2.88130465474767e-05
('dr.', 'bull', 'suddenly') 2.88130465474767e-05
('ever', 'since', 'syme') 2.88130465474767e-05
('fat', 'old', 'gentleman') 2.88130465474767e-05
('first', 'saw', 'sunday') 2.88130465474767e-05
('founded', 'upon', 'love') 2.88130465474767e-05
('heavy', 'iron', 'door') 2.88130465474767e-05
('horses', 'behind', 'us') 2.88130465474767e-05
('hundred', 'yards', 'farther') 2.88130465474767e-05
('let', 'us', 'go') 2.88130465474767e-05
('marquis', 'de', 'saint') 2.88130465474767e-05
('marquis', 'sprang', 'back') 2.88130465474767e-05
('mr.', 'gabriel', 'syme') 2.88130465474767e-05
('mr.', 'joseph', 'chamberlain') 2.88130465474767e-05
('mr.', 'lucian', 'gregory') 2.88130465474767e-05
('peaceable', 'french', 'town') 2.88130465474767e-05
('playing', 'blind', 'man') 2.88130465474767e-05
('said', 'inspector', 'ratcliffe') 2.88130465474767e-05
('said', 'mr.', 'buttons') 2.88130465474767e-05
('said', 'syme', 'firmly') 2.88130465474767e-05
('said', 'syme', 'reflectively') 2.88130465474767e-05
('said', 'syme', 'sardonically') 2.88130465474767e-05
('said', 'syme', 'shortly') 2.88130465474767e-05
('said', 'syme', 'slowly') 2.88130465474767e-05
('said', 'syme', 'thoughtfully') 2.88130465474767e-05
('shall', 'never', 'get') 2.88130465474767e-05
('six', 'men', 'going') 2.88130465474767e-05
('small', 'blue', 'card') 2.88130465474767e-05
('standing', 'quite', 'still') 2.88130465474767e-05
('supreme', 'anarchist', 'council') 2.88130465474767e-05
('syme', 'looked', 'straight') 2.88130465474767e-05
('syme', 'stood', 'staring') 2.88130465474767e-05
('thin', 'red', 'hair') 2.88130465474767e-05
('upper', 'waistcoat', 'pocket') 2.88130465474767e-05
('white', 'cloud', 'went') 2.88130465474767e-05
('within', 'twenty', 'miles') 2.88130465474767e-05
("'nor", 'public', 'flame') 1.440652327373835e-05
("'the", 'celebrated', 'mr.') 1.440652327373835e-05
("'the", 'marquis', 'de') 1.440652327373835e-05
("'why", 'leap', 'ye') 1.440652327373835e-05
("'you", 'prevail', 'like') 1.440652327373835e-05
```

- **For Julius Caesar:**

```
('let', 'vs', 'heare') 0.00011880717595342758
('mine', 'owne', 'part') 0.00011880717595342758
('answer', 'euery', 'man') 7.920478396895172e-05
('caesar', 'shall', 'go') 7.920478396895172e-05
('euery', 'one', 'doth') 7.920478396895172e-05
('great', 'caesar', 'caes') 7.920478396895172e-05
('mark', 'antony', 'ant') 7.920478396895172e-05
('marke', 'antony', 'ant') 7.920478396895172e-05
('noble', 'antony', 'ant') 7.920478396895172e-05
('shall', 'finde', 'time') 7.920478396895172e-05
('successe', 'hath', 'done') 7.920478396895172e-05
('trade', 'art', 'thou') 7.920478396895172e-05
('yet', 'brutus', 'sayes') 7.920478396895172e-05
("'em", 'stay', 'heere') 3.960239198447586e-05
('accents', 'yet', 'vnknowne') 3.960239198447586e-05
('accidentall', 'euils', 'bru') 3.960239198447586e-05
('ages', 'hence', 'shall') 3.960239198447586e-05
('alasse', 'good', 'soule') 3.960239198447586e-05
('ambitious', 'ocean', 'swell') 3.960239198447586e-05
('angry', 'spot', 'doth') 3.960239198447586e-05
('another', 'caesar', 'haue') 3.960239198447586e-05
('another', 'generall', 'shout') 3.960239198447586e-05
('antonio', 'send', 'word') 3.960239198447586e-05
('antony', 'come', 'downe') 3.960239198447586e-05
('antony', 'go', 'vp') 3.960239198447586e-05
('antony', 'haue', 'made') 3.960239198447586e-05
('antony', 'haue', 'spoke') 3.960239198447586e-05
('antony', 'may', 'safely') 3.960239198447586e-05
('antony', 'shall', 'say') 3.960239198447586e-05
('antony', 'shall', 'speake') 3.960239198447586e-05
('arguing', 'make', 'vs') 3.960239198447586e-05
('art', 'mighty', 'yet') 3.960239198447586e-05
('art', 'thou', 'downe') 3.960239198447586e-05
('art', 'thou', 'gone') 3.960239198447586e-05
('art', 'thou', 'heere') 3.960239198447586e-05
('art', 'thou', 'pindarus') 3.960239198447586e-05
('art', 'willing', 'luc') 3.960239198447586e-05
('asse', 'beares', 'gold') 3.960239198447586e-05
("assur'd", 'whether', 'yond') 3.960239198447586e-05
('astonish', 'vs', 'cassi') 3.960239198447586e-05
('auoyded', 'whose', 'end') 3.960239198447586e-05
('away', 'slight', 'man') 3.960239198447586e-05
('away', 'thy', 'face') 3.960239198447586e-05
('bad', 'ayre', 'cassi') 3.960239198447586e-05
('bad', 'soules', 'fla.') 3.960239198447586e-05
('bad', 'strokes', 'brutus') 3.960239198447586e-05
('bad', 'strokes', 'octauius') 3.960239198447586e-05
('bad', 'verses', 'cin') 3.960239198447586e-05
('barren', 'spirited', 'fellow') 3.960239198447586e-05
('base', 'spaniell', 'fawning') 3.960239198447586e-05
```

Mutual Information:

- **For** *The Man Who Was Thursday by G. K. Chesterton:*

```
(('saffron', 'park'), 12.471483550431378)
(('st.', 'eustache'), 12.138059816706186)
(('hundred', 'yards'), 11.328030760350256)
(('leicester', 'square'), 11.12872195212685)
(('de', 'worms'), 11.038524143155271)
(('inspector', 'ratcliffe'), 10.924488899909244)
(('straw', 'hat'), 10.873464896884776)
(('cheers', ')'), 10.760990167626364)
(('contrast', 'between'), 10.64995885523762)
(('supreme', 'council'), 10.190527236600321)
(('colonel', 'ducroix'), 9.897051717202391)
(('any', 'rate'), 9.737143425671993)
(('run', 'away'), 9.72536625789564)
(('blue', 'card'), 9.710053202401138)
(("'ll", 'tell'), 9.147458514708436)
(('police', 'station'), 9.120986303347243)
(('dr.', 'bull'), 9.061244219655356)
(('top', 'hat'), 9.025467990329826)
(('ca', "n't"), 9.021064934020504)
(('dr.', 'renard'), 9.008776799761218)
(('common', 'sense'), 8.912993261071415)
(('am', 'afraid'), 8.811455234609351)
(('red', 'hair'), 8.59106516618405)
(('low', 'voice'), 8.567218424229681)
(('anarchist', 'council'), 8.555441256453328)
(('white', 'road'), 8.263651532212961)
(('old', 'gentleman'), 8.208635455057053)
(('passed', 'through'), 8.200275213151883)
(('let', 'us'), 8.154646649428827)
(('comrade', 'gregory'), 8.128721952126853)
(('my', 'dear'), 7.949519137096527)
(('professor', 'de'), 7.916533618776661)
(('right', 'enough'), 7.860525841177276)
(('sat', 'down'), 7.8470635813707155)
(('dark', 'room'), 7.8134575875204995)
(('blue', 'eyes'), 7.75798767946484)
(('why', 'does'), 7.75021032887312)
(('fell', 'back'), 7.714848385981641)
(('led', 'them'), 7.693179383836741)
(('place', 'where'), 7.656653507811628)
(('quite', 'true'), 7.637489503275402)
(('lit', 'up'), 7.632531783479434)
(('each', 'other'), 7.575123622315029)
(('an', 'instant'), 7.547996444292163)
(('little', 'doctor'), 7.536023802626088)
(('white', 'hair'), 7.503602324933709)
(('found', 'himself'), 7.477438744452057)
(('been', 'broken'), 7.3645566363753705)
(('at', 'least'), 7.36295909365321)
(('do', "n't"), 7.349224982619381)
```

- **For** Julius Caesar:
- **Irregularities**: For MI list of Julius Caesar, I found that my stopwords list was not 100% accurate. The first pair includes “ ’ ” as a word which is not important for my analysis. Therefore I will add “ ’ ” as a stopword in my stopword dictionary.

```
(("'", 'l'), 10.100490946662747)
(('caius', 'ligarius'), 10.075616278023718)
(('metellus', 'cymber'), 10.075616278023718)
(('wee', "'"), 10.007381542271265)
(('mine', 'owne'), 9.027117760332528)
(('any', 'thing'), 8.94316598200007)
(('fell', 'downe'), 8.837456540828954)
(('mark', 'antony'), 8.288319008307367)
(('was', 'ambitious'), 7.879891807149351)
(('at', 'philippi'), 7.866940735441776)
(('marke', 'antony'), 7.658268618057672)
(('good', 'morrow'), 7.587879290166276)
(('most', 'noble'), 7.531295761799909)
(('be', 'satisfied'), 7.424380557883394)
(('what', 'trade'), 7.349791241462713)
(('too', 'much'), 7.181109406871032)
(('thou', 'hast'), 7.141132931160094)
(('honourable', 'men'), 7.138626075549519)
(("did'st", 'thou'), 7.100490946662747)
(('lou', "'d"), 7.0315958654516795)
(('mou', "'d"), 7.0315958654516795)
(('haue', 'seene'), 7.009343058604552)
(('our', 'selues'), 6.9473905678617935)
(('haue', 'beene'), 6.938953730713154)
(('caius', 'cassius'), 6.836150343328329)
(('my', 'lord'), 6.779190246888943)
(('offer', "'d"), 6.768561459617887)
(('enter', 'lucius'), 6.73096810663627)
(('euery', 'man'), 6.620300767858652)
(('let', 'vs'), 6.616324788087507)
(('thou', 'art'), 6.594138280637958)
(('haue', 'heard'), 6.576383651328445)
(('your', 'selues'), 6.530415264341457)
(('he', 'loues'), 6.478375447524126)
(('come', 'downe'), 6.476000081754952)
(('no', 'more'), 6.456862868998687)
(('his', 'body'), 6.436347357744644)
(("'t", 'is'), 6.408200630827961)
(('durst', 'not'), 6.3627414920910095)
(('art', 'thou'), 6.304631663442974)
(('good', 'night'), 6.2468423723312085)
(('ides', 'of'), 6.1564473526367625)
(('tell', 'them'), 6.116918988450124)
(('bring', 'me'), 6.077158442832124)
(('pardon', 'me'), 6.077158442832124)
(('my', 'selfe'), 5.953219646663992)
(('haue', 'done'), 5.9008186018263835)
(('shall', 'finde'), 5.892733871694697)
(('there', "'s"), 5.869165400556291)
(('put', 'it'), 5.802051204697753)
```

Difference between Top Bigrams list and Top PMI pointwise mutual information of the texts:
Bigrams just calculates most frequent pairs of consecutive words (same with trigrams)
PMI calculates the probability or likelihood of one words affecting likelihood of finding another word nearby. For example in case of Julius Caesar, Top bigram is ('let', 'vs') is most frequent pair out of all frequent pair of words. Whereas top PMI in the same document is ('caius', 'ligarius') which means that the likelihood of “ligarius” is increased if word “caius” has occurred in the text and this probability is highest of all other possible combinations in Julius Caesar.

Task 3:

Question for comparison: Explain how *The Man Who Was Thursday by G. K. Chesterton* & *Julius Caesar by William Shakespeare* **differ or are similar** in terms of word frequency, Bigrams, Trigrams, or mutual information between words in both these documents.

- **More importantly explain how the vocabulary in these documents differ because of their differences in creation in history an genres.**

I analyze the two texts: *The Man Who Was Thursday by G. K. Chesterton* & *Julius Caesar by William Shakespeare* **to see if I could point out differentiation between their style of writing and characterizing words and phrases.**

More importantly I tried to find an explaination of how the vocabulary in these documents differ because of their differences in creation in history an genres with help of top words, bigrams, trigrams and mutual information between words.

Similarities:

Despite being created centuries apart, both texts had the protagonist of the story as the highest repeated alphabetical word out of all.

For *The Man Who Was Thursday,* it was *Gabriel Syme, who is a poet and philosopher.*

For *Julius Caesar,* it was *Julius Caesar himself, who was the king who was going to be assassinated by his brother Brutus. The following screenshots of* ***frequency distribution of words*** *shows that both documents had repeatedly used name of their protagonist, and for the highest number of times as compared to other words.*

Most common words in chesterton-thursday.txt

```
In [189]: for pair in Thurs_topwords:
              print(pair)

('syme', 516)
('said', 507)
('man', 274)
('like', 258)
('one', 198)
('professor', 148)
```

Most common words in Julius Caesar by William Shakespeare

```
In [200]: caesar_fdist = FreqDist(caesar_cleaned)
          caesar_topwords = caesar_fdist.most_common(50)
```

```
In [201]: for pair in caesar_topwords:
              print(pair)
          totalwords2 = len(caesar_cleaned)
```

```
('caesar', 190)
('brutus', 161)
('bru', 153)
('haue', 147)
('shall', 125)
('thou', 115)
```

Differences:

1) **The two texts were a lot different in terms of salutation vocabulary used to greet people**
    a. For instance, in *Julius Caesar* Top 50 Bigrams search suggested that the words "good" and "morrow" as in **Good Morrow** which is an archaic greeting were commonly used in the past, especially during the Renaissance era. It essentially meant good morning or good day, a phrase of wishing someone to have a good day.

```
In [130]: for bscore in caesar_scored[:50]:
              print (bscore)
```

```
(('let', 'vs'), 0.0006336382717516138)
(('mark', 'antony'), 0.0005148310957981862)
(('marke', 'antony'), 0.00047522870381371034)
(('thou', 'art'), 0.00043562631182923446)
(('art', 'thou'), 0.00035642152786028277)
(('enter', 'brutus'), 0.00035642152786028277)
(('noble', 'brutus'), 0.00035642152786028277)
(('thou', 'hast'), 0.00035642152786028277)
(('caesar', 'caes'), 0.0003168191358758069)
(('good', 'morrow'), 0.0003168191358758069)
(('good', 'night'), 0.0003168191358758069)
```

    b. But in *The Man Who Was Thursday,* the standard greeting vocabulary included phrase **Good Evening, Good day and Good Morning** which is even used in modern era Nowadays.

2) The Two texts also differed in their writing style:
"The Man Who Was Thursday" is a narrative style combining mystery, satire and philosophical ideas.
"Julius Caesar" is a play written by Shakespeare and it uses Dramatic and poetic languages and dialogues.
    - These different writings styles between them can be explained by glancing at trigrams of both documents:
    - For "The Man Who Was Thursday", Trigram included top trigrams including narrative phrases in descriptive manner like ('dr.', 'bull', 'smiled') **(Dr. Bull smiled)** ('said', 'mr.', 'buttons') **(said Mr. Buttons). Example like these, of trigrams, being highest**

repeated suggest inclusion of narration by author.

Triagrams:

In [133]:
```
from nltk.collocations import TrigramCollocationFinder, TrigramAssocMeasures
thursday_trigramFinder = TrigramCollocationFinder.from_words(thursday_lc)

# Apply word filter using `cff` and `sw`
thursday_trigramFinder.apply_word_filter(cff)
thursday_trigramFinder.apply_word_filter(lambda w: w in sw)

# Score the trigrams
tm = TrigramAssocMeasures()
thursday_scored = thursday_trigramFinder.score_ngrams(tm.raw_freq)

# Print the scored trigrams
for trigram, score in thursday_scored[0:51]:
    print(trigram, score)
```

```
('said', 'dr.', 'bull') 0.0003457565585697204
('professor', 'de', 'worms') 0.0002449108956535519
('de', 'st.', 'eustache') 5.76260930949534e-05
('central', 'anarchist', 'council') 4.321956982121505e-05
('marquis', 'de', 'st.') 4.321956982121505e-05
('said', 'syme', 'impatiently') 4.321956982121505e-05
('said', 'syme', 'seriously') 4.321956982121505e-05
('ask', 'one', 'man') 2.88130465474767e-05
('british', 'police', 'force') 2.88130465474767e-05
('de', 'saint', 'eustache') 2.88130465474767e-05
('dr.', 'bull', 'smiled') 2.88130465474767e-05
('dr.', 'bull', 'suddenly') 2.88130465474767e-05
('ever', 'since', 'syme') 2.88130465474767e-05
('fat', 'old', 'gentleman') 2.88130465474767e-05
('first', 'saw', 'sunday') 2.88130465474767e-05
('founded', 'upon', 'love') 2.88130465474767e-05
('heavy', 'iron', 'door') 2.88130465474767e-05
('horses', 'behind', 'us') 2.88130465474767e-05
('hundred', 'yards', 'farther') 2.88130465474767e-05
('let', 'us', 'go') 2.88130465474767e-05
('marquis', 'de', 'saint') 2.88130465474767e-05
('marquis', 'sprang', 'back') 2.88130465474767e-05
('mr.', 'gabriel', 'syme') 2.88130465474767e-05
('mr.', 'joseph', 'chamberlain') 2.88130465474767e-05
('mr.', 'lucian', 'gregory') 2.88130465474767e-05
('peaceable', 'french', 'town') 2.88130465474767e-05
('playing', 'blind', 'man') 2.88130465474767e-05
('said', 'inspector', 'ratcliffe') 2.88130465474767e-05
('said', 'mr.', 'buttons') 2.88130465474767e-05
('said', 'syme', 'firmly') 2.88130465474767e-05
('said', 'syme', 'reflectively') 2.88130465474767e-05
('said', 'syme', 'sardonically') 2.88130465474767e-05
('said', 'syme', 'shortly') 2.88130465474767e-05
```

- For “Julius Caesar” Most repeated trigrams included phrases like **"art thou gone"** which is archaic form of “are you gone” or **“have you departed”**. These words were **Dialogues that were said again and again** by assassins of Caesar in the play. This shows that “Julius Caesar” was a dramatic play.

('art', 'thou', 'downe') 3.960239198447586e-05
('art', 'thou', 'gone') 3.960239198447586e-05
('art', 'thou', 'heere') 3.960239198447586e-05
('art', 'thou', 'pindarus') 3.960239198447586e-05
('art', 'willing', 'luc') 3.960239198447586e-05
('asse', 'beares', 'gold') 3.960239198447586e-05
("assur'd", 'whether', 'yond') 3.960239198447586e-05