Assignment 1

Dixitha Kasturi
dkasturi@syr.edu

For this assignment I chose 2 books, Harry Potter and the Sorcerer's Stone' book by J.K.Rowling and ‘The Tragedie of Macbeth’ by William Shakespeare 1603

The tragedie of Macbeth is available in the nltk Gutenberg corpus. I took the harry po tter book from a website in pdf format and converted that into txt file outside the pyth on script. I loaded these files into the python notebook for analysis.

For my analysis I am comparing an author from the 1600’s and an author from late 1 900’s or 2000’s, to see and understand how the usage of language changed over the decades. This is my question to be answered.

Initial text understanding :
Length of book Harry Potter and the sorcerers stone= 437161
number of tokens = 98998

Length of book Charlie and the chocolate factory = 100351
number of tokens = 22233

Transforming the tokens into lowercase to give equal priority t words that are the sa me but only vary in case. eg : The and the

To understand the writing style of an author, using frequency distributions would initi ally give a rough idea.

For harry potter there were 5985 unique key out of the 98998 tokens, this shows that there were a lot of repetitions in words usage. Which is around 94% decrease. This shows a lot of repetitions.

For The tragedies of macbeth there were 33554 unique key out of the 22233 tokens, this shows that there were a lot of repetitions in words usage. Which is around 75% decrease. This shows a lot of repetitions.

This in a way shows that J.K Rowling preferred reusing words more compared to Sh akespeare.

Common words :
The top 50 common words in both of these books were:

```
[(',', 5.712236610840623),
 ('.', 5.187983595628194),
 ('the', 3.6798723206529425),
 ("'", 2.570759005232429),
 ('``', 2.3222691367502373),
 ('and', 1.9576153053597043),
 ('to', 1.8737752277823796),
 ('he', 1.772763086122952),
```

```
 ('a', 1.707105194044324),
 ('harry', 1.3485120911533568),
 ('of', 1.271742863492192),
 ('was', 1.2707327420755974),
 ('it', 1.1889129073314613),
 ('you', 1.0101214165942747),
 ("'s", 0.9889088668457949),
 ('in', 0.9727469241802865),
 ('his', 0.9464837673488354),
 ('i', 0.9222408533505727),
 ('-', 0.882846118103396),
 ('said', 0.802036404775854),
 ("n't", 0.8000161619426654),
 ('?', 0.7666821551950544),
 ('had', 0.7626416695286773),
 ('that', 0.6929432917836724),
 ('they', 0.6929432917836724),
 ('on', 0.6424372209539586),
 ('at', 0.6313258853714216),
 ('as', 0.5303137437119942),
 ('him', 0.5060708297137316),
 ('but', 0.48788864421503464),
 ('!', 0.481827915715469),
 ('ron', 0.43334208771894384),
 ('with', 0.4202105093032183),
 ('all', 0.3989979595547385),
 ('what', 0.3818258954726358),
 ('out', 0.37576516697307016),
 ('for', 0.37374492413988164),
 ('hagrid', 0.37374492413988164),
 ('up', 0.37172468130669306),
 ('be', 0.3697044384735045),
 ("'", 0.338390674559082),
 ('were', 0.3343501888927049),
 ('them', 0.32929958180973357),
 ('have', 0.3222287318935736),
 ('there', 0.31313763914422515),
 ('do', 0.30909715347784805),
 ('could', 0.2969756964787l674),
 ('did', 0.28889472514596254),
 ('hermione', 0.27273278248045413),
 ('we', 0.2666720539808885)]
```

For macbeth:

```
[(',', 8.824720010794765),
 ('.', 5.482840822201233),
 ('the', 2.919084244141591),
 ('and', 2.451311114109657),
 (':', 2.1454594521656998),
 ('to', 1.7226645077137586),
 ('of', 1.5202626726037871),
 ('i', 1.4887779426977916),
 ('?', 1.0839742724778483),
 ('a', 1.0749786353618496),
 ('that', 1.0614851796878515),
 ('is', 0.9490397157378672),
 ('you', 0.922052804389871),
```

```
('my', 0.9130571672738721),
('in', 0.8995637115998741),
("'d", 0.8635811631358792),
('not', 0.8320964332298835),
('it', 0.7241487878378986),
('with', 0.6881662393739036),
('his', 0.656681509467908),
('be', 0.6162011424459137),
('macb', 0.6162011424459137),
("'s", 0.5757207754239194),
('your', 0.5667251383079206),
('our', 0.5532316826339225),
('haue', 0.5487338640759232),
('but', 0.5397382269599243),
('what', 0.5262447712859263),
('me', 0.5082534970539289),
('he', 0.5037556784959295),
('for', 0.4902622228219314),
('this', 0.4677731300319345),
('all', 0.44978185579993707),
('so', 0.4317905815679396),
('him', 0.40480367021994335),
('as', 0.40030585166194393),
('thou', 0.3913102145459453),
('we', 0.37331894031394774),
('enter', 0.364323303197949),
('which', 0.3598254846399496),
("'", 0.33283857329195343),
('are', 0.328340754733954),
('will', 0.32384293617595467),
('they', 0.3148472990599559),
('shall', 0.3058516619439572),
('no', 0.30135384338595783),
('then', 0.2833625691539603),
('do', 0.2833625691539603),
('macbeth', 0.278864750595961),
('their', 0.278864750595961)]
```

After looking at these theow frequency distributions, we see that from the harry potte r book, we do see the character names repeating quite frequently. The main charact ers are in the list. Whereas in macbeth, there are common words and sophisticated words like “thou” ,”haue”.

This shows that we need to filter out our data for further analysis. We do bigram and trigram analysis to understand the text better. While doing this we clean our data. For both bigram and trigram collocations, i did analysis using Raw frequencies and P MI.

For both books I did the following :

1) Bigram analysis
    a. Using Raw frequency
        i. without filtering
        ii. with filtering
    b. Using PMI :

i. without filtering
ii. with filtering

2) Trigram analysis
    a. Using Raw frequency
        i. without filtering
        ii. with filtering
    b. Using PMI :
        i. without filtering
        ii. with filtering

For harry potter book:

1) Bigram analysis :
    a. Raw frequency :
        i. without filters :

```
(('.', '``'), 0.015677084385543143)
((',', "'"), 0.007111254772823693)
(('.', "'"), 0.006727408634517869)
(("'", 'said'), 0.005646578718761995)
(('?', "'"), 0.005232428937958343)
(("'", '``'), 0.005060708297137316)
((',', 'and'), 0.004101092951372755)
(('.', 'he'), 0.0038889674538879573)
(('!', "'"), 0.0031717812481060225)
((',', 'but'), 0.0029697569647871673)
(('of', 'the'), 0.002878846037293683)
(('in', 'the'), 0.002717226610638599)
(('.', 'harry'), 0.0026162144689791714)
(("'", 'he'), 0.0026061132548132284)
(('``', 'i'), 0.0024444938281581448)
(("'", 'harry'), 0.002262671973171175)
(('it', 'was'), 0.002262671973171175)
(('on', 'the'), 0.0021616598315117477)
(('he', 'was'), 0.0021010525465160915)
((',', '``'), 0.0020101416190226066)
(('.', 'the'), 0.001868724620699408)
(('did', "n't"), 0.001868724620699408)
(('.', 'it'), 0.0017273076223762096)
(('to', 'the'), 0.0017172064082102668)
((',', 'i'), 0.0016161942665508394)
((',', 'he'), 0.0015050809107254692)
((',', 'harry'), 0.0014747772682276409)
(('harry', ','), 0.0014747772682276409)
(('it', "'s"), 0.0014747772682276409)
(('out', 'of'), 0.0014747772682276409)
(('at', 'the'), 0.0014242711973979272)
(('do', "n't"), 0.0013939675549000099)
(('he', "'d"), 0.0013636639124022707)
(('he', 'had'), 0.0013636639124022707)
(('said', 'harry'), 0.0013636639124022707)
(('.', 'they'), 0.0012828541990747289)
(('harry', '.'), 0.001232348128245015)
((',', 'the'), 0.0011919432715812441)
(('i', "'m"), 0.0011818420574153013)
(('him', '.'), 0.0011111133558253702)
```

```
(('said', 'ron'), 0.0011010323440877594)
(('to', 'be'), 0.0011010323440877594)
(('-', "'"), 0.0010909311299218166)
(('uncle', 'vernon'), 0.0010909311299218166)
(('``', 'you'), 0.001070728701589931)
(('he', 'said'), 0.001070728701589931)
(('in', 'a'), 0.0010606274874239883)
(('harry', "'s"), 0.0010505262732580457)
(("'", "'"), 0.001040425059092103)
(('and', 'the'), 0.0010202226307602174)
```

There are a lot of stop words and characters, so we have to filter these out. After doi ng this, words made better sense. From the highlighted words, we understand the im portant bigrams which are more contextual compared to the unfiltered ones. A lot of i t is conversational, is what we understand. No complicated words.

```
(('said', 'harry'), 0.0013636639124022707)
(('said', 'ron'), 0.0011010323440877594)
(('uncle', 'vernon'), 0.0010909311299218166)
(('professor', 'mcgonagall'), 0.0009495141315986181)
(('said', 'hagrid'), 0.000878805632437019)
(('aunt', 'petunia'), 0.0005252631366290229)
(('harry', 'potter'), 0.0004949594941311946)
(('said', 'hermione'), 0.00042425099496959536)
(('mr.', 'dursley'), 0.0003030364249782824)
(('said', 'dumbledore'), 0.00027273278248045416)
(('harry', 'looked'), 0.0002222267116507404)
(('professor', 'dumbledore'), 0.0002222267116507404)
(('looked', 'like'), 0.00021212549748479768)
(('common', 'room'), 0.00020202428331885492)
(('harry', 'felt'), 0.00020202428331885492)
(('first', 'years'), 0.00019192306915291217)
(('mrs.', 'dursley'), 0.00019192306915291217)
(('professor', 'quirrell'), 0.00019192306915291217)
(('hermione', 'granger'), 0.00017172064082102668)
(('said', 'professor'), 0.00017172064082102668)
(('stone', 'chapter'), 0.00017172064082102668)
(('harry', 'asked'), 0.00016161942665508393)
(('harry', 'thought'), 0.00016161942665508393)
(('mr.', 'ollivander'), 0.00016161942665508393)
(('privet', 'drive'), 0.00016161942665508393)
(('great', 'hall'), 0.0001515182124891412)
(('nimbus', 'two'), 0.0001515182124891412)
(('professor', 'flitwick'), 0.0001515182124891412)
(('first', 'time'), 0.00014141699832319844)
(('invisibility', 'cloak'), 0.00014141699832319844)
(('madam', 'pomfrey'), 0.00014141699832319844)
(('two', 'thousand'), 0.00014141699832319844)
(('get', 'past'), 0.00013131578415725572)
(('said', 'uncle'), 0.00013131578415725572)
(('harry', 'told'), 0.00012121456999131296)
(('madam', 'hooch'), 0.00012121456999131296)
(('mr.', 'potter'), 0.00012121456999131296)
(('mrs.', 'norris'), 0.00012121456999131296)
(('next', 'morning'), 0.00012121456999131296)
(('nicolas', 'flamel'), 0.00012121456999131296)
(('said', 'wood'), 0.00012121456999131296)
```

```
(('tell', 'yeh'), 0.00012121456999131296)
(('harry', 'said'), 0.0001111133558253702)
(('never', 'seen'), 0.0001111133558253702)
(('fifty', 'points'), 0.00010101214165942746)
(('go', 'back'), 0.00010101214165942746)
(('harry', 'saw'), 0.00010101214165942746)
(('harry', 'tried'), 0.00010101214165942746)
(('house', 'cup'), 0.00010101214165942746)
(('leaky', 'cauldron'), 0.00010101214165942746)
```

b. Using PMI

When we use PMI without filtering, these bigrams make more senses comp ared to raw frequencies. the first one is a spell, which is given more likelihood of occ urring together, while the words are more contextual, they are ordered and would req uire more refining

```
(("'hocus", 'pocus'), 16.595111759092934)
(('17', 'railview'), 16.595111759092934)
(('382', 'b.c'), 16.595111759092934)
(('adalbert', 'waffling'), 16.595111759092934)
(('african', 'prince'), 16.595111759092934)
(('alberic', 'grunnion'), 16.595111759092934)
(('amber', 'liquid'), 16.595111759092934)
(('amid', 'gales'), 16.595111759092934)
(('apple', 'pies'), 16.595111759092934)
(('arsenius', 'jigger'), 16.595111759092934)
(('bathilda', 'bagshot'), 16.595111759092934)
(('blackpool', 'pier'), 16.595111759092934)
(('bletchley', 'dives'), 16.595111759092934)
(('booming', 'barks'), 16.595111759092934)
(('burly', 'fifth-year'), 16.595111759092934)
(('butter', 'mellow'), 16.595111759092934)
(('canary-yellow', 'circus'), 16.595111759092934)
(('caput', 'draconis'), 16.595111759092934)
(('cases', 'glimmered'), 16.595111759092934)
(('chilled', 'steel'), 16.595111759092934)
(('circus', 'tent'), 16.595111759092934)
(('cranberry', 'sauce'), 16.595111759092934)
(('d', 'umbledore'), 16.595111759092934)
(('different-colored', 'bonnets'), 16.595111759092934)
(('dinky', 'duddydums'), 16.595111759092934)
(('directing', 'troops'), 16.595111759092934)
(('drift', 'lazily'), 16.595111759092934)
(('druidess', 'cliodna'), 16.595111759092934)
(('elastic', 'band'), 16.595111759092934)
(('emeric', 'switch'), 16.595111759092934)
(('english', 'muffins'), 16.595111759092934)
(('exact', 'art'), 16.595111759092934)
(('exchange', 'mystified'), 16.595111759092934)
(('existence', 'belongs'), 16.595111759092934)
(('fabulous', 'jewels'), 16.595111759092934)
(("figure's", 'prowling'), 16.595111759092934)
(('firecrackers', 'exploding'), 16.595111759092934)
(('firsthand', 'experience'), 16.595111759092934)
(('flanks', 'heaving'), 16.595111759092934)
```

```
(('flowered', 'bonnet'), 16.595111759092934)
(('frantic', 'scrabbling'), 16.595111759092934)
(('funeral', 'march'), 16.595111759092934)
(('gamekeeping', 'duties'), 16.595111759092934)
(('gazed', 'open-mouthed'), 16.595111759092934)
(('gentle', 'drip'), 16.595111759092934)
(('good-for-nothing', 'husband'), 16.595111759092934)
(('grand', 'sorc.'), 16.595111759092934)
(('grassy', 'slope'), 16.595111759092934)
(('greek', 'chappie'), 16.595111759092934)
(('grow-your-own-warts', 'kit'), 16.595111759092934)
```

On filtering out the stop words, non-alphabetic characters and setting a frequency thr eshold as 5, we get the bigrams below, these are more insight giving compared to th e older versions. The ones highlighted below have characters and some imaginary w ords, which give a sense of magical presence.

```
(('diagon', 'alley'), 13.595111759092934)
(('flavor', 'beans'), 13.425186757650623)
(('smelting', 'stick'), 13.273183664205572)
(('bloody', 'baron'), 12.840224256929467)
(('lee', 'jordan'), 12.813752045568274)
(('leaky', 'cauldron'), 12.425186757650621)
(('privet', 'drive'), 12.273183664205572)
(('chocolate', 'frogs'), 12.180074259814091)
(('seamus', 'finnigan'), 12.050791242869124)
(('fat', 'lady'), 11.880866241426812)
(('portrait', 'hole'), 11.787756837035332)
(('eleven', "o'clock"), 11.595111759092935)
(('nicolas', 'flamel'), 11.465828742147968)
(('madam', 'hooch'), 11.425186757650623)
(('madam', 'pomfrey'), 11.425186757650623)
(('nearly', 'headless'), 11.425186757650623)
(('madam', 'malkin'), 11.425186757650621)
(('vault', 'seven'), 11.398714546289431)
(('mrs.', 'figg'), 11.135680140455637)
(('platform', 'nine'), 11.10325866276326)
(('miss', 'granger'), 11.058170773234723)
(('mrs.', 'norris'), 11.020202923035702)
(('every', 'flavor'), 10.867191304529735)
(('invisibility', 'cloak'), 10.762221744928194)
(('fifty', 'points'), 10.757168517201906)
(('entrance', 'hall'), 10.658473820090363)
(('number', 'four'), 10.507648917842594)
(('fast', 'asleep'), 10.465828742147968)
(('aunt', 'petunia'), 10.440293650040829)
(('dark', 'arts'), 10.29743121045225)
(('mr.', 'ollivander'), 10.121406009473517)
(('seven', 'hundred'), 10.08505406738567)
(('third', 'floor'), 10.073511319369207)
(('stone', 'chapter'), 9.965755139013323)
(('common', 'room'), 9.901624801593607)
(('nimbus', 'two'), 9.872645734621843)
(('five', 'minutes'), 9.813752045568274)
(('ten', 'minutes'), 9.781330567875896)
(('weasley', 'twins'), 9.656512303757077)
```

```
(('green', 'light'), 9.61783183559302)
(('living', 'room'), 9.588685489933502)
(('past', 'fluffy'), 9.58388450366968)
(('mrs.', 'dursley'), 9.576252731841617)
(('uncle', 'vernon'), 9.561280928565946)
(('two', 'thousand'), 9.525182547627344)
(('draco', 'malfoy'), 9.477324380985797)
(('mr.', 'dursley'), 9.390866684466745)
(('quidditch', 'match'), 9.297431210452249)
(('fell', 'asleep'), 9.264194880978318)
(('minutes', 'later'), 9.261956408782318)
```

On moving to trigrams, the same repetitions in special characters and commas happ ened. So we filter them out. All of these trigrams are in simple English. Nothing too c omplicated besides the fancy names. More or less the same results were obtained th rough PMI.

```
(('said', 'professor', 'mcgonagall'), 0.00016161942665508393)
(('nimbus', 'two', 'thousand'), 0.00014141699832319844)
(('said', 'uncle', 'vernon'), 0.00013131578415725572)
(('get', 'past', 'fluffy'), 7.070849916159922e-05)
(('gryffindor', 'common', 'room'), 7.070849916159922e-05)
(('every', 'flavor', 'beans'), 6.060728499565648e-05)
(('vault', 'seven', 'hundred'), 6.060728499565648e-05)
(('got', 'ta', 'get'), 4.040485666377098e-05)
(('mr.', 'h.', 'potter'), 4.040485666377098e-05)
(('nearly', 'headless', 'nick'), 4.040485666377098e-05)
(('said', 'aunt', 'petunia'), 4.040485666377098e-05)
(('ter', 'tell', 'yeh'), 4.040485666377098e-05)
(('award', 'gryffindor', 'house'), 3.030364249782824e-05)
(('great', 'uncle', 'algie'), 3.030364249782824e-05)
(('grubby', 'little', 'package'), 3.030364249782824e-05)
(('one', 'thousand', 'magical'), 3.030364249782824e-05)
(('professor', 'mcgonagall', 'told'), 3.030364249782824e-05)
(('said', 'fred', 'weasley'), 3.030364249782824e-05)
(('said', 'harry', 'anxiously'), 3.030364249782824e-05)
(('said', 'mr.', 'ollivander'), 3.030364249782824e-05)
(('thousand', 'magical', 'herbs'), 3.030364249782824e-05)
(('uncle', 'vernon', 'made'), 3.030364249782824e-05)
(('yelled', 'uncle', 'vernon'), 3.030364249782824e-05)
```

When we move to analysing Macbeth, we see difference in the language

Bigrams using raw frequencies :

unfiltered text :

```
((',', 'and'), 0.017991274231997482)
(('macb', '.'), 0.006162011424459137)
((',', 'that'), 0.004272927630099402)
(('.', 'enter'), 0.0032834075473395403)
(('.', 'i'), 0.0031484729905995592)
((',', 'the'), 0.003058516619439572)
((',', 'to'), 0.002698691134799622)
(('macd', '.'), 0.0026087347636396347)
```

```
((',', 'i'), 0.0024738002068996536)
(('.', 'what'), 0.0022489092789996852)
((',', 'as'), 0.0018441056087797419)
((',', 'which'), 0.0018441056087797419)
(('i', 'haue'), 0.0018441056087797419)
(('lady', '.'), 0.0018441056087797419)
((',', 'but'), 0.0017991274231997482)
(('rosse', '.'), 0.0017991274231997482)
(("'d", ','), 0.0017541492376197544)
((':', 'and'), 0.0017091710520397607)
(("'t", 'is'), 0.0013943237529798049)
(('.', 'the'), 0.0013943237529798049)
((',', 'for'), 0.0013043673818198174)
(('.', 'exeunt'), 0.0013043673818198174)
(('exeunt', '.'), 0.0013043673818198174)
(('i', 'am'), 0.0013043673818198174)
((',', 'or'), 0.0012593891962398237)
(('of', 'the'), 0.0012144110106598299)
(('to', 'the'), 0.0012144110106598299)
((',', 'my'), 0.0011694328250798362)
((':', 'i'), 0.0011694328250798362)
(('in', 'the'), 0.0011694328250798362)
(('th', "'"), 0.0011694328250798362)
((',', 'with'), 0.0011244546394998426)
(('.', 'but'), 0.0011244546394998426)
(('mal', '.'), 0.0011244546394998426)
(('banq', '.'), 0.001079476453919849)
(('macbeth', ','), 0.001079476453919849)
((':', 'but'), 0.0010344982683398551)
((',', 'when'), 0.0009895200827598615)
(('and', 'the'), 0.0009895200827598615)
(('my', 'lord'), 0.0009895200827598615)
(('.', 'macb'), 0.0009445418971798677)
(('.', 'scena'), 0.0009445418971798677)
(('1', '.'), 0.0009445418971798677)
((':', 'the'), 0.0009445418971798677)
(('?', 'macb'), 0.0009445418971798677)
(('la', '.'), 0.0009445418971798677)
(('me', ','), 0.0009445418971798677)
(('.', 'my'), 0.0008995637115998741)
(('can', 'not'), 0.0008995637115998741)
(('doct', '.'), 0.0008995637115998741)
```

filtered text :

```
(('enter', 'macbeth'), 0.0007196509692798992)
(('thou', 'art'), 0.0004048036702199433)
(('good', 'lord'), 0.0003598254846399496)
(('haue', 'done'), 0.0003598254846399496)
(('let', 'vs'), 0.00031484729905995593)
(('lord', 'macb'), 0.00031484729905995593)
(('enter', 'lady'), 0.00026986911347996224)
(('enter', 'malcolme'), 0.00022489092789996853)
(('euery', 'one'), 0.00022489092789996853)
(('make', 'vs'), 0.00022489092789996853)
(('mine', 'eyes'), 0.00022489092789996853)
(('mine', 'owne'), 0.00022489092789996853)
```

```
(('scena', 'secunda'), 0.00022489092789996853)
(('three', 'witches'), 0.00022489092789996853)
(('thy', 'selfe'), 0.00022489092789996853)
(('worthy', 'thane'), 0.00022489092789996853)
(('enter', 'banquo'), 0.0001799127423199748)
(('enter', 'macduffe'), 0.0001799127423199748)
(('enter', 'rosse'), 0.0001799127423199748)
(('haile', 'king'), 0.0001799127423199748)
(('haile', 'macbeth'), 0.0001799127423199748)
(('hath', 'made'), 0.0001799127423199748)
(('old', 'man'), 0.0001799127423199748)
(('scena', 'prima'), 0.0001799127423199748)
(('see', 'thee'), 0.0001799127423199748)
(('ten', 'thousand'), 0.0001799127423199748)
(('thy', 'face'), 0.0001799127423199748)
(('woman', 'borne'), 0.0001799127423199748)
(('art', 'thou'), 0.00013493455673998112)
(('byrnane', 'wood'), 0.00013493455673998112)
(('cauldron', 'bubble'), 0.00013493455673998112)
(('enter', 'king'), 0.00013493455673998112)
(('fill', 'vp'), 0.00013493455673998112)
(('fire', 'burne'), 0.00013493455673998112)
(('get', 'thee'), 0.00013493455673998112)
(('giue', 'thee'), 0.00013493455673998112)
(('hath', 'beene'), 0.00013493455673998112)
(('haue', 'beene'), 0.00013493455673998112)
(('haue', 'knowne'), 0.00013493455673998112)
(('haue', 'seene'), 0.00013493455673998112)
(('haue', 'thee'), 0.00013493455673998112)
(('ile', 'doe'), 0.00013493455673998112)
(('macbeth', 'shall'), 0.00013493455673998112)
(('night', 'lady'), 0.00013493455673998112)
(('scena', 'quarta'), 0.00013493455673998112)
(('scena', 'quinta'), 0.00013493455673998112)
(('scena', 'tertia'), 0.00013493455673998112)
(('shall', 'neuer'), 0.00013493455673998112)
(('sir', 'macb'), 0.00013493455673998112)
(('thee', 'still'), 0.00013493455673998112)
```

Sophisticated words like thou, thee, art, shall, which are not commonly used anymor e were occurring in the bigrams. This partly shows the old traditional way of writing.

On using PMI :
without filtering: the adjectives used are very rich in vocabulary. eg : dire, antidote, c areless, dauntlesse etc.

```
(('accompany', 'old-age'), 14.440415010465204)
(('accounted', 'dangerous'), 14.440415010465204)
(('acheron', 'meete'), 14.440415010465204)
(('actuall', 'performances'), 14.440415010465204)
(('adders', 'forke'), 14.440415010465204)
(('afterwards', 'seale'), 14.440415010465204)
(('alarme', 'excite'), 14.440415010465204)
(('all-thing', 'vnbecomming'), 14.440415010465204)
(('alter', 'fauor'), 14.440415010465204)
```

```
(('anoynted', 'temple'), 14.440415010465204)
(('antidote', 'cleanse'), 14.440415010465204)
(('augure', 'hole'), 14.440415010465204)
(('barren', 'scepter'), 14.440415010465204)
(('bladed', 'corne'), 14.440415010465204)
(('blaspheming', 'iew'), 14.440415010465204)
(('blinde-wormes', 'sting'), 14.440415010465204)
(('bonelesse', 'gummes'), 14.440415010465204)
(('boundlesse', 'intemperance'), 14.440415010465204)
(('breefe', 'candle'), 14.440415010465204)
(('cannons', 'ouer-charg'), 14.440415010465204)
(('carelesse', 'trifle'), 14.440415010465204)
(('charnell', 'houses'), 14.440415010465204)
(('childrens', 'ghosts'), 14.440415010465204)
(('clamorous', 'harbingers'), 14.440415010465204)
(('climbe', 'vpward'), 14.440415010465204)
(('colmes', 'ynch'), 14.440415010465204)
(('compunctious', 'visitings'), 14.440415010465204)
(('continent', 'impediments'), 14.440415010465204)
(('craues', 'composition'), 14.440415010465204)
(('create', 'soldiours'), 14.440415010465204)
(('dauntlesse', 'temper'), 14.440415010465204)
(('deadly', 'greefe'), 14.440415010465204)
(('deaths', 'counterfeit'), 14.440415010465204)
(('deerest', 'cooz'), 14.440415010465204)
(('desolate', 'shade'), 14.440415010465204)
(('digestion', 'waite'), 14.440415010465204)
(('dire', 'distresses'), 14.440415010465204)
(('direfull', 'thunders'), 14.440415010465204)
(('direnesse', 'familiar'), 14.440415010465204)
(('direst', 'crueltie'), 14.440415010465204)
(('discouery', 'erre'), 14.440415010465204)
(('disloyall', 'traytor'), 14.440415010465204)
(('doubly', 'redoubled'), 14.440415010465204)
(('downfall', 'birthdome'), 14.440415010465204)
(('dread', 'exploits'), 14.440415010465204)
(('dunnest', 'smoake'), 14.440415010465204)
(('dwarfish', 'theefe'), 14.440415010465204)
(('dyre', 'combustion'), 14.440415010465204)
(('eighth', 'appeares'), 14.440415010465204)
(('especially', 'prouoke'), 14.440415010465204)
```

on filtering out stop words, special characters and setting a frequency threshold of 5, I got the following bigrams, which were a little limited.

```
(('thou', 'art'), 6.997471514616473)
(('good', 'lord'), 6.6075249963004605)
(('enter', 'macbeth'), 6.146368697193703)
(('let', 'vs'), 6.048097587686444)
(('haue', 'done'), 5.38039465595735)
(('enter', 'lady'), 5.100565007580579)
(('lord', 'macb'), 4.901810336118695)
```

For trigram analysis, I used the same approach with raw frequencies and PMI

Raw frequencies without filtering:

```
(('.', 'exeunt', '.'), 0.0012593891962398237)
(('.', 'macb', '.'), 0.0009445418971798677)
(('?', 'macb', '.'), 0.0009445418971798677)
((',', 'and', 'the'), 0.0008545855260198804)
(('exeunt', '.', 'scena'), 0.0006746727836999055)
(('.', 'enter', 'macbeth'), 0.0006296945981199119)
(('my', 'lord', ','), 0.0006296945981199119)
(('macb', '.', 'i'), 0.0005847164125399181)
(('thane', 'of', 'cawdor'), 0.0005847164125399181)
(('?', 'rosse', '.'), 0.0005397382269599245)
((',', 'and', 'yet'), 0.0004947600413799307)
(('.', 'i', ','), 0.0004947600413799307)
(('.', 'i', 'haue'), 0.0004947600413799307)
((',', 'that', 'i'), 0.00044978185579993706)
(('.', "'t", 'is'), 0.0004048036702199433)
(('.', 'what', "'s"), 0.0004048036702199433)
(('?', 'lady', '.'), 0.0004048036702199433)
(("'d", ',', 'and'), 0.0003598254846399496)
((',', 'my', 'lord'), 0.0003598254846399496)
(('.', 'what', 'is'), 0.0003598254846399496)
(('?', 'macd', '.'), 0.0003598254846399496)
(('my', 'good', 'lord'), 0.0003598254846399496)
(('the', 'thane', 'of'), 0.0003598254846399496)
((',', 'and', 'a'), 0.00031484729905995593)
((',', 'and', 'to'), 0.00031484729905995593)
(('.', 'enter', '.'), 0.00031484729905995593)
(('.', 'i', 'am'), 0.00031484729905995593)
(('.', 'macd', '.'), 0.00031484729905995593)
(('enter', 'macbeth', '.'), 0.00031484729905995593)
(('i', 'pray', 'you'), 0.00031484729905995593)
(('lord', 'macb', '.'), 0.00031484729905995593)
(('wee', "'", 'l'), 0.00031484729905995593)
((',', 'and', 'my'), 0.00026986911347996224)
((',', 'and', 'with'), 0.00026986911347996224)
((',', 'i', 'haue'), 0.00026986911347996224)
((',', 'i', 'will'), 0.00026986911347996224)
((',', 'lenox', ','), 0.00026986911347996224)
((',', 'my', 'good'), 0.00026986911347996224)
((',', 'to', 'make'), 0.00026986911347996224)
(('.', 'enter', 'lady'), 0.00026986911347996224)
(('.', 'lady', '.'), 0.00026986911347996224)
(('.', 'mal', '.'), 0.00026986911347996224)
(('.', 'there', "'s"), 0.00026986911347996224)
(('.', 'well', ','), 0.00026986911347996224)
(((':', 'i', 'haue'), 0.00026986911347996224)
(('?', 'wife', '.'), 0.00026986911347996224)
(('can', 'not', 'be'), 0.00026986911347996224)
(('enter', 'macbeth', ','), 0.00026986911347996224)
(('knock', ',', 'knock'), 0.00026986911347996224)
(('macb', '.', 'if'), 0.00026986911347996224)
```

using raw frequencies with filtering : The trigrams solidify out conclusion that a very s ophisticated language was used, which is not used in todays date anymore.

```
('good', 'lord', 'macb'), 0.00013493455673998112)
(('god', 'blesse', 'vs'), 8.99563711599874e-05)
(('see', 'thee', 'still'), 8.99563711599874e-05)
(('till', 'byrnane', 'wood'), 8.99563711599874e-05)
(("'gainst", 'nature', 'still'), 4.49781855799937e-05)
(("'twould", 'haue', 'anger'), 4.49781855799937e-05)
(('aboue', 'deale', 'betweene'), 4.49781855799937e-05)
(('account', 'thy', 'loue'), 4.49781855799937e-05)
(('accounted', 'dangerous', 'folly'), 4.49781855799937e-05)
(('alarum', 'enter', 'macbeth'), 4.49781855799937e-05)
(('alas', 'poore', 'countrey'), 4.49781855799937e-05)
(('aliue', 'till', 'famine'), 4.49781855799937e-05)
(('all-thing', 'vnbecomming', 'macb'), 4.49781855799937e-05)
(("and't", 'please', 'heauen'), 4.49781855799937e-05)
(('angry', 'god', 'macd'), 4.49781855799937e-05)
(('approach', 'old', 'seyward'), 4.49781855799937e-05)
(('approach', 'thou', 'like'), 4.49781855799937e-05)
(('arme', "'gainst", 'arme'), 4.49781855799937e-05)
(('art', 'thou', "affear'd"), 4.49781855799937e-05)
(('assistance', 'doe', 'make'), 4.49781855799937e-05)
(('auarice', 'stickes', 'deeper'), 4.49781855799937e-05)
(('ayde', 'doth', 'seeme'), 4.49781855799937e-05)
(('babes', 'sauagely', 'slaughter'), 4.49781855799937e-05)
(('banquets', 'bloody', 'kniues'), 4.49781855799937e-05)
(('banquo', 'smiles', 'vpon'), 4.49781855799937e-05)
(('banquo', 'sticke', 'deepe'), 4.49781855799937e-05)
(('bat', 'hath', 'flowne'), 4.49781855799937e-05)
(('beare', 'thy', 'prayses'), 4.49781855799937e-05)
(('bene', 'shed', 'ere'), 4.49781855799937e-05)
(('beside', 'vs', 'sey'), 4.49781855799937e-05)
(('better', 'health', 'attend'), 4.49781855799937e-05)
(('better', 'thee', 'without'), 4.49781855799937e-05)
(('bid', 'god-eyld', 'vs'), 4.49781855799937e-05)
(('bid', 'thee', 'ioyne'), 4.49781855799937e-05)
(('bid', 'thy', 'mistresse'), 4.49781855799937e-05)
(('bid', 'vs', 'welcome'), 4.49781855799937e-05)
(('bin', 'strangely', 'borne'), 4.49781855799937e-05)
(('bird', 'hath', 'made'), 4.49781855799937e-05)
(('birnane', 'forrest', 'come'), 4.49781855799937e-05)
(('black', 'heccats', 'summons'), 4.49781855799937e-05)
(('bleed', 'poore', 'country'), 4.49781855799937e-05)
(('blesse', 'vs', 'lady'), 4.49781855799937e-05)
(('blessing', 'may', 'soone'), 4.49781855799937e-05)
(('blood', 'hath', 'bene'), 4.49781855799937e-05)
(('blood', 'vpon', 'thy'), 4.49781855799937e-05)
(('borne', 'shall', 'harme'), 4.49781855799937e-05)
(('bosomes', 'empty', 'macd'), 4.49781855799937e-05)
(('bought', 'golden', 'opinions'), 4.49781855799937e-05)
(('bounteous', 'nature', 'hath'), 4.49781855799937e-05)
(('boyle', 'thou', 'first'), 4.49781855799937e-05)
```

To see if PMI betters the previous results:
without filtering :

```
('forge', 'quarrels', 'vniust'), 28.880830020930407)
(('grim', 'alarme', 'excite'), 28.880830020930407)
(('lated', 'traueller', 'apace'), 28.880830020930407)
(('lifes', 'fitfull', 'feuer'), 28.880830020930407)
(('minutely', 'reuolts', 'vpbraid'), 28.880830020930407)
(('multitudinous', 'seas', 'incarnardine'), 28.880830020930407)
(('obliuious', 'antidote', 'cleanse'), 28.880830020930407)
(('quoth', 'i.', 'aroynt'), 28.880830020930407)
(('saint', 'colmes', 'ynch'), 28.880830020930407)
(('william', 'shakespeare', '1603'), 28.880830020930407)
(('accounted', 'dangerous', 'folly'), 27.880830020930407)
(('auarice', 'stickes', 'deeper'), 27.880830020930407)
(('choppie', 'finger', 'laying'), 27.880830020930407)
(('doubly', 'redoubled', 'stroakes'), 27.880830020930407)
(('hideous', 'trumpet', 'calls'), 27.880830020930407)
(('humane', 'statute', 'purg'), 27.880830020930407)
(('iourney', 'soundly', 'inuite'), 27.880830020930407)
(('mothers', 'womb', 'vntimely'), 27.880830020930407)
(('ruines', 'wastfull', 'entrance'), 27.880830020930407)
(('rumpe-fed', 'ronyon', 'cryes'), 27.880830020930407)
(('sad', 'bosomes', 'empty'), 27.880830020930407)
(('sore', 'labors', 'bath'), 27.880830020930407)
(('trebble', 'scepters', 'carry'), 27.880830020930407)
(('weightie', 'reasons', '2.murth'), 27.880830020930407)
(('womb', 'vntimely', 'ript'), 27.880830020930407)
(('yesty', 'waues', 'confound'), 27.880830020930407)
(('castles', 'gently', 'rendred'), 27.29586752020925)
(('dismall', 'treatise', 'rowze'), 27.29586752020925)
(('euen-handed', 'iustice', 'commends'), 27.29586752020925)
(('hard', 'iourney', 'soundly'), 27.29586752020925)
(('interim', 'hauing', 'weigh'), 27.29586752020925)
(('m.', 'gods', 'benyson'), 27.29586752020925)
(('neptunes', 'ocean', 'wash'), 27.29586752020925)
(('salt', 'sea', 'sharke'), 27.29586752020925)
(('sometime', 'accounted', 'dangerous'), 27.29586752020925)
(('stout', 'norweyan', 'rankes'), 27.29586752020925)
(('sundry', 'weightie', 'reasons'), 27.29586752020925)
(('valued', 'file', 'distinguishes'), 27.29586752020925)
(("you'le", 'sweat', "for't"), 27.29586752020925)
(('mortals', 'cheefest', 'enemie'), 26.880830020930407)
(('mowsing', 'owle', 'hawkt'), 26.880830020930407)
(('n', ']', 'miserable'), 26.880830020930407)
(('naked', 'frailties', 'hid'), 26.880830020930407)
(('natio', '[', 'n'), 26.880830020930407)
(('shakespeare', '1603', ']'), 26.880830020930407)
(('solely', 'soueraigne', 'sway'), 26.880830020930407)
(('sweet', 'obliuious', 'antidote'), 26.880830020930407)
(('sweltred', 'venom', 'sleeping'), 26.880830020930407)
(('tarquins', 'rauishing', 'sides'), 26.880830020930407)
(('wicked', 'dreames', 'abuse'), 26.880830020930407)
```

with filtering :
when the frequency limit was set to 5, I did not get any results. When I set it to 2 I got the following, which aren't any better than the previous.

```
(('till', 'byrnane', 'wood'), 20.099470307405745)
(('god', 'blesse', 'vs'), 18.788072880010553)
(('see', 'thee', 'still'), 14.74796885953706)
(('good', 'lord', 'macb'), 12.534870424526293)
```

The question I was trying to answer was if there was a difference in language usage from 1600's to 1900's and general context of the books,

From bigram and trigram analysis, we see that there is a lot of difference in how words are used, while common English was used in harry potter like you, she ,he, his, her etc, in macbeth words like thee, thy, thou were used which were more sophisticated. Some complex vocabulary was used in macbeth like dire, antidote, careless, dauntlesse as adjectives, whereas in harry potter no complex adjectives were used.
The harry potter book showed magical and imaginary collocations like chocolate frog, invisibility cloak etc where as macbeth was more empathetic/ sorrowful in a way eg : dire, distresses.
In harry potter, PMI gave better results for bigrams and trigrams. where as for macbe th raw frequencies made sense.

Conclusion :
With most frequent words ,bigram and trigram analysis of a text corpus, we see word s which were used in old ages compared to centuries later. This overview tells us whi ch era a certain text corpus could belong to. Harry potter was more conversational co mpared to macbeth which seems very narrative. Harry potter uses simple yet mystica l language where as macbeth uses old sophisticated language. Both books are really famous.

Reporting :
I tried out different combinations of filters and frequency thresholds to finally settle do wn on 5 and 2 which were mentioned above. For some instances raw frequencies ga ve better insights compared to PMI and vice versa.