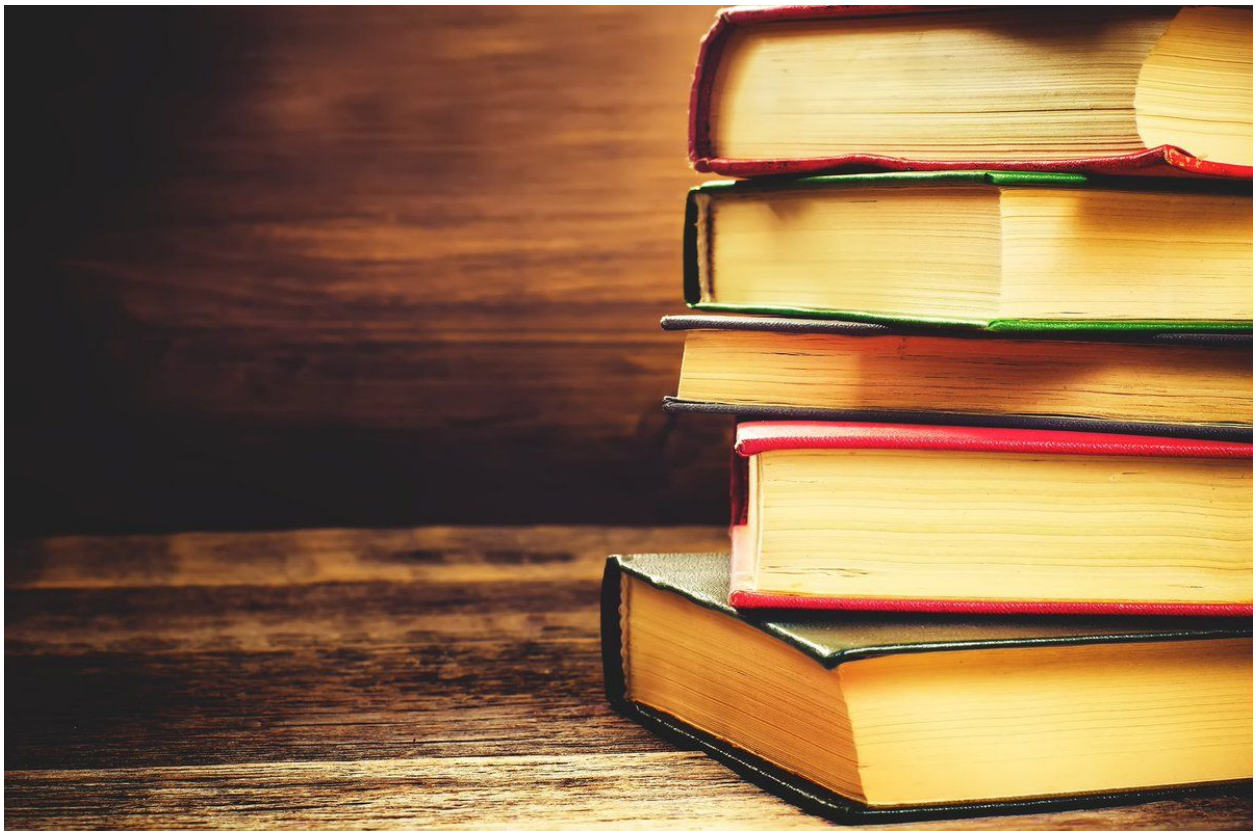
ALI HO & KENDRA OSBURN | 4-24-19 | IST NLP | HW1 | TOKENS & BIGRAMS

COMPARING HARRYS



Introduction

Natural language processing breaks down conversational language into points of data that can be analyzed. When a layperson thinks of data, they likely imagine a spreadsheet or list of numbers. In reality, however, data comes in many different formats. One of the most important data formats is spoken or written language.

Of course, language is subjective, with differences in tone and meaning that allow a variety of interpretations. In order to effectively analyze language, we must first turn it into something that can be measured.

Turning language into information that we can process begins by taking each word and breaking it down into a “token,” or quantifiable data point. In turn, tokens can be combined

into larger units, like bigrams (two words) or trigrams (three words). By converting words to tokens, these new data points can answer a variety of useful questions.

For example, natural language processing can be used to learn more about the subject matter of a given text. By comparing token, bigram and trigram frequencies, we can illuminate trends character development and even identify potential villains. Additionally, we can identify trends in tone and sentiment, helping determine if the book is lighter or darker in subject matter.

Methodology & Results

Methodology

Data Collection

The text of all seven Harry Potter books were found online in pdf form. They were all downloaded and converted to .txt documents before being saved and imported into the attached project directory. Python was used to write the program. Python's Natural Language Toolkit (NLTK) was used for the textual analysis. Within NLTK, Frequency Distribution was used as well as Collocations.

Data processing

Single Word Frequency

Before processing all seven books, the first book was used as a template. The first analysis began simply with Harry Potter book 1. The text was imported, read and the first 150 words were examined. The length was determined and stored for later. The NLTK tokenizer that was trained on news articles (word_tokenize) was used. This separated the text by white space and special characters but left words like the salutation "Mr." together. The length of the tokenized text was determined and stored for later. The tokens were all converted to lowercase and stopwords were removed. The initial analysis used only the 'english' stopwords included in `nltk.corpus.stopwords('english')`. Over the course of the processing, additional stopwords were added. First, stylistic halfwords spoken by one of the main characters (Hagrid) were removed. Eventually, more stopwords were added to the list.

One stopword was removed and readded (said). The final list of stopwords removed from the document (in addition to the nltk english stopwords) were:

```
["'re", "'ve", "'ll", "'d", "'s", "'t", "chapter", "'m","\"", "1"]
```

Then a frequency distribution was applied to all the words. The 50 most common words were pulled -- with their frequency count -- for analysis. A normalized frequency count was also determined by dividing each frequency by the total number of tokens. Initially, punctuation was included in the frequency distribution. This wasn't useful to our analysis so punctuation was removed by selecting only alphanumeric characters using:

```
if word.isalpha()
```

Bigram Frequency

A list of brigrams was created using nltk.bigrams. The frequency distribution was applied to the bigrams. The 50 most common bigrams were pulled -- with their frequency count-- for analysis. Initially, stopwords were added back to the dataset to get an accurate bigram count. This produced a haphazard and fairly meaningless list of bigrams, so the stopwords were again removed, as well as the punctuation.

Bigram Score

To get the Bigram score, nltk.collocations.BigramAssocMeasures() was used. Then, the BigramCollocationFinder (also from nltk) was used. To get the Bigram score, this line was used:

```
scored = finder.score_ngrams(bigram_measures.raw_freq)
```

Bigram PMI

The BigramCollocationFinder was again used to get the Bigram PMI.

```
scored = finder.score_ngrams(bigram_measures.pmi)
```

In order to get the most meaningful bigram scores, a minimum frequency of 5 was applied, as a high PMI score does not show the frequency of the bigrams occurring but rather shows how likely the two words appear together. It does not add value if the words "Openous Maximus" appeared together every time either word appeared if the words only

appeared 2 or 3 times. Therefore, it was decided that a minimum frequency for the bigrams of 5 needed to be applied.

```
finder3.apply_freq_filter(5)
scored = finder2.score_ngrams(bigram_measures.pmi)
```

Trigrams

After pulling the bigrams, the "said professor" pattern prompted us to wonder "which professor was doing the saying." So we decided to introduce trigrams as well. The aforementioned processes were repeated with trigrams using:

```
Nltk.trigrams
```

Results

Question 1:

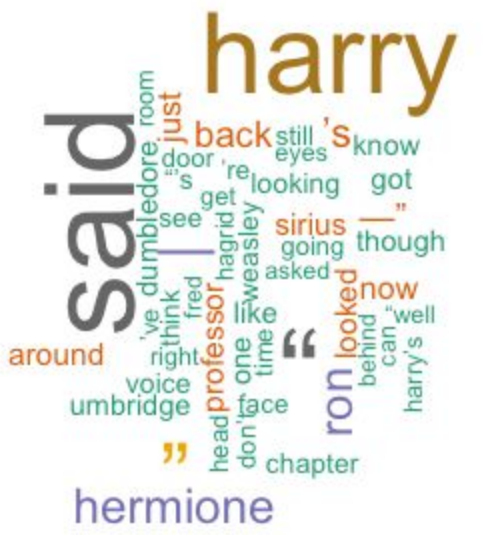
The text of all seven Harry Potter books were found online in pdf form. They were all downloaded and converted to .txt documents before being saved and imported into the attached project directory. Python was used to write the program. Python's Natural Language Toolkit (NLTK) was used for the textual analysis. Within NLTK, Frequency Distribution was used as well as Collocations. The original pdfs can be found at https://305digitallibrary.weebly.com/harry-potter-series.html

Question 2:

Tokenization was used to turn the words into data points. The tokenizer trained on news articles (the one that removes white spaces and special characters while still leaving words like salutations) was used for this project. We chose this tokenizer because it was most applicable to the texts we were analyzing. For the analysis, it was decided that it did not matter if words appeared at the beginning of the sentence and therefore were capitalized. A complete count of individual words, regardless of capitalization was desired. Therefore, the words were all transformed to **lowercase** to ensure all words were treated equally, regardless of being a proper noun or starting a sentence. This wasn't useful to our analysis so punctuation was eventually removed by selecting only alphanumeric characters using:

```
if word.isalpha()
```

Stop words were used to remove excess noise from the text that might detract from analysis. It was decided that traditional stopwords from the nltk stopwords corpus were appropriate to remove, as they were deemed not to provide useful information when analyze the text. Some examples of the nltk built in stopwords are: I, me, my, we, our, you, he, her, a if, and or. These words by themselves do not aid in text analysis, and therefore



removed. Word clouds were created for each Harry Potter text and as a result "said" was initially removed. However, after further thought and discussion it was decided that it might be interesting to see which character talked the most throughout the books and "said" was reintroduced. Through tokenization tokens were created that needed to be removed. The additional stop words that were deemed not to provide any insightful information and appear in a large quantity were "'re", "'ve", "'ll", "'d", "'s", and "'t". Finally "chapter" and "rowling" were removed.

Lemmatization was not utilized in the processing of words. It was decided that having the words in whole was important for understanding and analysis of the books.

Top 50 Word By Frequency - Normalized by the Length of the Document

```
NORMALIZED FREQUENCY
Harry Potter 1
('harry', 0.032679250913974486)
('said', 0.019746822850605587)
('ron', 0.01066925315227934)
('hagrid', 0.00912730980626228)
('could', 0.0073117958665970305)
('hermione', 0.0067149145713646195)
```

```
('back', 0.006491084085652466)
('one', 0.006366733815812381)
('got', 0.005123231117411525)
('like', 0.00482479046979532)
('get', 0.004799920415827302)
('know', 0.004675570145987217)
('see', 0.00447660971424308)
('professor', 0.00447660971424308)
('snape', 0.004227909174562909)
('looked', 0.004203039120594891)
('dumbledore', 0.0039294685269467035)
('would', 0.003854858365042652)
('around', 0.0035315476634584296)
('dudley', 0.0034071973936183443)
('going', 0.0033574572856823097)
('go', 0.003332587231714293)
('something', 0.0032828471237782587)
('never', 0.003133626799970156)
('look', 0.003133626799970156)
('right', 0.003108756746002139)
('think', 0.0030590166380661046)
('uncle', 0.0030341465840980873)
('malfoy', 0.0030092765301300704)
('time', 0.0029595364221940363)
('vernon', 0.0028849262602899848)
('yeh', 0.0028849262602899848)
('neville', 0.0028351861523539506)
('first', 0.002760575990449899)
('quirrell', 0.0027357059364818822)
('door', 0.002636225720609814)
('well', 0.002636225720609814)
('even', 0.0026113556666417966)
('eyes', 0.0026113556666417966)
('potter', 0.0025367455047377455)
('mcgonagall', 0.002511875450769728)
('two', 0.002487005396801711)
('head', 0.002462135342833694)
('looking', 0.0024372652888656767)
('people', 0.00241239523489766)
('thought', 0.0023875251809296425)
('next', 0.0023875251809296425)
('come', 0.0023626551269616256)
```

```
('way', 0.0023377850729936083)
('told', 0.0023129150190255915)
HP2.txt
HP2.txt
NORMALIZED FREQUENCY - Harry Potter 2
('harry', 0.036156124760411225)
('said', 0.026507231224952084)
('ron', 0.015533681826102109)
('hermione', 0.0069480745774525175)
('back', 0.0060115002613695765)
('lockhart', 0.005314514723819481)
('one', 0.0048571179648022305)
('malfoy', 0.004813556368705349)
('could', 0.004595748388220944)
('professor', 0.004181913225300575)
('got', 0.004094790033106813)
('like', 0.0038552012545739677)
('around', 0.003768078062380206)
('weasley', 0.0037245164662833246)
('dobby', 0.0036373932740895627)
('know', 0.003550270081895801)
('hagrid', 0.0035067084857989196)
('dumbledore', 0.003463146889702039)
('looked', 0.003397804495556717)
('go', 0.003332462101411396)
('see', 0.003245338909217634)
('would', 0.003245338909217634)
('eyes', 0.0032017773131207527)
('think', 0.0032017773131207527)
('time', 0.003158215717023872)
('potter', 0.0030493117267816694)
('get', 0.003027530928733229)
('riddle', 0.0029186269384910263)
('right', 0.002831503746297 2644)
('still', 0.0027661613521519426)
('door', 0.0026790381599581807)
('look', 0.0026572573619097406)
('face', 0.0026136957658128594)
('head', 0.002570134169715978)
('slytherin', 0.002548353371667538)
('come', 0.002548353371667538)
('ginny', 0.002548353371667538)
```

```
('never', 0.0025265725736190975)
('looking', 0.0024830109775222163)
('long', 0.002461230179473776)
('though', 0.002461230179473776)
('voice', 0.002461230179473776)
('something', 0.0023741069872800138)
('school', 0.002330545391183133)
('wand', 0.002330545391183133)
('well', 0.002330545391183133)
('going', 0.002330545391183133)
('let', 0.0023087645931346924)
('chamber', 0.002243422198989371)
('fred', 0.002243422198989371)
HP3.txt
HP3.txt
NORMALIZED FREQUENCY - Harry Potter 3
('harry', 0.035364714834157644)
('said', 0.025524850413308892)
('ron', 0.013644381420122436)
('hermione', 0.011828589215923634)
('professor', 0.0072977553349704285)
('lupin', 0.0070383564486563134)
('black', 0.006588731712378515)
('back', 0.006104520457925501)
('one', 0.0051360979490194724)
('hagrid', 0.004617300176391242)
('snape', 0.004427074326427558)
('around', 0.004254141735551482)
('looked', 0.004184968699201051)
('like', 0.004150382181025836)
('could', 0.003994742849237367)
('see', 0.0038909832947117214)
('got', 0.003856396776536506)
('get', 0.0034240652993463147)
('know', 0.003303012485733061)
('still', 0.003251132708470238)
('malfoy', 0.003251132708470238)
('would', 0.003112786635769377)
('go', 0.003078200117594162)
('eyes', 0.003060906858506554)
('though', 0.0028879742676304774)
('face', 0.0028879742676304774)
```

```
('time', 0.002836094490367655)
('looking', 0.002836094490367655)
('going', 0.002818801231280047)
('right', 0.0028015079721924394)
('dumbledore', 0.0028015079721924394)
('think', 0.0026458686404039707)
('well', 0.0025075225677031092)
('door', 0.0024383495313526787)
('come', 0.002421056272265071)
('saw', 0.002421056272265071)
('sirius', 0.0024037630131774636)
('head', 0.0023864697540898557)
('voice', 0.002369176495002248)
('look', 0.0023345899768270327)
('away', 0.002317296717739425)
('toward', 0.0023000034586518176)
('gryffindor', 0.002230830422301387)
('something', 0.002178950645038564)
('heard', 0.0020924843496005255)
('behind', 0.0020924843496005255)
('last', 0.002040604572337703)
('hand', 0.0020060180541624875)
('ever', 0.0020060180541624875)
('turned', 0.00198872479507488)
HP4.txt
HP4.txt
NORMALIZED FREQUENCY - Harry Potter 4
('harry', 0.03146821438287839)
('said', 0.026530838447044125)
('ron', 0.010418870852352307)
('hermione', 0.008705903690940419)
('back', 0.005934927400421189)
('dumbledore', 0.005894622290740908)
('could', 0.005824088348800419)
('would', 0.004957528490674405)
('around', 0.004937375935834265)
('one', 0.004897070826153985)
('looked', 0.004866841993893776)
('like', 0.00432272301321)
('though', 0.00423203651642937)
('got', 0.0042219602390093005)
('weasley', 0.003909595638987132)
```

```
('hagrid', 0.0037584514776860837)
('moody', 0.003587154761544895)
('know', 0.003567002206704755)
('see', 0.0034964682647642656)
('crouch', 0.0032546376066825873)
('looking', 0.0032546376066825873)
('eyes', 0.0031841036647420975)
('right', 0.003022883226020979)
('get', 0.003022883226020979)
('well', 0.002952349284080489)
('professor', 0.0029221204518202794)
('still', 0.0029120441744002095)
('face', 0.0028012051227794404)
('wand', 0.0027911288453593705)
('going', 0.0027709762905192307)
('time', 0.002740747458259021)
('look', 0.002690366071158671)
('go', 0.0026601372388984613)
('cedric', 0.0026601372388984613)
('come', 0.002559374464697762)
('voice', 0.0025492981872776922)
('snape', 0.0024787642453372024)
('potter', 0.002398154025976643)
('saw', 0.002398154025976643)
('bagman', 0.002307467529196014)
('head', 0.002297391251775944)
('think', 0.0022671624195157343)
('sirius', 0.0022268573098354542)
('told', 0.0021966284775752445)
('voldemort', 0.0021764759227351047)
('krum', 0.0020958657033745454)
('room', 0.0020757131485344056)
('thought', 0.002045484316274196)
('way', 0.0020152554840139857)
('something', 0.001974950374333706)
HP5.txt
HP5.txt
NORMALIZED FREQUENCY - Harry Potter 5
('harry', 0.030369862605430214)
('said', 0.029386343993085567)
('ron', 0.009701069949035853)
('hermione', 0.009701069949035853)
```

```
('back', 0.005819151789705839)
('dumbledore', 0.004932494858879981)
('sirius', 0.00478347688731261)
('well', 0.004768575090155873)
('could', 0.004746222394420767)
('professor', 0.004418382856972551)
('umbridge', 0.004291717581140286)
('would', 0.004083092420945966)
('around', 0.004068190623789229)
('looked', 0.004023485232319018)
('know', 0.003934074449378595)
('one', 0.0039191726522218575)
('like', 0.0038744672607516468)
('got', 0.00375525288349775)
('though', 0.0036583912019789587)
('weasley', 0.0036285876076654843)
('hagrid', 0.003487020534676482)
('looking', 0.003323100765952374)
('voice', 0.0032485917801686882)
('see', 0.00318898459154174)
('right', 0.003174082794385003)
('think', 0.0030921229100229486)
('still', 0.003025064822817632)
('time', 0.0029431049384555778)
('snape', 0.0029356540398772094)
('get', 0.0028983995469853666)
('face', 0.002868595952671892)
('going', 0.0028611450540935238)
('head', 0.002853694155515155)
('door', 0.002823890561201681)
('room', 0.0025854618066938873)
('look', 0.0025184037194885705)
('eyes', 0.002488600125175096)
('fred', 0.002436443835126516)
('come', 0.0023395821536077253)
('asked', 0.0023172294578726192)
('wand', 0.0022874258635591452)
('behind', 0.002279974964980777)
('thought', 0.002242720472088934)
('george', 0.002138407891991774)
('go', 0.0021235060948350372)
('oh', 0.002093702500521563)
```

```
('much', 0.00205644800762972)
('something', 0.0020489971090513516)
('neville', 0.002041546210472983)
('potter', 0.002026644413316246)
HP6.txt
HP6.txt
NORMALIZED FREQUENCY - Harry Potter 6
('harry', 0.031572076405112)
('said', 0.027941917456827448)
('dumbledore', 0.011703540836425266)
('ron', 0.010043058036736751)
('hermione', 0.008027575466080345)
('could', 0.00551967385827493)
('would', 0.005210480509367414)
('back', 0.004752416288763685)
('one', 0.004706609866703312)
('slughorn', 0.004695158261188218)
('snape', 0.004363061701250516)
('well', 0.004317255279190142)
('malfoy', 0.00423709404058449)
('know', 0.004088223168888279)
('like', 0.004030965141312812)
('think', 0.003859191058586414)
('looked', 0.003813384636526041)
('see', 0.003549997709678897)
('time', 0.0035270944986487106)
('though', 0.003446933260043058)
('around', 0.0034011268379826854)
('got', 0.003194997938711007)
('professor', 0.003137739911135541)
('still', 0.003023223855984609)
('room', 0.0029316110118638634)
('voldemort', 0.002874352984288397)
('thought', 0.0028285465622280245)
('looking', 0.0027827401401676515)
('little', 0.002668224085016719)
('look', 0.002668224085016719)
('ginny', 0.0026567724795016263)
('right', 0.0026338692684714396)
('hagrid', 0.0026338692684714396)
('asked', 0.0025651596353808803)
('yes', 0.0025537080298657874)
```

```
('face', 0.0025308048188356007)
('hand', 0.0024735467912601347)
('get', 0.0024277403691997618)
('weasley', 0.002393385552654482)
('eyes', 0.0023590307361092025)
('seemed', 0.002347579130594109)
('dark', 0.0022559662864733636)
('voice', 0.0022330630754431773)
('good', 0.0022101598644129906)
('much', 0.0021987082588978977)
('going', 0.0021872566533828043)
('wand', 0.002175805047867711)
('behind', 0.0021529018368375247)
('come', 0.0021529018368375247)
('oh', 0.0021414502313224313)
HP7.txt
HP7.txt
NORMALIZED FREQUENCY - Harry Potter 7
('harry', 0.03083913189174272)
('said', 0.019504839250162623)
('hermione', 0.0119946383867852)
('ron', 0.011580690307701405)
('could', 0.00636691569257456)
('dumbledore', 0.005962823520135618)
('wand', 0.005775561293883424)
('back', 0.005223630521771698)
('know', 0.004553428869921744)
('like', 0.004474581616762926)
('would', 0.004435157990183517)
('one', 0.00440559027024896)
('voldemort', 0.004356310737024699)
('looked', 0.004257751670576176)
('around', 0.003538270485501961)
('still', 0.0034101436991188817)
('think', 0.003370720072539473)
('eyes', 0.0032327373795115414)
('death', 0.0030454751532593485)
('snape', 0.002946916086810826)
('got', 0.002838501113717451)
('see', 0.002730086140624076)
('get', 0.002730086140624076)
('asked', 0.0027202302339792237)
```

```
('time', 0.002690662514044667)
('knew', 0.002680806607399815)
('saw', 0.002680806607399815)
('voice', 0.002621671167530701)
('face', 0.002611815260885849)
('seemed', 0.002601959354240997)
('room', 0.0025526798210167355)
('thought', 0.002542823914371883)
('little', 0.002532968007727031)
('right', 0.002434408941278508)
('us', 0.002424553034633656)
('well', 0.0024048412213439514)
('away', 0.00235556168811969)
('never', 0.00235556168811969)
('door', 0.002276714434960872)
('though', 0.00226685852831602)
('looking', 0.0022570026216711674)
('hand', 0.002197867181802054)
('potter', 0.0021485876485777926)
('look', 0.002128875835288088)
('come', 0.002128875835288088)
('going', 0.002119019928643236)
('go', 0.002119019928643236)
('two', 0.0020993081153535312)
('felt', 0.002079596302063827)
('head', 0.0020303167688395657)
```

Top 50 Bigrams by Frequencies

```
HP1.txt
BIGRAM FREQUENCY - Harry Potter 1
(('said', 'harry'), 0.003556417717426447)
(('said', 'ron'), 0.0027854460444179164)
(('uncle', 'vernon'), 0.002685965828545848)
(('professor', 'mcgonagall'), 0.0023626551269616256)
(('said', 'hagrid'), 0.00223830485712154)
(('aunt', 'petunia'), 0.0012932428063368898)
(('harry', 'ron'), 0.0012186326444328383)
(('harry', 'could'), 0.0011937625904648213)
(('said', 'hermione'), 0.0010445422666567187)
(('ron', 'hermione'), 0.0008953219428486159)
(('could', 'see'), 0.0007958417269765476)
```

```
(('harry', 'potter'), 0.0007461016190405133)
(('said', 'dumbledore'), 0.0006963615111044791)
(('harry', 'felt'), 0.0005968812952324106)
(('harry', 'looked'), 0.0005968812952324106)
(('professor', 'dumbledore'), 0.0005471411872963764)
(('harry', 'hermione'), 0.0005222711333283594)
(('harry', 'said'), 0.0005222711333283594)
(('looked', 'like'), 0.0005222711333283594)
(('common', 'room'), 0.0004974010793603423)
(('first', 'years'), 0.0004725310253923251)
(('hagrid', 'said'), 0.0004725310253923251)
(('crabbe', 'goyle'), 0.00044766097142430797)
(('professor', 'quirrell'), 0.00044766097142430797)
(('yes', 'said'), 0.00044766097142430797)
(('fred', 'george'), 0.0004227909174562909)
(('harry', 'asked'), 0.0004227909174562909)
(('hermione', 'granger'), 0.0004227909174562909)
(('said', 'professor'), 0.0004227909174562909)
(('harry', 'thought'), 0.0003979208634882738)
(('privet', 'drive'), 0.0003979208634882738)
(('great', 'hall'), 0.00037305080952025664)
(('nimbus', 'two'), 0.00037305080952025664)
(('professor', 'flitwick'), 0.00037305080952025664)
(('first', 'time'), 0.00034818075555223955)
(('harry', 'harry'), 0.00034818075555223955)
(('harry', 'told'), 0.00034818075555223955)
(('invisibility', 'cloak'), 0.00034818075555223955)
(('ron', 'harry'), 0.00034818075555223955)
(('two', 'thousand'), 0.00034818075555223955)
(('back', 'harry'), 0.00032331070158422246)
(('get', 'past'), 0.00032331070158422246)
(('madam', 'pomfrey'), 0.00032331070158422246)
(('nicolas', 'flamel'), 0.00032331070158422246)
(('right', 'said'), 0.00032331070158422246)
(('said', 'uncle'), 0.00032331070158422246)
(('sorcerer', 'stone'), 0.00032331070158422246)
(('could', 'hear'), 0.0002984406476162053)
(('harry', 'knew'), 0.0002984406476162053)
(('harry', 'never'), 0.0002984406476162053)
HP2.txt
BIGRAM FREQUENCY - Harry Potter 2
(('said', 'harry'), 0.005031364349189754)
```

```
(('said', 'ron'), 0.004573967590172504)
(('harry', 'ron'), 0.002069175814601847)
(('harry', 'potter'), 0.0020038334204565255)
(('professor', 'mcgonagall'), 0.0018731486321658826)
(('said', 'hermione'), 0.0014810942672939538)
(('harry', 'said'), 0.0013504094790033107)
(('chamber', 'secrets'), 0.0013286286809548703)
(('ron', 'hermione'), 0.0012415054887611081)
(('fred', 'george'), 0.0009801359121798222)
(('gilderoy', 'lockhart'), 0.0009365743160829413)
(('uncle', 'vernon'), 0.0008494511238891793)
(('said', 'lockhart'), 0.0008276703258407388)
(('said', 'malfoy'), 0.0008276703258407388)
(('harry', 'could'), 0.0008058895277922983)
(('said', 'dumbledore'), 0.0008058895277922983)
(('harry', 'looked'), 0.0007405471336469769)
(('said', 'professor'), 0.0007405471336469769)
(('headless', 'nick'), 0.0007187663355985363)
(('nearly', 'headless'), 0.0007187663355985363)
(('said', 'riddle'), 0.0007187663355985363)
(('said', 'weasley'), 0.0006534239414532148)
(('madam', 'pomfrey'), 0.0006316431434047744)
(('could', 'see'), 0.0005663007492594528)
(('said', 'fred'), 0.0005663007492594528)
(('heir', 'slytherin'), 0.0005445199512110123)
(('moaning', 'myrtle'), 0.0005227391531625719)
(('professor', 'sprout'), 0.0005227391531625719)
(('said', 'george'), 0.0005227391531625719)
(('common', 'room'), 0.0004791775570656909)
(('said', 'hagrid'), 0.0004791775570656909)
(('aunt', 'petunia'), 0.0004573967590172504)
(('let', 'go'), 0.0004573967590172504)
(('lucius', 'malfoy'), 0.0004573967590172504)
(('polyjuice', 'potion'), 0.0004573967590172504)
(('could', 'hear'), 0.0004356159609688099)
(('harry', 'harry'), 0.0004356159609688099)
(('harry', 'saw'), 0.0004356159609688099)
(('right', 'said'), 0.0004356159609688099)
(('great', 'hall'), 0.0004138351629203694)
(('sorting', 'hat'), 0.0004138351629203694)
(('whomping', 'willow'), 0.0004138351629203694)
(('c', 'h'), 0.0003920543648719289)
```

```
(('dueling', 'club'), 0.0003920543648719289)
(('e', 'r'), 0.0003920543648719289)
(('h', 'p'), 0.0003920543648719289)
(('harry', 'felt'), 0.0003920543648719289)
(('hospital', 'wing'), 0.0003920543648719289)
(('p', 'e'), 0.0003920543648719289)
(('draco', 'malfoy'), 0.00037027356682348844)
HP3.txt
BIGRAM FREQUENCY - Harry Potter 3
(('said', 'harry'), 0.004582713658216027)
(('said', 'ron'), 0.0031646664130322)
(('said', 'hermione'), 0.0022481236813889946)
(('professor', 'lupin'), 0.0021097776086881335)
(('ron', 'hermione'), 0.0018676719814616262)
(('harry', 'ron'), 0.0017120326496731573)
(('professor', 'mcgonagall'), 0.001677446131497942)
(('professor', 'trelawney'), 0.001677446131497942)
(('said', 'lupin'), 0.001677446131497942)
(('harry', 'said'), 0.0013834607270086122)
(('uncle', 'vernon'), 0.0012451146543077508)
(('said', 'professor'), 0.0011586483588697126)
(('harry', 'hermione'), 0.0010548888043440667)
(('aunt', 'marge'), 0.0010030090270812437)
(('sirius', 'black'), 0.0008992494725555978)
(('said', 'hagrid'), 0.0007263168816795213)
(('common', 'room'), 0.0007090236225919137)
(('ron', 'said'), 0.0006571438453290907)
(('said', 'dumbledore'), 0.0006571438453290907)
(('madam', 'pomfrey'), 0.0006398505862414831)
(('fred', 'george'), 0.0005879708089786602)
(('harry', 'looked'), 0.0005706775498910525)
(('said', 'snape'), 0.0005706775498910525)
(('said', 'fudge'), 0.0005533842908034448)
(('crabbe', 'goyle'), 0.0005360910317158371)
(('could', 'see'), 0.0005187977726282295)
(('harry', 'felt'), 0.0005187977726282295)
(('said', 'fred'), 0.0005187977726282295)
(('harry', 'could'), 0.0005015045135406219)
(('harry', 'saw'), 0.0005015045135406219)
(('marauder', 'map'), 0.0005015045135406219)
(('yes', 'said'), 0.0005015045135406219)
(('said', 'black'), 0.00048421125445301423)
```

```
(('harry', 'potter'), 0.00046691799536540654)
(('professor', 'snape'), 0.00046691799536540654)
(('hagrid', 'said'), 0.0004496247362777989)
(('fat', 'lady'), 0.00043233147719019127)
(('hermione', 'said'), 0.00043233147719019127)
(('knight', 'bus'), 0.00043233147719019127)
(('looked', 'around'), 0.00043233147719019127)
(('aunt', 'petunia'), 0.00041503821810258363)
(('expecto', 'patronum'), 0.00041503821810258363)
(('invisibility', 'cloak'), 0.00041503821810258363)
(('leaky', 'cauldron'), 0.00041503821810258363)
(('ministry', 'magic'), 0.00041503821810258363)
(('dark', 'arts'), 0.00039774495901497594)
(('c', 'h'), 0.0003804516999273683)
(('defense', 'dark'), 0.0003804516999273683)
(('e', 'r'), 0.0003804516999273683)
(('h', 'p'), 0.0003804516999273683)
HP4.txt
BIGRAM FREQUENCY - Harry Potter 4
(('said', 'harry'), 0.004403333232570559)
(('said', 'ron'), 0.0030833408905413984)
(('said', 'hermione'), 0.0021663996453150348)
(('harry', 'said'), 0.0016424332194713985)
(('harry', 'ron'), 0.0015315941678506293)
(('ron', 'hermione'), 0.0015215178904305594)
(('said', 'dumbledore'), 0.0012696109549288112)
(('said', 'weasley'), 0.0011990770129883216)
(('harry', 'could'), 0.0010781616839474824)
(('harry', 'potter'), 0.0009471700774865734)
(('madame', 'maxime'), 0.0009471700774865734)
(('could', 'see'), 0.0008262547484457342)
(('professor', 'mcgonagall'), 0.0008262547484457342)
(('uncle', 'vernon'), 0.0008061021936055944)
(('fred', 'george'), 0.0007960259161855244)
(('rita', 'skeeter'), 0.0007758733613453846)
(('harry', 'looked'), 0.0007053394194048951)
(('harry', 'saw'), 0.0007053394194048951)
(('death', 'eaters'), 0.0006348054774644056)
(('great', 'hall'), 0.0006146529226242657)
(('looked', 'around'), 0.0006146529226242657)
(('hermione', 'said'), 0.0005844240903640559)
(('world', 'cup'), 0.0005844240903640559)
```

```
(('said', 'fred'), 0.0005642715355239161)
(('said', 'moody'), 0.0005642715355239161)
(('ludo', 'bagman'), 0.0005441189806837762)
(('said', 'hagrid'), 0.0005239664258436363)
(('harry', 'felt'), 0.0005138901484235664)
(('ron', 'said'), 0.0004735850387432867)
(('right', 'said'), 0.00046350876132321677)
(('could', 'hear'), 0.0004534324839031468)
(('entrance', 'hall'), 0.0004433562064830769)
(('harry', 'thought'), 0.0004433562064830769)
(('yeah', 'said'), 0.0004433562064830769)
(('common', 'room'), 0.000423203651642937)
(('triwizard', 'tournament'), 0.000423203651642937)
(('aunt', 'petunia'), 0.0004030510968027972)
(('yes', 'said'), 0.0004030510968027972)
(('said', 'george'), 0.00039297481938272723)
(('said', 'sirius'), 0.00039297481938272723)
(('professor', 'dumbledore'), 0.0003828985419626573)
(('magical', 'eye'), 0.00036274598712251744)
(('said', 'bagman'), 0.00036274598712251744)
(('daily', 'prophet'), 0.00035266970970244753)
(('looked', 'though'), 0.00035266970970244753)
(('quidditch', 'world'), 0.00035266970970244753)
(('viktor', 'krum'), 0.00035266970970244753)
(('looking', 'around'), 0.0003425934322823776)
(('goblet', 'fire'), 0.00033251715486230766)
(('looked', 'like'), 0.00033251715486230766)
HP5.txt
BIGRAM FREQUENCY - Harry Potter 5
(('said', 'harry'), 0.005200727207701249)
(('said', 'hermione'), 0.0031368283014931602)
(('said', 'ron'), 0.0025929127052722557)
(('harry', 'said'), 0.0014454743242034989)
(('professor', 'umbridge'), 0.0014454743242034989)
(('professor', 'mcgonagall'), 0.0011325365839120198)
(('ron', 'hermione'), 0.0011176347867552826)
(('said', 'sirius'), 0.0010356749023932286)
(('fred', 'george'), 0.001013322206658123)
(('harry', 'ron'), 0.0010058713080797544)
(('said', 'hagrid'), 0.0010058713080797544)
(('said', 'dumbledore'), 0.0008419515393556463)
(('harry', 'could'), 0.0007897952493070665)
```

```
(('said', 'weasley'), 0.0007897952493070665)
(('said', 'professor'), 0.0007823443507286979)
(('said', 'fred'), 0.0007674425535719607)
(('uncle', 'vernon'), 0.0007599916549935923)
(('yes', 'said'), 0.0006780317706315381)
(('yeah', 'said'), 0.0006482281763180639)
(('looked', 'around'), 0.0006258754805829583)
(('harry', 'looked'), 0.0006035227848478526)
(('well', 'said'), 0.0006035227848478526)
(('death', 'eaters'), 0.0005960718862694841)
(('could', 'see'), 0.0005662682919560099)
(('said', 'george'), 0.0005588173933776413)
(('harry', 'felt'), 0.0005439155962209042)
(('ministry', 'magic'), 0.0005439155962209042)
(('said', 'umbridge'), 0.0005066611033290615)
(('harry', 'asked'), 0.0004992102047506929)
(('harry', 'hermione'), 0.0004694066104372l874)
(('aunt', 'petunia'), 0.0004470539147021131)
(('harry', 'potter'), 0.0004470539147021131)
(('right', 'said'), 0.0004470539147021131)
(('said', 'snape'), 0.0004470539147021131)
(('know', 'said'), 0.0004396301612374454)
(('harry', 'saw'), 0.000432152117545376)
(('professor', 'trelawney'), 0.000432152117545376)
(('could', 'hear'), 0.0004172503203886389)
(('common', 'room'), 0.00039489762465353323)
(('department', 'mysteries'), 0.00039489762465353323)
(('potter', 'said'), 0.00039489762465353323)
(('ron', 'said'), 0.00039489762465353323)
(('dark', 'arts'), 0.0003725449289184276)
(('hermione', 'said'), 0.0003725449289184276)
(('said', 'lupin'), 0.0003725449289184276)
(('great', 'hall'), 0.0003501922331833219)
(('defense', 'dark'), 0.00034274133460495337)
(('grimmauld', 'place'), 0.0003352904360265848)
(('said', 'ginny'), 0.0003352904360265848)
(('said', 'luna'), 0.00032783953744821626)
HP6.txt
BIGRAM FREQUENCY - Harry Potter 6
(('said', 'harry'), 0.006115157345059777)
(('said', 'dumbledore'), 0.0033324172048921257)
(('said', 'ron'), 0.0022330630754431773)
```

```
(('said', 'hermione'), 0.0021529018368375247)
(('harry', 'said'), 0.001271128212175347)
(('ron', 'hermione'), 0.0012138701845998809)
(('prime', 'minister'), 0.001122257340479135)
(('harry', 'could'), 0.0009848380742980166)
(('said', 'slughorn'), 0.0009619348632678302)
(('harry', 'ron'), 0.0008703220191470844)
(('professor', 'mcgonagall'), 0.0008245155970867115)
(('death', 'eaters'), 0.0007329027529659658)
(('yes', 'said'), 0.0006756447253904997)
(('said', 'weasley'), 0.0006527415143603133)
(('dark', 'lord'), 0.0006412899088452201)
(('said', 'professor'), 0.0006412899088452201)
(('said', 'snape'), 0.0006298383033301269)
(('common', 'room'), 0.000584031881269754)
(('looked', 'around'), 0.0005725802757546608)
(('lord', 'voldemort'), 0.0005611286702395676)
(('right', 'said'), 0.0005611286702395676)
(('asked', 'harry'), 0.0005496770647244744)
(('harry', 'saw'), 0.0005496770647244744)
(('harry', 'potter'), 0.0005267738536942879)
(('said', 'hagrid'), 0.0004924190371490083)
(('death', 'eater'), 0.0004809674316339151)
(('harry', 'looked'), 0.0004695158261188218)
(('said', 'ginny'), 0.0004695158261188218)
(('felix', 'felicis'), 0.0004466126150886354)
(('sir', 'said'), 0.0004466126150886354)
(('well', 'said'), 0.0004466126150886354)
(('dark', 'arts'), 0.0004351610095735422)
(('harry', 'thought'), 0.0004351610095735422)
(('could', 'see'), 0.000423709404058449)
(('fred', 'george'), 0.000423709404058449)
(('harry', 'felt'), 0.0004008061930282626)
(('invisibility', 'cloak'), 0.0004008061930282626)
(('professor', 'trelawney'), 0.0004008061930282626)
(('professor', 'slughorn'), 0.0003779029819980761)
(('dumbledore', 'said'), 0.0003664513764829829)
(('harry', 'asked'), 0.0003549997709678897)
(('madam', 'pomfrey'), 0.0003549997709678897)
(('yeah', 'said'), 0.0003549997709678897)
(('c', 'h'), 0.0003435481654527965)
(('e', 'r'), 0.0003435481654527965)
```

```
(('h', 'p'), 0.0003435481654527965)
(('p', 'e'), 0.0003435481654527965)
(('room', 'requirement'), 0.0003435481654527965)
(('harry', 'knew'), 0.0003320965599377033)
(('tom', 'riddle'), 0.0003320965599377033)
HP7.txt
BIGRAM FREQUENCY - Harry Potter 7
(('said', 'harry'), 0.003991642191165165)
(('said', 'ron'), 0.00235556168811969)
(('said', 'hermione'), 0.0022175789950917586)
(('ron', 'hermione'), 0.0018627663558770772)
(('death', 'eaters'), 0.0013798269302793165)
(('harry', 'could'), 0.0009757347578403737)
(('harry', 'said'), 0.0008771756913918511)
(('harry', 'potter'), 0.0008180402515227376)
(('harry', 'ron'), 0.0007391929983639195)
(('harry', 'saw'), 0.0007391929983639195)
(('harry', 'looked'), 0.0006702016518499537)
(('elder', 'wand'), 0.0006406339319153969)
(('death', 'eater'), 0.0006209221186256924)
(('invisibility', 'cloak'), 0.0006012103053359879)
(('godric', 'hollow'), 0.0005814984920462833)
(('harry', 'felt'), 0.0005814984920462833)
(('could', 'see'), 0.0005617866787565788)
(('albus', 'dumbledore'), 0.000532218958822022)
(('harry', 'knew'), 0.0005223630521771698)
(('dark', 'lord'), 0.0005026512388874653)
(('harry', 'hermione'), 0.0005026512388874653)
(('professor', 'mcgonagall'), 0.0005026512388874653)
(('asked', 'harry'), 0.0004730835189529085)
(('looked', 'around'), 0.0004730835189529085)
(('deathly', 'hallows'), 0.00045337170566320397)
(('said', 'dumbledore'), 0.00045337170566320397)
(('harry', 'thought'), 0.0004435157990183517)
(('said', 'lupin'), 0.0004435157990183517)
(('said', 'voldemort'), 0.0004040921724389427)
(('right', 'said'), 0.00038438035914923816)
(('c', 'h'), 0.0003548126392146814)
(('could', 'hear'), 0.0003548126392146814)
(('e', 'r'), 0.0003548126392146814)
(('h', 'p'), 0.0003548126392146814)
(('p', 'e'), 0.0003548126392146814)
```

```
(('first', 'time'), 0.0003449567325698291)
(('phineas', 'nigellus'), 0.0003252449192801246)
(('uncle', 'vernon'), 0.0003252449192801246)
(('bill', 'fleur'), 0.0003153890126352723 4)
(('could', 'feel'), 0.0003153890126352723 4)
(('hermione', 'said'), 0.0003153890126352723 4)
(('ron', 'said'), 0.0003153890126352723 4)
(('asked', 'hermione'), 0.0003055331059904201)
(('asked', 'ron'), 0.0003055331059904201)
(('know', 'said'), 0.0003055331059904201)
(('lord', 'voldemort'), 0.0003055331059904201)
(('yeah', 'said'), 0.0003055331059904201)
(('fred', 'george'), 0.0002956771993455678)
(('hermione', 'looked'), 0.0002956771993455678)
(('said', 'snape'), 0.0002956771993455678)
```

Top 50 Bigrams by Mutual Information Scores, with a Minimum Frequency of 5

```
HP1.txt
BIGRAM PMI
(('adalbert', 'waffling'), 15.295230836230358)
(('african', 'prince'), 15.295230836230358)
(('alberic', 'grunnion'), 15.295230836230358)
(('alchemist', 'opera'), 15.295230836230358)
(('amber', 'liquid'), 15.295230836230358)
(('amid', 'gales'), 15.295230836230358)
(('angrier', 'spoken'), 15.295230836230358)
(('apple', 'pies'), 15.295230836230358)
(('arsenius', 'jigger'), 15.295230836230358)
(('art', 'potionmaking'), 15.295230836230358)
(('awaits', 'sin'), 15.295230836230358)
(('balloons', 'kit'), 15.295230836230358)
(('bandages', 'blasted'), 15.295230836230358)
(('basket', 'propped'), 15.295230836230358)
(('bathilda', 'bagshot'), 15.295230836230358)
(('bedtime', 'trot'), 15.295230836230358)
(('beyond', 'bedtime'), 15.295230836230358)
(('bicycle', 'carousel'), 15.295230836230358)
(('bid', 'freedom'), 15.295230836230358)
(('blackpool', 'pier'), 15.295230836230358)
```

```
(('bletchley', 'dives'), 15.295230836230358)
(('blinded', 'staggered'), 15.295230836230358)
(('blubber', 'oddment'), 15.295230836230358)
(('booming', 'barks'), 15.295230836230358)
(('bread', 'english'), 15.295230836230358)
(('briefcase', 'pecked'), 15.295230836230358)
(('brocklehurst', 'mandy'), 15.295230836230358)
(('bulstrode', 'millicent'), 15.295230836230358)
(('butter', 'mellow'), 15.295230836230358)
(('caput', 'draconis'), 15.295230836230358)
(('cases', 'glimmered'), 15.295230836230358)
(('catcalling', 'bulstrode'), 15.295230836230358)
(('ceased', 'tottered'), 15.295230836230358)
(('chances', 'deserves'), 15.295230836230358)
(('chf', 'warlock'), 15.295230836230358)
(('chilled', 'steel'), 15.295230836230358)
(('chipolatas', 'tureens'), 15.295230836230358)
(('circe', 'paracelsus'), 15.295230836230358)
(('code', 'conduct'), 15.295230836230358)
(('coffee', 'sardine'), 15.295230836230358)
(('conduct', 'uprising'), 15.295230836230358)
(('correcting', 'grips'), 15.295230836230358)
(('countercurses', 'bewitch'), 15.295230836230358)
(('cows', 'sheep'), 15.295230836230358)
(('cranberry', 'sauce'), 15.295230836230358)
(('curl', 'thinnest'), 15.295230836230358)
(('curtain', 'ivy'), 15.295230836230358)
(('daisies', 'butter'), 15.295230836230358)
(('dennis', 'malcolm'), 15.295230836230358)
(('detest', 'unlike'), 15.295230836230358)
HP2.txt
BIGRAM PMI
(('abyssinian', 'shrivelfigs'), 15.48658365906727)
(('accept', 'huntsmen'), 15.48658365906727)
(('accepted', 'combative'), 15.48658365906727)
(('acne', 'clears'), 15.48658365906727)
(('activities', 'horseback'), 15.48658365906727)
(('acts', 'nature'), 15.48658365906727)
(('admirable', 'sentiments'), 15.48658365906727)
(('adventures', 'martin'), 15.48658365906727)
(('affair', 'bungled'), 15.48658365906727)
(('aftermath', 'duels'), 15.48658365906727)
```

```
(('airborne', 'menace'), 15.48658365906727)
(('angus', 'fleet'), 15.48658365906727)
(('animals', 'surrounding'), 15.48658365906727)
(('ants', 'villages'), 15.48658365906727)
(('apothecary', 'skulkin'), 15.48658365906727)
(('artist', 'imagined'), 15.48658365906727)
(('auntie', 'mabel'), 15.48658365906727)
(('badger', 'resisting'), 15.48658365906727)
(('battering', 'ram'), 15.48658365906727)
(('bee', 'bonnet'), 15.48658365906727)
(('befall', 'whilst'), 15.48658365906727)
(('believable', 'verifiable'), 15.48658365906727)
(('berserk', 'squirted'), 15.48658365906727)
(('bishops', 'wrestled'), 15.48658365906727)
(('blonde', 'pigtails'), 15.48658365906727)
(('blotched', 'tearstained'), 15.48658365906727)
(('bold', 'braved'), 15.48658365906727)
(('bookshelf', 'dozens'), 15.48658365906727)
(('borrow', 'hermes'), 15.48658365906727)
(('brawling', 'public'), 15.48658365906727)
(('bronze', 'jangling'), 15.48658365906727)
(('bucket', 'leaping'), 15.48658365906727)
(('buckets', 'mops'), 15.48658365906727)
(('budgies', 'arguing'), 15.48658365906727)
(('bun', 'placing'), 15.48658365906727)
(('bunch', 'grapes'), 15.48658365906727)
(('burn', 'flogging'), 15.48658365906727)
(('bus', 'stops'), 15.48658365906727)
(('buttons', 'produced'), 15.48658365906727)
(('cannonball', 'ninth'), 15.48658365906727)
(('causes', 'devastation'), 15.48658365906727)
(('cemented', 'jaws'), 15.48658365906727)
(('cheeriness', 'hints'), 15.48658365906727)
(('chess', 'heaven'), 15.48658365906727)
(('chewed', 'greedily'), 15.48658365906727)
(('chink', 'knives'), 15.48658365906727)
(('clauses', 'fishing'), 15.48658365906727)
(('clench', 'unclench'), 15.48658365906727)
(('cleverly', 'widest'), 15.48658365906727)
(('clove', 'garlic'), 15.48658365906727)
HP3.txt
BIGRAM PMI
```

```
(('aaaaaaaaaaaaaarrrrrrrrrrrrgggggghhhhh', 'nooooooooooooooooo'),
15.819430689525769)
(('absence', 'aim'), 15.819430689525769)
(('absently', 'rooftops'), 15.819430689525769)
(('absurdity', 'statement'), 15.819430689525769)
(('accomplishing', 'task'), 15.819430689525769)
(('admirers', 'resemblance'), 15.819430689525769)
(('adopted', 'oily'), 15.819430689525769)
(('advertise', 'firebolts'), 15.819430689525769)
(('advise', 'entrust'), 15.819430689525769)
(('aerodynamic', 'perfection'), 15.819430689525769)
(('agriculture', 'fisheries'), 15.819430689525769)
(('aine', 'kiely'), 15.819430689525769)
(('allies', 'deliver'), 15.819430689525769)
(('american', 'edition'), 15.819430689525769)
(('angela', 'biola'), 15.819430689525769)
(('anglesey', 'aberdeen'), 15.819430689525769)
(('apothecary', 'replenish'), 15.819430689525769)
(('apron', 'nightshirt'), 15.819430689525769)
(('aunts', 'tally'), 15.819430689525769)
(('bandage', 'unraveled'), 15.819430689525769)
(('banners', 'slogans'), 15.819430689525769)
(('based', 'type'), 15.819430689525769)
(('bath', 'buns'), 15.819430689525769)
(('bathilda', 'bagshot'), 15.819430689525769)
(('beau', 'iful'), 15.819430689525769)
(('beauties', 'proprietor'), 15.819430689525769)
(('benches', 'squeezing'), 15.819430689525769)
(('binder', 'clips'), 15.819430689525769)
(('biola', 'mike'), 15.819430689525769)
(('bitter', 'sidelong'), 15.819430689525769)
(('blasted', 'alf'), 15.819430689525769)
(('blazing', 'lizards'), 15.819430689525769)
(('boom', 'suggestions'), 15.819430689525769)
(('bordering', 'reverence'), 15.819430689525769)
(('bred', 'bulldogs'), 15.819430689525769)
(('broomtail', 'honed'), 15.819430689525769)
(('brutality', 'absorbed'), 15.819430689525769)
(('buckets', 'mops'), 15.819430689525769)
(('buildings', 'benches'), 15.819430689525769)
(('burped', 'richly'), 15.819430689525769)
(('bury', 'hatchet'), 15.819430689525769)
```

```
(('bushes', 'wastebaskets'), 15.819430689525769)
(('cageful', 'pixies'), 15.819430689525769)
(('calming', 'breaths'), 15.819430689525769)
(('caterpillars', 'cauldrons'), 15.819430689525769)
(('changes', 'lineup'), 15.819430689525769)
(('characters', 'elements'), 15.819430689525769)
(('cherry', 'syrup'), 15.819430689525769)
(('chooses', 'divulge'), 15.819430689525769)
(('chris', 'mas'), 15.819430689525769)
HP4.txt
BIGRAM PMI
(('abbot', 'swapping'), 16.598677726425358)
(('absorbed', 'needlework'), 16.598677726425358)
(('acclaimed', 'chairwizard'), 16.598677726425358)
(('accurate', 'title'), 16.598677726425358)
(('acknowledged', 'jovial'), 16.598677726425358)
(('administered', 'fatal'), 16.598677726425358)
(('adorned', 'lions'), 16.598677726425358)
(('advise', 'marketing'), 16.598677726425358)
(('agatha', 'timms'), 16.598677726425358)
(('ancestors', 'abided'), 16.598677726425358)
(('apollyon', 'pringle'), 16.598677726425358)
(('approved', 'flashy'), 16.598677726425358)
(('artichoke', 'buttonhole'), 16.598677726425358)
(('artifact', 'registry'), 16.598677726425358)
(('astonishing', 'gymnastics'), 16.598677726425358)
(('attachments', 'unlock'), 16.598677726425358)
(('audience', 'assembled'), 16.598677726425358)
(('balancing', 'teetering'), 16.598677726425358)
(('banana', 'fritters'), 16.598677726425358)
(('bangended', 'scoots'), 16.598677726425358)
(('bangings', 'scrapings'), 16.598677726425358)
(('baskets', 'lunascopes'), 16.598677726425358)
(('bass', 'drum'), 16.598677726425358)
(('beautifully', 'proportioned'), 16.598677726425358)
(('beggars', 'vagrants'), 16.598677726425358)
(('belladonna', 'piling'), 16.598677726425358)
(('bewitchments', 'charmes'), 16.598677726425358)
(('birdbath', 'sundial'), 16.598677726425358)
(('blackened', 'log'), 16.598677726425358)
(('bleak', 'forbidding'), 16.598677726425358)
(('blotts', 'bookshop'), 16.598677726425358)
```

```
(('blunders', 'culprits'), 16.598677726425358)
(('bode', 'croaker'), 16.598677726425358)
(('boggart', 'riddikulus'), 16.598677726425358)
(('brambles', 'branches'), 16.598677726425358)
(('branstone', 'eleanor'), 16.598677726425358)
(('brows', 'unknitted'), 16.598677726425358)
(('buildup', 'earwax'), 16.598677726425358)
(('bulky', 'backpacks'), 16.598677726425358)
(('buses', 'trains'), 16.598677726425358)
(('bush', 'pruning'), 16.598677726425358)
(('buttocks', 'waddled'), 16.598677726425358)
(('buzz', 'chatter'), 16.598677726425358)
(('calluses', 'blisters'), 16.598677726425358)
(('campsites', 'raucous'), 16.598677726425358)
(('caps', 'hinkypunks'), 16.598677726425358)
(('cart', 'piled'), 16.598677726425358)
(('cauldwell', 'owen'), 16.598677726425358)
(('cello', 'bagpipes'), 16.598677726425358)
(('cheese', 'grated'), 16.598677726425358)
HP5.txt
BIGRAM PMI
(('accusations', 'levelled'), 17.034154144235064)
(('accuses', 'subheading'), 17.034154144235064)
(('acknowledge', 'superiority'), 17.034154144235064)
(('actress', 'divorce'), 17.034154144235064)
(('adolescent', 'agonizing'), 17.034154144235064)
(('adventure', 'substantial'), 17.034154144235064)
(('advisory', 'cuse'), 17.034154144235064)
(('aided', 'abetted'), 17.034154144235064)
(('airy', 'overly'), 17.034154144235064)
(('angela', 'biola'), 17.034154144235064)
(('annoyingly', 'buoyant'), 17.034154144235064)
(('apathetic', 'phases'), 17.034154144235064)
(('apparating', 'disapparating'), 17.034154144235064)
(('aquavirius', 'maggots'), 17.034154144235064)
(('arabella', 'doreen'), 17.034154144235064)
(('araminta', 'meliflua'), 17.034154144235064)
(('ardently', 'desires'), 17.034154144235064)
(('assisting', 'demonstration'), 17.034154144235064)
(('assumption', 'airiness'), 17.034154144235064)
(('authorities', 'unregistered'), 17.034154144235064)
(('avid', 'anticipation'), 17.034154144235064)
```

```
(('balding', 'redhaired'), 17.034154144235064)
(('barred', 'rightful'), 17.034154144235064)
(('barrel', 'eels'), 17.034154144235064)
(('barry', 'ryan'), 17.034154144235064)
(('basset', 'hound'), 17.034154144235064)
(('beaklike', 'protuberance'), 17.034154144235064)
(('bears', 'hallmarks'), 17.034154144235064)
(('belching', 'sooty'), 17.034154144235064)
(('bends', 'hedgerows'), 17.034154144235064)
(('berk', 'tailin'), 17.034154144235064)
(('besieged', 'requests'), 17.034154144235064)
(('blanched', 'yoooou'), 17.034154144235064)
(('blinkered', 'fettered'), 17.034154144235064)
(('border', 'sligh'), 17.034154144235064)
(('boxes', 'inscribed'), 17.034154144235064)
(('brad', 'walrod'), 17.034154144235064)
(('brand', 'fertilizer'), 17.034154144235064)
(('bravest', 'boldest'), 17.034154144235064)
(('breached', 'measures'), 17.034154144235064)
(('breeds', 'abraxan'), 17.034154144235064)
(('brownish', 'smock'), 17.034154144235064)
(('bull', 'elephants'), 17.034154144235064)
(('bullfrogs', 'cawing'), 17.034154144235064)
(('bumped', 'lintel'), 17.034154144235064)
(('buttering', 'crumpet'), 17.034154144235064)
(('candy', 'wrappers'), 17.034154144235064)
(('capacity', 'delude'), 17.034154144235064)
(('caradoc', 'dearborn'), 17.034154144235064)
(('carries', 'defying'), 17.034154144235064)
HP6.txt
BIGRAM PMI
(('aches', 'pains'), 16.4140905960897)
(('admirably', 'succinct'), 16.4140905960897)
(('admonitions', 'helpers'), 16.4140905960897)
(('advisability', 'agreeing'), 16.4140905960897)
(('affably', 'snapping'), 16.4140905960897)
(('aged', 'monkey'), 16.4140905960897)
(('ambrosius', 'flume'), 16.4140905960897)
(('american', 'edition'), 16.4140905960897)
(('amulet', 'seller'), 16.4140905960897)
(('amusin', 'outsmarted'), 16.4140905960897)
(('ancestor', 'nonsense'), 16.4140905960897)
```

```
(('anecdotes', 'illustrious'), 16.4140905960897)
(('angela', 'biola'), 16.4140905960897)
(('animatedly', 'proprietor'), 16.4140905960897)
(('apple', 'cores'), 16.4140905960897)
(('appraising', 'possibilities'), 16.4140905960897)
(('aproned', 'helper'), 16.4140905960897)
(('arabella', 'figg'), 16.4140905960897)
(('arkie', 'philpott'), 16.4140905960897)
(('associated', 'logos'), 16.4140905960897)
(('atop', 'curls'), 16.4140905960897)
(('attractive', 'barmaid'), 16.4140905960897)
(('au', 'revoir'), 16.4140905960897)
(('austere', 'efficiency'), 16.4140905960897)
(('avery', 'yaxley'), 16.4140905960897)
(('balder', 'grayer'), 16.4140905960897)
(('banish', 'usurping'), 16.4140905960897)
(('barty', 'crouch'), 16.4140905960897)
(('bench', 'eavesdrop'), 16.4140905960897)
(('bes', 'wiz'), 16.4140905960897)
(('besieged', 'requests'), 16.4140905960897)
(('bibble', 'buggins'), 16.4140905960897)
(('bids', 'freedom'), 16.4140905960897)
(('blatant', 'wizardishness'), 16.4140905960897)
(('blissful', 'relaxation'), 16.4140905960897)
(('boils', 'blackheads'), 16.4140905960897)
(('bolstered', 'comforted'), 16.4140905960897)
(('boyfriend', 'fermenting'), 16.4140905960897)
(('brad', 'walrod'), 16.4140905960897)
(('brand', 'reasoned'), 16.4140905960897)
(('bravado', 'awkwardness'), 16.4140905960897)
(('british', 'overcook'), 16.4140905960897)
(('browne', 'continuity'), 16.4140905960897)
(('bubbled', 'sluggishly'), 16.4140905960897)
(('burglar', 'alarms'), 16.4140905960897)
(('cages', 'au'), 16.4140905960897)
(('calamity', 'disaster'), 16.4140905960897)
(('canary', 'islands'), 16.4140905960897)
(('carpenters', 'builder'), 16.4140905960897)
(('categorically', 'untruthfully'), 16.4140905960897)
HP7.txt
BIGRAM PMI
(('abate', 'overnight'), 16.63057997843407)
```

```
(('abilities', 'attainment'), 16.63057997843407)
(('abstract', 'furls'), 16.63057997843407)
(('accomplished', 'household'), 16.63057997843407)
(('acknowledgement', 'flexed'), 16.63057997843407)
(('adalbert', 'waffling'), 16.63057997843407)
(('adept', 'producing'), 16.63057997843407)
(('adjoining', 'cubicles'), 16.63057997843407)
(('advance', 'publicity'), 16.63057997843407)
(('advisory', 'bureau'), 16.63057997843407)
(('alchemical', 'conference'), 16.63057997843407)
(('alicia', 'spinnet'), 16.63057997843407)
(('alleged', 'sightings'), 16.63057997843407)
(('anchors', 'immortality'), 16.63057997843407)
(('angela', 'biola'), 16.63057997843407)
(('angelina', 'johnson'), 16.63057997843407)
(('anthony', 'goldstein'), 16.63057997843407)
(('antioch', 'cadmus'), 16.63057997843407)
(('arcus', 'livius'), 16.63057997843407)
(('arkie', 'alderton'), 16.63057997843407)
(('armando', 'dippet'), 16.63057997843407)
(('armfuls', 'venomous'), 16.63057997843407)
(('arouses', 'passions'), 16.63057997843407)
(('assaulted', 'screeches'), 16.63057997843407)
(('assister', 'vous'), 16.63057997843407)
(('attainment', 'awards'), 16.63057997843407)
(('au', 'revoir'), 16.63057997843407)
(('authorities', 'attributing'), 16.63057997843407)
(('babysit', 'cubs'), 16.63057997843407)
(('bamboozled', 'buttons'), 16.63057997843407)
(('barnabas', 'deverill'), 16.63057997843407)
(('barnabus', 'finkley'), 16.63057997843407)
(('barty', 'crouch'), 16.63057997843407)
(('bashed', 'wee'), 16.63057997843407)
(('basset', 'hound'), 16.63057997843407)
(('batch', 'homemade'), 16.63057997843407)
(('batches', 'canapés'), 16.63057997843407)
(('batting', 'eyelashes'), 16.63057997843407)
(('beleaguered', 'blackmailed'), 16.63057997843407)
(('bell', 'angelina'), 16.63057997843407)
(('bernie', 'pillsworth'), 16.63057997843407)
(('blades', 'imbibe'), 16.63057997843407)
(('bleak', 'depressing'), 16.63057997843407)
```

```
(('blibbering', 'humdinger'), 16.63057997843407)
(('blustery', 'april'), 16.63057997843407)
(('boar', 'fox'), 16.63057997843407)
(('bolognese', 'tinned'), 16.63057997843407)
(('bolster', 'credentials'), 16.63057997843407)
(('bonnet', 'cracker'), 16.63057997843407)
(('bosom', 'stands'), 16.63057997843407)
```

Top 50 Trigrams

```
HP1.txt
TRIGRAM FREQUENCY - Harry Potter 1
(('said', 'professor', 'mcgonagall'), 16)
(('nimbus', 'two', 'thousand'), 14)
(('said', 'uncle', 'vernon'), 13)
(('harry', 'could', 'see'), 10)
(('harry', 'ron', 'hermione'), 10)
(('hagrid', 'said', 'harry'), 7)
(('gryffindor', 'common', 'room'), 7)
(('get', 'past', 'fluffy'), 7)
(('said', 'hagrid', 'harry'), 6)
(('vault', 'seven', 'hundred'), 6)
(('seven', 'hundred', 'thirteen'), 6)
(('yes', 'said', 'harry'), 6)
(('fred', 'george', 'weasley'), 6)
(('every', 'flavor', 'beans'), 6)
(('harry', 'shook', 'head'), 5)
(('portrait', 'fat', 'lady'), 5)
(('said', 'ron', 'harry'), 5)
(('said', 'aunt', 'petunia'), 4)
(('first', 'time', 'life'), 4)
(('aunt', 'petunia', 'dudley'), 4)
(('said', 'harry', 'trying'), 4)
(('vernon', 'aunt', 'petunia'), 4)
(('said', 'harry', 'hagrid'), 4)
(('school', 'witchcraft', 'wizardry'), 4)
(('ter', 'tell', 'yeh'), 4)
(('got', 'ta', 'get'), 4)
(('harry', 'could', 'help'), 4)
(('harry', 'never', 'seen'), 4)
```

```
(('platforms', 'nine', 'ten'), 4)
(('said', 'one', 'twins'), 4)
(('said', 'harry', 'ron'), 4)
(('bott', 'every', 'flavor'), 4)
(('nearly', 'headless', 'nick'), 4)
(('malfoy', 'crabbe', 'goyle'), 4)
(('hundred', 'fifty', 'points'), 4)
(('right', 'said', 'ron'), 4)
(('corner', 'privet', 'drive'), 3)
(('professor', 'mcgonagall', 'turned'), 3)
(('lily', 'james', 'potter'), 3)
(('yes', 'said', 'dumbledore'), 3)
(('yeah', 'said', 'hagrid'), 3)
(('good', 'luck', 'harry'), 3)
(('said', 'harry', 'well'), 3)
(('could', 'ever', 'remember'), 3)
(('harry', 'uncle', 'vernon'), 3)
(('know', 'said', 'harry'), 3)
(('flash', 'green', 'light'), 3)
(('thanks', 'said', 'harry'), 3)
(('next', 'morning', 'harry'), 3)
(('yelled', 'uncle', 'vernon'), 3)
HP2.txt
TRIGRAM FREQUENCY - Harry Potter 2
(('nearly', 'headless', 'nick'), 33)
(('said', 'professor', 'mcgonagall'), 21)
(('c', 'h', 'p'), 18)
(('h', 'p', 'e'), 18)
(('p', 'e', 'r'), 18)
(('harry', 'ron', 'hermione'), 18)
(('harry', 'could', 'see'), 14)
(('moaning', 'myrtle', 'bathroom'), 13)
(('said', 'harry', 'quickly'), 10)
(('said', 'uncle', 'vernon'), 9)
(('harry', 'said', 'ron'), 9)
(('said', 'harry', 'ron'), 9)
(('defense', 'dark', 'arts'), 9)
(('harry', 'said', 'lockhart'), 9)
(('right', 'said', 'harry'), 8)
(('nimbus', 'two', 'thousand'), 8)
(('harry', 'potter', 'said'), 8)
(('harry', 'potter', 'must'), 8)
```

```
(('fifty', 'years', 'ago'), 8)
(('hogwarts', 'harry', 'potter'), 7)
(('gryffindor', 'common', 'room'), 7)
(('said', 'nearly', 'headless'), 7)
(('slytherin', 'common', 'room'), 7)
(('hogwarts', 'school', 'witchcraft'), 6)
(('famous', 'harry', 'potter'), 6)
(('said', 'ron', 'looking'), 6)
(('said', 'professor', 'sprout'), 6)
(('e', 'r', 'e'), 6)
(('e', 'e', 'n'), 6)
(('harry', 'potter', 'chamber'), 5)
(('potter', 'chamber', 'secrets'), 5)
(('chamber', 'secrets', 'opened'), 5)
(('school', 'witchcraft', 'wizardry'), 5)
(('ron', 'fred', 'george'), 5)
(('harry', 'never', 'seen'), 5)
(('harry', 'told', 'ron'), 5)
(('fred', 'george', 'weasley'), 5)
(('chamber', 'secrets', 'said'), 5)
(('justin', 'nearly', 'headless'), 5)
(('sir', 'said', 'riddle'), 5)
(('sir', 'said', 'dobby'), 4)
(('back', 'onto', 'bed'), 4)
(('said', 'harry', 'angrily'), 4)
(('yes', 'said', 'harry'), 4)
(('harry', 'backed', 'away'), 4)
(('back', 'said', 'ron'), 4)
(('harry', 'looked', 'around'), 4)
(('e', 'r', 'f'), 4)
(('dark', 'arts', 'teacher'), 4)
(('course', 'said', 'ron'), 4)
HP3.txt
TRIGRAM FREQUENCY - Harry Potter 3
(('harry', 'ron', 'hermione'), 46)
(('said', 'professor', 'mcgonagall'), 31)
(('said', 'professor', 'lupin'), 23)
(('c', 'h', 'p'), 22)
(('h', 'p', 'e'), 22)
(('p', 'e', 'r'), 22)
(('defense', 'dark', 'arts'), 21)
(('talons', 'tea', 'leaves'), 15)
```

```
(('professor', 'trelawney', 'prediction'), 13)
(('flight', 'fat', 'lady'), 12)
(('wormtail', 'padfoot', 'prongs'), 12)
(('servant', 'lord', 'voldemort'), 12)
(('yes', 'said', 'harry'), 12)
(('said', 'harry', 'quickly'), 11)
(('care', 'magical', 'creatures'), 11)
(('gryffindor', 'versus', 'ravenclaw'), 10)
(('cat', 'rat', 'dog'), 10)
(('yeah', 'said', 'harry'), 10)
(('said', 'ron', 'looking'), 10)
(('said', 'madam', 'pomfrey'), 10)
(('aunt', 'marge', 'big'), 9)
(('marge', 'big', 'mistake'), 9)
(('said', 'harry', 'hermione'), 9)
(('malfoy', 'crabbe', 'goyle'), 9)
(('disposal', 'dangerous', 'creatures'), 9)
(('nimbus', 'two', 'thousand'), 8)
(('harry', 'said', 'hermione'), 8)
(('said', 'professor', 'trelawney'), 8)
(('hogwarts', 'harry', 'potter'), 7)
(('moony', 'wormtail', 'padfoot'), 7)
(('said', 'harry', 'suddenly'), 7)
(('harry', 'said', 'ron'), 7)
(('harry', 'looked', 'around'), 7)
(('hagrid', 'said', 'hermione'), 7)
(('said', 'dumbledore', 'quietly'), 7)
(('know', 'said', 'harry'), 7)
(('harry', 'could', 'see'), 7)
(('said', 'madam', 'rosmerta'), 7)
(('hagrid', 'said', 'harry'), 7)
(('expecto', 'patronum', 'expecto'), 7)
(('e', 'e', 'n'), 7)
(('uncle', 'vernon', 'aunt'), 6)
(('vernon', 'aunt', 'petunia'), 6)
(('said', 'uncle', 'vernon'), 6)
(('right', 'said', 'harry'), 6)
(('said', 'harry', 'looking'), 6)
(('said', 'aunt', 'marge'), 6)
(('okay', 'said', 'harry'), 6)
(('said', 'harry', 'ron'), 6)
(('ron', 'said', 'hermione'), 6)
```

```
HP4.txt
TRIGRAM FREQUENCY - Harry Potter 4
(('harry', 'ron', 'hermione'), 74)
(('quidditch', 'world', 'cup'), 35)
(('harry', 'could', 'see'), 31)
(('yeah', 'said', 'harry'), 27)
(('harry', 'looked', 'around'), 24)
(('care', 'magical', 'creatures'), 18)
(('put', 'name', 'goblet'), 17)
(('right', 'said', 'harry'), 16)
(('okay', 'said', 'harry'), 15)
(('said', 'professor', 'mcgonagall'), 15)
(('harry', 'could', 'hear'), 14)
(('said', 'hermione', 'looking'), 14)
(('defense', 'dark', 'arts'), 14)
(('malfoy', 'crabbe', 'goyle'), 14)
(('harry', 'ever', 'seen'), 13)
(('said', 'ron', 'looking'), 13)
(('said', 'harry', 'ron'), 13)
(('nearly', 'headless', 'nick'), 13)
(('said', 'madame', 'maxime'), 13)
(('said', 'cold', 'voice'), 12)
(('harry', 'said', 'dumbledore'), 12)
(('said', 'rita', 'skeeter'), 12)
(('ron', 'said', 'hermione'), 11)
(('harry', 'could', 'tell'), 10)
(('yes', 'said', 'harry'), 10)
(('er', 'said', 'harry'), 9)
(('said', 'ron', 'hermione'), 9)
(('yeah', 'said', 'ron'), 9)
(('moody', 'magical', 'eye'), 9)
(('said', 'harry', 'slowly'), 9)
(('harry', 'felt', 'though'), 8)
(('international', 'magical', 'cooperation'), 8)
(('said', 'harry', 'grinning'), 8)
(('harry', 'never', 'seen'), 8)
(('right', 'said', 'ron'), 8)
(('harry', 'said', 'hermione'), 8)
(('bill', 'charlie', 'percy'), 8)
(('harry', 'could', 'make'), 8)
(('hermione', 'said', 'ron'), 8)
(('said', 'professor', 'trelawney'), 8)
```

```
(('karkaroff', 'madame', 'maxime'), 8)
(('thanks', 'said', 'harry'), 8)
(('magical', 'games', 'sports'), 7)
(('yeah', 'right', 'said'), 7)
(('said', 'ron', 'harry'), 7)
(('department', 'regulation', 'control'), 7)
(('copy', 'daily', 'prophet'), 7)
(('harry', 'said', 'ron'), 7)
(('said', 'nearly', 'headless'), 7)
(('back', 'gryffindor', 'tower'), 7)
HP5.txt
TRIGRAM FREQUENCY - Harry Potter 5
(('yeah', 'said', 'harry'), 45)
(('defense', 'dark', 'arts'), 45)
(('said', 'professor', 'mcgonagall'), 44)
(('harry', 'ron', 'hermione'), 41)
(('c', 'h', 'p'), 38)
(('h', 'p', 'e'), 38)
(('p', 'e', 'r'), 38)
(('said', 'professor', 'umbridge'), 35)
(('yes', 'said', 'harry'), 29)
(('said', 'hermione', 'looking'), 24)
(('harry', 'looked', 'around'), 23)
(('educational', 'decree', 'number'), 22)
(('harry', 'could', 'see'), 22)
(('twelve', 'grimmauld', 'place'), 21)
(('said', 'uncle', 'vernon'), 21)
(('harry', 'said', 'hermione'), 21)
(('number', 'twelve', 'grimmauld'), 19)
(('said', 'ron', 'looking'), 19)
(('said', 'harry', 'quietly'), 18)
(('mungo', 'hospital', 'magical'), 17)
(('hospital', 'magical', 'maladies'), 17)
(('magical', 'maladies', 'injuries'), 17)
(('know', 'said', 'harry'), 17)
(('said', 'harry', 'quickly'), 17)
(('hogwarts', 'high', 'inquisitor'), 16)
(('right', 'said', 'harry'), 16)
(('yeah', 'said', 'ron'), 16)
(('said', 'harry', 'angrily'), 16)
(('snape', 'worst', 'memory'), 15)
(('well', 'said', 'harry'), 15)
```

```
(('oh', 'yeah', 'said'), 15)
(('harry', 'said', 'ron'), 15)
(('said', 'harry', 'ron'), 15)
(('nearly', 'headless', 'nick'), 15)
(('noble', 'ancient', 'house'), 14)
(('ancient', 'house', 'black'), 14)
(('christmas', 'closed', 'ward'), 14)
(('second', 'war', 'begins'), 14)
(('potter', 'said', 'snape'), 14)
(('said', 'harry', 'loudly'), 13)
(('harry', 'could', 'hear'), 13)
(('sorting', 'hat', 'new'), 12)
(('hat', 'new', 'song'), 12)
(('e', 'r', 'w'), 12)
(('r', 'h', 'r'), 12)
(('said', 'harry', 'well'), 12)
(('gryffindor', 'common', 'room'), 12)
(('well', 'said', 'hermione'), 12)
(('yes', 'said', 'hermione'), 12)
(('must', 'tell', 'lies'), 12)
HP6.txt
TRIGRAM FREQUENCY - Harry Potter 6
(('c', 'h', 'p'), 30)
(('h', 'p', 'e'), 30)
(('p', 'e', 'r'), 30)
(('harry', 'ron', 'hermione'), 29)
(('said', 'professor', 'mcgonagall'), 29)
(('defense', 'dark', 'arts'), 27)
(('harry', 'said', 'dumbledore'), 26)
(('sir', 'said', 'harry'), 25)
(('yes', 'said', 'dumbledore'), 20)
(('harry', 'said', 'hermione'), 19)
(('said', 'prime', 'minister'), 18)
(('right', 'said', 'harry'), 18)
(('yeah', 'said', 'harry'), 17)
(('harry', 'could', 'see'), 17)
(('said', 'harry', 'quickly'), 16)
(('yes', 'said', 'harry'), 16)
(('said', 'harry', 'well'), 15)
(('harry', 'could', 'tell'), 15)
(('said', 'professor', 'trelawney'), 14)
(('lord', 'voldemort', 'request'), 13)
```

```
(('e', 'r', 'w'), 12)
(('hermione', 'helping', 'hand'), 11)
(('said', 'dumbledore', 'quietly'), 11)
(('said', 'ron', 'looking'), 11)
(('said', 'hermione', 'looking'), 11)
(('r', 'w', 'e'), 11)
(('oh', 'yes', 'said'), 10)
(('harry', 'looked', 'around'), 10)
(('said', 'dumbledore', 'smiling'), 10)
(('harry', 'said', 'nothing'), 10)
(('w', 'e', 'n'), 10)
(('lightning', 'struck', 'tower'), 9)
(('yeah', 'said', 'ron'), 9)
(('professor', 'said', 'harry'), 9)
(('yes', 'said', 'hermione'), 9)
(('eyes', 'fixed', 'upon'), 8)
(('said', 'harry', 'looking'), 8)
(('think', 'said', 'dumbledore'), 8)
(('dunno', 'said', 'harry'), 8)
(('dumbledore', 'said', 'harry'), 8)
(('hermione', 'said', 'harry'), 8)
(('said', 'harry', 'got'), 8)
(('said', 'dumbledore', 'harry'), 8)
(('hogwarts', 'harry', 'potter'), 7)
(('said', 'harry', 'could'), 7)
(('right', 'said', 'hermione'), 7)
(('neither', 'ron', 'hermione'), 7)
(('know', 'said', 'harry'), 7)
(('said', 'ron', 'hermione'), 7)
(('course', 'said', 'harry'), 7)
HP7.txt
TRIGRAM FREQUENCY - Harry Potter 7
(('harry', 'ron', 'hermione'), 44)
(('c', 'h', 'p'), 36)
(('h', 'p', 'e'), 36)
(('p', 'e', 'r'), 36)
(('said', 'professor', 'mcgonagall'), 19)
(('harry', 'could', 'see'), 18)
(('life', 'lies', 'albus'), 14)
(('lies', 'albus', 'dumbledore'), 14)
(('tale', 'three', 'brothers'), 14)
(('harry', 'looked', 'around'), 14)
```

```
(('yeah', 'said', 'harry'), 14)
(('harry', 'said', 'hermione'), 14)
(('e', 'r', 'w'), 12)
(('pulled', 'invisibility', 'cloak'), 12)
(('sacking', 'severus', 'snape'), 11)
(('said', 'ron', 'looking'), 11)
(('said', 'ron', 'harry'), 11)
(('said', 'hermione', 'looking'), 11)
(('tales', 'beedle', 'bard'), 11)
(('harry', 'could', 'hear'), 11)
(('r', 'w', 'e'), 11)
(('r', 'h', 'r'), 10)
(('right', 'said', 'harry'), 10)
(('harry', 'could', 'help'), 10)
(('ron', 'said', 'hermione'), 10)
(('w', 'e', 'n'), 10)
(('e', 'r', 'h'), 9)
(('know', 'said', 'harry'), 9)
(('ron', 'hermione', 'looked'), 9)
(('said', 'ron', 'hermione'), 9)
(('harry', 'could', 'tell'), 9)
(('pouch', 'around', 'neck'), 9)
(('one', 'death', 'eaters'), 8)
(('harry', 'could', 'feel'), 8)
(('yeah', 'said', 'ron'), 8)
(('said', 'harry', 'dumbledore'), 8)
(('harry', 'opened', 'eyes'), 8)
(('said', 'auntie', 'muriel'), 8)
(('right', 'said', 'hermione'), 8)
(('hogwarts', 'harry', 'potter'), 7)
(('dark', 'lord', 'ascending'), 7)
(('final', 'hiding', 'place'), 7)
(('said', 'uncle', 'vernon'), 7)
(('well', 'said', 'harry'), 7)
(('took', 'deep', 'breath'), 7)
(('harry', 'said', 'ron'), 7)
(('sorry', 'said', 'ron'), 7)
(('harry', 'saw', 'hermione'), 7)
(('e', 'e', 'n'), 7)
(('beneath', 'invisibility', 'cloak'), 7)
```

Question 2 B

Potential issues for the lists and bigrams that were found are that some of the words on their surface do not provide information about the text. The word “looked” is prevalent in all seven of the books, however “looked” on its own does not provide meaningful information. This arises the question who looked? The bigrams answer this question with “harry looked” and “hagrid looked” both being included in the top 50 bigrams for book 1. This information proves more meaningful and justifies “looked” being included and not removed. There are multiple words that for the top 50 words by frequency might not seem useful, however are potentially useful when included in bigrams and trigrams.

The top 50 bigrams were different when using PMI and frequency measures. This is due to the fact that PMI measures the likelihood that those words appear together when they appear. For example the bigram 'adalbert' waffling' has a PMI of 16.23, this means that practically every time “adalbert” is in the text “waffling” will directly follow. The top bigram by frequency is “uncle vernon”. This shows that the bigram “uncle vernon” was the most common bigram, but “uncle” did not necessarily mean that “vernon” would follow and “vernon” did not necessarily mean that “uncle” would precede it. There is a big difference in the bigrams scored by frequency and the top bigrams using PMI. The top 50 bigrams scored by frequency use words that are more common they are easy to understand and typically include a character or a verb in part of the bigram. A couple examples are “said dumbledore”, “ harry potter”, “common room”, and “harry looked”. Quite a few of the words used in the bigrams are also words that appear in the list of the top 50 words by frequency. This is not the case for the bigrams scored by PMI. The bigrams scored by PMI do not use common words or characters they are more obscure. A couple examples are “adalbert waffling”, “chipolatas tureens”, “chewed greedily”, and “aquavirius maggots”. These are words that are not as commonly used.

Question 3

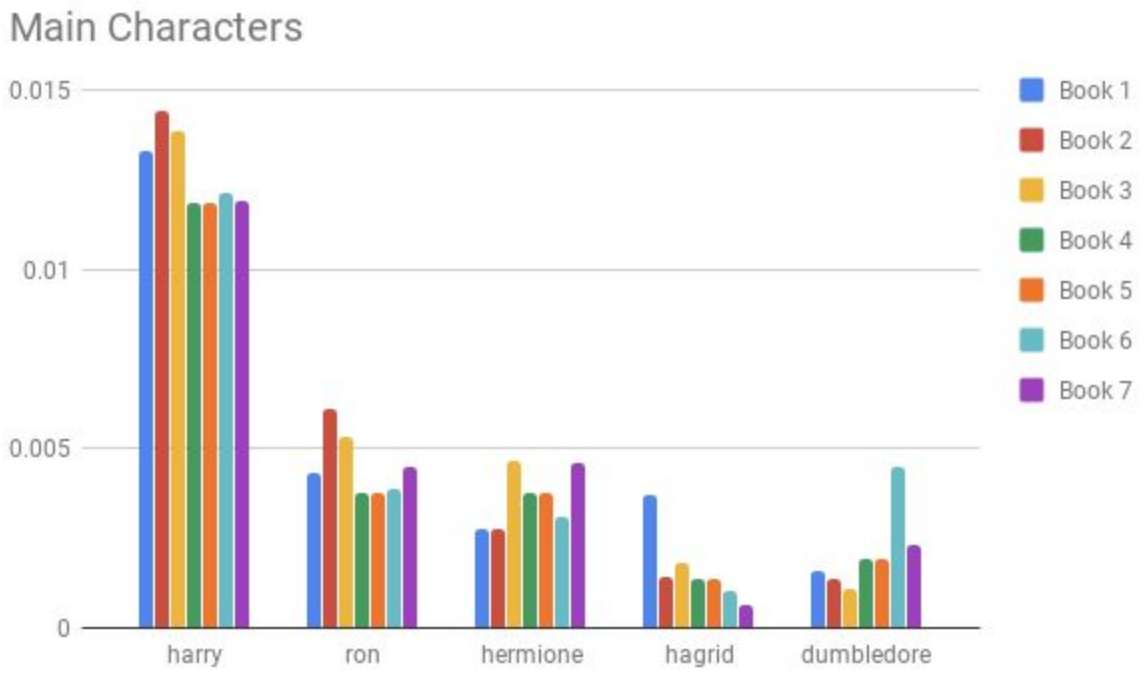
Questions to address:

1. **Main characters - do they change and how?**

2. **Can we predict who the villains are based on the frequency lists?**
3. **Do the books get darker over time**

Question 1: Main Characters - Do they change and how?

We wanted to take a look at the main characters and see how they evolve in the text. Do characters' importance wax and wean? Are certain characters consistent throughout the book?



By analyzing the top 50 words for main characters, it was enlightening to see how the main characters changed throughout the books. Hagrid play a major role in book 1, a lesser role in books 2, 3, 4, 5 and 6 and then dropped off of the list of the top 50 words for book 7. Another interesting aspect was the evolution of Hermione's importance in the books. In book 1 the top three characters by frequency were Harry, Ron and Hagrid. Hermione's name did not appear until the 6 word, however in books 2 -5 Hermione was the third most frequent character. In book 6 Dumbledore plays a major role and knocks both Ron and Hermione down a step in the top frequency list. However, Hermione barely edges out Ron in the top frequency list for book 7 and becomes the second most frequent character.

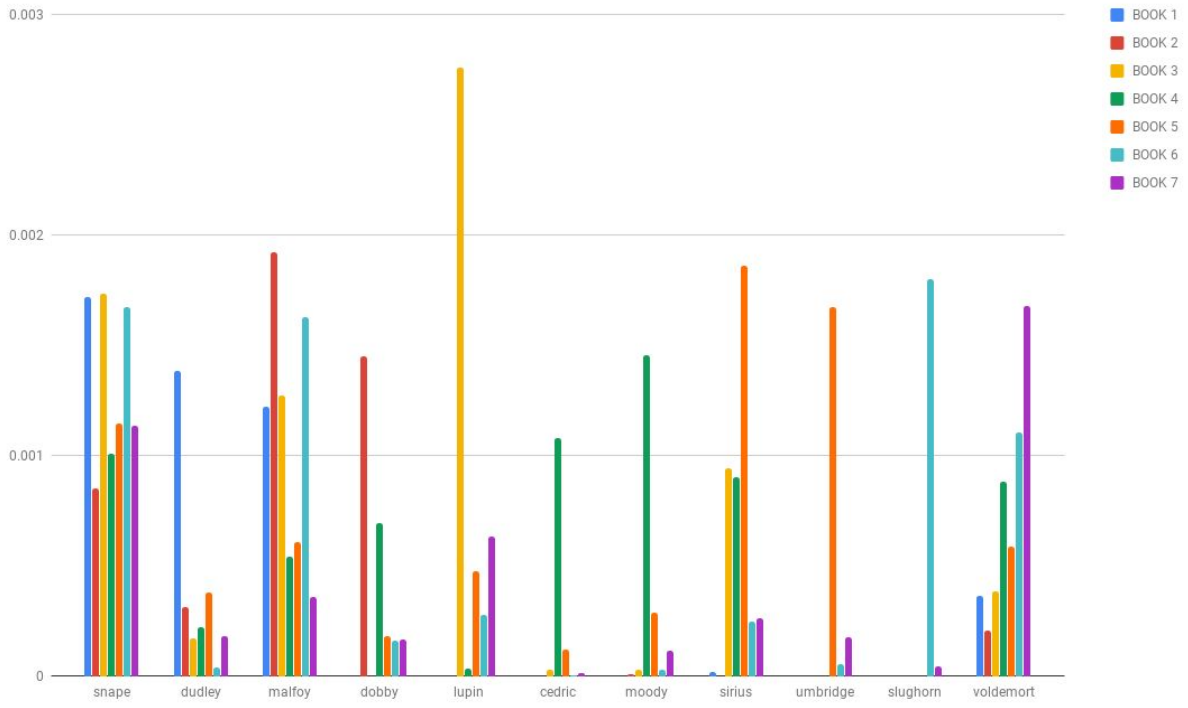
Hermione's evolution is also prevalent in the top 50 bigrams. The word said was included to enable the researchers to determine which character or characters spoke the most throughout the book. In books 1 – 4 the top two bigrams are "said harry" and "said ron". This shows that Harry and Ron played an active role in the book. The bigram "said hermione" was the 9th most frequent in book 1, the 6th most frequent in book 2, and the 3rd most frequent in books 3 & 4. In book 5 "said hermione" became the second highest bigram with a normalized frequency of .00314 and "said ron" dropped to the third with a normalized frequency of .00259. The top 50 bigrams for book 6, confirm the information gained from the top 50 word list in book 6, that Dumbledore becomes a major character. The bigram "said dumbledore" surpassed the bigrams "said ron" and "said hermione" and was the second most frequent bigram at .0033 with Ron and Hermione's bigrams having a frequency of .0022. For book 7, "said ron" has a slightly higher normalized frequency score than "said hermione". The normalized frequencies for "ron" and "hermione" in the list of top 50 words in book 7 are very close with the normalized frequency score for "hermione" at .01199 and "ron" .01158. Ultimately the three main characters are Harry, Ron and Hermione in books 2, 3, 4, 5 and 7. In book 1 Hagrid eclipsed Hermione as a top 3 character and in book 1 and then remained an important character in books 2 - 5, before dropping of the list of the top 50 most frequent words. Dumbledore was a consistent character of importance throughout all of the books and appears to have had a more significant role in book 6.

It is important to note that the bigrams by frequency do give a better understanding of the important characters in the books and the roles that they played. "Harry" is paired with a lot of action words such as "saw", "looked", and "thought" this shows that he was an active character throughout the books. Other characters whose names did not appear on the list of top 50 words, do appear in the list of top 50 bigrams.

Through analyzing the trigrams Professor McGonagall is shown to have played an important part throughout all of the books with the trigram "said professor mcgonagall" appearing at least 15 times per book. It appears that Professor McGonagall played a heavier role in books 3, 5 and 6 with the trigram appearing 31, 44, and 29 times respectively.

Question 2: Can we predict who the villains are based on the frequency lists?

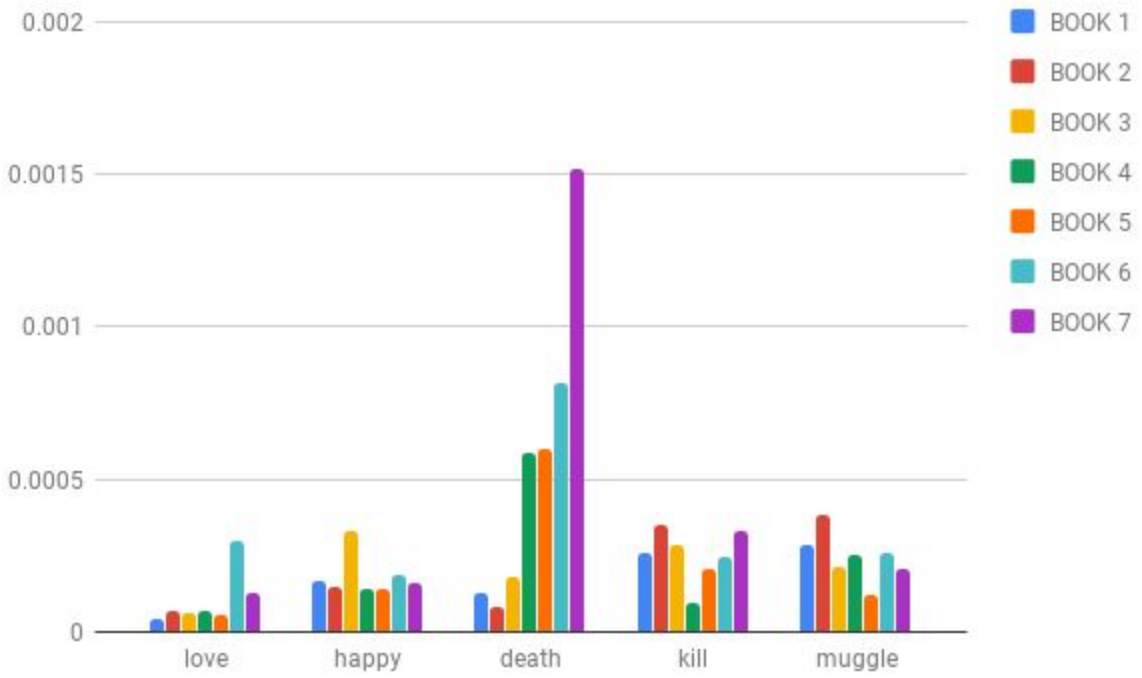
We wanted to determine if it was possible to identify the villains based on the top 50 word frequency list, without using our prior knowledge.



This question proved extremely difficult to answer. Without prior knowledge it is difficult to determine the villains based strictly on frequency lists. The characters who made the list of top 50 words by normalized frequency was compiled and used to try to determine villains. It is interesting to see the evolution of Voldemort throughout the books, his presence is amplified in books 6 and 7. Ultimately, we were able to compile a list of possible villains, but could not determine which characters are villains based on solely the information from word frequencies.

Question 3: Do the books get darker over time?

We wanted to determine if it was possible to tell if the books increased in darkness by looking at words we deemed to be polarizing.



To answer this question word frequencies for the following polarizing words were compiled: love, happy, death, kill and muggle. The books clearly got darker over time, as the word death increased in frequency throughout the books. The bigrams also displayed a move towards darkness with the bigram “death eaters” . The bigram first appears in the list of top 50 bigrams by frequency in book 5. The frequency of the bigram continually increases throughout the remaining books. “Death Eaters” is a particularly dark term and displays change in darkness over the series of the books. It is interesting that why the books get darker, love becomes more prevalent.