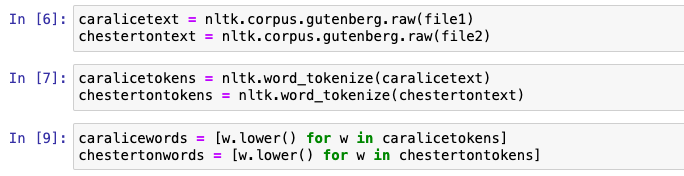
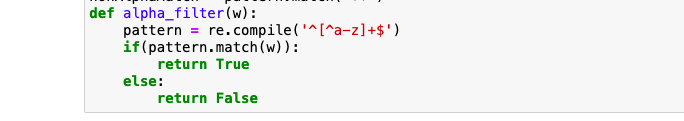
Zane Alderfer

Prof. Larche

IST 664

October 24, 2023

Homework 1

1. I used two documents from the Gutenberg: carroll-alice.txt and chesterton-ball.txt. To get these two documents ready to develop the top 50 words, bigram frequencies, I first created two variables for each document using the raw function the nltk library and then separated the documents into tokens using the word\_tokenize function and then finally made all the words in the documents using the lower() function as seen below.
2. Before I collected the top 50 words and bigram frequencies, I found out the top 50 words that didn’t involve “stoppedwords” or “alphawords” which was a function created as seen below. This function was created in the lab so I reused it here. 

After creating two new variables that collect the words of each document with this added function, and added “stoppedwords” as well, this is the top 50 words by frequency for each document.

**TOP 50 words for Carroll Alice:**

('said', 462)

('alice', 396)

("n't", 204)

("'s", 194)

('little', 128)

('one', 99)

('would', 90)

('know', 88)

('could', 86)

('like', 85)

('went', 83)

('queen', 75)

('thought', 74)

('time', 68)

('see', 67)

('king', 62)

('began', 58)

('turtle', 58)

("'and", 56)

("'m", 56)

('hatter', 56)

('mock', 56)

('quite', 55)

("'it", 55)

('gryphon', 54)

('think', 53)

('way', 53)

("'ll", 52)

('much', 51)

('say', 51)

('first', 50)

('head', 50)

("'you", 50)

('thing', 49)

('go', 48)

('voice', 48)

('rabbit', 47)

('looked', 45)

('never', 45)

('got', 45)

('get', 44)

('must', 44)

('mouse', 42)

('duchess', 42)

('round', 41)

('came', 40)

('tone', 40)

('dormouse', 40)

('great', 39)

("'ve", 39)

**Top 50 words for Chesterton Ball:**

('said', 652)

('turnbull', 544)

('macian', 425)

('man', 336)

('like', 326)

('one', 316)

("'s", 285)

("n't", 178)

('two', 177)

('us', 162)

('quite', 144)

('would', 143)

('even', 140)

('evan', 140)

('know', 139)

('god', 136)

('face', 134)

('little', 129)

('say', 127)

('seemed', 121)

('could', 120)

('see', 119)

('really', 116)

('thing', 113)

('world', 112)

('something', 112)

('still', 105)

('looked', 104)

('come', 104)

('made', 103)

('men', 102)

('voice', 102)

('long', 100)

('well', 98)

('cried', 97)

('may', 97)

('upon', 95)

('went', 95)

('back', 94)

('almost', 94)

('think', 94)

('old', 92)

('saw', 92)

('head', 91)

('came', 90)

('time', 87)

('eyes', 87)

('garden', 86)

('first', 86)

('mr.', 85)

After finding the top words for each document, I went through a similar process of applying the alpha filter and stoppedwords to the bigrams. I then used the function score\_ngrams from nltk to find the bigrams with the highest frequencies and got these results for each document:

**TOP 50 bigram frequencies for Carroll Alice**

(('said', 'the'), 0.006180211381142891)

(('of', 'the'), 0.0038215799844748314)

(('said', 'alice'), 0.003433450767301606)

(('in', 'a'), 0.0028960410819848332)

(('and', 'the'), 0.00235863139666806)

(('in', 'the'), 0.0023287753030393505)

(('it', 'was'), 0.0021794948348958024)

(('the', 'queen'), 0.0020600704603809636)

(('to', 'the'), 0.0020600704603809636)

(('as', 'she'), 0.0018212217113512867)

(('the', 'king'), 0.0018212217113512867)

(('at', 'the'), 0.0017913656177225771)

(('she', 'had'), 0.0017913656177225771)

(('a', 'little'), 0.0017615095240938676)

(('i', "'m"), 0.0016719412432077386)

(('she', 'was'), 0.0016719412432077386)

(('mock', 'turtle'), 0.001642085149579029)

(('and', 'she'), 0.0015823729623216098)

(('the', 'mock'), 0.0015823729623216098)

(('do', "n't"), 0.0015525168686929003)

(('the', 'gryphon'), 0.0015525168686929003)

(('the', 'hatter'), 0.0015525168686929003)

(('to', 'be'), 0.0015226607750641907)

(('went', 'on'), 0.0014330924941780617)

(('to', 'herself'), 0.001343524213291933)

(('you', 'know'), 0.0012838120260345136)

(('the', 'duchess'), 0.0011942437451483848)

(('said', 'to'), 0.0011643876515196752)

(('out', 'of'), 0.0011046754642622559)

(('i', 'do'), 0.0010748193706335463)

(('there', 'was'), 0.0010449632770048367)

(('on', 'the'), 0.0010151071833761271)

(('she', 'said'), 0.0010151071833761271)

(('the', 'dormouse'), 0.0010151071833761271)

(('she', 'could'), 0.0009852510897474175)

(('with', 'the'), 0.0009852510897474175)

(('i', "'ve"), 0.0009553949961187078)

(('it', "'s"), 0.0009553949961187078)

(('that', 'she'), 0.0009553949961187078)

(('and', 'then'), 0.0009255389024899983)

(('march', 'hare'), 0.0009255389024899983)

(('was', 'a'), 0.0009255389024899983)

(('the', 'other'), 0.0008956828088612886)

(('she', 'went'), 0.000865826715232579)

(('so', 'she'), 0.000865826715232579)

(('the', 'march'), 0.000865826715232579)

(('did', 'not'), 0.0008359706216038693)

(('the', 'mouse'), 0.0008359706216038693)

(('to', 'her'), 0.0008359706216038693)

(('i', "'ll"), 0.0008061145279751597)

**TOP 50 bigram frequencies for Chesteron Ball:**

(('of', 'the'), 0.006324297643042244)

(('in', 'the'), 0.004133931142281272)

(('with', 'a'), 0.0022932006087779196)

(('and', 'the'), 0.001994981695528773)

(('it', 'was'), 0.0019641314631236887)

(('in', 'a'), 0.001953848052321994)

(('it', 'is'), 0.0018407305335033524)

(('to', 'the'), 0.0018201637118999629)

(('of', 'a'), 0.0016247789066677636)

(('he', 'said'), 0.0015322282094525112)

(('he', 'had'), 0.001501377977047427)

(('he', 'was'), 0.0014910945662457325)

(('with', 'the'), 0.0014191106906338694)

(('at', 'the'), 0.001367693636625396)

(('on', 'the'), 0.001295709761013533)

(('the', 'other'), 0.0012854263502118384)

(('to', 'be'), 0.0012854263502118384)

(('i', 'am'), 0.0012751429394101436)

(('like', 'a'), 0.0012545761178067542)

(('said', 'turnbull'), 0.0012340092962033647)

(('said', 'the'), 0.0012237258854016701)

(('all', 'the'), 0.0011003249557813336)

(('of', 'his'), 0.001018057669367776)

(('as', 'if'), 0.0009872074369626919)

(('and', 'he'), 0.000976924026160997)

(('is', 'a'), 0.0009357903829542183)

(('in', 'his'), 0.0009049401505491342)

(('was', 'a'), 0.0009049401505491342)

(('a', 'man'), 0.0008843733289457447)

(('into', 'the'), 0.0008843733289457447)

(('out', 'of'), 0.0008843733289457447)

(('you', 'are'), 0.0008843733289457447)

(('that', 'the'), 0.00087408991814405)

(('said', 'macian'), 0.0008535230965406606)

(('by', 'the'), 0.0008432396857389659)

(('do', "n't"), 0.0008432396857389659)

(('if', 'you'), 0.0008432396857389659)

(('i', 'have'), 0.000771255810127103)

(('from', 'the'), 0.0007404055777220188)

(('is', 'the'), 0.0007404055777220188)

(('the', 'same'), 0.0007404055777220188)

(('do', 'you'), 0.0007301221669203242)

(('for', 'the'), 0.0007301221669203242)

(('the', 'world'), 0.0007301221669203242)

(('sort', 'of'), 0.0007198387561186294)

(('that', 'he'), 0.0007198387561186294)

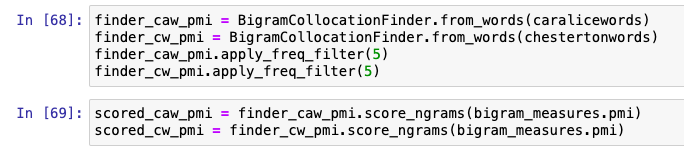
(('and', 'i'), 0.0007095553453169347)

(('one', 'of'), 0.0007095553453169347)

(('the', 'man'), 0.00069927193451524)

(('they', 'had'), 0.00069927193451524)

Finally for top 50 bigrams by their Mutual Information scores, I set the minimum frequency for these scores to be 5 as shown below. I then ran some for loops to get the top 30 bigrams based on pmi score.



**TOP 50 bigrams by Mutual Info scores in Carroll Alice:**

(('golden', 'key'), 11.639297636056998)

(('play', 'croquet'), 11.768580653001967)

(('kid', 'gloves'), 11.572183440198463)

(('few', 'minutes'), 10.987220939477307)

(("'of", 'course'), 10.262227986977177)

(('white', 'kid'), 10.124724463227242)

(('beautiful', 'soup'), 9.987220939477307)

(('three', 'gardeners'), 9.972721369782192)

(('march', 'hare'), 9.94415221758542)

(('good', 'deal'), 9.669044979451051)

(('any', 'rate'), 9.613762543949864)

(('cheshire', 'cat'), 9.598655651559655)

(('trembling', 'voice'), 9.183618152280811)

(('their', 'slates'), 9.166544638921868)

(("'d", 'better'), 9.165366447724587)

(('mock', 'turtle'), 9.147638855175243)

(('next', 'witness'), 9.124724463227242)

(('feet', 'high'), 9.105615640279538)

(('\*', '\*'), 9.050723881783464)

(('white', 'rabbit'), 9.029567230186903)

(('your', 'majesty'), 8.999193581143384)

(('great', 'hurry'), 8.87174372205737)

(('your', 'pardon'), 8.86169005739345)

(('right', 'size'), 8.746212839973513)

(('beg', 'your'), 8.709686963948398)

(('offended', 'tone'), 8.709686963948398)

(('their', 'heads'), 8.622224122698057)

(('oh', 'dear'), 8.572183440198462)

(('its', 'mouth'), 8.461759450504811)

(("'m", 'afraid'), 8.446652558114604)

(('other', 'side'), 8.207186623419213)

(('minute', 'or'), 8.074513017273473)

(('another', 'moment'), 7.93991522469895)

(("'ve", 'tried'), 7.820213421417289)

(('are', 'old'), 7.814061194435693)

(('left', 'off'), 7.786854824470858)

(('no', 'use'), 7.76482851814086)

(('my', 'dear'), 7.738833309607916)

(('same', 'thing'), 7.700698180721142)

(('let', 'me'), 7.664044298392685)

(('very', 'politely'), 7.6496395802784605)

(("'off", 'with'), 7.547799281571505)

(('little', 'golden'), 7.5461882316655196)

(('please', 'your'), 7.539761962506086)

(('more', 'than'), 7.478305759384694)

(('sat', 'down'), 7.441651877056239)

(('or', 'two'), 7.37285429184322)

(('ca', "n't"), 7.359189716864265)

(('wo', "n't"), 7.359189716864265)

(('came', 'upon'), 7.331175340694669)

**Top 50 bigrams based on Mutual Info Scores in Chesterton Ball:**

(('literary', 'archive'), 13.98435911671854)

(('eighteenth', 'century'), 13.569321617439696)

(('archive', 'foundation'), 13.399396615997386)

(('count', 'gregory'), 12.984359116718538)

(('\_the', 'atheist\_'), 12.569321617439696)

(('fleet', 'street'), 12.177004194660935)

(('st.', 'paul'), 11.925465427664971)

(('public', 'domain'), 11.868881899298604)

(('gutenberg', 'literary'), 11.846855592968605)

(('dr.', 'quayle'), 11.818893763466928)

(('cumberland', 'vane'), 11.809988210245036)

(('ludgate', 'hill'), 11.624463171632158)

(('project', 'gutenberg'), 11.44003860049473)

(('project', 'gutenberg-tm'), 11.440038600494729)

(('small', 'print'), 11.109889998802396)

(('ten', 'minutes'), 11.089541353410596)

(('madeleine', 'durand'), 10.732820349722575)

(('professor', 'lucifer'), 10.732820349722575)

(('dr.', 'hutton'), 10.643322198883475)

(('few', 'moments'), 10.476564476519842)

(('lunatic', 'asylum'), 10.45384440001976)

(('flying', 'ship'), 10.355002496638928)

(('difference', 'between'), 10.018574832056451)

(('few', 'yards'), 9.954611773324489)

(('south', 'sea'), 9.841401162876497)

(("'m", 'afraid'), 9.772741166976267)

(('mr.', 'wilkinson'), 9.74489318202315)

(('three', 'sides'), 9.68179634669811)

(('great', 'deal'), 9.643322198883475)

(('public', 'house'), 9.595863404892189)

(('three', 'days'), 9.594333505447771)

(('any', 'rate'), 9.473397197441162)

(('swung', 'round'), 9.454538170189846)

(('\*', '\*'), 9.427103545167228)

(('good', 'lord'), 9.399396615997386)

(("'m", 'sure'), 9.332828999058396)

(('greater', 'than'), 9.251004776104715)

(('ca', "n't"), 9.093588186473298)

(('wo', "n't"), 9.093588186473296)

(('young', 'woman'), 9.088733347205837)

(('beg', 'your'), 8.96940877525257)

(('whole', 'universe'), 8.946223987831768)

(('my', 'dear'), 8.849199533436904)

(('mr.', 'cumberland'), 8.781419058048266)

(('your', 'pardon'), 8.706374369418777)

(('print', '!'), 8.604980749647277)

(('young', 'lady'), 8.5001595929323)

(('mr.', 'james'), 8.496965668579564)

(('excuse', 'me'), 8.487933290599042)

(('paul', "'s"), 8.414503508387591)

**a.)** In terms of choosing the processing options I did, I essentially used whatever we used in the labs to get the data where I wanted it to be to develop these lists. What we used in the lab in terms of tokenizing, adding the alpha filter and eliminating “stoppedwords” seemed like the most effective way to do this.

**b.)** For the top 50 bigrams list, for some reason adding the alpha function and removing the stopped words didn’t seem to work but you’ll see in my python file that I applied them. So, in terms of accuracy, it’s a correct bigram frequency list but it doesn’t tell us much in terms of meaningful information regarding the documents.

**c.)** Well as I mentioned in step b, I couldn’t get the bigrams to a point that delivered meaningful words but the regular bigram frequency and the PMI score delivers a distinct difference between the amount of times they’re frequently together and the association between them. The bigrams with a higher PMI score indicates some level of association between the two words while bigrams with a high frequency of occurrence together doesn’t necessarily suggest an association between the two words.

**d.)** On top of the stop words list, I added the alpha filter function that we created in the lab that used a regex to get rid of tokens that were essentially just punctuation.

**e.)** I did not add trigrams for this exercise.

1. My question is: **“Based on PMI scores and types of words frequently used in each document, is it possible to tell what the topic of each document is?”**

In *The Ball and the Cross* by Chesterton, you can tell by bigrams with high PMI scores like (('\_the', 'atheist\_'), 12.569321617439696) or (('st.', 'paul'), 11.925465427664971) or (('whole', 'universe'), 8.946223987831768) that this is a document very much fueled by religious talk or explanations pertaining to the universe. You can also tell by the high frequency of “young lady”, “young woman”, and high frequency of using the titles Mr. or Dr. that this is more than likely a lecture style format.

Then you can see the familiar word associations that you’d expect from *Alice in Wonderland* by Lewis Carroll with bigrams such as (('golden', 'key'), 11.639297636056998), (('white', 'rabbit'), 9.029567230186903), and (('cheshire', 'cat'), 9.598655651559655). It is pretty amazing that even if I didn’t know the title of this document and by just using bigrams and their PMI scores, I’d probably be able to guess that this is *Alice in Wonderland*.

In conclusion, a valuable lesson that presents itself here is that by using PMI scores and bigram frequencies, you can unlock all sorts of interesting ideas from previous century writers such as style of writing, syntax, and even maybe be able to discover what it is exactly that made these writers so unique and popular.