**IST 664 Homework 1:**

## **Corpus Statistics and Mutual Information**

**Student: Patrick Walsh**

**Instructor: Professor Larche**

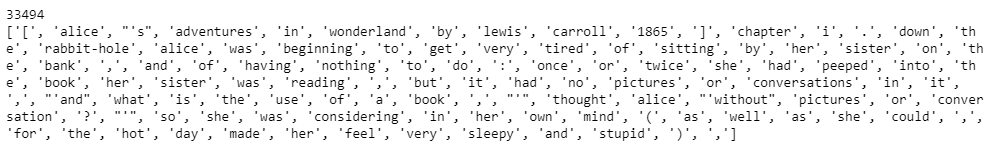
**Date: 10/23/2023**

1. **Choosing the data: Choose existing large documents from NLTK or the Gutenberg collection on the web.**

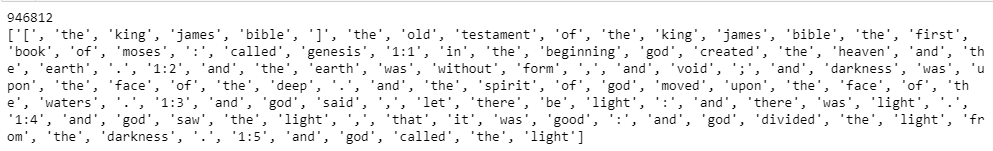
The two documents I chose for this analysis are Alice in Wonderland, and the KJV Bible, both from the Gutenberg collection.

I tokenized each document and set the tokens to lowercase. Alice in Wonderland is 33,494 tokens in length (see Figure 1), while the KJV Bible is 946,812 tokens in length (see Figure 2).

**Figure 1**



**Figure 2**



1. **Examine the text in the documents that you chose and decide how to process the words, i.e. decide on tokenization and whether to use all lowercase, stopwords, or lemmatization. Using the process developed in the lab:**
   1. **list the top 50 words by frequency (normalized by the length of the document)**

Before additional preprocessing, the corpus was analyzed for word frequency distribution. The top 50 words were examined for both documents. The results are seen below. Notice that the words include non-alphabetic tokens such as punctuation marks.

**Alice in Wonderland's top 50 words:**

, 2418

the 1616

' 1309

. 975

and 810

to 720

a 631

she 544

it 539

i 533

of 499

said 462

! 450

alice 396

was 366

in 359

you 356

that 284

-- 264

as 256

her 248

: 233

at 209

n't 204

? 202

's 194

; 194

on 191

had 184

with 179

all 178

be 148

for 146

so 144

very 139

not 135

they 135

but 131

this 131

little 128

do 125

he 117

is 113

out 113

what 103

down 102

one 99

up 97

his 95

about 94

**KJV Bible’s top 50 words:**

, 70573

the 64023

and 51696

of 34670

. 26202

to 13580

that 12912

: 12706

in 12667

he 10419

; 10139

shall 9838

unto 8997

for 8971

i 8854

his 8473

a 8177

lord 7944

they 7376

be 7013

is 6989

not 6780

him 6659

them 6430

it 6129

with 6012

all 5620

thou 5474

thy 4600

was 4522

god 4467

which 4413

my 4368

me 4096

said 3999

but 3992

ye 3983

their 3942

have 3904

will 3836

thee 3826

from 3642

as 3520

? 3297

are 2950

when 2834

this 2785

out 2775

were 2772

upon 2748

Among the top words in each document is comma (“,”), “the”, and “and.” As we do not want to include common stopwords or non-alphabetic characters in our word frequency distribution, we will perform additional preprocessing to remove stopwords as well as non-alphabetic characters.

Here are the updated top 50 words after additional preprocessing on each document:

**Alice in Wonderland's top 50 words (after additional preprocessing):**

said 462

alice 396

little 128

one 99

know 88

like 85

went 83

queen 75

thought 74

time 68

see 67

king 62

began 58

turtle 58

'and 56

hatter 56

mock 56

quite 55

'it 55

gryphon 54

think 53

way 53

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

mouse 42

duchess 42

round 41

came 40

tone 40

dormouse 40

great 39

'but 39

'what 38

well 37

back 37

two 37

cat 36

march 34

large 33

**Bible top 50 words (after additional preprocessing):**

shall 9838

unto 8997

lord 7944

thou 5474

thy 4600

god 4467

said 3999

ye 3983

thee 3826

upon 2748

man 2728

israel 2571

king 2506

son 2388

hath 2264

people 2143

came 2093

house 2024

come 1971

one 1969

children 1818

also 1769

day 1740

land 1718

men 1673

shalt 1616

let 1511

go 1492

hand 1466

us 1448

saying 1445

made 1405

went 1400

even 1393

behold 1326

saith 1262

therefore 1237

every 1236

things 1162

father 1111

sons 1090

hast 1070

david 1058

make 1056

say 1056

may 1027

earth 987

jesus 983

great 962

name 955

* 1. **list the top 50 bigrams by frequencies**

NLTK’s bigram collection finder was used to generate the top 50 bigram frequencies in each document. Similarly to the word frequency distribution, the results contained many non-alphabetic characters as well as stopwords, as seen below:

**Alice in Wonderland**

<class 'list'>

((',', 'and'), 0.013733803069206425)

((',', "'"), 0.012808264166716427)

(("'", 'said'), 0.009852510897474175)

(('!', "'"), 0.008449274496924822)

(('.', "'"), 0.00782229653072192)

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

(("'", 'i'), 0.005045679823251926)

(('?', "'"), 0.00468740669970741)

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

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

(("'", 'the'), 0.003224458111900639)

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

((',', 'i'), 0.002418343583925479)

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

(('alice', ','), 0.0023287753030393505)

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

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

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

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

((',', 'but'), 0.0018510778049799965)

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

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

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

(("'", 'she'), 0.0017913656177225771)

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

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

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

(('it', ','), 0.001731653430465158)

(('\*', '\*'), 0.0017017973368364484)

(("'", 'alice'), 0.0016719412432077386)

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

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

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

((',', 'you'), 0.0016122290559503194)

((';', 'and'), 0.0016122290559503194)

(('alice', '.'), 0.0016122290559503194)

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

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

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

(('--', "'"), 0.0015525168686929003)

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

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

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

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

(("'", "'"), 0.0014928046814354809)

(('.', 'the'), 0.0014928046814354809)

((',', 'that'), 0.0014629485878067713)

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

(('.', 'alice'), 0.0013733803069206425)

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

**KJV Bible**

<class 'list'>

((',', 'and'), 0.026345251221995495)

(('of', 'the'), 0.01218932586405749)

(('the', 'lord'), 0.007410129994127662)

(('and', 'the'), 0.006616941906101739)

(('in', 'the'), 0.005312564690772825)

((';', 'and'), 0.0033966616392694642)

((':', 'and'), 0.0031991567491751268)

((',', 'that'), 0.0031590220656265446)

(('and', 'he'), 0.002946730713172203)

((',', 'the'), 0.0026013611994778266)

(('shall', 'be'), 0.0025992488477121116)

(('to', 'the'), 0.002272890499909169)

(('all', 'the'), 0.002258104037549165)

(('and', 'they'), 0.0022031828916405792)

(('him', ','), 0.0021514302733805653)

(('unto', 'the'), 0.0021461493939662784)

(('i', 'will'), 0.0020225768156719604)

((',', 'which'), 0.0018937233579633549)

(('lord', ','), 0.0018050045838033317)

(('of', 'israel'), 0.0017902181214433277)

(('said', ','), 0.0017743754832004663)

(('for', 'the'), 0.001765926076137607)

(('said', 'unto'), 0.0017363531514175993)

((':', 'for'), 0.0017236790408233103)

(('the', 'king'), 0.001713117281994736)

((',', 'i'), 0.001700443171400447)

(('them', ','), 0.0016983308196347321)

(('son', 'of'), 0.0016909375884547303)

(('out', 'of'), 0.0015863761760518456)

(('the', 'son'), 0.0015842638242861307)

(('the', 'children'), 0.0014997697536575372)

(('children', 'of'), 0.0014680844771718146)

(('saying', ','), 0.0014057701000832266)

((',', 'saying'), 0.0013445118988774964)

(('the', 'land'), 0.00133078161240035)

(('and', 'i'), 0.0013255007329860628)

(('thou', 'shalt'), 0.0013202198535717756)

(('the', 'people'), 0.0012906469288517677)

(('thee', ','), 0.0012800851700231937)

(('me', ','), 0.0012631863558974748)

(('the', 'house'), 0.0012378381347088968)

(('from', 'the'), 0.0012219954964660354)

((',', 'he'), 0.0012167146170517485)

((',', 'to'), 0.0012019281546917446)

(('unto', 'him'), 0.0011913663958631703)

(('behold', ','), 0.0011850293405660257)

(('of', 'his'), 0.0011723552299717367)

(('i', 'have'), 0.0011206026117117232)

(('and', 'all'), 0.001107928501117434)

(('and', 'his'), 0.0011068723252345767)

To improve the results, non-alphabetic characters and stopwords were removed from the bigram list. Here are the updated results after additional preprocessing:

**Alice in Wonderland**

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

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

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

(('thought', 'alice'), 0.0007762584343464501)

(('white', 'rabbit'), 0.0006568340598316116)

(('alice', 'thought'), 0.00035827312354451543)

(("'of", 'course'), 0.0003284170299158058)

(('alice', 'said'), 0.0003284170299158058)

(('poor', 'alice'), 0.0003284170299158058)

(('alice', 'replied'), 0.00026870484265838657)

(('alice', 'looked'), 0.00023884874902967696)

(('king', 'said'), 0.00023884874902967696)

(('little', 'thing'), 0.00023884874902967696)

(('poor', 'little'), 0.00023884874902967696)

(('alice', 'began'), 0.00020899265540096732)

(('cried', 'alice'), 0.00020899265540096732)

(('good', 'deal'), 0.00020899265540096732)

(('oh', 'dear'), 0.00020899265540096732)

(('beautiful', 'soup'), 0.00017913656177225771)

(('golden', 'key'), 0.00017913656177225771)

(('great', 'hurry'), 0.00017913656177225771)

(('little', 'door'), 0.00017913656177225771)

(('said', 'nothing'), 0.00017913656177225771)

(('three', 'gardeners'), 0.00017913656177225771)

(('alice', 'felt'), 0.0001492804681435481)

(('alice', 'went'), 0.0001492804681435481)

(('another', 'moment'), 0.0001492804681435481)

(('came', 'upon'), 0.0001492804681435481)

(('cheshire', 'cat'), 0.0001492804681435481)

(('feet', 'high'), 0.0001492804681435481)

(('kid', 'gloves'), 0.0001492804681435481)

(('little', 'golden'), 0.0001492804681435481)

(('next', 'witness'), 0.0001492804681435481)

(('offended', 'tone'), 0.0001492804681435481)

(('play', 'croquet'), 0.0001492804681435481)

(('right', 'size'), 0.0001492804681435481)

(('trembling', 'voice'), 0.0001492804681435481)

(('white', 'kid'), 0.0001492804681435481)

(('alice', 'hastily'), 0.00011942437451483848)

(('alice', 'ventured'), 0.00011942437451483848)

(('come', 'back'), 0.00011942437451483848)

(('father', 'william'), 0.00011942437451483848)

(('hare', 'said'), 0.00011942437451483848)

(('inches', 'high'), 0.00011942437451483848)

(('lobster', 'quadrille'), 0.00011942437451483848)

(('low', 'voice'), 0.00011942437451483848)

(('never', 'heard'), 0.00011942437451483848)

(('old', 'fellow'), 0.00011942437451483848)

(('one', 'finger'), 0.00011942437451483848)

(('ootiful', 'soo'), 0.00011942437451483848)

**KJV Bible**

(('said', 'unto'), 0.0017363531514175993)

(('thou', 'shalt'), 0.0013202198535717756)

(('thou', 'hast'), 0.000811143078034499)

(('ye', 'shall'), 0.0007995251433230673)

(('lord', 'god'), 0.0005756158561572942)

(('unto', 'thee'), 0.0005291441173115676)

(('thus', 'saith'), 0.0004689420919886947)

(('say', 'unto'), 0.00042035800137725336)

(('shall', 'come'), 0.00041824564961153853)

(('thy', 'god'), 0.0003770547901800991)

(('thou', 'art'), 0.0003358639307486597)

(('every', 'one'), 0.0003326954031000875)

(('lord', 'thy'), 0.00032318982015437067)

(('every', 'man'), 0.0003094595336772242)

(('lord', 'hath'), 0.0003041786542629371)

(('spake', 'unto'), 0.00028622366425436093)

(('shalt', 'thou'), 0.00027249337777721447)

(('lord', 'said'), 0.00023341487011148992)

(('let', 'us'), 0.0002270778148143454)

(('unto', 'moses'), 0.0002165160559857712)

(('jesus', 'christ'), 0.00020912282480576926)

(('burnt', 'offering'), 0.00019433636244576538)

(('lord', 'shall'), 0.00019222401068005052)

(('came', 'unto'), 0.0001911678347971931)

(('saith', 'unto'), 0.0001911678347971931)

(('god', 'hath'), 0.000185886955382906)

(('thy', 'servant'), 0.00018483077950004858)

(('pray', 'thee'), 0.00018166225185147633)

(('right', 'hand'), 0.00017638137243718922)

(('shall', 'ye'), 0.0001742690206714744)

(('neither', 'shall'), 0.0001711004930229021)

(('hast', 'thou'), 0.00016793196537432986)

(('even', 'unto'), 0.0001594825583114705)

(('priest', 'shall'), 0.00015631403066289823)

(('thine', 'hand'), 0.0001552578547800408)

(('lord', 'spake'), 0.00015208932713146856)

(('come', 'unto'), 0.00014575227183432404)

(('years', 'old'), 0.00014575227183432404)

(('shall', 'go'), 0.00014363992006860918)

(('unto', 'us'), 0.0001394152165371795)

(('thy', 'father'), 0.00013730286477146465)

(('unto', 'thy'), 0.00013730286477146465)

(('thy', 'name'), 0.00013624668888860723)

(('thy', 'people'), 0.00013624668888860723)

(('speak', 'unto'), 0.00013307816124003498)

(('shall', 'eat'), 0.0001299096335914627)

(('meat', 'offering'), 0.0001288534577086053)

(('wilt', 'thou'), 0.00012674110594289046)

(('lord', 'came'), 0.00012568493006003304)

(('lord', 'jesus'), 0.00012462875417717563)

* 1. **list the top 50 bigrams by their Mutual Information scores (using min frequency 5)**

Pointwise Mutual Information (PMI) was used to generate pairs of words (bigrams) that are likely to show up together. Below are the results:

**Alice in Wonderland**

(("'any", 'shrimp'), 15.03161505883576)

(("'cheshire", 'puss'), 15.03161505883576)

(("'orange", 'marmalade'), 15.03161505883576)

(("'ou", 'est'), 15.03161505883576)

(("'rule", 'forty-two'), 15.03161505883576)

(("'seven", 'jogged'), 15.03161505883576)

(("'than", 'waste'), 15.03161505883576)

(("'with", 'extras'), 15.03161505883576)

(('abide', 'figures'), 15.03161505883576)

(('barking', 'hoarsely'), 15.03161505883576)

(('bathing', 'machines'), 15.03161505883576)

(('bright-eyed', 'terrier'), 15.03161505883576)

(('buttered', 'toast'), 15.03161505883576)

(('canvas', 'bag'), 15.03161505883576)

(('carroll', '1865'), 15.03161505883576)

(('crocodile', 'improve'), 15.03161505883576)

(('daresay', "it's"), 15.03161505883576)

(('deepest', 'contempt'), 15.03161505883576)

(('draggled', 'feathers'), 15.03161505883576)

(('edgar', 'atheling'), 15.03161505883576)

(('energetic', 'remedies'), 15.03161505883576)

(('exact', 'shape'), 15.03161505883576)

(('fair', 'warning'), 15.03161505883576)

(('feather', 'flock'), 15.03161505883576)

(('graceful', 'zigzag'), 15.03161505883576)

(('hundred', 'pounds'), 15.03161505883576)

(('immediate', 'adoption'), 15.03161505883576)

(('immense', 'length'), 15.03161505883576)

(('labelled', "'orange"), 15.03161505883576)

(('latin', 'grammar'), 15.03161505883576)

(('lewis', 'carroll'), 15.03161505883576)

(('lodging', 'houses'), 15.03161505883576)

(('meeting', 'adjourn'), 15.03161505883576)

(('muscular', 'strength'), 15.03161505883576)

(('never-ending', 'meal'), 15.03161505883576)

(('occasional', 'exclamation'), 15.03161505883576)

(('positively', 'refused'), 15.03161505883576)

(('prosecute', 'you.'), 15.03161505883576)

(('reasonable', 'pace'), 15.03161505883576)

(('red-hot', 'poker'), 15.03161505883576)

(('riper', 'years'), 15.03161505883576)

(('roast', 'turkey'), 15.03161505883576)

(('saucepan', 'flew'), 15.03161505883576)

(('shedding', 'gallons'), 15.03161505883576)

(('sneezed', 'occasionally'), 15.03161505883576)

(('splendidly', 'dressed'), 15.03161505883576)

(('thank', 'ye'), 15.03161505883576)

(('tide', 'rises'), 15.03161505883576)

(('tinkling', 'sheep-bells'), 15.03161505883576)

(('tittered', 'audibly'), 15.03161505883576)

**KJV Bible**

(('anathema', 'maranatha'), 19.852718465501717)

(('appii', 'forum'), 19.852718465501717)

(('ashteroth', 'karnaim'), 19.852718465501717)

(('barbed', 'irons'), 19.852718465501717)

(('changeable', 'suits'), 19.852718465501717)

(('dromedary', 'traversing'), 19.852718465501717)

(('equally', 'distant'), 19.852718465501717)

(('eubulus', 'greeteth'), 19.852718465501717)

(('fared', 'sumptuously'), 19.852718465501717)

(('grandmother', 'lois'), 19.852718465501717)

(("herod's", 'jurisdiction'), 19.852718465501717)

(('infallible', 'proofs'), 19.852718465501717)

(('je', 'hoshaphat'), 19.852718465501717)

(('monthly', 'prognosticators'), 19.852718465501717)

(('narrowed', 'rests'), 19.852718465501717)

(('sergius', 'paulus'), 19.852718465501717)

(('sticketh', 'closer'), 19.852718465501717)

(('talitha', 'cumi'), 19.852718465501717)

(('tottering', 'fence'), 19.852718465501717)

(('unequally', 'yoked'), 19.852718465501717)

(('warreth', 'entangleth'), 19.852718465501717)

(('119:9', 'wherewithal'), 18.852718465501717)

(('15:52', 'arab'), 18.852718465501717)

(('15:58', 'halhul'), 18.852718465501717)

(('15:60', 'kirjathbaal'), 18.852718465501717)

(('57:5', 'enflaming'), 18.852718465501717)

(('83:7', 'gebal'), 18.852718465501717)

(('83:8', 'assur'), 18.852718465501717)

(('bloomed', 'blossoms'), 18.852718465501717)

(('dissolve', 'doubts'), 18.852718465501717)

(('doubtful', 'disputations'), 18.852718465501717)

(('exile', 'hasteneth'), 18.852718465501717)

(('harpers', 'harping'), 18.852718465501717)

(('heinous', 'crime'), 18.852718465501717)

(('homeborn', 'slave'), 18.852718465501717)

(('inhabiteth', 'eternity'), 18.852718465501717)

(('lama', 'sabachthani'), 18.852718465501717)

(('prosperously', 'effected'), 18.852718465501717)

(('shameful', 'spewing'), 18.852718465501717)

(('speedy', 'riddance'), 18.852718465501717)

(('well-beloved', 'epaenetus'), 18.852718465501717)

(('48:41', 'kerioth'), 18.267755964780562)

(('68:34', 'ascribe'), 18.267755964780562)

(('bible', ']'), 18.267755964780562)

(('chooseth', 'strangling'), 18.267755964780562)

(('confidently', 'affirmed'), 18.267755964780562)

(('damnable', 'heresies'), 18.267755964780562)

(('fallow', 'deer'), 18.267755964780562)

(('imperious', 'whorish'), 18.267755964780562)

(('impose', 'toll'), 18.267755964780562)

Finally, a minimum frequency filter of 5 was applied to the PMI analysis. The results can be seen below:

**Alice in Wonderland**

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

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

(('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)

**KJV Bible**

(('untempered', 'morter'), 16.39328684686442)

(('cock', 'crew'), 16.26775596478056)

(('gray', 'hairs'), 15.945827869893197)

(('cock', 'crow'), 15.782329137610319)

(('filthy', 'lucre'), 15.502221218417583)

(('skins', 'dyed'), 14.782329137610319)

(('ill', 'favoured'), 14.72343544855675)

(('judas', 'iscariot'), 14.518817728948278)

(('curious', 'girdle'), 14.28286285717077)

(('brook', 'kidron'), 14.277809629444484)

(('measuring', 'reed'), 14.156247649563568)

(('divers', 'colours'), 14.058302599151611)

(('mary', 'magdalene'), 13.972300081254389)

(('wreathen', 'chains'), 13.643265099872766)

(('dyed', 'red'), 13.639371183768276)

(('fiery', 'furnace'), 13.623899775005835)

(('committeth', 'adultery'), 13.604790952058131)

(('earthen', 'vessel'), 13.592190915278495)

(('golden', 'spoon'), 13.545289940309468)

(('bright', 'spot'), 13.520806282041733)

(('tenth', 'deals'), 13.512868462617092)

(('bottomless', 'pit'), 13.393286846864417)

(('familiar', 'spirits'), 13.329156509444704)

(('solemn', 'feasts'), 13.316665565261506)

(('walketh', 'uprightly'), 13.232132055049838)

(('linen', 'breeches'), 13.152278747360624)

(('twined', 'linen'), 13.152278747360624)

(('straitly', 'charged'), 13.042789599780285)

(('dearly', 'beloved'), 13.032539503086529)

(('wave', 'breast'), 13.004721558946766)

(('flour', 'mingled'), 12.968265259012952)

(('fine', 'twined'), 12.96215017519513)

(('deal', 'kindly'), 12.945827869893197)

(('sweet', 'savour'), 12.769208215876876)

(('dealt', 'treacherously'), 12.666191496722274)

(('burning', 'fiery'), 12.623899775005835)

(('new', 'moons'), 12.614313726176636)

(('parched', 'corn'), 12.595330622809064)

(('green', 'herb'), 12.56916704232741)

(('tender', 'mercies'), 12.567316246639468)

(('solemn', 'assembly'), 12.549952627701249)

(('shittim', 'wood'), 12.478322950720216)

(('unleavened', 'cakes'), 12.448049939606419)

(('due', 'season'), 12.4130953279446)

(('must', 'needs'), 12.393286846864417)

(('commit', 'adultery'), 12.389434521368813)

(('fine', 'flour'), 12.344483509401268)

(('summer', 'fruits'), 12.290476041280643)

(('sing', 'praises'), 12.269844708508513)

(('nether', 'parts'), 12.26775596478056)

**You may wish to modify the stop word list, based on your question in Task 3. To complete this part:**

**a) Briefly state why you chose the processing options that you did.**

I chose to perform additional processing options in the form of removing stopwords and non-alphabetic characters to the word frequency distribution and bigram analysis. I did this because punctuation marks and common words such as “of”, “the”, “and”, etc., failed to return meaningful results in the output. By removing non-alphabetic characters and stopwords, the language model achieved results that were more meaningful to understanding the text.

**b) Are there any problems with the word or bigram lists that you found? Could you get a better list of bigrams?**

Before additional text preprocessing steps were performed, the word and bigram lists did not return many useful results, as the top results were tokens such as “,”, “and”, and “of”, along with common bigrams such as “to the”.

**c) How are the top 50 bigrams by frequency different from the top 50 bigrams scored by Mutual Information?**

The top 50 bigrams returned pairs of words that frequently occurred together, but the PMI analysis went a step further and showed words that were likely to occur together as a matter of association. This level of analysis yielded bigrams that are uniquely characteristic of the text, such as “march hare” in Alice in Wonderland, and “fiery furnace” in the KJV Bible.

**d) If you modify the stop word list, or expand the methods of filtering, describe that here.**

I added additional words to the stopword list, including parts of contractions that may be counted as tokens in words like “wouldn’t” or “hasn’t”. For the PMI analysis, I also applied a minimum frequency filter of 5 so that the model returns only bigrams that occur 5 or more times. Applying this minimum frequency filter prevented sets of numbers from showing up in the PMI bigram results of the KJV Bible, which may not be particularly useful when analyzing the text.

**e) You may choose to also run top trigram lists, and include them in the analysis in part 3.**

1. Describe a problem or question that is based on the difference between the two documents. In the case of literary works, for example, this could be how to characterize the style between two authors or two works of different classes. Another example would be to compare the informal text in blogs with more formal text. Or you can do a topic-related comparison that selects words (as in the SOTU speeches example). You could also make a comparison of similar text but at two different times.

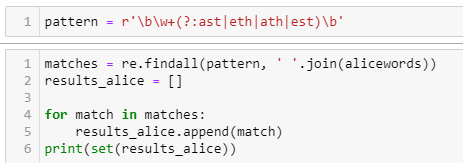
The writing styles of Alice in Wonderland and the KJV Bible could be examined to classify them into one or more categories. One method proposed here is to examine suffixes ending in ‘est’, ‘eth’, ‘ath’, and ‘ast’. By using REGEX to examine the words found in each document and looking at unique occurrences of words containing these suffixes, a particular writing style may emerge. The hypothesis is defined below:

***Hypothesis****: One aspect of the writing styles of Alice in Wonderland and the KJV Bible is words ending in 'eth', 'est', 'ath', and 'eth'. Alice in Wonderland, being a more contemporary piece of literature, will contain more common words ending in these suffixes such as "best", "smallest", or "last", while the KJV Bible will contain more archaic terms such as "ruleth", "hast", and "lovest".*

**Methodology**

A REGEX pattern was developed to identify all words in each document ending in one of the aforementioned suffixes. The results were stored into a list and then cast into a set to display only unique instances of these words. See Figure 3.

**Figure 3**



**Results**

The hypothesis was not disproven, as the results show the analysis of suffixes reveals a difference in writing style between the two documents. Moreover, the length of the lists is quite different, as Alice in Wonderland had much shorter results compared to the KJV Bible. Alice in Wonderland contains words that are common to contemporary English, such as “honest”, “last”, and “best”, while the KJV Bible contains many archaic and outdated words such as “withholdeth”, “boasteth”, and “transgressest”. Additionally, the KJV Bible contains many non-English words such as “tikvath”, “tabbath”, and “ashtoreth”. These differences are to be expected given the fact that these documents were written at different times in history by different authors and follow drastically different writing conventions.

**Alice in Wonderland**

{'last', 'breath', 'best', 'twentieth', 'coast', 'least', 'conquest', 'driest', 'death', 'lest', 'nest', 'honest', 'slightest', 'highest', 'roast', 'stupidest', 'rest', 'toast', 'largest', 'deepest', 'fast', 'queerest', 'oldest', 'loveliest', 'interest', 'teeth', 'past', 'beast', 'smallest', 'underneath'}

**KJV Bible**

{'harosheth', 'afflictest', 'withholdeth', 'shaketh', 'counteth', 'dasheth', 'jerubbesheth', 'mayest', 'reproveth', 'wanderest', 'mustereth', 'committeth', 'scarest', 'hindereth', 'twentieth', 'rejoiceth', 'holdeth', 'callest', 'weakeneth', 'shameth', 'consulteth', 'riphath', 'tikvath', 'happeneth', 'drewest', 'selleth', 'findeth', 'watereth', 'possessest', 'rescueth', 'boasteth', 'pleaseth', 'spoileth', 'tilleth', 'meetest', 'requireth', 'provideth', 'bewrayeth', 'settlest', 'satest', 'scaleth', 'repentest', 'appertaineth', 'dishonourest', 'partakest', 'lackest', 'bringeth', 'slumbereth', 'eldest', 'turneth', 'readest', 'fadeth', 'earneth', 'baptizest', 'croucheth', 'tabbath', 'searchest', 'writest', 'delayeth', 'burneth', 'rewardeth', 'saluteth', 'feareth', 'meddleth', 'prevaileth', 'upholdest', 'gathereth', 'shihorlibnath', 'striveth', 'believest', 'broughtest', 'sewest', 'restest', 'causeth', 'walketh', 'markest', 'trembleth', 'delivereth', 'grieveth', 'sufferest', 'paweth', 'chinnereth', 'refuseth', 'understandest', 'ordaineth', 'desirest', 'ruleth', 'executeth', 'drinketh', 'payeth', 'appeareth', 'prevailest', 'fretteth', 'travailest', 'ashtoreth', 'tanhumeth', 'shineth', 'approvest', 'compasseth', 'liest', 'worketh', 'confirmeth', 'buildest', 'trieth', 'revilest', 'oppresseth', 'leddest', 'feignest', 'aiath', 'taketh', 'blindeth', 'warmeth', 'coveteth', 'acknowledgeth', 'ascendeth', 'deviseth', 'purgeth', 'flieth', 'dishonest', 'readeth', 'requirest', 'smiteth', 'hast', 'request', 'exceedest', 'lowest', 'sittest', 'transgressest', 'yieldeth', 'badest', 'sticketh', 'standeth', 'brakest', 'plucketh', 'slayeth', 'owest', 'honoureth', 'overpast', 'persuadest', 'thoughtest', 'cumbereth', 'lovest', 'fastest', 'sendeth', 'sinnest', 'openeth', 'prayeth', 'constraineth', 'lightest', 'adorneth', 'defileth', 'respecteth', 'intendest', 'studieth', 'lieth', 'satisfieth', 'trusteth', 'mephibosheth', 'upholdeth', 'sheath', 'teachest', 'stablisheth', 'west', 'sleepeth', 'shouteth', 'slewest', 'detest', 'performeth', 'diggeth', 'greeteth', 'cleanseth', 'bashemath', 'panteth', 'striketh', 'abhorreth', 'chastiseth', 'mephaath', 'knoweth', 'wipeth', 'delightest', 'publisheth', 'leth', 'concerneth', 'subdueth', 'heardest', 'succeedest', 'fast', 'hearest', 'passest', 'soundeth', 'filleth', 'committest', 'becamest', 'clappeth', 'renderest', 'leadest', 'pertaineth', 'troublest', 'runneth', 'path', 'sibboleth', 'preservest', 'revengeth', 'corrupteth', 'asenath', 'forsaketh', 'hidest', 'last', 'ishbosheth', 'putteth', 'reigneth', 'staggereth', 'whirleth', 'dippeth', 'hemath', 'faileth', 'sawest', 'rebellest', 'mourneth', 'provoketh', 'halteth', 'fluttereth', 'repeateth', 'increaseth', 'establisheth', 'forgettest', 'hatcheth', 'careth', 'availeth', 'goest', 'wanteth', 'betrayest', 'coast', 'thresheth', 'sentest', 'maarath', 'purifieth', 'fellest', 'upbraideth', 'treasurest', 'girdeth', 'anath', 'goath', 'compellest', 'falleth', 'enticeth', 'plotteth', 'coverest', 'shewest', 'changest', 'boweth', 'zoheleth', 'slideth', 'beatest', 'reproacheth', 'commandest', 'forest', 'barest', 'pleadeth', 'prepareth', 'escapeth', 'wotteth', 'praiseth', 'fallest', 'draweth', 'swalloweth', 'runnest', 'testifieth', 'waterest', 'waketh', 'dealest', 'gendereth', 'stumbleth', 'marketh', 'stilleth', 'cheweth', 'winketh', 'chooseth', 'forsookest', 'rebuketh', 'coucheth', 'dureth', 'sufficeth', 'nameth', 'protest', 'increasest', 'meaneth', 'rageth', 'prophesieth', 'sweareth', 'beateth', 'needest', 'rusheth', 'alameth', 'talketh', 'feedeth', 'winneth', 'loweth', 'stoopeth', 'tellest', 'mocketh', 'disannulleth', 'threwest', 'wast', 'excellest', 'satisfiest', 'marrieth', 'accuseth', 'contendeth', 'fitteth', 'reserveth', 'cheereth', 'correcteth', 'heath', 'travaileth', 'perceivest', 'resisteth', 'forbiddeth', 'camest', 'boastest', 'spendeth', 'consumeth', 'woundeth', 'breatheth', 'biddeth', 'speakest', 'baalath', 'waiteth', 'speaketh', 'causest', 'obeyeth', 'cast', 'glorifieth', 'swimmeth', 'dreameth', 'overcometh', 'seest', 'remembereth', 'askest', 'remeth', 'dresseth', 'hammoleketh', 'inhabiteth', 'foreseeth', 'changeth', 'breast', 'hath', 'rabbath', 'deserveth', 'maintainest', 'calveth', 'sheddeth', 'getteth', 'escheweth', 'kindleth', 'scornest', 'procureth', 'obtaineth', 'defendest', 'revealeth', 'overfloweth', 'observest', 'sorroweth', 'rentest', 'describeth', 'choicest', 'decayeth', 'condemneth', 'loveth', 'stretcheth', 'travelleth', 'beholdeth', 'doubteth', 'formeth', 'voweth', 'healeth', 'withdraweth', 'blast', 'standest', 'beneath', 'devoureth', 'floweth', 'sowest', 'preventest', 'scourgeth', 'overturneth', 'sophereth', 'stedfast', 'criest', 'oughtest', 'triest', 'addeth', 'clothest', 'tillest', 'understandeth', 'straitest', 'bechorath', 'holdest', 'erreth', 'fleeth', 'hammath', 'crowneth', 'exhorteth', 'asketh', 'covereth', 'seirath', 'shutteth', 'chaseth', 'priest', 'encampeth', 'containeth', 'exalteth', 'believeth', 'fleddest', 'breaketh', 'bendeth', 'heapeth', 'slandereth', 'lighteneth', 'recompensest', 'cleaveth', 'meshullemeth', 'uttereth', 'entangleth', 'gaddest', 'withheldest', 'seeth', 'laboureth', 'tempest', 'slanderest', 'opposeth', 'excelleth', 'hazarmaveth', 'withereth', 'parteth', 'despiseth', 'modest', 'frameth', 'edifieth', 'calleth', 'lifteth', 'curseth', 'singeth', 'stealeth', 'reignest', 'sojourneth', 'buyest', 'rest', 'timnath', 'stayeth', 'gropeth', 'deckest', 'servest', 'bath', 'belongest', 'directeth', 'eightieth', 'madest', 'guest', 'tempteth', 'planteth', 'licketh', 'blotteth', 'jehoshabeath', 'fighteth', 'gnasheth', 'lettest', 'converteth', 'warreth', 'gath', 'hideth', 'helkath', 'liketh', 'knewest', 'descendeth', 'comest', 'settest', 'fewest', 'greatest', 'slippeth', 'steppeth', 'hareth', 'fortieth', 'lest', 'bringest', 'aileth', 'swarest', 'crownest', 'usest', 'profiteth', 'diblath', 'preachest', 'weavest', 'deferreth', 'soweth', 'exacteth', 'nest', 'mightiest', 'liveth', 'weepeth', 'roast', 'layest', 'commendeth', 'fearest', 'hearkeneth', 'doest', 'goodliest', 'suffereth', 'rolleth', 'coupleth', 'heth', 'enrichest', 'shimrath', 'preacheth', 'sharpeneth', 'rulest', 'tendeth', 'deliverest', 'spendest', 'loadeth', 'helpeth', 'approveth', 'perceiveth', 'vilest', 'pondereth', 'humbleth', 'overthroweth', 'earnest', 'japheth', 'stoppeth', 'washest', 'preserveth', 'seeketh', 'rendereth', 'reckoneth', 'jahath', 'hurleth', 'endeth', 'lacketh', 'riseth', 'groaneth', 'foundest', 'hardeneth', 'proclaimeth', 'spreadest', 'anointest', 'stinketh', 'foameth', 'shimeath', 'springeth', 'rejecteth', 'dabbasheth', 'biteth', 'pibeseth', 'youngest', 'willeth', 'groweth', 'bereaveth', 'keepest', 'handleth', 'answerest', 'wreath', 'inheriteth', 'justifieth', 'scattereth', 'saveth', 'dwellest', 'ahuzzath', 'sanctifieth', 'roareth', 'forgavest', 'looseth', 'choosest', 'behaveth', 'numberest', 'awakest', 'belongeth', 'returneth', 'smootheth', 'nourisheth', 'divineth', 'forgetteth', 'closest', 'blasphemest', 'troubleth', 'leaneth', 'oweth', 'sigheth', 'peleth', 'receiveth', 'tasteth', 'sayest', 'vanisheth', 'killeth', 'observeth', 'proveth', 'enlargeth', 'ministereth', 'beast', 'languisheth', 'retaineth', 'genubath', 'underneath', 'breakest', 'carest', 'elisabeth', 'tibhath', 'baptizeth', 'borroweth', 'thirsteth', 'bewaileth', 'separateth', 'contendest', 'sitteth', 'shouldest', 'sheweth', 'naarath', 'bindeth', 'presseth', 'wearieth', 'owneth', 'judgest', 'lodgeth', 'restrainest', 'wrest', 'confesseth', 'compoundeth', 'hathath', 'tahath', 'chastenest', 'maath', 'needeth', 'mispereth', 'departeth', 'blessest', 'visitest', 'backbiteth', 'feedest', 'goeth', 'turnest', 'fainteth', 'ordereth', 'reapeth', 'enviest', 'elath', 'offereth', 'straiteneth', 'basmath', 'kohath', 'loatheth', 'appeaseth', 'bloweth', 'girdest', 'zephath', 'multiplieth', 'accepteth', 'valuest', 'thinkest', 'alemeth', 'seweth', 'ploweth', 'makest', 'stillest', 'prolongeth', 'sheth', 'openest', 'eatest', 'leaveth', 'affecteth', 'delighteth', 'hateth', 'pineth', 'leaveneth', 'advantageth', 'lusteth', 'entreateth', 'desireth', 'taphath', 'rattleth', 'forbeareth', 'roasteth', 'seekest', 'forgiveth', 'buildeth', 'foldeth', 'lendeth', 'betrayeth', 'disappointeth', 'reproachest', 'listeth', 'gloriest', 'setteth', 'wasteth', 'mithredath', 'manahath', 'telleth', 'kirhareseth', 'hopeth', 'thinketh', 'couldest', 'honourest', 'snuffeth', 'basest', 'flattereth', 'redeemeth', 'pardoneth', 'opposest', 'graveth', 'hamath', 'watcheth', 'begetteth', 'strongest', 'spreadeth', 'hasteth', 'nazareth', 'death', 'wrath', 'destroyest', 'regardeth', 'witnesseth', 'frustrateth', 'stoodest', 'mightest', 'remaineth', 'writeth', 'fiftieth', 'fattest', 'zoheth', 'craveth', 'persuadeth', 'east', 'risest', 'knowest', 'favourest', 'envieth', 'zarephath', 'wouldest', 'anaharath', 'compassest', 'sleepest', 'seth', 'proceedeth', 'passeth', 'smallest', 'boast', 'repayeth', 'poorest', 'alloweth', 'tookest', 'spareth', 'appointeth', 'chargest', 'pursueth', 'walkest', 'convinceth', 'considerest', 'persecutest', 'trustest', 'endureth', 'differeth', 'possesseth', 'ramath', 'transgresseth', 'letteth', 'poureth', 'wakeneth', 'weareth', 'castest', 'layeth', 'altereth', 'flourisheth', 'condemnest', 'resteth', 'repliest', 'deceiveth', 'oath', 'devourest', 'worshippeth', 'quickeneth', 'aboundeth', 'catcheth', 'fillest', 'searcheth', 'ladeth', 'esteemeth', 'decketh', 'cuttest', 'azmaveth', 'agreeth', 'exerciseth', 'renewest', 'vomiteth', 'treadeth', 'continueth', 'ceaseth', 'favoureth', 'perisheth', 'mockest', 'savourest', 'dealeth', 'teeth', 'moveth', 'refresheth', 'savest', 'longeth', 'disguiseth', 'refraineth', 'outcast', 'benzoheth', 'chanceth', 'breath', 'shibboleth', 'baketh', 'chest', 'waxeth', 'best', 'comforteth', 'gibeath', 'lodgest', 'puttest', 'laugheth', 'casteth', 'lotheth', 'sabbath', 'abideth', 'begettest', 'scorneth', 'daberath', 'weepest', 'preparest', 'followeth', 'divideth', 'hunteth', 'createth', 'extendeth', 'leadeth', 'bozkath', 'overtaketh', 'contemneth', 'honest', 'talkest', 'seduceth', 'maketh', 'meanest', 'buyeth', 'givest', 'considereth', 'terrifiest', 'profaneth', 'serveth', 'regardest', 'wandereth', 'melteth', 'entereth', 'lingereth', 'manifest', 'sprinkleth', 'cometh', 'puffeth', 'forecast', 'pierceth', 'removeth', 'lappeth', 'mast', 'spakest', 'valiantest', 'awaketh', 'vaunteth', 'fashioneth', 'occupieth', 'blasphemeth', 'destroyeth', 'killest', 'feast', 'reapest', 'trimmest', 'intermeddleth', 'playeth', 'tarrieth', 'smelleth', 'crieth', 'hirest', 'droppeth', 'beginnest', 'executest', 'giveth', 'sacrificeth', 'rememberest', 'despisest', 'boscath', 'wentest', 'ephrath', 'goliath', 'ariseth', 'leftest', 'consisteth', 'chiefest', 'stirreth', 'challengeth', 'kirharaseth', 'bethazmaveth', 'hasteneth', 'remainest', 'useth', 'winnoweth', 'tarriest', 'limiteth', 'bridleth', 'fairest', 'blesseth', 'gatherest', 'darkeneth', 'chasteneth', 'kattath', 'dishonoureth', 'driveth', 'kenath', 'weigheth', 'avengeth', 'pochereth', 'acceptest', 'commandeth', 'visiteth', 'smotest', 'bethanath', 'smitest', 'hangeth', 'hottest', 'drieth', 'enquirest', 'rejoicest', 'sinneth', 'meeteth', 'spoilest', 'keepeth', 'discerneth', 'diddest', 'shooteth', 'glorieth', 'uncovereth', 'imputeth', 'relieveth', 'hatest', 'mahalath', 'perverteth', 'thirtieth', 'sighest', 'toucheth', 'imagineth', 'concealeth', 'vowest', 'wavereth', 'zereth', 'sellest', 'huntest', 'knocketh', 'ashvath', 'cherisheth', 'mahath', 'looketh', 'purposeth', 'highest', 'repenteth', 'teacheth', 'befalleth', 'wrongeth', 'climbeth', 'rakkath', 'becometh', 'trickleth', 'takest', 'signifieth', 'judgeth', 'pisseth', 'pitieth', 'stingeth', 'distributeth', 'carrieth', 'raiseth', 'past', 'sealest', 'stonest', 'sendest', 'faintest', 'lookest', 'denieth', 'exceedeth', 'wroughtest', 'shallecheth', 'liftest', 'stretchest', 'livest', 'strengtheneth', 'zererath', 'dieth', 'prospereth', 'dissembleth', 'loseth', 'nahath', 'holiest', 'fetcheth', 'prayest', 'ginath', 'seemeth', 'supplieth', 'thrusteth', 'thundereth', 'topheth', 'beholdest', 'least', 'restoreth', 'exaltest', 'lighteth', 'reacheth', 'dissolvest', 'withdrawest', 'gavest', 'sealeth', 'heareth', 'cursest', 'cutteth', 'answereth', 'creepeth', 'findest', 'discovereth', 'harvest', 'bearest', 'declareth', 'moreshethgath', 'diest', 'teareth', 'kirjath', 'declineth', 'inhabitest', 'inclineth', 'pacifieth', 'tebeth', 'robbeth', 'abhorrest', 'rideth', 'dwelleth', 'emboldeneth', 'beareth', 'doeth', 'quieteth', 'eateth', 'jotbath', 'approacheth', 'carriest', 'abodest', 'swimmest', 'heweth', 'finest', 'jetheth'}