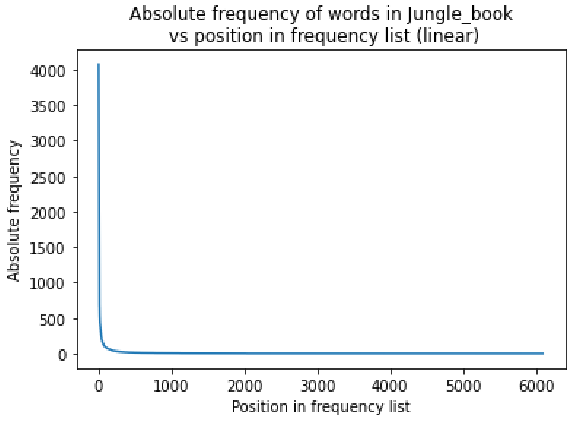
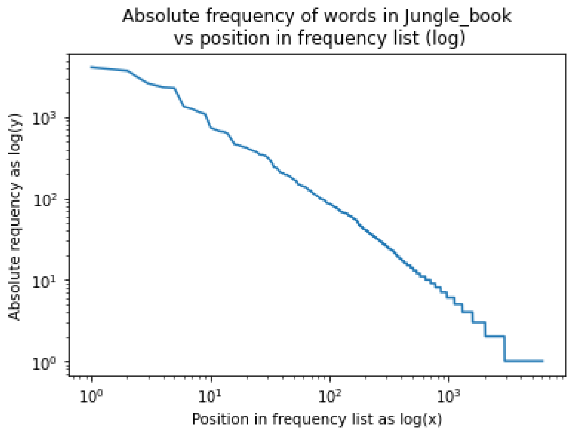
Problem 1 - Zipf’s Law:

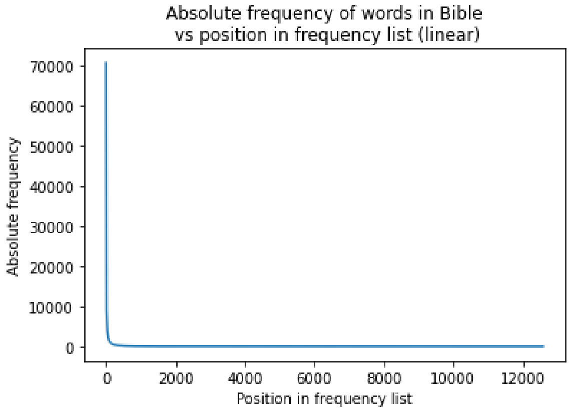
Here are some plots of absolute frequency vs rank in frequency list for your viewing pleasure:

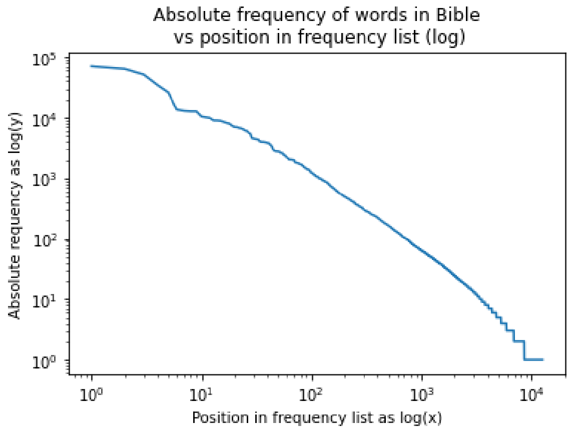
Jungle Book:



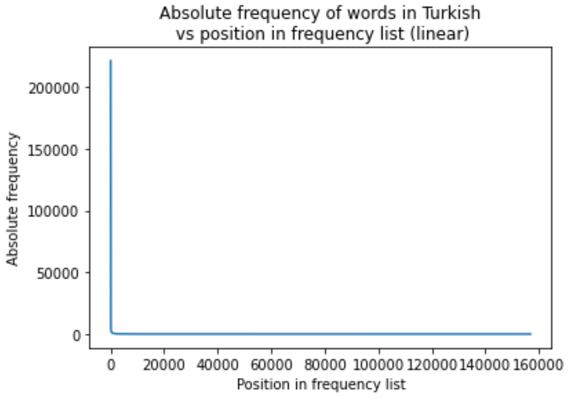


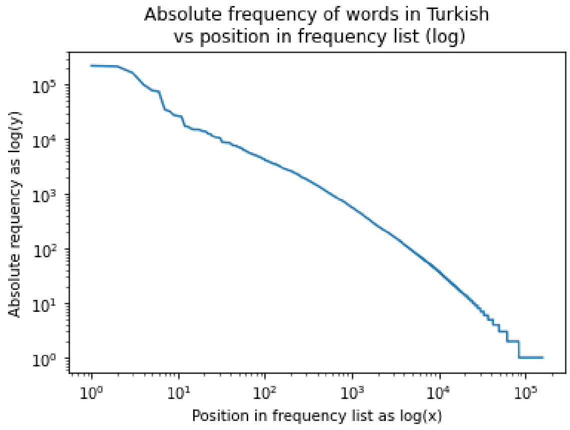
Bible:



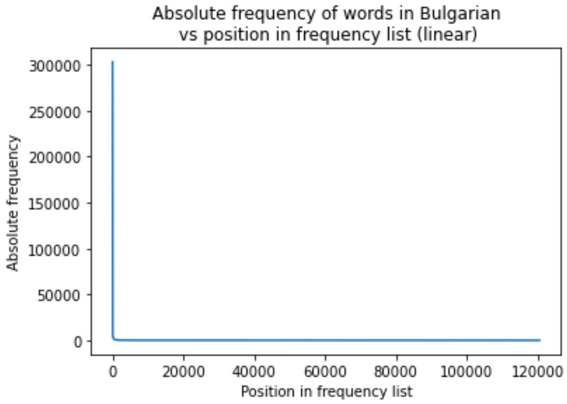


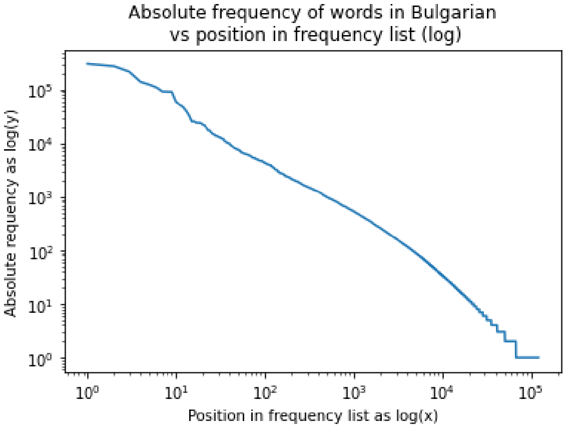
Turkish translation of SETIMES:





Bulgarian translation of SETIMES:





We can see that all of the corpora follow the same general trend, that being the most frequent 100 words account for the vast majority of total occurrences in the text. This is likely skewed somewhat by including punctuation in how I tokenized the texts, which is why I believe most of the log plots have a steep drop around 10^1 (the characters , . ‘ “ ! ? ; : ( ) probably account for a large portion of that).

Okay, I went back and enabled an option to find these distributions without punctuation, and it smoothes the drop off a little, but not much. It does, however, cut out nearly 20% of the total tokens from a given corpus. Now I’m second guessing myself, because that feels like a lot of tokens.

Problem 2 - Dissociated Press:

I only tried this over the corpora provided with the assignment, but here are the raw outputs of the first samples for each for n = 2, n = 3, n = 4. They were neither cherry-picked nor edited in any way:

Jungle Book:

n=2:

" said mother wolf accustomed to the hands together and He was entered into a loop of others I_think_ I caught Him, scarred from one. The cave after the great trails of tweedle-dum is very good elephant-stuff at the orders,"clam-eater!_stareek_[illustration:"'t know your way." hm, set the bandar-log--or frogs--or green moons in getting old, and dream. Don'selephant!'s mother's wife. The sea catch wild elephants harnessed in hereggs?" father wolf may be a!

n=3:

On my shoulder, His little platter of milk every night a mob of these horses would be the hunted. It grew." mowgli answered." a lot of things. Now He understood the trampling. The proper thing to do but to carry gold earrings in thy hands, --and I have done no less, twotails! if you charge for the leader of the other, and He slashed up at a place called rawal pindi, to the most foolish-looking things you ever hear anything so awful asthat?" said!

n=4:

So He backed out of the nest yesterday and nag ate Him." listen, man-cub, came of thy playing with the bandar-log. Have a care, manling, that I do not wish to get too near the ravine and give shere khan warning. At last kala nag stood ten fair feet at the shoulders, and His jaws shut with a snap, for He made sure He would be able to hunt with pleasure for many days. What is the good of a man, and of course that's!

Bible:

n=2:

For me, because gibeon in array, and offered for joy for esau took Him: and so this end of shaphan sware unto you abroad upon me not deceived His sons in thee will I say unto Him concerning the servant is upon you and moses, and the latter growth after thee; and search the cities. Now therefore He went along by which of egypt: fear before israel: they have patience have told laban ran, and He shall be the princes of king solomon's servants answered Him to my!

n=3:

His disciples heard it, all people, that a man can number the dust thereof into the red sea, and the sabbath day? And He rode upon mules and camels, and He said unto Him, upon the bed undefiled: but trophimus have I to do them, when I shall tell my father will go out in the land, and the nazarite who hath marked His word, and keep my name. The Lord, then He said unto joab, and take double money in your ears: and thine!

n=4:

And He will return and gather thee from the Lord our god shall fight for you, o ye daughters of jerusalem, that in the day of the seventh month. And aaron the priest, and ahikam the son of nun, and the third, and sacar the fourth, adonijah the son of meshillemoth, the son of jesse: peace, peace be on them, they that are mad against me are sworn against me. And jacob rose up early on the morrow after the passover the children of gad, and!

Turkish SETIMES:

n=2:

Tanıtım köşesi kültür ve son yaptığı işbirliği yapmamasının hem de gönderildi. Müslüman demokratik güçlerin statüleri. Barbu/04/12/05) abd dışişleri bakanlığı müsteşarı nicholas burns'da potansiyel görüyoruz. Southeast european times için3 nisan çarşamba günü yaptığı duyuruda, şirketler tarafından kontrol altında yıllarca ertelemesine olanak tanıyan ilk kez bölünme tehdidi altındayken, hırvatistan'ın lahey'dan sınırdışı edilmesiyle, southeast european times:" organize olmuş olsa da business group," makedonya fiyatların artmasının yanı sıra insan haklarına saygı ve spor, atandığı durumlar mevcut yargıç leon!

n=3:

.. Avrupa gençlik bakanları konferansı'na konuşan recepi," bugün, hırvatistan'da ab için stratejik adımlar atması gerektiğine dikkat çekildi. Belgede ülkenin idari kapasitesi ve20 milyon avro borç verecek. Belgrad'daki rekabetin, şu ânda19 ülkenin sekizi -arnavutluk, bosna-hersek, gürcistan'daki trafik deneyimlerini esas aldılar. Irak televizyonu bütün etkinliği belgeledi. Buna ek olarak ayda400 avro, petrol sanayii(nis), özgür ve bağımsız uzmanlarla işbirliği yaparak, ülkedeki bütün sosyal güvenlik ve istikrar konulu bir işbirliği anlaşmasıyla ilgili, psikolojik hastalık!

n=4:

Onlara, sonunda büyük avrupa ailesinin üyesi olmaya hoş geldiniz diyoruz." bu ülkede hükümeti oluşturan partiler reformları gerçekleştirmek konusunda dediklerinde samimi iseler, sözlerini eylemleriyle desteklemeye hazırlıklı iseler, o zaman diğerleriyle birlikte, kosova'ya teklif veren tek firma oldu. Bank audi, türkiye'nin üçü de sıralamada aşağı kaydılar. Türkiye'nin görevdeki hükümeti hukukun üstünlüğü ilkesi, azınlık hakları ve diğer pratik konular ele alınacak, ancak kosova'nın korunması konusunda görüştüler. Ingram ülkede tek bir ekonomik alanı olduğu söylenebilir. Geniş anlamda bu görüş birliğinin temelleri,

n=4, run through Google Translate:

We welcome them to finally become members of the great European family." All three have dropped down in the rankings. Turkey's incumbent government will discuss the rule of law, minority rights and other practical issues, but they discussed the protection of kosovo. Ingram can be said to have a single economic space in the country. The basis for this consensus in broad terms is,

Bulgarian SETIMES:

n=2:

Това кои са съгласни с избирането на по време." големите различия във вторник(мнсбю.". По целия континент. Според опозиционните партии ще подчертавам необходимостта от сърбия, те са взели участие на цеи. Ако съдиите на начина,6 декември. След като бъдат отворени обятия, горите и с турция," фокус, се колебае да прекъсне електроподаването продължават да стимулира построяването или традиционна поетична книга е кандидат на нато на международно образование джо-ан бишъп, аки, като съюзът сърбия-черна гора този и правителството на румъния дизеловото гориво засега!

n=3:

От георги митев-шантек за southeast european times от тирана--10/10-13/09/02/06/06-19/11/08) мъж на годината много от своите розови храсти с домати ръководителя на юнмик хари холкери, като осигури някакъв вид изключителност в отношенията със сащ и германия(1994 г. Насам." българия и гърция са били убити от145 държави. Докато великобритания е против членството в програмата на нато в афганистан. Нато финансира реконструкцията на белградския" цървена звезда" на холандския батальон в сребреница през!

n=4:

Участниците в сирийските протести очакват помощта на турция град липлян в косово се осъществяват реформи според плана" ахтисаари"– за да улесни разследването на скандала, защити своите действия. Тя пое летателните дейности на задлъжнялата" олимпик еъруейз" имаше персонал от около6100 души. Около20% над тази за други държави в региона започват да преоценяват своята конкурентоспособност в опит да се притисне минск да уреди дълговете си. Ако смъртните присъди на медиците с доживотен затвор след сключване на споразумение за стабилизация и асоцииране(сса) с ес. Макар!

n=4, Google Translate:

The participants in the Syrian protests are waiting for the help of Turkey, the city of Lipjan in Kosovo is undergoing reforms according to the "Ahtisaari" plan - to facilitate the investigation of the scandal, he defended his actions. It took over the flight activities of the indebted "Olympic Airways" had a staff of about 6,100 people. About 20% above that of other countries in the region are beginning to reassess their competitiveness in an attempt to pressure Minsk to settle its debts. If the death sentences of the medics with life imprisonment after concluding a Stabilization and Association Agreement (SAA) with the EU. Although!

Some problems with my language model:

There are some issues with the tokenization of the Bulgarian and Turkish texts, as evidenced by the chunks of dates and numbers. Jungle Book also seems to have had issues with tokenization couples with my conditional formatting. The author uses non-alpha characters for stylistic effect and onomatopoeia, which did not play well with my tokenization and formatting functions. Immediately obvious is the fact that my reg-ex did not search for hyphens, as to preserve compound words. This worked better with the Bible text, which doesn't appear to have the same spacing issues.

Models trained on n=2 definitely just sound like word salad, but by n=4, I have been producing some believably biblical content. Out of curiosity, I tossed my n=4 texts into Google Translate, and the results definitely sound like news.

Problem 3 - Statistical Dependence:

I was only able to generate data for Jungle Book and the Bible, since the runtimes of my program on the two larger corpora were prohibitively long. I calculated PMI for all word pairs, but since it seemed boring to have pair lists stuffed with punctuation, I made top-20 and bottom-20 lists both with and without punctuation included in the calculations. One interesting observation is that the bottom 20 (no punctuation) is largely populated by closed-class words (is, and, that, for, to), while the top 20 is generally open-class, often in (adjective, noun) pairs.

As far as the independence assumption of unigram models is concerned, these data demonstrate that many words are, in fact, highly dependent upon their context. The highest PMI calculated among my data is for the word pair (ill, favored), with a PMI of 10.17. I have not gone back to check for occurrences of “favored” without “ill”, but with such strong association, these words form a sort of set phrase, I would even be tempted to hyphenate them (ie. “ill-favored”) in my own writing. Therefore, we cannot assume independence such that our language model is based on only the relative frequency of a word in a given corpus. As in all things, context matters.

The following pages show the top-20 and bottom-20 results for the Jungle Book and the Bible:

Jungle Book:

```
Most frequent 20 word pairs are as follows:
Word pair (w1,w2): ('united', 'states'), pmi = 8.46
Word pair (w1,w2): ('literary', 'archive'), pmi = 8.39
Word pair (w1,w2): ('cold', 'lairs'), pmi = 7.85
Word pair (w1,w2): ('archive', 'foundation'), pmi = 7.74
Word pair (w1,w2): ('stretched', 'myself'), pmi = 7.59
Word pair (w1,w2): ('petersen', 'sahib'), pmi = 7.57
Word pair (w1,w2): ('gutenberg', 'literary'), pmi = 7.52
Word pair (w1,w2): ('fore', 'paws'), pmi = 7.32
Word pair (w1,w2): ('hind', 'legs'), pmi = 7.29
Word pair (w1,w2): ('twenty', 'yoke'), pmi = 7.25
Word pair (w1,w2): ('electronic', 'works'), pmi = 7.11
Word pair (w1,w2): ('paragraph', '1'), pmi = 7.09
Word pair (w1,w2): ('hind', 'flippers'), pmi = 7.09
Word pair (w1,w2): ('[', 'illustration'), pmi = 7.07
Word pair (w1,w2): ('bring', 'news'), pmi = 7.03
Word pair (w1,w2): ('moon', 'rose'), pmi = 6.95
Word pair (w1,w2): ('council', 'rock'), pmi = 6.88
Word pair (w1,w2): ('black', 'panther'), pmi = 6.85
Word pair (w1,w2): ('whole', 'line'), pmi = 6.85
Word pair (w1,w2): ('villagers', 'lived'), pmi = 6.83

Least frequent 20 word pairs are as follows:
Word pair (w1,w2): ('united', 'states'), pmi = 8.46
Word pair (w1,w2): ('.', ','), pmi = -5.07
Word pair (w1,w2): ('of', '.'), pmi = -3.97
Word pair (w1,w2): ('of', 'and'), pmi = -3.86
Word pair (w1,w2): ('a', '.'), pmi = -3.81
Word pair (w1,w2): ('to', '"'), pmi = -3.77
Word pair (w1,w2): ('"', ','), pmi = -3.77
Word pair (w1,w2): ('the', '"'), pmi = -3.75
Word pair (w1,w2): ('his', ','), pmi = -3.7
Word pair (w1,w2): ('a', '"'), pmi = -3.68
Word pair (w1,w2): ('i', 'the'), pmi = -3.55
Word pair (w1,w2): ('with', ','), pmi = -3.28
Word pair (w1,w2): ('as', ','), pmi = -3.26
Word pair (w1,w2): ('.', '.'), pmi = -3.23
Word pair (w1,w2): ('s', ','), pmi = -3.21
Word pair (w1,w2): ('that', 'and'), pmi = -3.16
Word pair (w1,w2): ('his', '"'), pmi = -3.12
Word pair (w1,w2): ('"', '.'), pmi = -3.1
Word pair (w1,w2): ('of', ','), pmi = -3.03
Word pair (w1,w2): ('at', ','), pmi = -2.98
```

Jungle Book, punctuation removed:

```
Most frequent 20 word pairs, with punctuation removed, are as follows:
Word pair (w1,w2): ('united', 'states'), pmi = 8.27
Word pair (w1,w2): ('literary', 'archive'), pmi = 8.21
Word pair (w1,w2): ('cold', 'lairs'), pmi = 7.67
Word pair (w1,w2): ('archive', 'foundation'), pmi = 7.55
Word pair (w1,w2): ('stretched', 'myself'), pmi = 7.41
Word pair (w1,w2): ('petersen', 'sahib'), pmi = 7.38
Word pair (w1,w2): ('gutenberg', 'literary'), pmi = 7.33
Word pair (w1,w2): ('fore', 'paws'), pmi = 7.13
Word pair (w1,w2): ('hind', 'legs'), pmi = 7.11
Word pair (w1,w2): ('twenty', 'yoke'), pmi = 7.07
Word pair (w1,w2): ('electronic', 'works'), pmi = 6.92
Word pair (w1,w2): ('hind', 'flippers'), pmi = 6.91
Word pair (w1,w2): ('bring', 'news'), pmi = 6.84
Word pair (w1,w2): ('moon', 'rose'), pmi = 6.76
Word pair (w1,w2): ('council', 'rock'), pmi = 6.69
Word pair (w1,w2): ('black', 'panther'), pmi = 6.66
Word pair (w1,w2): ('whole', 'line'), pmi = 6.66
Word pair (w1,w2): ('villagers', 'lived'), pmi = 6.64
Word pair (w1,w2): ('master', 'words'), pmi = 6.62
Word pair (w1,w2): ('brown', 'bear'), pmi = 6.6

Least frequent 20 word pairs, with punctuation removed, are as follows:
Word pair (w1,w2): ('of', 'and'), pmi = -4.04
Word pair (w1,w2): ('i', 'the'), pmi = -3.74
Word pair (w1,w2): ('that', 'and'), pmi = -3.34
Word pair (w1,w2): ('had', 'the'), pmi = -2.97
Word pair (w1,w2): ('for', 'and'), pmi = -2.97
Word pair (w1,w2): ('is', 'and'), pmi = -2.96
Word pair (w1,w2): ('and', 'is'), pmi = -2.96
Word pair (w1,w2): ('and', 'of'), pmi = -2.94
Word pair (w1,w2): ('it', 'and'), pmi = -2.82
Word pair (w1,w2): ('on', 'and'), pmi = -2.77
Word pair (w1,w2): ('there', 'the'), pmi = -2.57
Word pair (w1,w2): ('one', 'the'), pmi = -2.54
Word pair (w1,w2): ('were', 'the'), pmi = -2.44
Word pair (w1,w2): ('he', 'and'), pmi = -2.44
Word pair (w1,w2): ('this', 'the'), pmi = -2.41
Word pair (w1,w2): ('is', 'of'), pmi = -2.41
Word pair (w1,w2): ('would', 'the'), pmi = -2.4
Word pair (w1,w2): ('for', 'to'), pmi = -2.36
Word pair (w1,w2): ('to', 'as'), pmi = -2.27
Word pair (w1,w2): ('them', 'the'), pmi = -2.26
```

```
Bible:
Most frequent 20 word pairs are as follows:
Word pair (w1,w2): ('ill', 'favoured'), pmi = 10.17
Word pair (w1,w2): ('judas', 'iscariot'), pmi = 10.03
Word pair (w1,w2): ('curious', 'girdle'), pmi = 9.87
Word pair (w1,w2): ('brook', 'kidron'), pmi = 9.86
Word pair (w1,w2): ('poureth', 'contempt'), pmi = 9.82
Word pair (w1,w2): ('measuring', 'reed'), pmi = 9.78
Word pair (w1,w2): ('persecution', 'ariseth'), pmi = 9.72
Word pair (w1,w2): ('divers', 'colours'), pmi = 9.71
Word pair (w1,w2): ('mary', 'magdalene'), pmi = 9.65
Word pair (w1,w2): ('overflowing', 'scourge'), pmi = 9.54
Word pair (w1,w2): ('wreathen', 'chains'), pmi = 9.43
Word pair (w1,w2): ('fiery', 'furnace'), pmi = 9.41
Word pair (w1,w2): ('sharp', 'sickle'), pmi = 9.41
Word pair (w1,w2): ('committeth', 'adultery'), pmi = 9.4
Word pair (w1,w2): ('earthen', 'vessel'), pmi = 9.39
Word pair (w1,w2): ('perpetual', 'desolations'), pmi = 9.39
Word pair (w1,w2): ('golden', 'spoon'), pmi = 9.36
Word pair (w1,w2): ('bright', 'spot'), pmi = 9.34
Word pair (w1,w2): ('tenth', 'deals'), pmi = 9.33
Word pair (w1,w2): ('cunning', 'workman'), pmi = 9.32

Least frequent 20 word pairs are as follows:
Word pair (w1,w2): ('ill', 'favoured'), pmi = 10.17
Word pair (w1,w2): ('of', 'to'), pmi = -6.24
Word pair (w1,w2): ('for', 'and'), pmi = -6.23
Word pair (w1,w2): ('of', 'he'), pmi = -5.97
Word pair (w1,w2): ('all', 'and'), pmi = -5.76
Word pair (w1,w2): ('the', 'said'), pmi = -5.63
Word pair (w1,w2): ('of', 'is'), pmi = -5.58
Word pair (w1,w2): ('with', ','), pmi = -5.44
Word pair (w1,w2): ('will', 'and'), pmi = -5.38
Word pair (w1,w2): ('when', ','), pmi = -5.38
Word pair (w1,w2): ('to', 'in'), pmi = -5.23
Word pair (w1,w2): ('shall', 'of'), pmi = -5.22
Word pair (w1,w2): ('the', 'israel'), pmi = -5.19
Word pair (w1,w2): ('of', ','), pmi = -5.18
Word pair (w1,w2): ('of', 'in'), pmi = -5.07
Word pair (w1,w2): ('this', 'and'), pmi = -5.06
Word pair (w1,w2): (',', 'me'), pmi = -5.06
Word pair (w1,w2): ('into', ','), pmi = -5.04
Word pair (w1,w2): ('to', ';'), pmi = -5.01
Word pair (w1,w2): (',', 'thee'), pmi = -4.99
```

Bible, punctuation removed:

```
Most frequent 20 word pairs, with punctuation removed, are as follows:
Word pair (w1,w2): ('ill', 'favoured'), pmi = 10.03
Word pair (w1,w2): ('judas', 'iscariot'), pmi = 9.88
Word pair (w1,w2): ('brook', 'kidron'), pmi = 9.72
Word pair (w1,w2): ('curious', 'girdle'), pmi = 9.72
Word pair (w1,w2): ('poureth', 'contempt'), pmi = 9.67
Word pair (w1,w2): ('measuring', 'reed'), pmi = 9.63
Word pair (w1,w2): ('divers', 'colours'), pmi = 9.57
Word pair (w1,w2): ('persecution', 'ariseth'), pmi = 9.57
Word pair (w1,w2): ('mary', 'magdalene'), pmi = 9.51
Word pair (w1,w2): ('overflowing', 'scourge'), pmi = 9.39
Word pair (w1,w2): ('wreathen', 'chains'), pmi = 9.28
Word pair (w1,w2): ('fiery', 'furnace'), pmi = 9.26
Word pair (w1,w2): ('sharp', 'sickle'), pmi = 9.26
Word pair (w1,w2): ('committeth', 'adultery'), pmi = 9.25
Word pair (w1,w2): ('earthen', 'vessel'), pmi = 9.24
Word pair (w1,w2): ('perpetual', 'desolations'), pmi = 9.24
Word pair (w1,w2): ('golden', 'spoon'), pmi = 9.21
Word pair (w1,w2): ('tenth', 'deals'), pmi = 9.19
Word pair (w1,w2): ('bright', 'spot'), pmi = 9.19
Word pair (w1,w2): ('cunning', 'workman'), pmi = 9.17

Least frequent 20 word pairs, with punctuation removed, are as follows:
Word pair (w1,w2): ('of', 'to'), pmi = -6.39
Word pair (w1,w2): ('for', 'and'), pmi = -6.37
Word pair (w1,w2): ('of', 'he'), pmi = -6.12
Word pair (w1,w2): ('all', 'and'), pmi = -5.91
Word pair (w1,w2): ('the', 'said'), pmi = -5.78
Word pair (w1,w2): ('of', 'is'), pmi = -5.72
Word pair (w1,w2): ('will', 'and'), pmi = -5.52
Word pair (w1,w2): ('to', 'in'), pmi = -5.38
Word pair (w1,w2): ('shall', 'of'), pmi = -5.37
Word pair (w1,w2): ('the', 'israel'), pmi = -5.34
Word pair (w1,w2): ('of', 'in'), pmi = -5.22
Word pair (w1,w2): ('this', 'and'), pmi = -5.2
Word pair (w1,w2): ('of', 'will'), pmi = -5.12
Word pair (w1,w2): ('from', 'of'), pmi = -5.07
Word pair (w1,w2): ('and', 'son'), pmi = -5.05
Word pair (w1,w2): ('there', 'and'), pmi = -5.01
Word pair (w1,w2): ('in', 'for'), pmi = -4.97
Word pair (w1,w2): ('the', 'he'), pmi = -4.94
Word pair (w1,w2): ('his', 'in'), pmi = -4.91
Word pair (w1,w2): ('and', 'house'), pmi = -4.88
```