Empricial Laws of the Corpus

July 26, 2023

Outline

- Characteristics of a corpus
- Type-to-Token Ratio
- Zipf's Law
- Heaps' Law

Characteristics of a Language Corpus

- Collection of text in digital form
- Comprises a large fraction of the dictionary
- Incremental - New text samples are added regularly
- Corpus is gigantic - billions of words normally
- Covers all fields of text of a language
- Numerous areas are evenly distributed in the text

Function Words Vs. Context Words

Function Words

Function words include determiners, conjunctions, prepositions, pronouns, auxiliary verbs, modals, qualifiers, and question words.

Content Words

Content words are words with specific meanings, such as nouns, adjectives, adverbs, and main verbs (those without helping verbs.)

Function Words Vs. Context Words

Function words

- provides structure
- No addition/deletion in language
- Close class

Content words

- provides meaning
- Addition/deletion in language
- Open class

An Example

"The sly brown fox jumped gracefully over the lazy dog and cat."

Function words are:

- the (determiner)
- over (preposition)
- and (conjunction)

Contents words are:

- fox, dog, and cat (nouns)
- sly, brown, and lazy (adjectives)
- gracefully (adverb)
- jumped (main verb)

Most Common Words in Brown Corpus I

Table: Brown Corpus most common words

#	Word	Frequency	POS	Description
1	the	62288	AT	Article
2	,	58153	,	,
3	.	48812	.	.
4	of	34864	IN	preposition
5	and	27676	CC	coordinating conjunction
6	a	21824	AT	Article
7	in	18975	IN	preposition
8	to	14679	TO	TO

Most Common Words in Brown Corpus II

9	to	10903	IN	preposition
10	is	9976	BEZ	verb be, present tense
11	was	9768	BEDZ	verb be, past tense
12	“	8837	“	“
13	”	8789	”	”
14	for	8363	IN	preposition
15	with	6993	IN	preposition

Type-Token Ratio (TTR)

Definition

A type-token ratio (TTR) is the total number of unique words (types) divided by the total number of words (tokens) in a given the text collection.

- TTR indicates the lexical richness of collection, or variety in the vocabulary
- In written language, does the same word repeates over and over, or variety of different words
- As TTR ratio approaches to 1, the frequency of appearance of new words gets higher.

Type-Token Ratio (TTR) - An Exampel

```
import nltk
tokens = []
sentences = ['Natural Language Processing with NLTK',
             'Natural Language gets developed over the time',
             'Language Processing is very important',
             'Natural Language Processing is an important
scientific field',
             'Natural Language Analysis',
             'Natural Language Processing',
             'Natural Language Understanding',
             'Natural Language Understanding']
for sentence in sentences:
    tokens.extend([word for word in nltk.word_tokenize(sentence)])

print('Tokens: ', tokens, '\n')
print('Types: ',set(tokens),'\n')

print('Total No. of tokens:',len(tokens))
print('Types of tokens:',len(set(tokens)))
```

Figure: Python Script to find TTR

Type-Token Ratio (TTR) - An Exampel

```
Tokens:  ['Natural', 'Language', 'Processing', 'with', 'NLTK',
'Natural', 'Language', 'gets', 'developed', 'over', 'the', 'tim
e', 'Language', 'Processing', 'is', 'very', 'important', 'Natur
al', 'Language', 'Processing', 'is', 'an', 'important', 'scient
ific', 'field', 'Natural', 'Language', 'Analysis', 'Natural', '
Language', 'Processing', 'Natural', 'Language', 'Understanding'
, 'Natural', 'Language', 'Understanding']

Types:  {'is', 'Analysis', 'developed', 'NLTK', 'Understanding'
, 'Language', 'time', 'Processing', 'very', 'over', 'scientific
', 'the', 'important', 'gets', 'an', 'Natural', 'field', 'with'
}

Total No. of tokens: 37
Types of tokens: 18
TTR 0.4864864864864865
```

Figure: Python Script to find TTR - Output

Python Script - TTR for different domains

```
# block of code to find TTR for differen cateogry of the brown corpus
import nltk
category = []
category = brown.categories()
print("Category_Name", "Word Count","Vocabulary_Size","TTR")
for item in category:
        print(item,len(brown.words(categories=item)),
              len(set(brown.words(categories=item))),
              len(brown.words(categories=item))/len(set(brown.words(categories=item))))
```

Figure: Python Script to find TTR for different domains

TTR for different domains

Table: TTR for different domains

Category	**Word Count**	**Vocabulary**	**TTR**
fiction	68488	9302	0.1358
news	100554	14394	0.1431
government	70117	8181	0.1166
education	181888	16859	0.0926

TTR for different domains

Table: TTR for different domains

Category	Word Count	Vocabulary	TTR
fiction	68488	9302	0.1358
news	100554	14394	0.1431
government	70117	8181	0.1166
education	181888	16859	0.0926

- The TTR value indicates how often a new 'Word' type appears in the corpus.
- The low TTR means tendency to use the same word (news)
- The high TTR means tendency to use the different word (education)

TTR - What knowledge it conveys

Interpretation of TTR

- TTR = 0.10 indicates that words occurs on average 10 times each.
- However, Words are distributed unevenly.

Major fraction of vocabulary is rare

- "Hapax Legomena" - A greek words means 'Read only once'
- This phenomena is found here too as 45% words occurs only once

Running TTR for 'Fiction' Category

Table: Running TTR (Cumulative Vocabulary Size at Every 1000 Tokens)

# tokens	# vocab	% of tokens	# tokens	# vocab	% of tokens
1000	438	0.44	15000	3354	0.22
2000	710	0.36	20000	4003	0.20
3000	972	0.32	25000	4683	0.19
4000	1239	0.31	30000	5429	0.18
5000	1511	0.30	35000	6050	0.17
6000	1750	0.29	40000	6726	0.17
7000	1939	0.28	45000	7183	0.16
8000	2169	0.27	50000	7764	0.16
9000	2310	0.26	55000	8220	0.15
10000	2467	0.25	60000	8572	0.14

Running TTR for various category

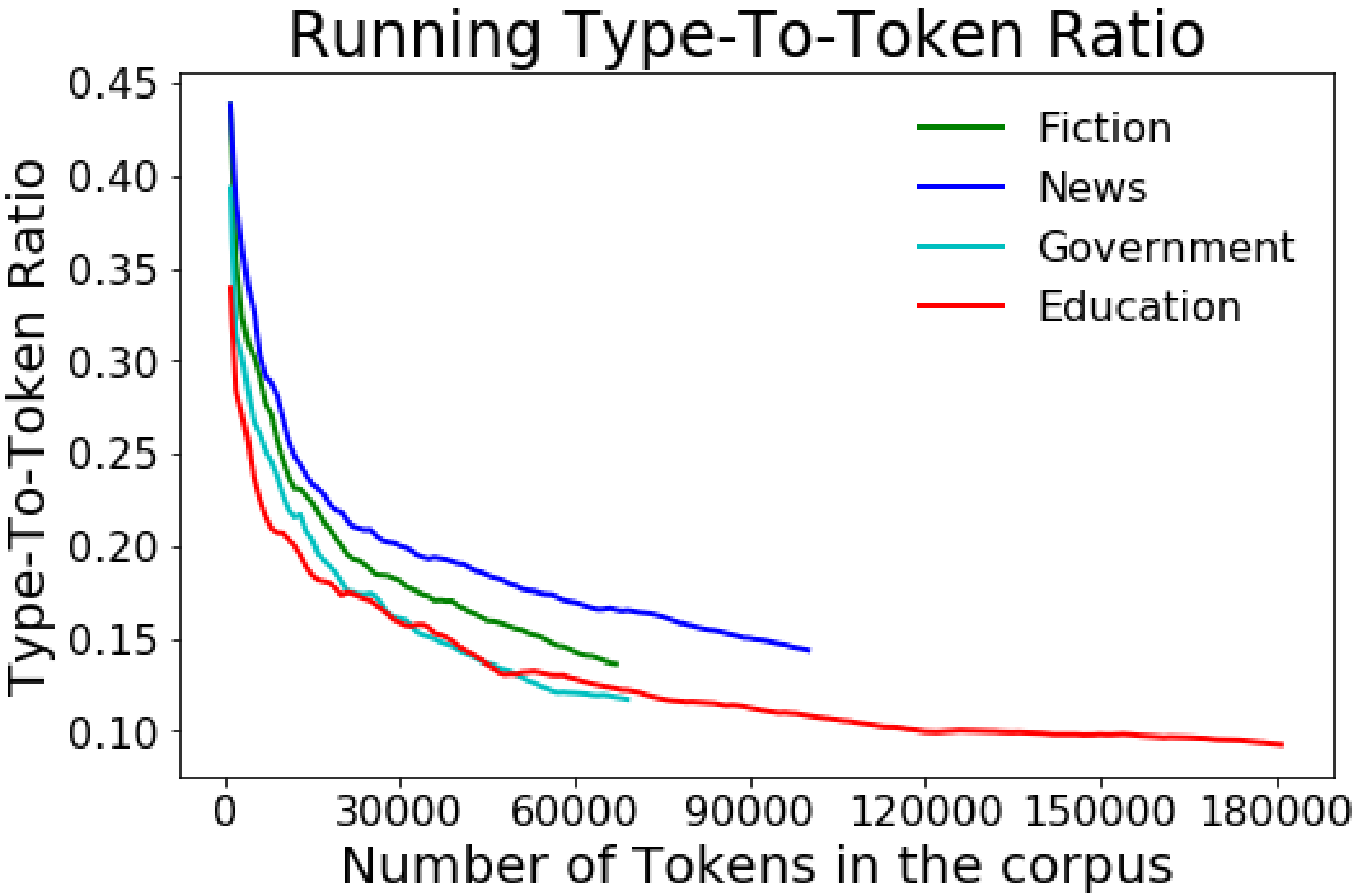


Figure: Running TTR (TTR at every 1000 tokens cumulative)

Lexical Diversity of NLTK Corpuses

```
from nltk.corpus import
genesis,udhr,inaugural,gutenberg,brown,floresta,abc
genesis = genesis.words()
udhr = udhr.words()
inaugural = inaugural.words()
gutenberg = gutenberg.words()
brown = brown.words()
floresta = floresta.words()
abc = abc.words()
for corpus in [genesis,udhr,inaugural,gutenberg,brown,floresta,abc]:
    word_count,vocab_size,diversity_score,corpus =
lexical_diversity(corpus)
    print('corpus name: ',corpus)
    print('word count: ',word_count)
    print('vocab_size: ',vocab_size)
    print('diversity score: ',diversity_score)
```

Figure: Python Script to find Lexical Diversity

Lexical Diversity of NLTK Corpuses

Table: TTR for NLTK corpuses

Corpus Name	Word Count	Vocabulary Size	TTR
abc	766811	31885	0.042
brown	1161192	56057	0.048
floresta	211852	29421	0.139
genesis	315268	25841	0.082
gutenberg	2621613	51156	0.020
inaugural	149797	9913	0.066
udhr	585510	113846	0.194

Zipf's law

Zipf's law - An empirical law

Zipf's law states that given a large sample of words used, the frequency(f) of any word is inversely proportional to its rank(r) in the frequency table (Also known as Zipfian distribution or zeta distribution).

$$f \propto \frac{1}{r} \tag{1}$$

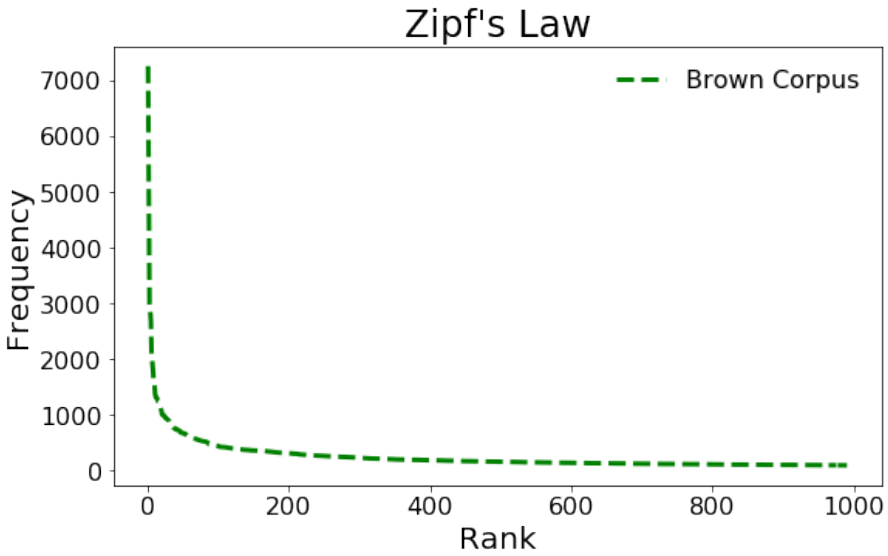
Or, there is a constant k such that,

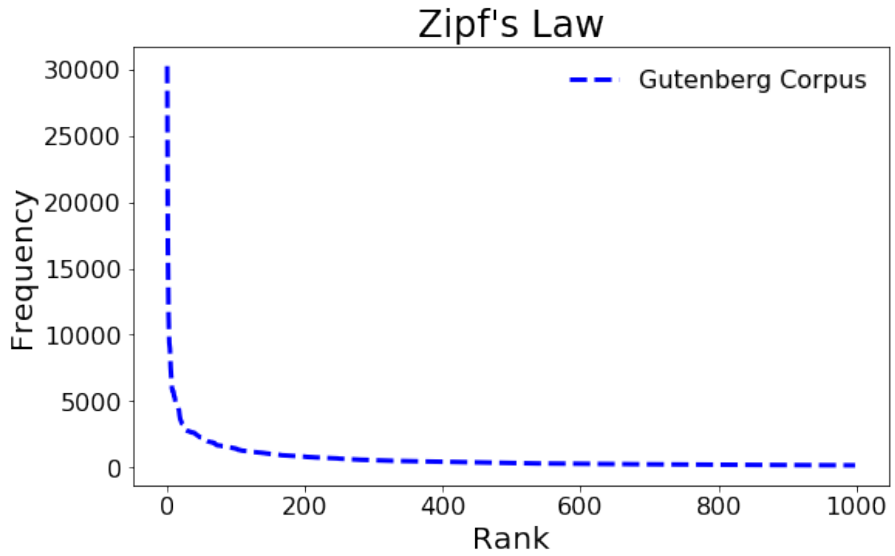
$$f \cdot r = k \tag{2}$$

Thus the most frequent word will occur about twice as often as the second most frequent word, three times as often as the third most frequent word, etc.

word	freq	rank	f.r	word	freq	rank	f.r
The	7258	1	7258	first	1242	15	18630
I	5161	2	10322	like	1237	16	19792
He	2982	3	8946	This	1179	17	20043
one	2873	4	11492	man	1151	18	20718
would	2677	5	13385	made	1122	19	21318
It	2037	6	12222	new	1060	20	21200
said	1943	7	13601	must	1003	21	21063
In	1801	8	14408	also	999	22	21978
could	1580	9	14220	Af	995	23	22885
time	1556	10	15560	even	985	24	23640
But	1374	11	15114	back	950	25	23750
A	1314	12	15768	years	943	26	24518
two	1311	13	17043	And	938	27	25326
may	1292	14	18088	many	925	28	25900
first	1242	15	18630	She	911	29	26419
like	1237	16	19792	much	900	30	27000

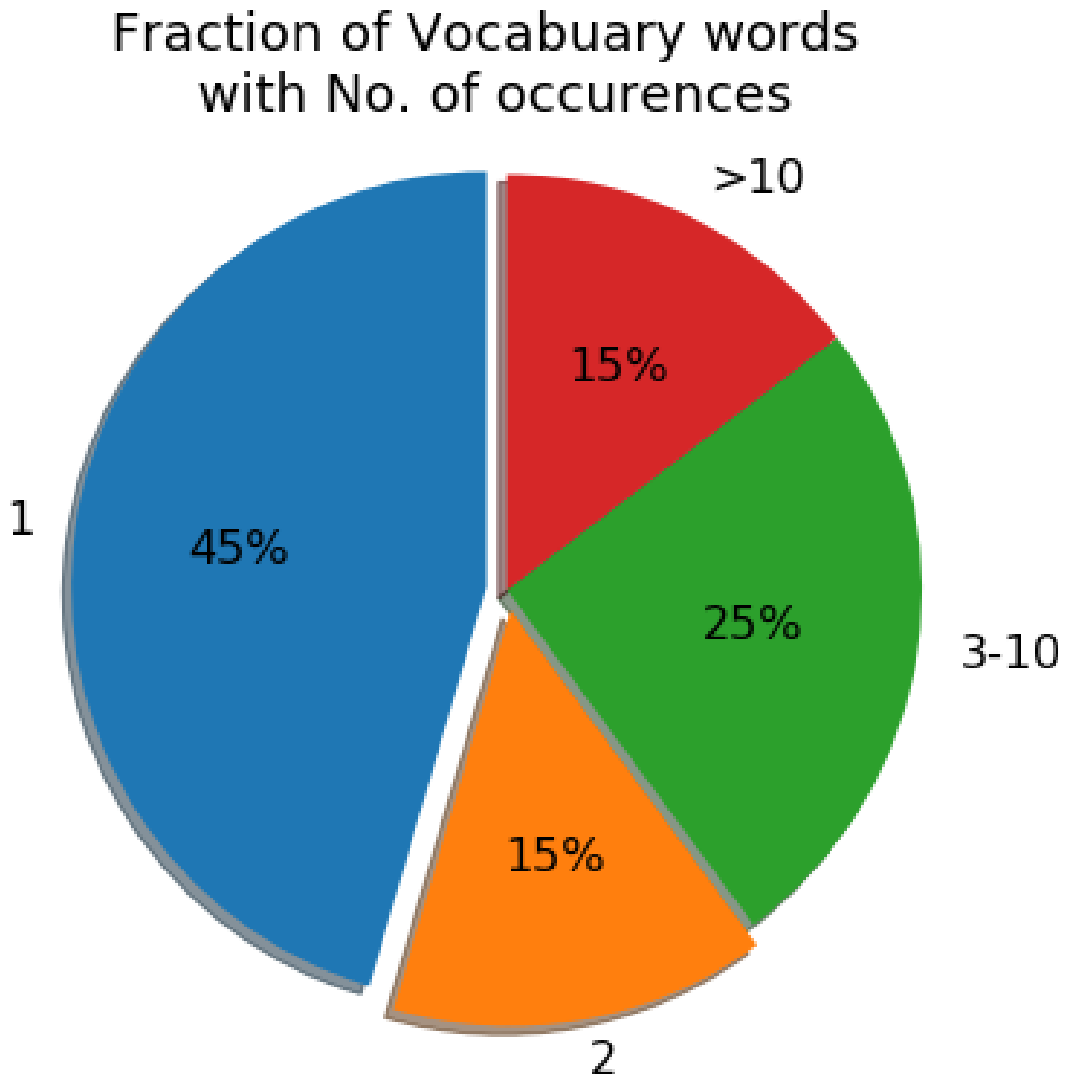
Plotting Zipfs frequency distribution





Frequency of frequencies

Word Frequency	Frequency of frequencies
1	137649
2	16170
3	7196
4	4251
5	2280
6	1447
7	1737
8	619
9	896
10	328
11 to 50	4973
51 to 100	658
>100	470



Zipf's Other Laws

Correlation: Number of meanings and word frequency

The number of meanings m of a word obeys the law:

$$m \propto \sqrt{f} \tag{3}$$

Given the first law,

$$m \propto \frac{1}{\sqrt{r}} \tag{4}$$

Zipf's Other law

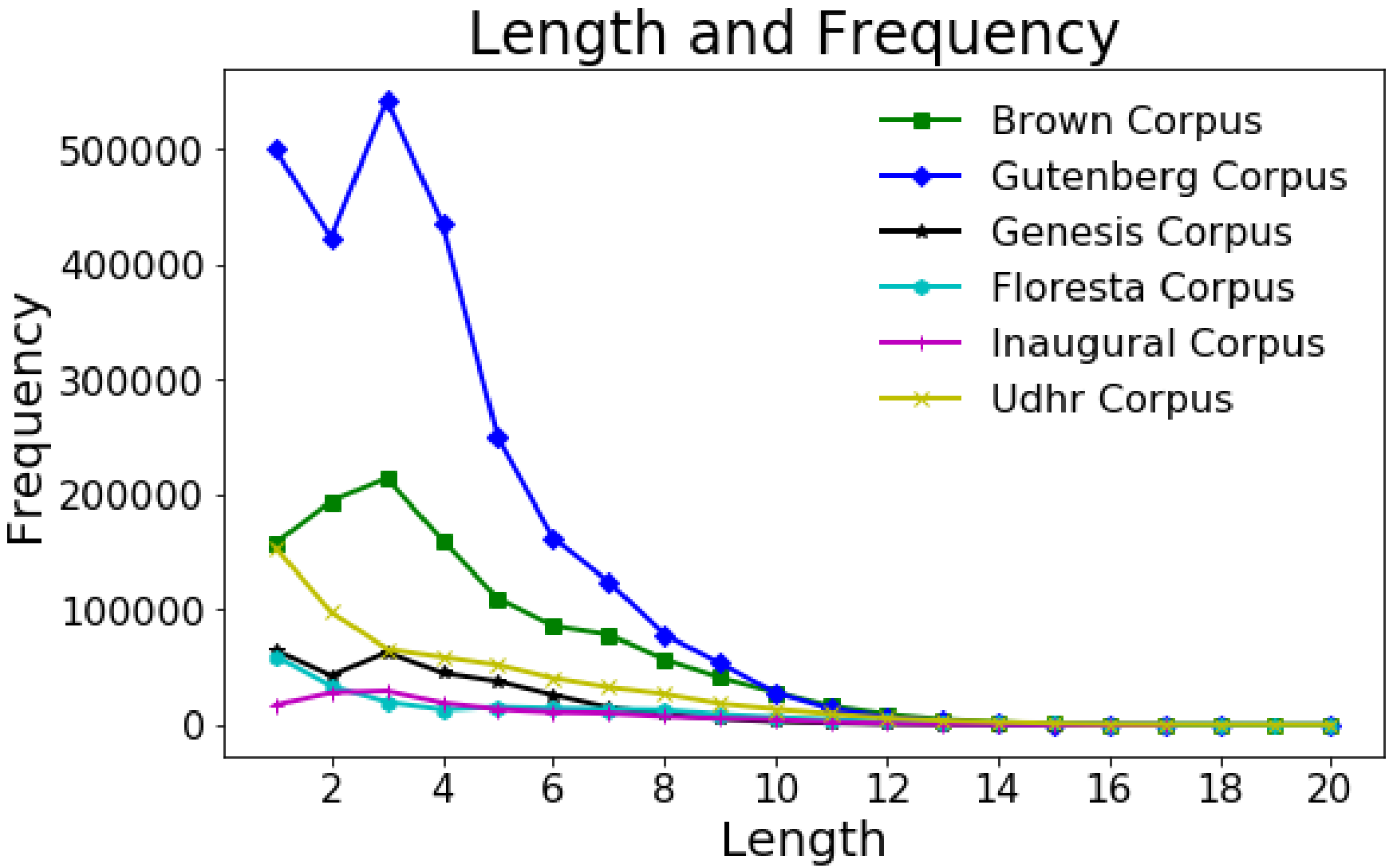
Correlation: Word length and word frequency

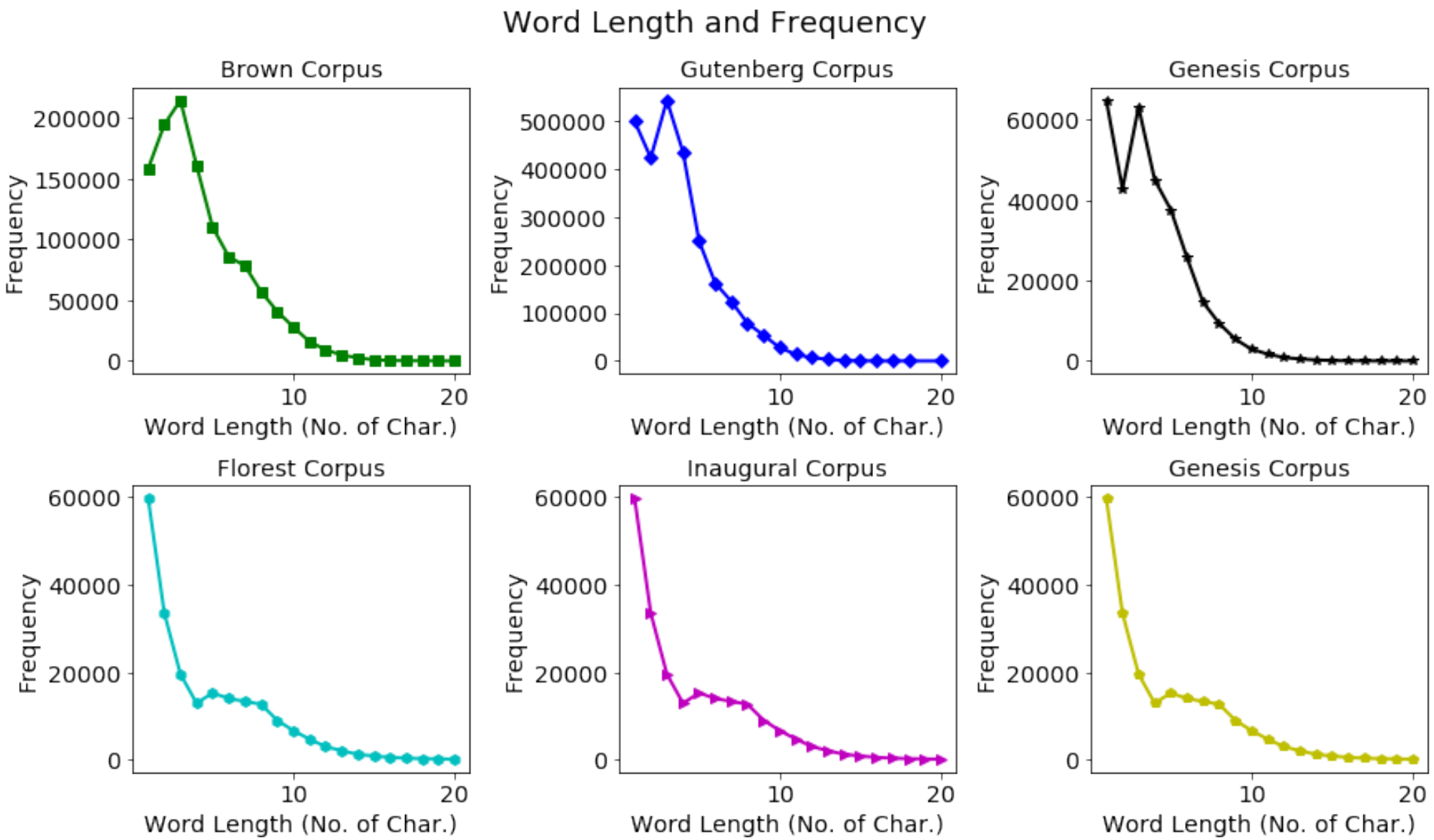
Word frequency is inversely proportional to their length.

Word Length and Freqeuncy

Word length	Gutenberg Corpus	Floresta Corpus
1	499543	59693
2	422962	33530
3	543212	19579
4	435309	12945
5	249601	15296
6	162092	14010
7	124059	13366
8	78349	12699
9	53623	9066
10	27548	6692

Word Frequency and Length




Word Length and Frequency
Brown Corpus
Gutenberg Corpus
Genesis Corpus
Florest Corpus
Inaugural Corpus
Genesis Corpus
Frequency
Word Length (No. of Char.)
0
50000
100000
150000
200000
300000
400000
500000
20000
40000
60000
10
20

Heap's law

Heap's law - Vocabulary and Number of tokens

In the information retrieval field, Heaps' law is an empirical law that describes the number of distinct type of words as the function of the total number of tokens in the corpus.

Heap's law

Heap's law - Vocabulary and Number of tokens

In the information retrieval field, Heaps' law is an empirical law that describes the number of distinct type of words as the function of the total number of tokens in the corpus.

The relationship

$$\begin{aligned} \mid V \mid &\propto n^{\beta} \\ \mid V \mid &= K n^{\beta} \end{aligned} \tag{5}$$

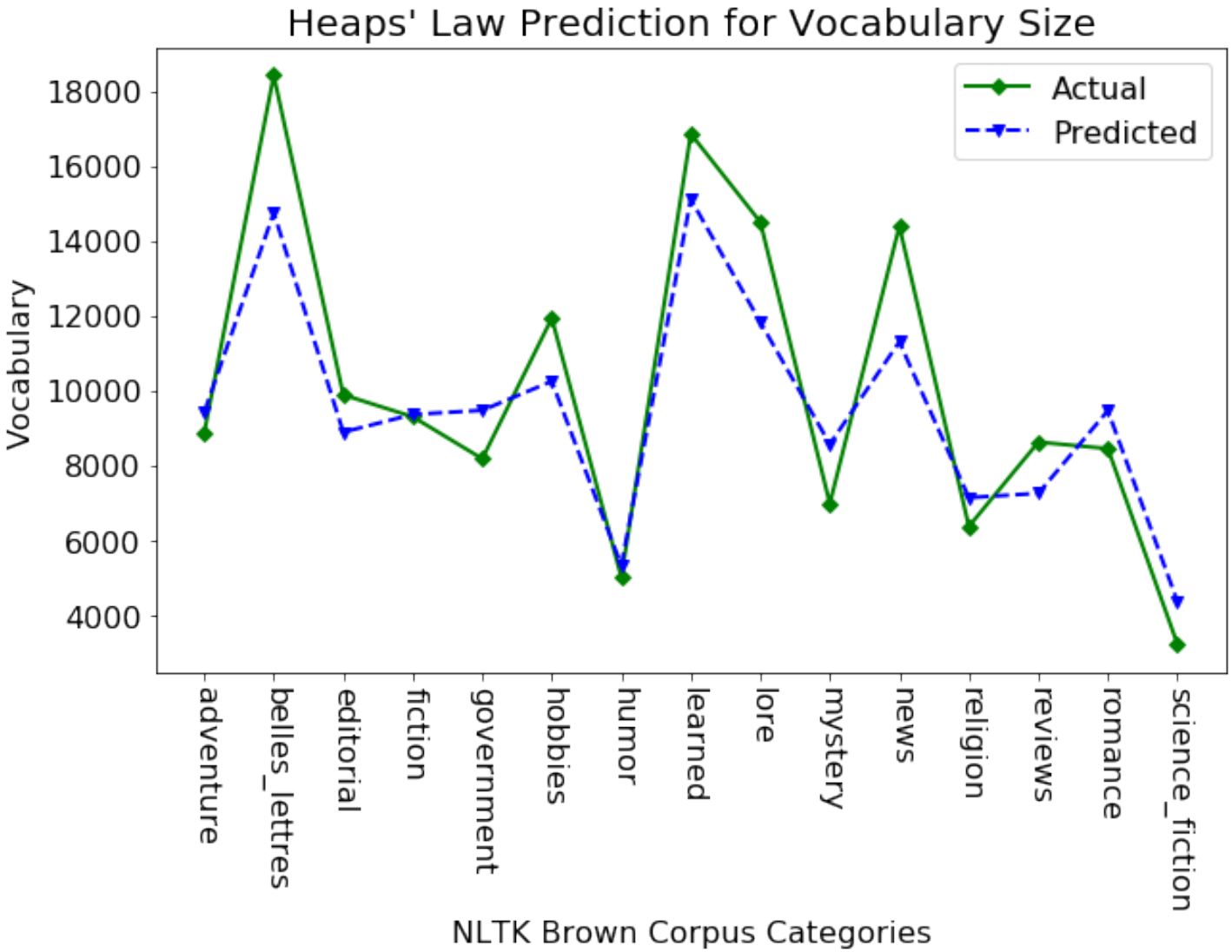
Where,
1) $\mid V \mid$ is the number of distinct words in the corpus
2) n is the total number of tokens in the corpus
3) Typically, K is between 30 and 100 and β is between 0.4 to 0.6.

Heaps law - Vocabulary size prediction

Table: Vocabulary size prediction (K=40, $\beta = 0.49$)

Category	**Word Count**	**Vocabulary**	**Predicted**
adventure	69342	8874	9422
belles_lettres	173096	18421	14751
editorial	61604	9890	8891
fiction	68488	9302	9365
government	70117	8181	9474
hobbies	82345	11935	10250
humor	21695	5017	5332
learned	181888	16859	15113
lore	110299	14503	11828
mystery	57169	6982	8572
news	100554	14394	11304
religion	39399	6373	7142
reviews	40704	8626	7257

Heaps law - Vocabulary size prediction



Questions?

Thank You All.....

References I