Word Frequency and Zipf's Law

Abby Sussman

What Is a Word?

Lemma: the canonical definition of a set of word forms

- Run, runs, ran, running all have the same lemma “run”

Token: an individual word

- The sentence “I was running while he ran” has 6 tokens and 5 lemmas

Most Common Words (written)

According to the Oxford English Corpus, the most common words in English are:

This uses lemmas. So, “be” includes am, is, are, was, etc.

1) the
2) be
3) to
4) of
5) and
6) a
7) in

Most Common Words (written)

According to the Oxford English Corpus, the most common words in English are:

This uses lemmas. So, “be” includes am, is, are, was, etc.

1) the
2) be
3) to
4) of
5) and
6) a
7) in
8) that
9) have
10) I

Is this the case for every text?

My approach

Stop Words

Stop words are common words that don't add substantial meaning to a sentence.

The first list of stop words was made in 1959 and consisted of just "a, an, and, as, at, by, for, from, if, in, of, on, or, the, to, with" .

Removing these words from your analysis gives a more accurate representation of the meaning of a text.

However, for frequency analysis, we will not be removing any stop words.

My Program (counts tokens)

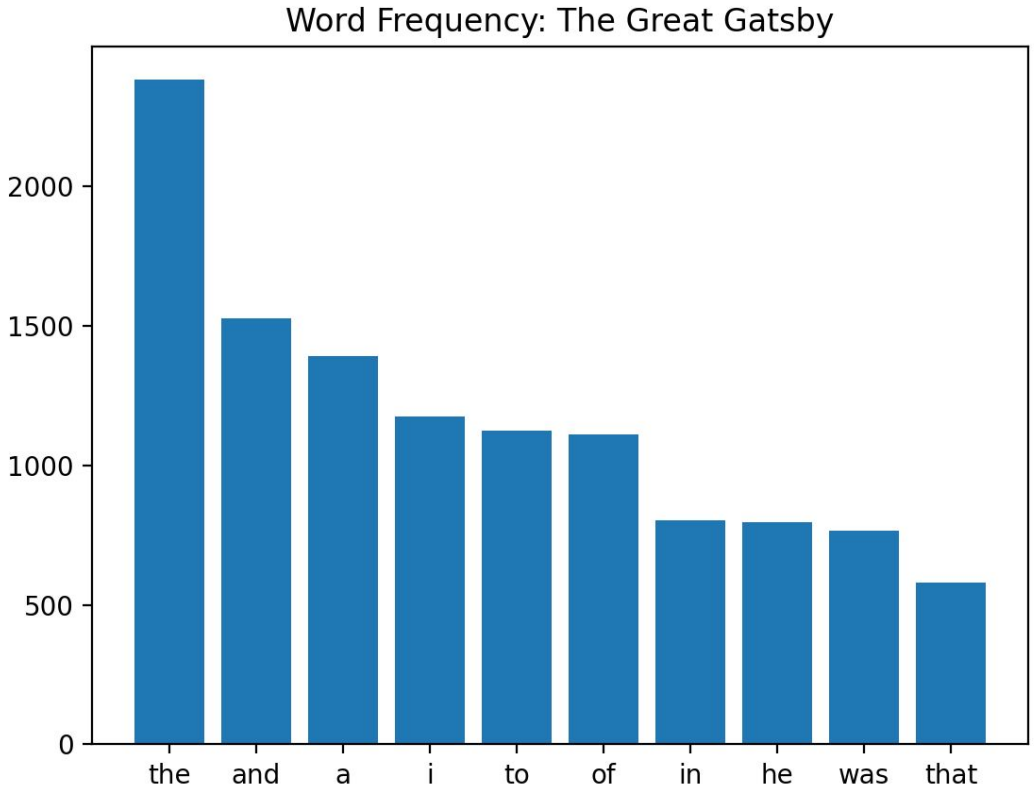
```
text = open("Great.Gatsby.txt","r")
freq_dict = dict()
for line in text:
    line = line.strip()
    if not line:
        continue
    line = line.lower()
    line = line.translate(line.maketrans("", "", string.punctuation))
    words = line.split(" ")
    for word in words:
        if word in freq_dict:
            freq_dict[word] = freq_dict[word] + 1
        else:
            freq_dict[word] = 1
```

Counts frequency of tokens and adds it to a dictionary

```
sorted_dict = sorted(freq_dict.items(), key=lambda x:x[1], reverse=True)
ordered_dict = dict(sorted_dict)
topten_dict = {key: ordered_dict[key]

for key in list(ordered_dict.keys())[:10]}

for key in list(topten_dict.keys()):
    print(key, ":", topten_dict[key])

names = list(topten_dict.keys())
values = list(topten_dict.values())

plt.bar(range(len(topten_dict)), values, tick_label=names)
plt.title('Word Frequency: The Great Gatsby')
plt.show()
```

Orders words by frequency and graphs the 10 most used words

the : 2,380
and : 1,526
a : 1,392
i : 1,176
to : 1,125
of : 1,109
in : 801
he : 795
was : 764
that : 581



48,346 tokens with 6,408 unique words

What is Zipf’s Law?

George Kingsley Zipf (1902-50)

American linguist and philologist

Studied statistical occurrences in languages

Studied and lectured at Harvard University

Zipf's Law

States that the relative frequency of a word is inversely proportional to its rank

$$\text{word frequency} \propto \frac{1}{\text{word rank}}$$

So, the 2nd most used word is used ½ as much as the 1st

- The 3rd most used word is used ⅓ as much as the 1st
- The 4th most used word is used ¼ as much as the 1st
- cont...

$$f(r) \propto \frac{1}{r^{\alpha}} \qquad \text{with } \alpha \approx 1$$

Plotting the top 1,000 words from a million-word collection of English writings:

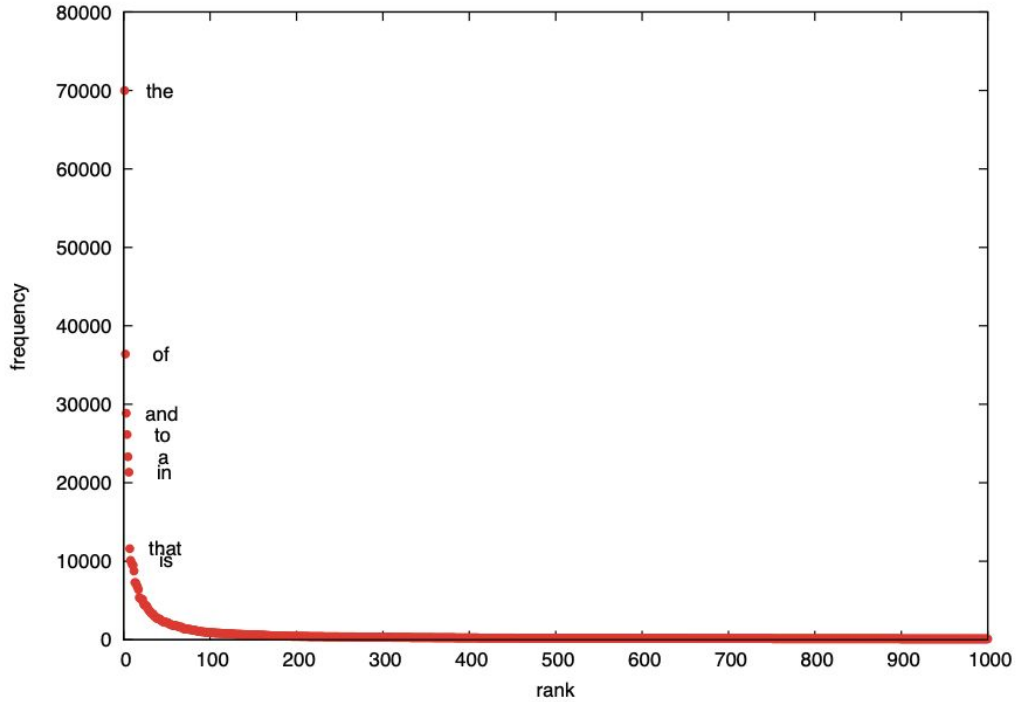


Figure 1. In a million words of writing in English, the word "the" appears 70 000 times, "of" appears about half as often, and most words occur just a few times or only once

"the" appears 70,000 times

"of" appears 36,000 times

Now for the log-log graph:

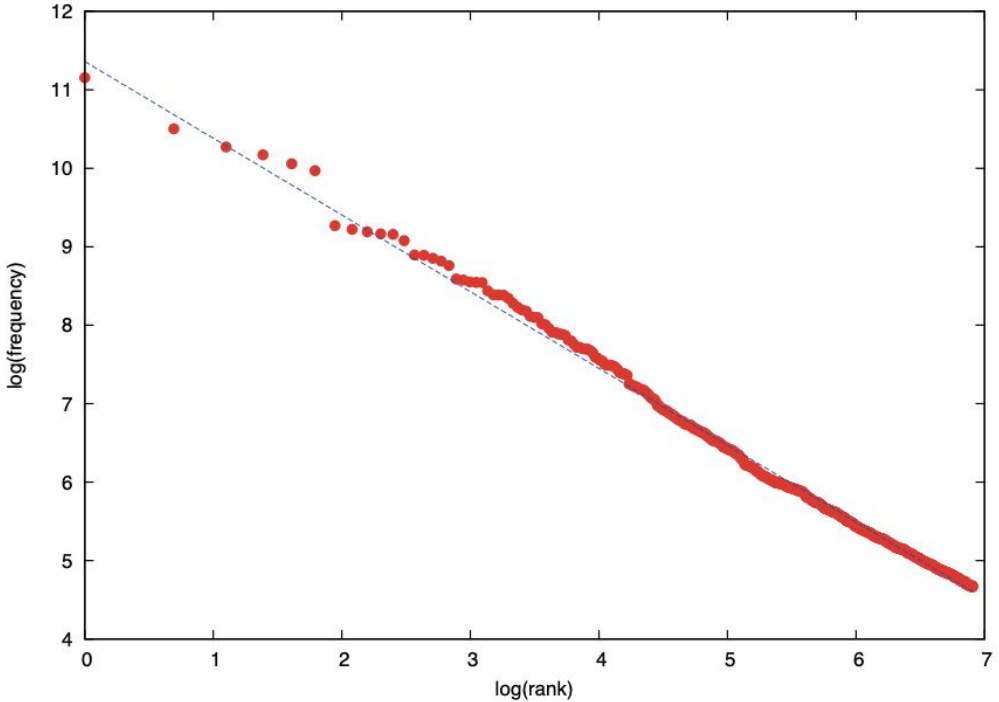
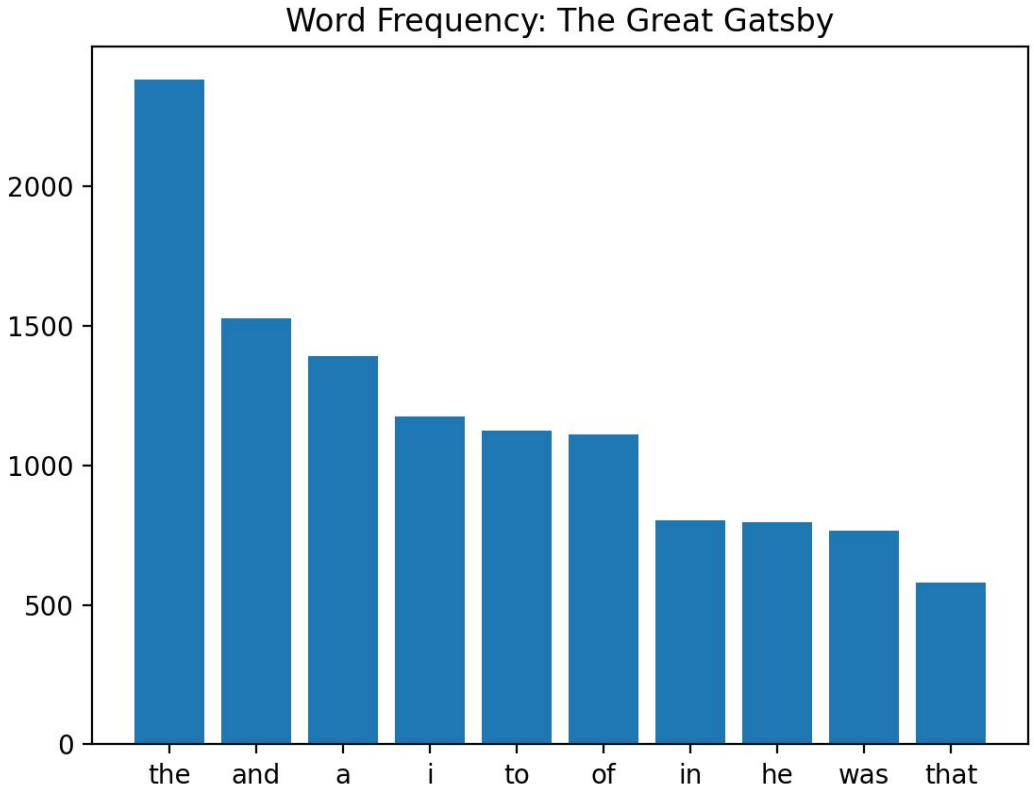


Figure 2. The word frequencies of Figure 1 plotted logarithmically

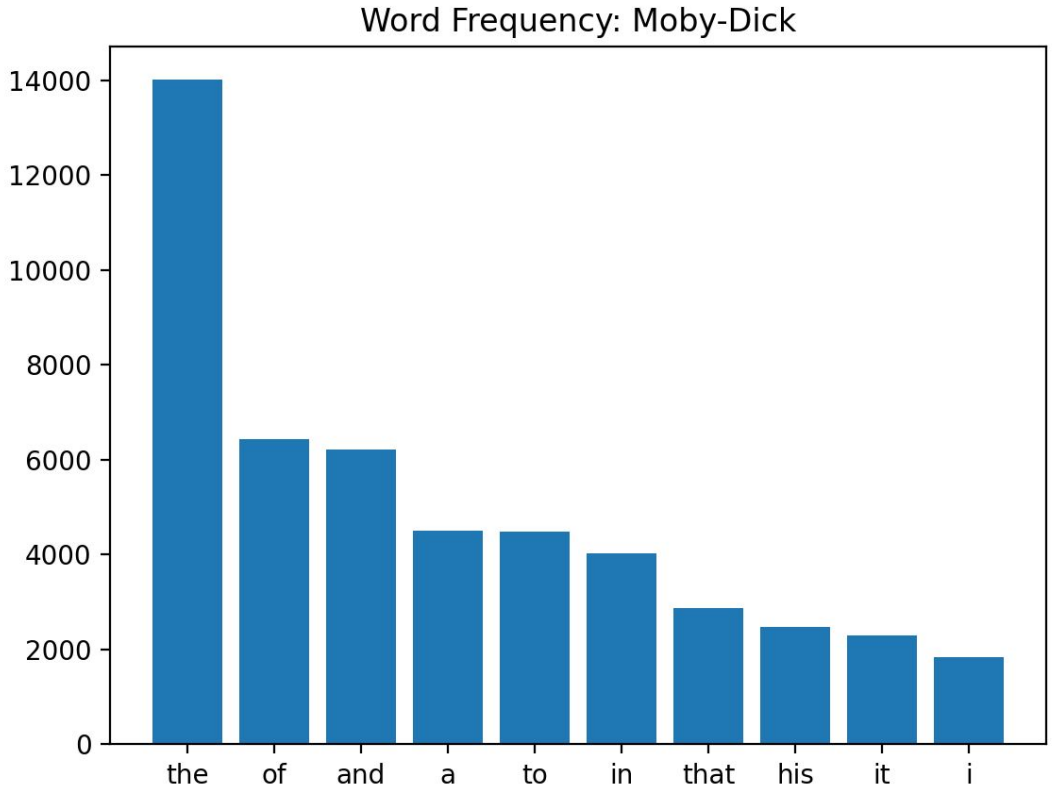
Forms a straight line with a slope roughly equal to -1

the : 2,380
and : 1,526
a : 1,392
i : 1,176
to : 1,125
of : 1,109
in : 801
he : 795
was : 764
that : 581



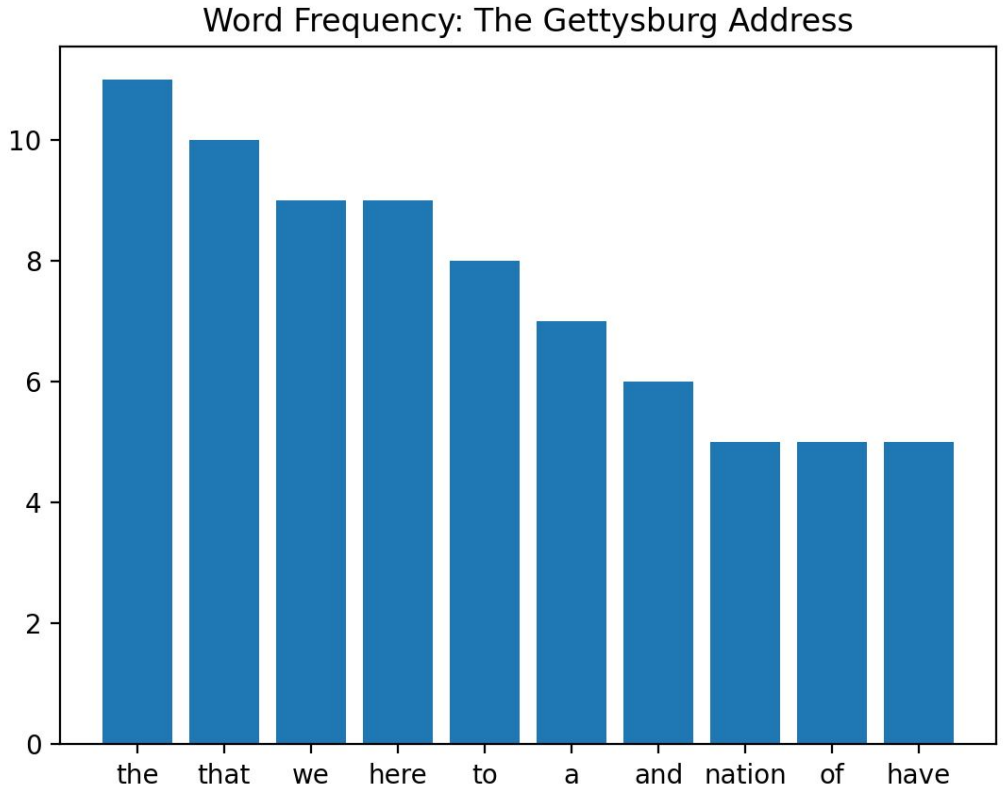
48,346 tokens with 6,408 unique words

the : 14,005
of : 6,435
and : 6,222
a : 4,512
to : 4,485
in : 4,033
that : 2,864
his : 2,479
it : 2,287
i : 1,835



208,459 tokens with 21,024 unique words

the : 11
that : 10
we : 9
here : 9
to : 8
a : 7
and : 6
nation : 5
of : 5
have : 5



262 tokens with 139 unique words

Pareto Principle (again)

As we remember from last week, the Pareto Principle states that 80% of the effects come from 20% of the causes.

The Great Gatsby: 20% of the words are used 85% of the time

Moby-Dick: 20% of the words are used 87% of the time

The Gettysburg Address: 20% of the words are used 49% of the time

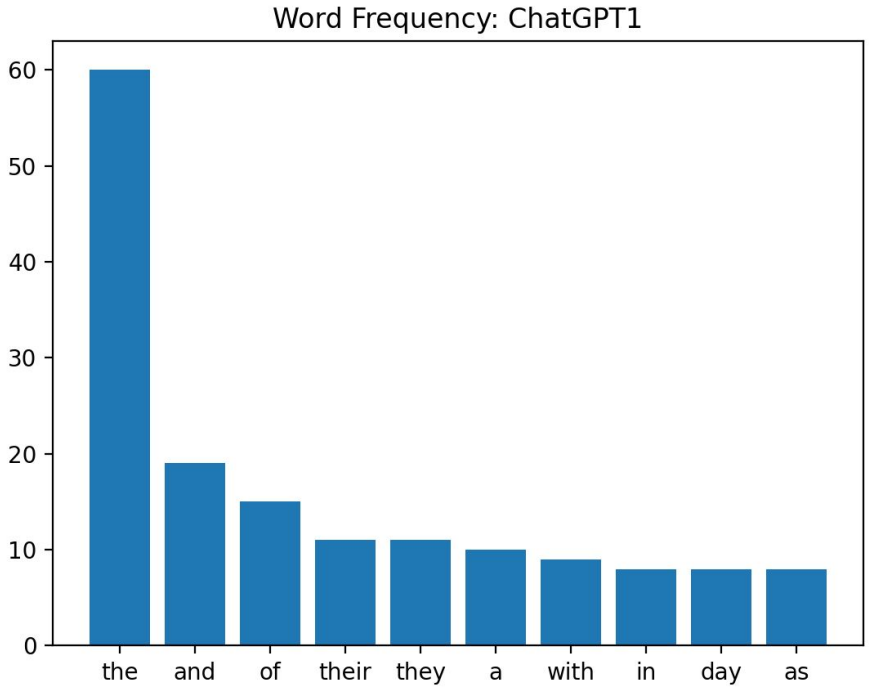
Other Applications (kind of)

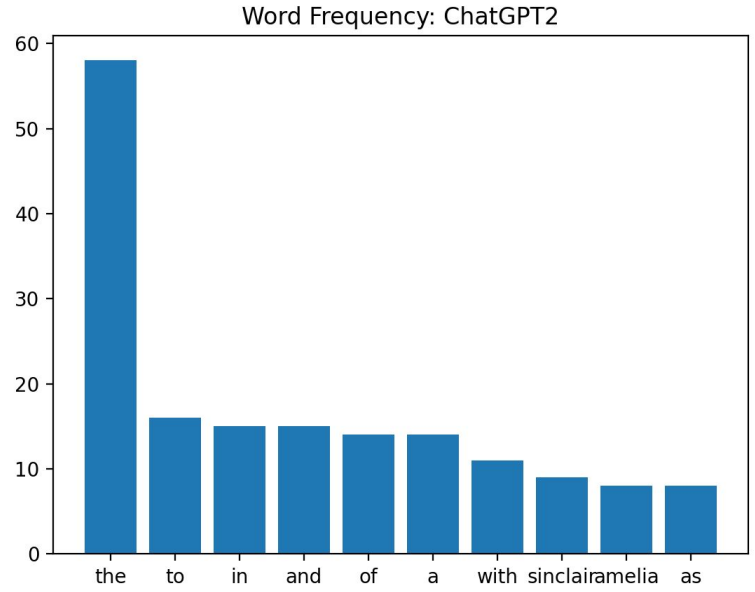
AI Generated Text

Told ChatGPT to write some stories

Definitely does not produce a Zipfian distribution

449 tokens
230 unique words
20% of the words are used 58% of the time

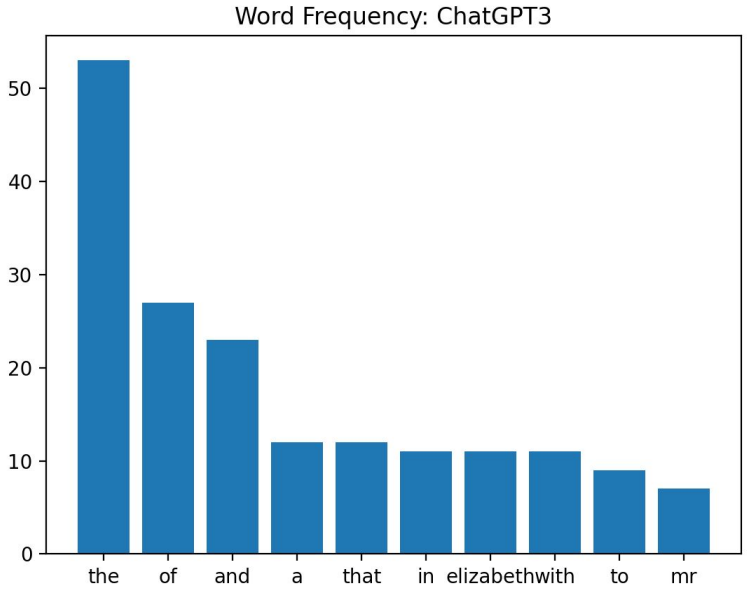




526 tokens

298 unique words

20% of the words are used 54% of the time



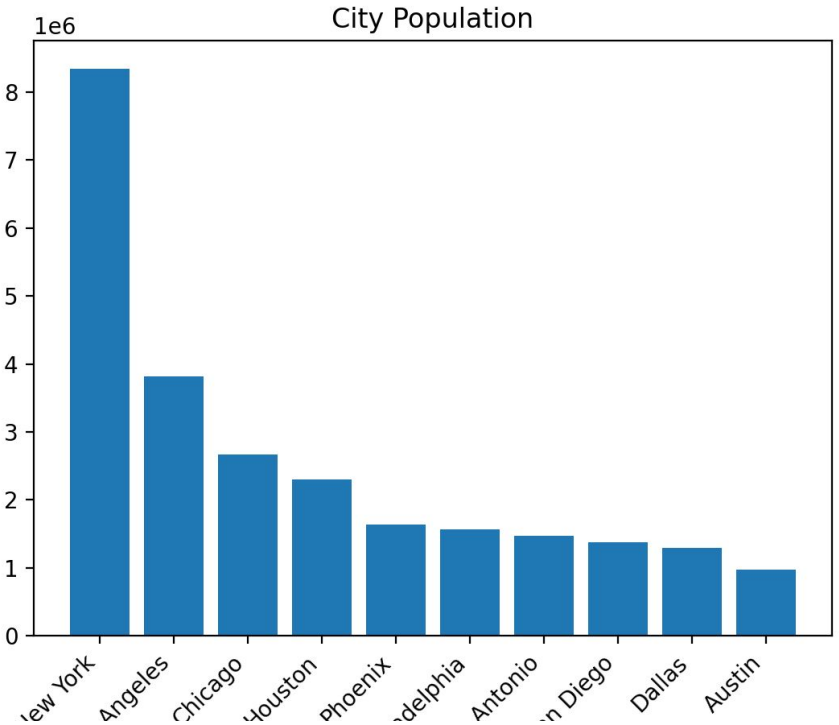
523 tokens

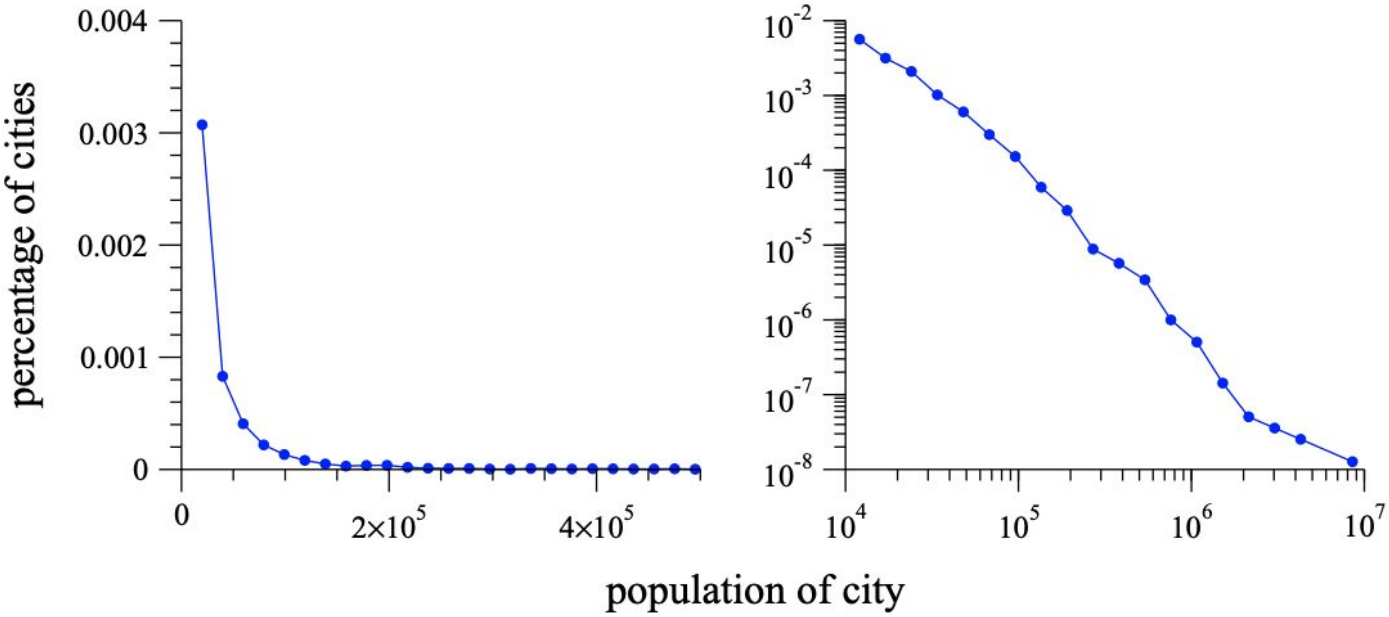
263 unique words

20% of the words are used 57% of the time

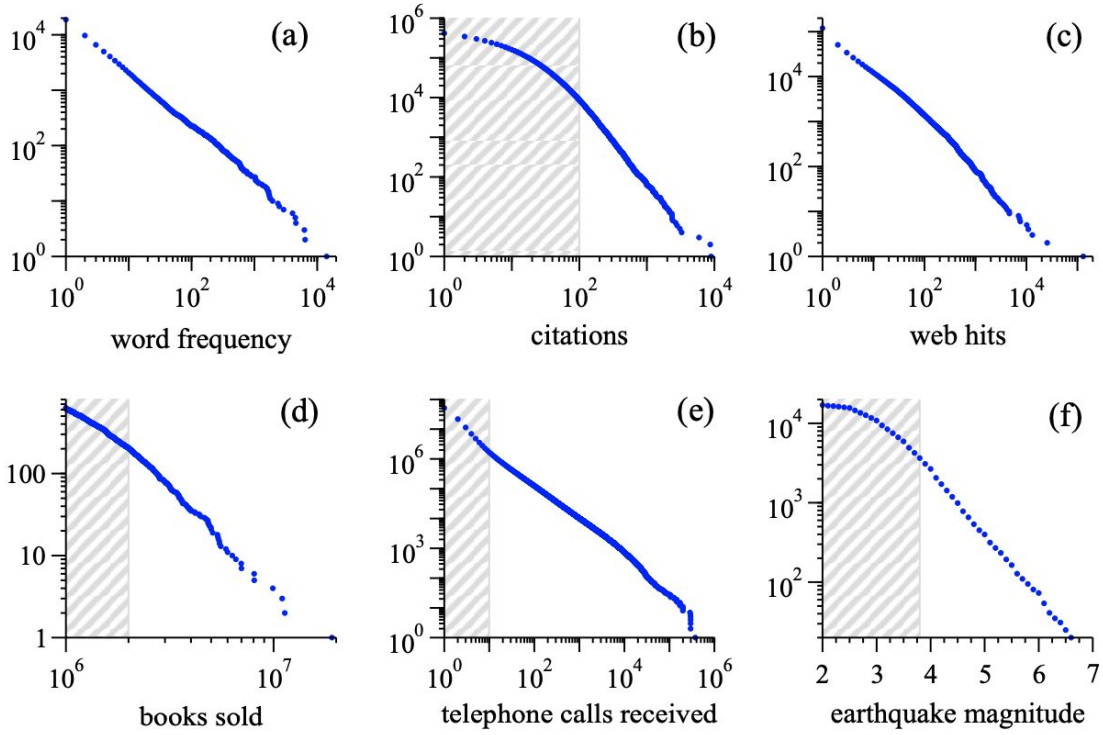
City Population

1	*New York*[c]	NY	8,335,897
2	*Los Angeles*	CA	3,822,238
3	*Chicago*	IL	2,665,039
4	*Houston*	TX	2,302,878
5	***Phoenix***	AZ	1,644,409
6	*Philadelphia*[d]	PA	1,567,258
7	San Antonio	TX	1,472,909
8	San Diego	CA	1,381,162
9	Dallas	TX	1,299,544
10	**Austin**	TX	974,447

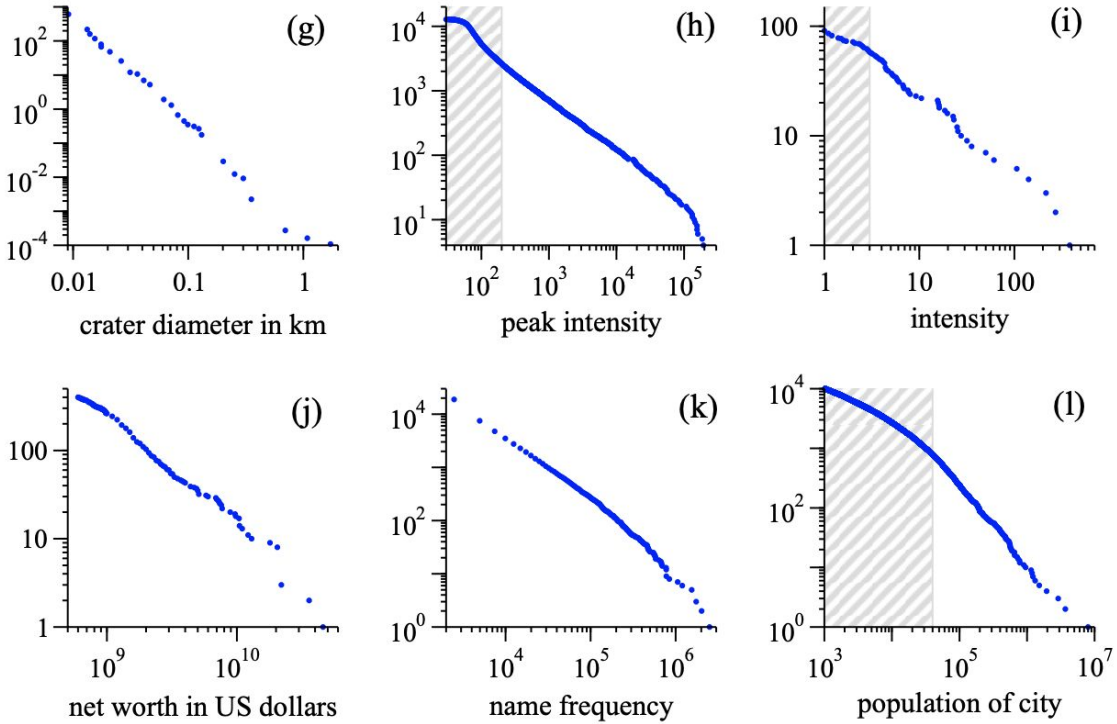



percentage of cities
0.004
0.003
0.002
0.001
0
0
2×10^5
4×10^5
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
10^-8
10^4
10^5
10^6
10^7
population of city

Other Fun Ones

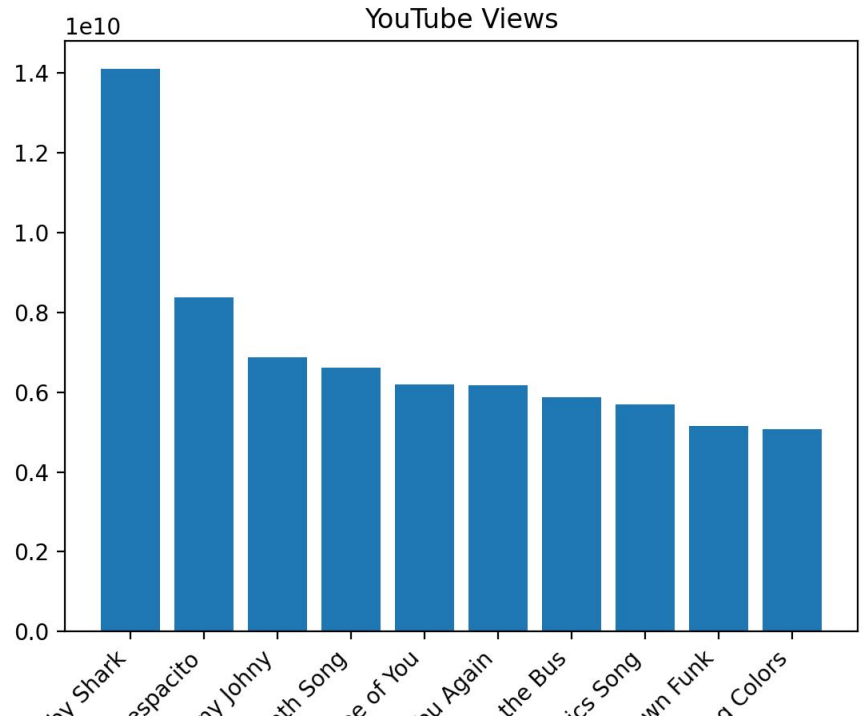


Other Fun Ones



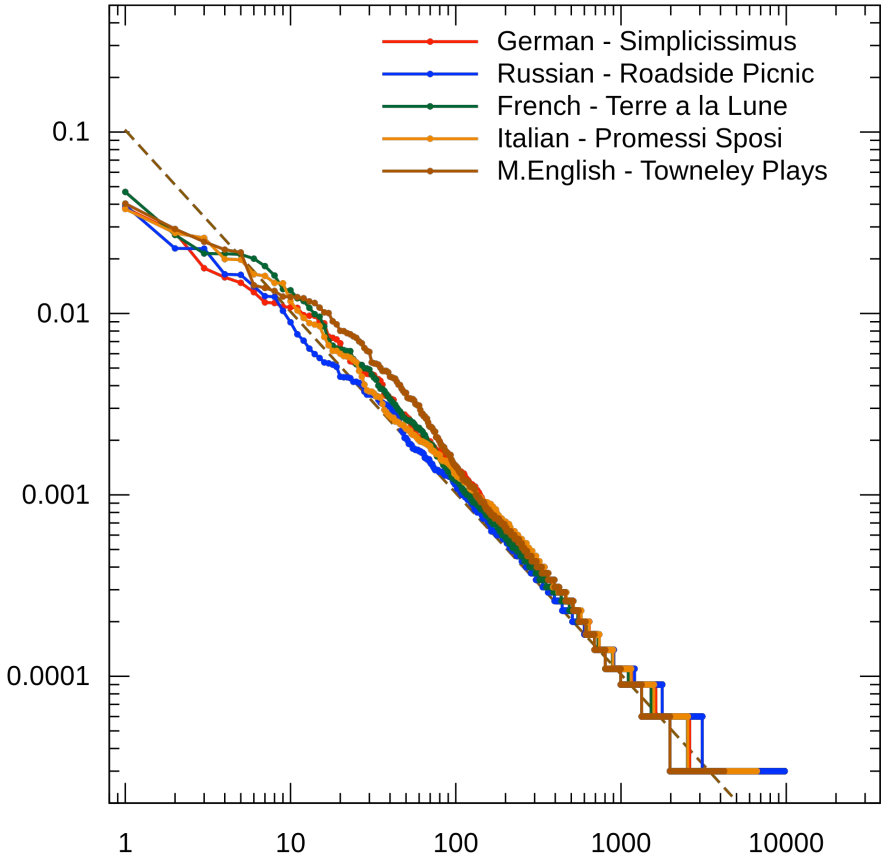
Youtube Views

		Views (billions)
1	"Baby Shark Dance"[6]	14.09
2	"Despacito"[9]	8.38
3	"Johny Johny Yes Papa"[17]	6.87
4	"Bath Song"[18]	6.62
5	"Shape of You"[19]	6.20
6	"See You Again"[22]	6.17
7	"Wheels on the Bus"[27]	5.88
8	"Phonics Song with Two Words"[28]	5.70
9	"Uptown Funk"[29]	5.15
10	"Learning Colors – Colorful Eggs on a Farm"[30]	5.07



Different Languages

Zipf's Law appears in other languages besides English



Alien Language

Lawrence Doyle (UC Davis) has been studying species that are dependent on acoustic communication (dolphins, whale, monkeys).

As it turns out, bottlenose dolphin whistles also obey Zipf's law.

Although we don't know what they're saying, we know they have a communication style that has similar complexities to human language.

Now, Doyle is analyzing microwave telescope data and keeping an eye out for signals that seem to obey Zipf's law.

Random Language

The Zipf Mystery 8:07-11:09

References

Doyle, L. R., & Mao, T. (2016, November 18). *Why alien language would stand out among all the noise of the universe*. Cosmos on Nautilus. https://web.archive.org/web/20200729120031/http://cosmos.nautil.us/feature/54/listening-for-extraterrestrial-blah-blah

Fitzgerald. (2024, February 2). *The Great Gatsby by F. Scott Fitzgerald*. Project Gutenberg. https://www.gutenberg.org/ebooks/64317

The Gettysburg Address. (n.d.). https://rmc.library.cornell.edu/gettysburg/good_cause/transcript.htm

Melville. (2021, August 18). *Moby Dick; or, the whale by Herman Melville*. Project Gutenberg. https://www.gutenberg.org/ebooks/2701

Newman, M. E. J. (2006). (publication). *Power laws, Pareto distributions and Zipf's law*.

Piantadosi, S. T. (2014, October). *Zipf's word frequency law in natural language: A critical review and future directions*. Psychonomic bulletin & review. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4176592/

Rosenberg, Daniel. "Stop, Words." *Representations*, vol. 127, no. 1, 2014, pp. 83–92. *JSTOR*, https://doi.org/10.1525/rep.2014.127.1.83. Accessed 20 Feb. 2024.

The Royal Statistical Society. (2013, December). Who's afraid of George Kingsley Zipf? https://www.ling.upenn.edu/~ycharles/sign708.pdf

Wikimedia Foundation. (2023a, September 11). *Lemma (morphology)*. Wikipedia. https://en.wikipedia.org/wiki/Lemma_(morphology)

Wikimedia Foundation. (2023b, November 9). *George Kingsley Zipf*. Wikipedia. https://en.wikipedia.org/wiki/George_Kingsley_Zipf#cite_note-1

Wikimedia Foundation. (2024a, January 21). *Most common words in English*. Wikipedia. https://en.wikipedia.org/wiki/Most_common_words_in_English

Wikimedia Foundation. (2024b, February 14). *List of United States cities by population*. Wikipedia. https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

Wikimedia Foundation. (2024c, February 19). *List of most-viewed YouTube videos*. Wikipedia. https://en.wikipedia.org/wiki/List_of_most-viewed_YouTube_videos