N-Gram Language Models

LIN 313 Language and Computers

UT Austin Fall 2025

Instructor: Gabriella Chronis

Today

Monday 9/8

- grammars/models
- What is a distribution
- Unigram language models
- Zipf's law

Wednesday 9/10

- Conditional probability (again)
- Bigram language models

Friday 9/12

- building our own language model
- text mashups

What is a Grammar?

What is a Grammar?

Description of a language that ACCEPTS all the utterances that are part of the language and REJECTS all the utterances that are not part of the language

What is a Grammar?

Description of a language that ACCEPTS all the utterances that are part of the language and REJECTS all the utterances that are not part of the language

- "I dwell in possibility." (Emily Dickinson)
- "In dwell possibility I." (Lemonsy Dinicky)

What is a Grammar?

Description of a language that ACCEPTS all the utterances that are part of the language and REJECTS all the utterances that are not part of the language

"I dwell in possibility." (Emily Dickinson)

"In dwell possibility I." (Lemonsy Dinicky)

Mental Grammar

Procedure that PRODUCES all the utterances that are part of the language and DOES NOT PRODUCE any utterance that are not part of the language

What is a *model* of language?

You can think of a language model as the generative version of a grammar.

A language model is a procedure that PRODUCES all the utterances that are part of the language and DOES NOT PRODUCE any utterance that are not part of the language

What are some models of language we have seen so far?

Corpora

In practice, we represent a language as a limited dataset of text called a corpus (plural *corpora*)

Corpora

In practice, we represent a language as a limited dataset of text called a corpus (plural *corpora*)

Today's Corpus:

It's one sentence.

"The cat gave the dog the fig."

Today's Corpus

It's one sentence.

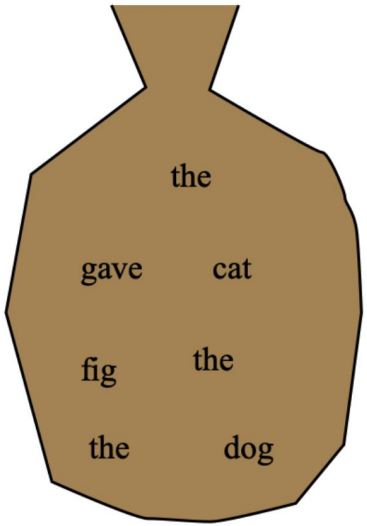
"The cat gave the dog the fig."

Bag of Words

Imagine you write "The cat gave the dog the fig" on a piece of paper, cut the words up and put them in a bag.

Reach into the bag and pull out one of the slips of paper.

What is the probability that the word written on paper is "the"?



Unigram Modeling Steps

1. list all N-grams (for unigram LMs this is equivalent to listing each word token
2. count the frequencies of each N-gram (for unigram LMs this is equivalent to listing each word *type* along with its *token* **frequency count**)
3. (options) visualize in a **histogram**
4. convert to a **probability distribution**
    a. we want to go from raw frequency counts to probabilities
    b. all of our N-gram probabilities should sum to 1!
5. generation
    a. we generate text by **sampling** from the distribution

Probability Distributions

A probability distribution is gives the probabilities of occurrence of **possible outcomes** for an **experiment**. The values always add up to one.

Experiment: An activity whose outcomes are not known

Possible Outcome: The list of all the outcomes in an experiment can be referred to as possible outcomes. In tossing a coin, the possible outcomes are heads or tails.

Sample Space: the set of outcomes of all the trials in an experiment. In tossing a coin, the sample space is the set {H, T}

Language as a distribution

How might we break up a string of text into discrete **events?**

Language as a distribution

How might we break up a string of text into discrete **events?**

- words
- characters
- sub-word "tokens" (this is what LLMs do)

Tokenization

The first step in language processing is **tokenization**. Our corpus starts out as one big long string of characters.

Tokenization is how we split the corpus into discrete **events.** We will use word tokens today.

Type-Token Distinction

type: a word in the abstract

token: any given occurrence of that word in speech or text.

A corpus has fewer types than tokens, because many/most words occur more than once.

Distributions

Distributions are also models

- We can make assumptions about how our data is distributed
- Once we have a distribution, we can "draw" or "simulate" from it, i.e use it to generate new data
- This is what a language model does when it predicts!

Unigram Distribution

The next simplest thing is to hypothesize that the language represented by our corpus is generated by a unigram distribution.

The unigram distribution assumes all words are independent events. The probability of a word is equal to its **observed probability** in the corpus.

The first step in building the unigram probability distribution is counting the token frequencies of each word type.

Histogram

A **histogram** is the way to visualize observed frequencies of events. Types go on the x-axis, and frequency counts on the y-axis.

Trials and Experiments

Experiment: An activity whose outcomes are not known

Trial: The numerous attempts in the process of an experiment. In other words, any particular performance of a random experiment is called a trial. For example, tossing a coin is a trial.

Outcome: This is the result of a trial.

Event: A trial with a clearly defined outcome is an event. For example, getting a tail when tossing a coin is termed as an event.

Uniform Language Model

Let's make a reaaally simple model based off of our corpus. We give all events (words) an equal probability.

https://gchronis.github.io/dice-roller.html

Unigram Language Model

Let's make slightly more informed but still reaaally dumb model based off of our corpus. We give all events (words) a probability proportional to their frequency in the corpus.

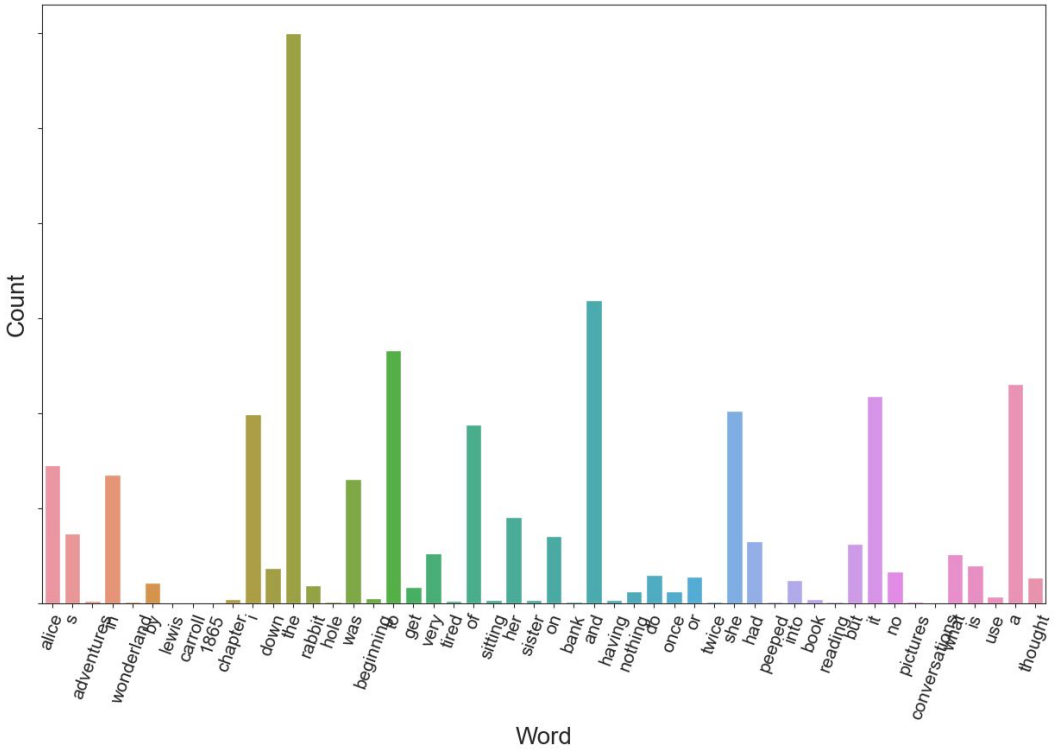
https://gchronis.github.io/dice-roller.html

Law of Large Numbers

Jakob Bernoulli (*Ars Conjectandi, 1713)*

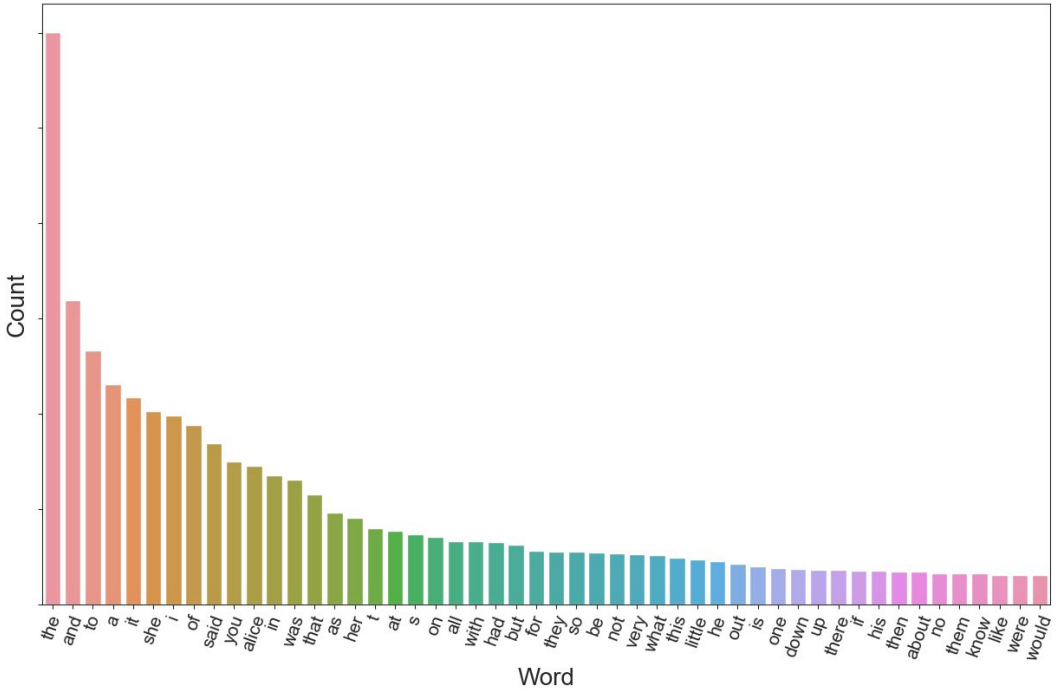
The average of the results obtained from a large number of independent random samples converges to the true value, if it exists.

Let's take a look at some real world distributions



Alice's Adventures in Wonderland

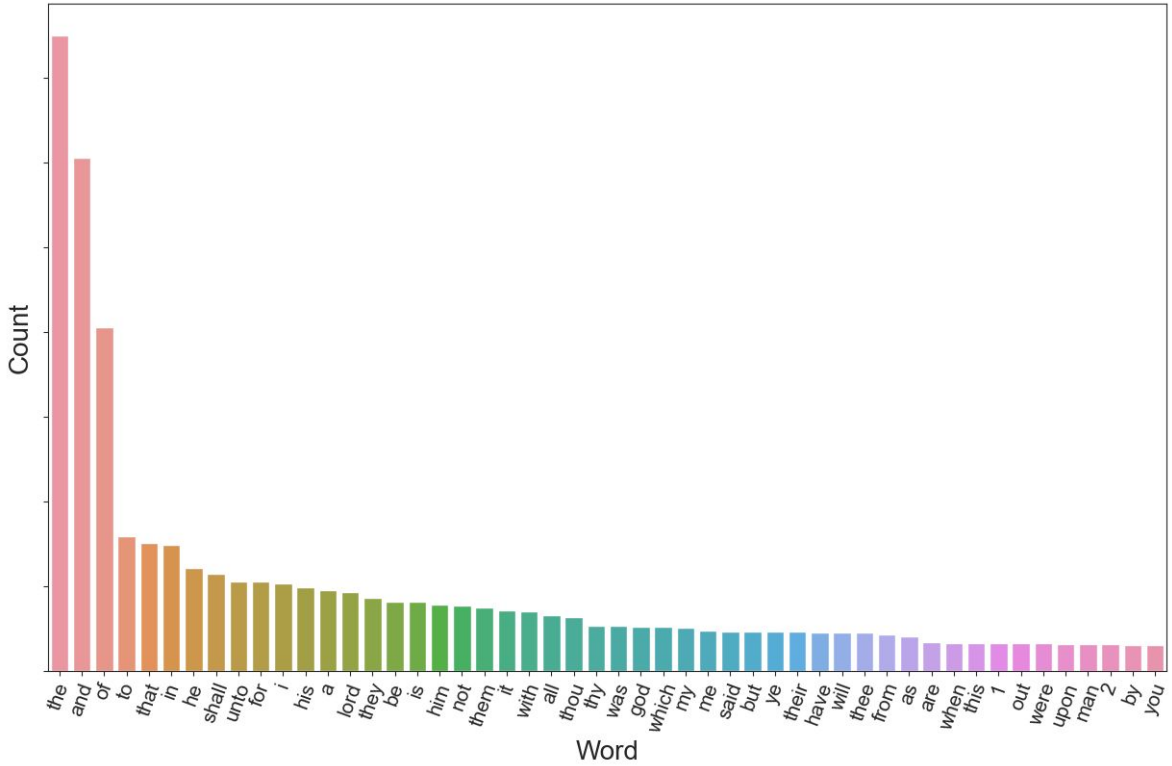
Let's take a look at some real world distributions



Alice's Adventures in Wonderland

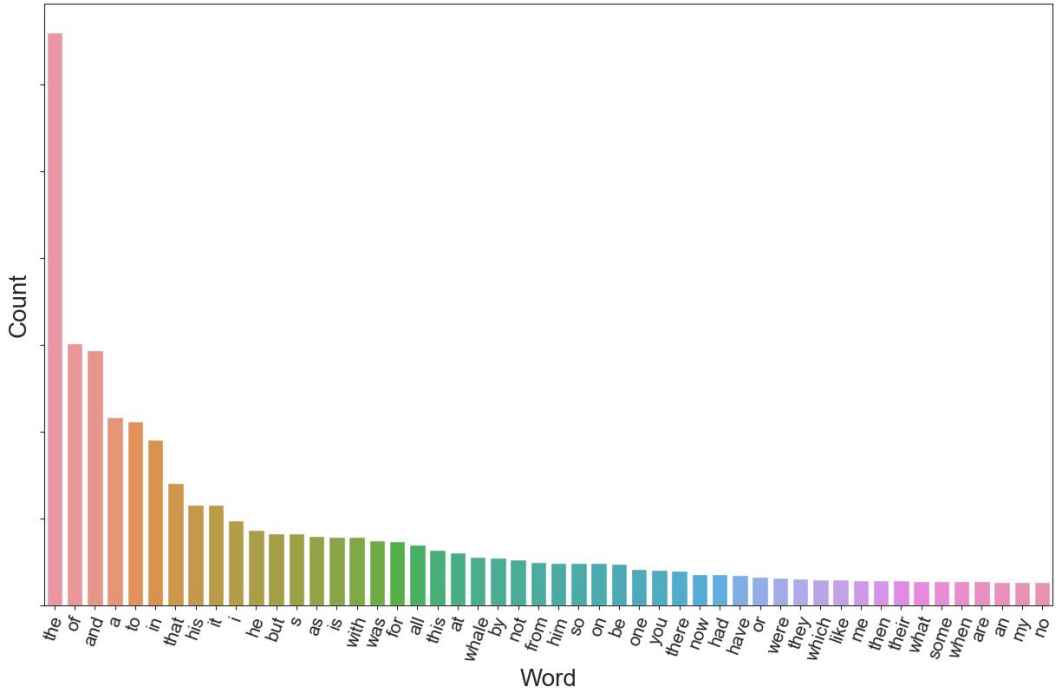
and when we order by frequency

Let's take a look at some real world distributions



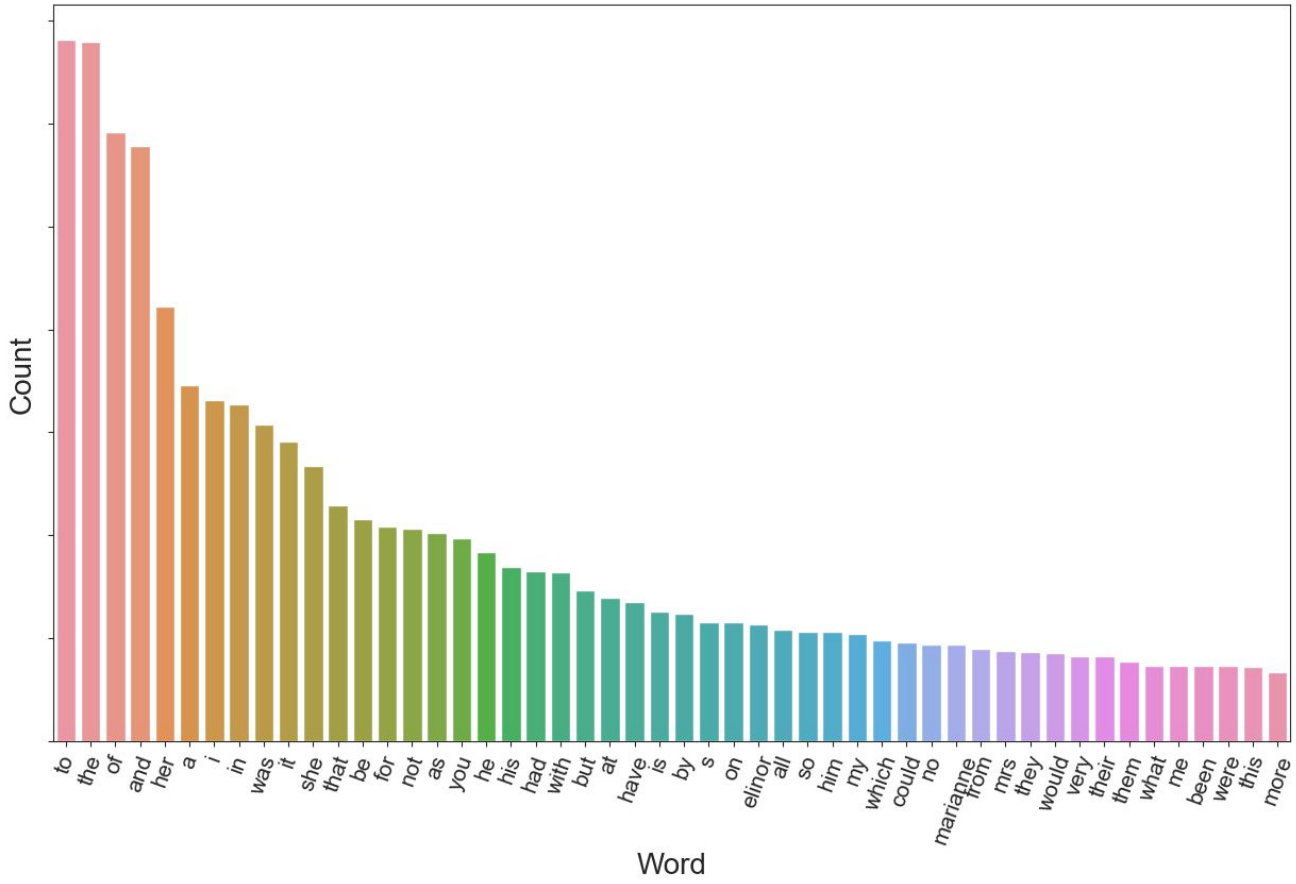
King James Bible

Let's take a look at some real world distributions



Herman Melville, *Moby Dick*

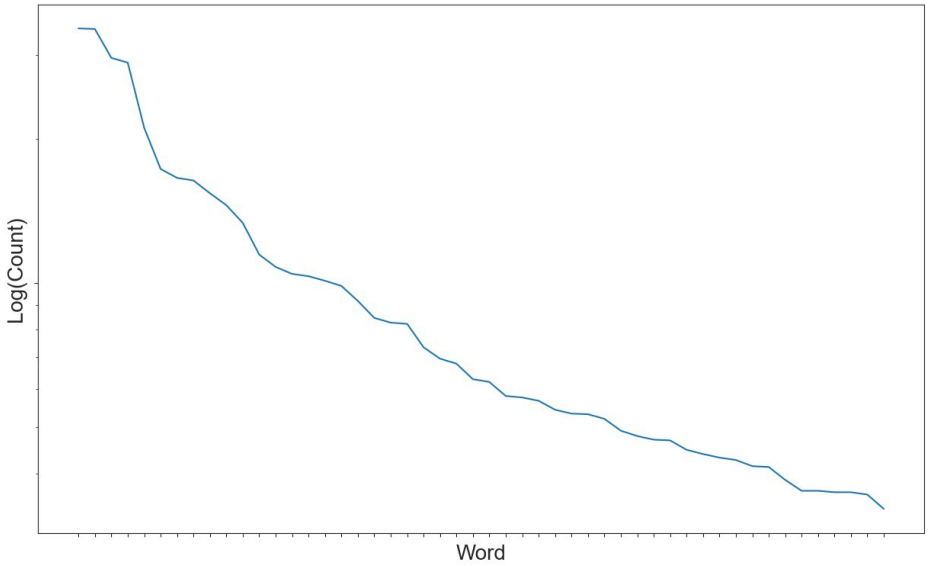
Let's take a look at some real world distributions



Jane Austen,
Sense and Sensibility

That graph looks exponential

When you plot on a logarithmic scale, it turns into a (more) straight(er) line



Alice's Adventures in Wonderland

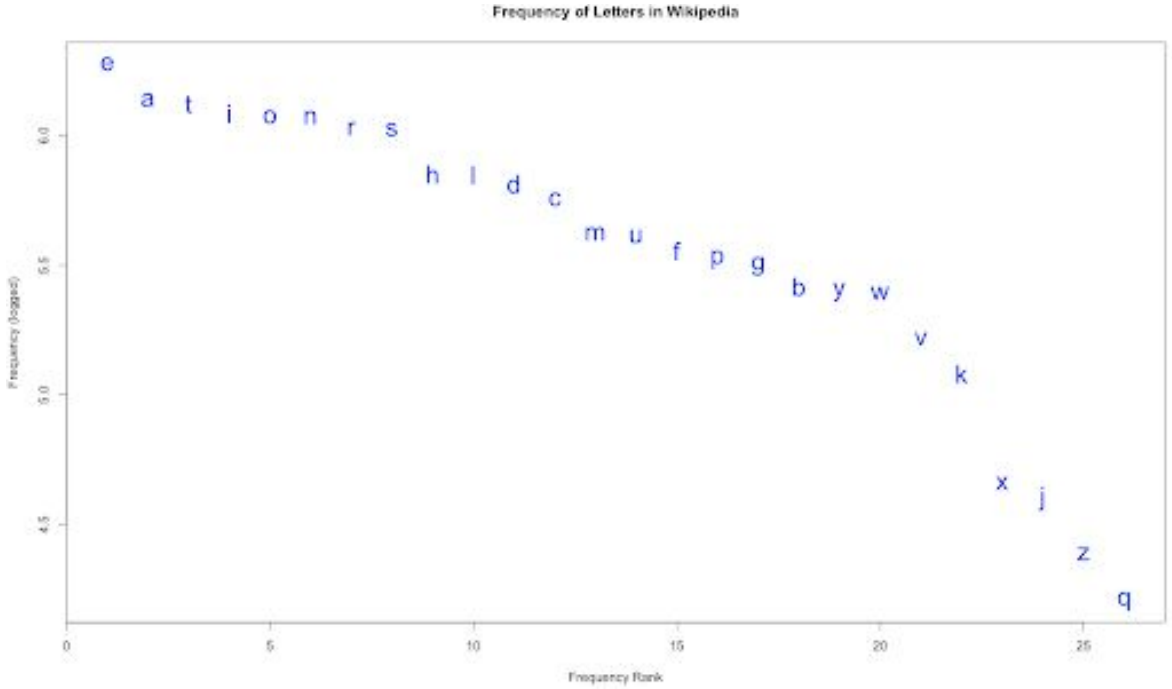
Zipf's law

$$P_n \sim 1/n^a$$

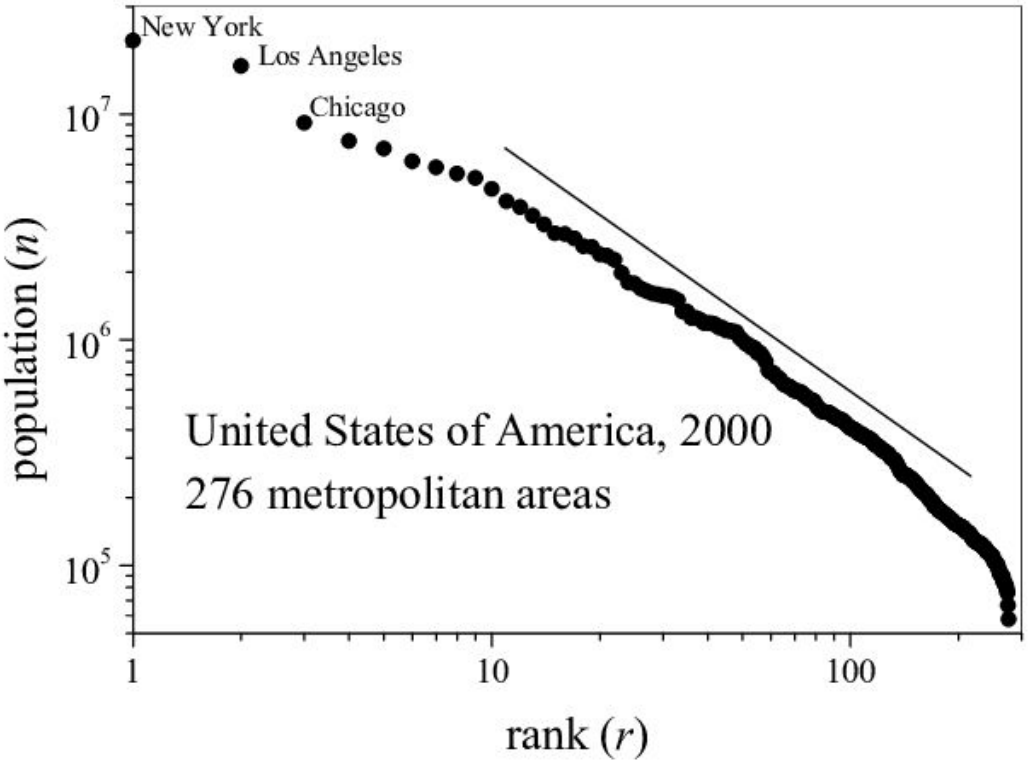
where P is the frequency of a word ranked nth and the exponent a is almost 1. This means that the second item occurs approximately 1/2 as often as the first, and the third item 1/3 as often as the first, and so on.

Given a large corpus of natural language occurrences, the **frequency** of any word is **inversely proportional** to its **rank** in frequency table.

Letters in Wikipedia



United States Cities



Company Size

REPORTS

Zipf Distribution of U.S. Firm Sizes

Robert L. Axtell

Analyses of firm sizes have historically used data that included limited samples of small firms, data typically described by lognormal distributions. Using data on the entire population of tax-paying firms in the United States, I show here that the Zipf distribution characterizes firm sizes: the probability a firm is larger than size s is inversely proportional to s. These results hold for data from multiple years and for various definitions of firm size.