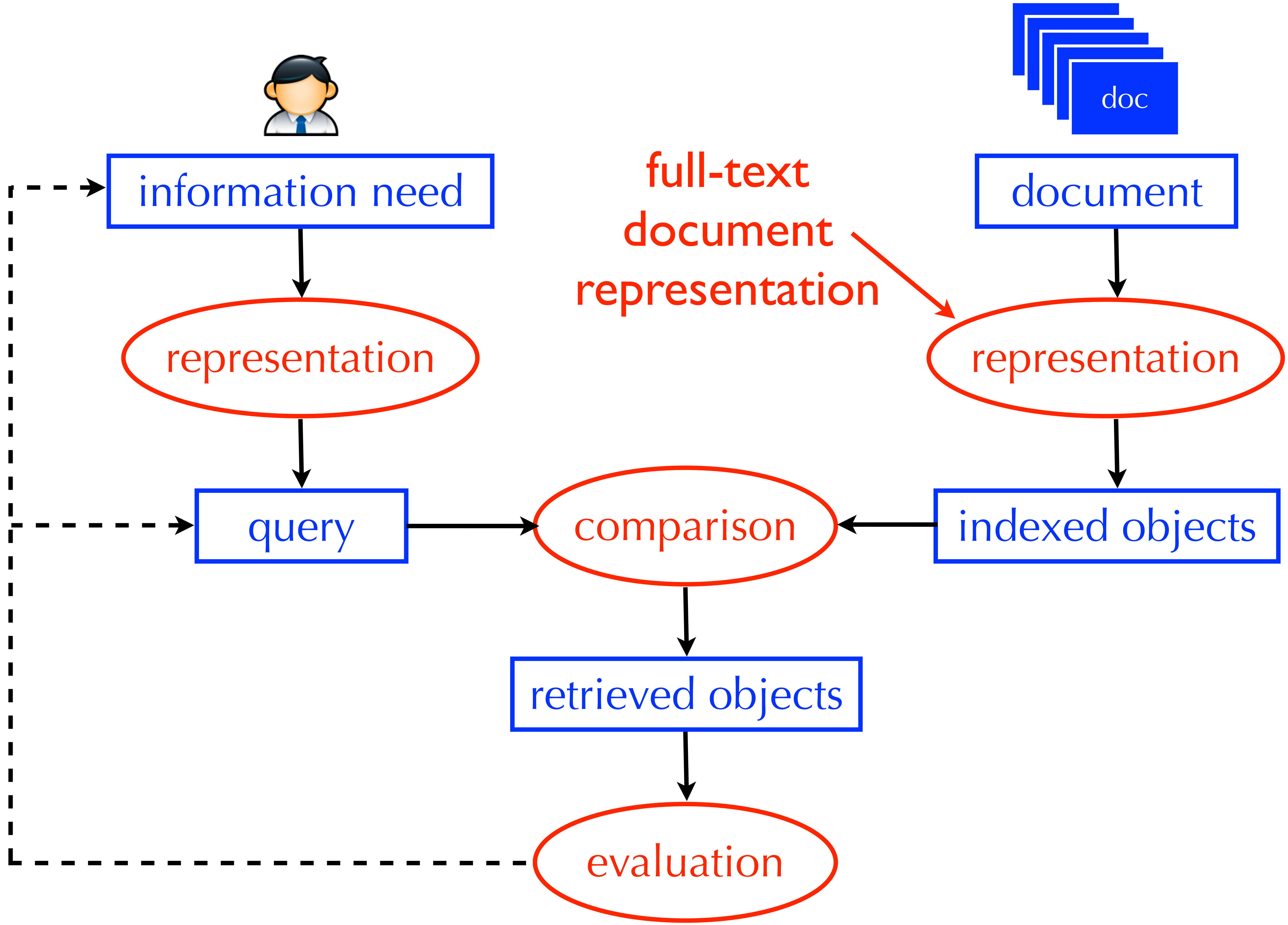
Statistical Properties of Text

Jaime Arguello
INLS 509: Information Retrieval
jarguell@email.unc.edu

September 11, 2013

The Basic IR Process



Text-Processing

<p>Gerard Salton (8 March 1927 in Nuremberg - 28 August 1995), also known as Gerry Salton, was a Professor of Computer Science at Cornell University. Salton was perhaps the leading computer scientist working in the field of information retrieval during his time. His group at Cornell developed the SMART Information Retrieval System, which he initiated when he was at Harvard.</p>

- Mark-up removal
- Down-casing
- Tokenization

Text-Processing

gerard salton 8 march 1978 in nuremberg 28 august 1995 also know as gerry salton was professor of computer science at cornell university salton was perhaps the leading computer scientist working in the field of information retrieval during his time his group at cornell developed the smart information retrieval system which he initiated when he was at harvard

- Our goal is to describe content using content
- Are all these words equally descriptive?
- What are the most descriptive words?
- How might a computer identify these?

Statistical Properties of Text

- We know that language use if very varied
- There are <u>many</u> ways to convey the same information (which makes IR difficult)
- But, are there statistical properties of word usage that are predictable? Across languages? Across modalities? Across genres?

IMDB Corpus

internet movie database

- Each document corresponds to a movie, a plot description, and a list of artists and their roles
    - number of documents: **230,721**
    - number of term occurrences (tokens): **36,989,629**
    - number of unique terms (token-types): **424,035**

http://www.imdb.com/

IMDB Corpus

term-frequencies

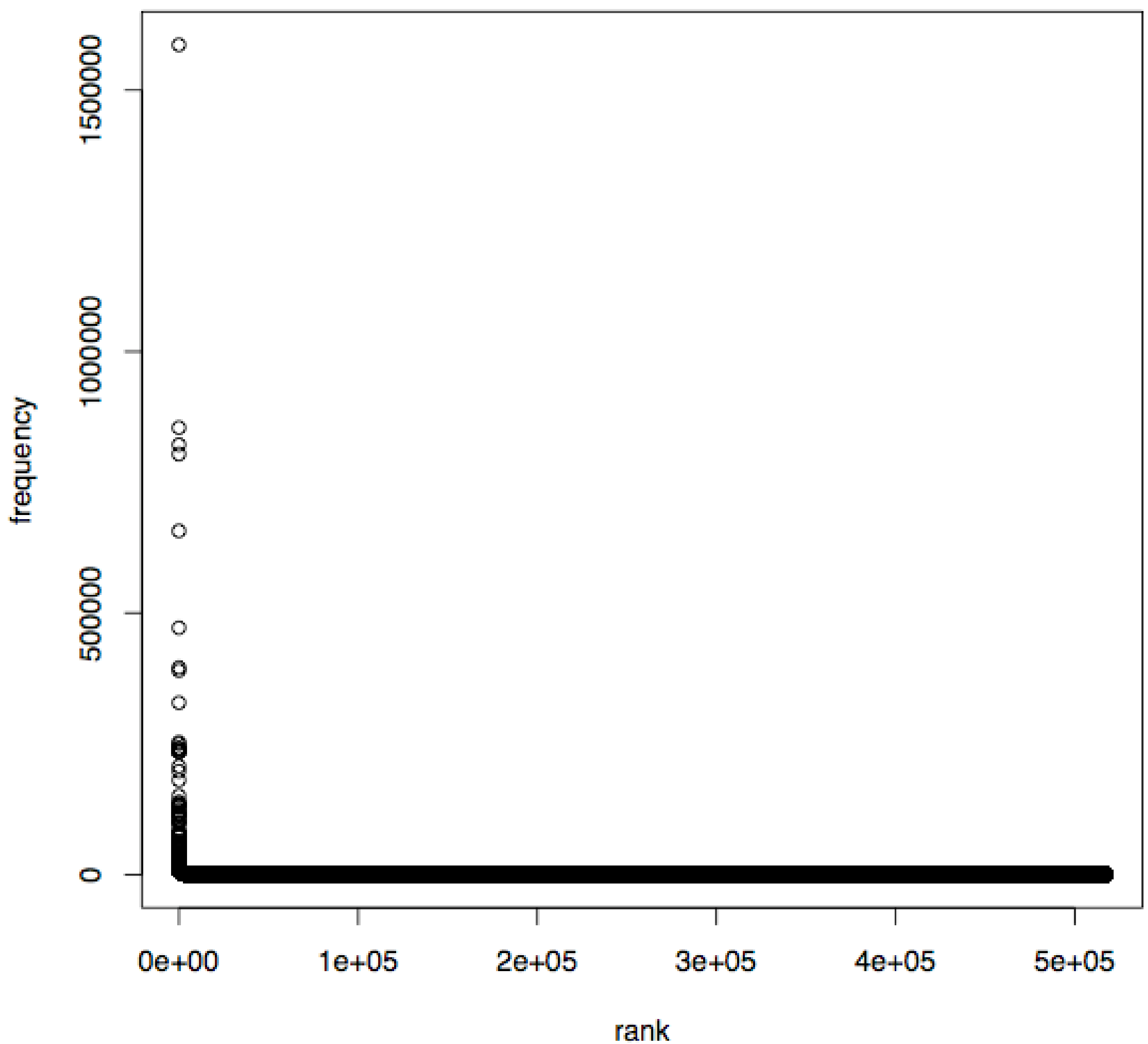
rank	term	frequency	rank	term	frequency
1	the	1586358	11	year	250151
2	a	854437	12	he	242508
3	and	822091	13	movie	241551
4	to	804137	14	her	240448
5	of	657059	15	artist	236286
6	in	472059	16	character	234754
7	is	395968	17	cast	234202
8	i	390282	18	plot	234189
9	his	328877	19	for	207319
10	with	253153	20	that	197723

IMDB Corpus

term-frequencies

rank	term	frequency	rank	term	frequency
21	on	180760	31	their	116803
22	as	150721	32	they	116113
23	by	138580	33	has	113336
24	himself	138214	34	him	112589
25	but	134017	35	when	106723
26	she	132237	36	I	100475
27	who	132151	37	are	99544
28	an	129717	38	it	98455
29	from	122086	39	man	87115
30	at	118190	40	ii	80583

IMDB Corpus
term-frequencies



term frequency decreases rapidly as a function of rank!

George Kingsley Zipf

Zipf's Law

- Term-frequency decreases rapidly as a function of rank
- How rapidly?
- Zipf's Law:

$$f_t = \frac{k}{r_t}$$

- $\mathbf{f_t}$ = frequency (number of times term **t** occurs)
- $\mathbf{r_t}$ = frequency-based rank of term **t**
- **k** = constant
- To gain more intuition, let's divide both sides by **N**, the total term-occurrences in the collection

Zipf's Law

$$\frac{1}{N} \times f_t = \frac{1}{N} \times \frac{k}{r_t}$$

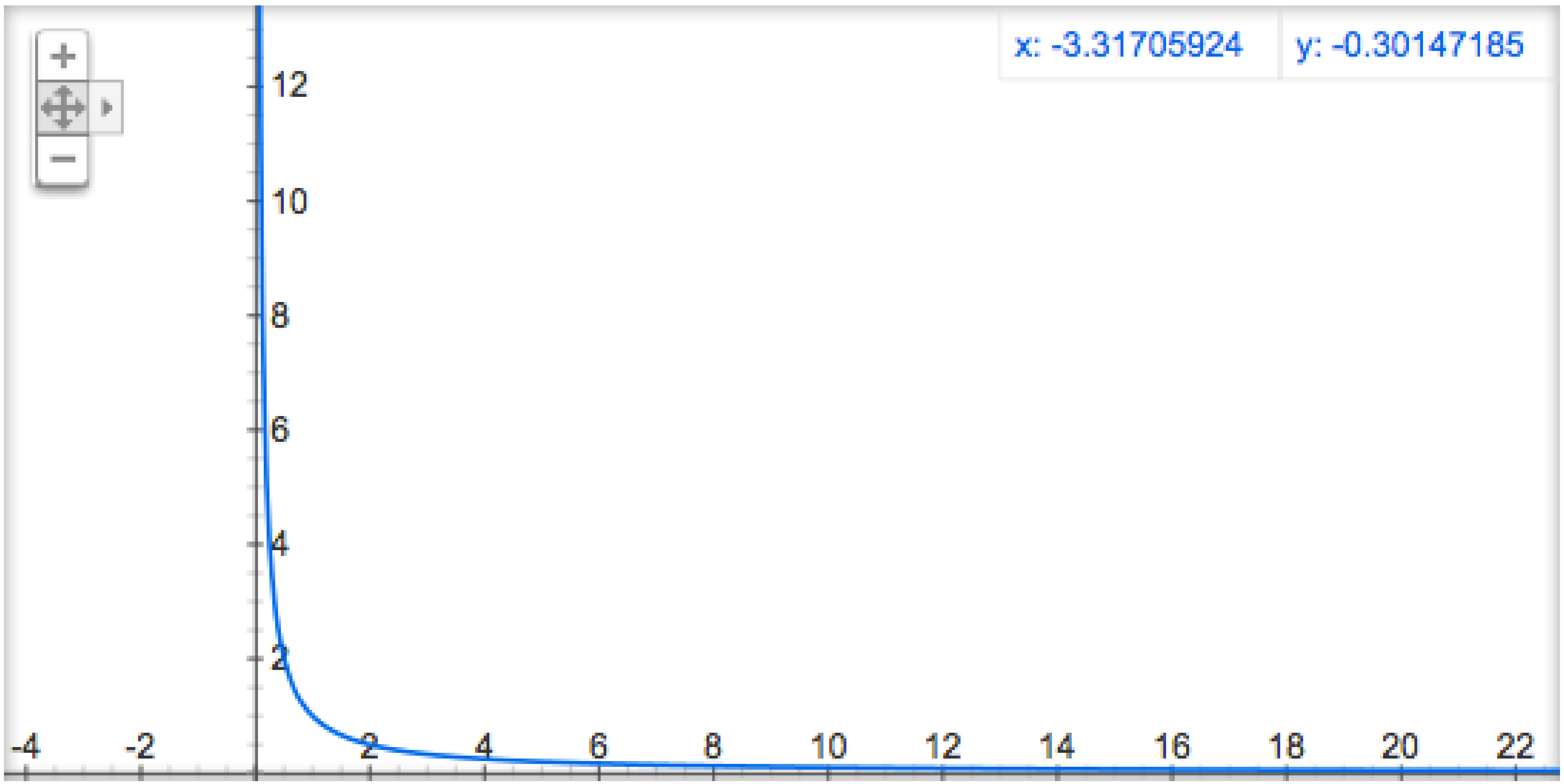
$$P_t = \frac{c}{r_t}$$

- **P_t** = proportion of the collection corresponding to term **t**
- **c** = constant
- For English **c** = 0.1 (more or less)
- What does this mean?

Zipf's Law

$$P_t = \frac{c}{r_t}$$

Graph for 1/x



Zipf's Law

$$P_t = \frac{c}{r_t}$$

c = 0.1

- The most frequent term accounts for 10% of the text
- The second most frequent term accounts for 5%
- The third most frequent term accounts for about 3%
- Together, the top 10 account for about 30%
- Together, the top 20 account for about 36%
- Together, the top 50 account for about 45%
  - that's nearly half the text!
- What else does Zipf's law tell us?

Zipf's Law

- With some crafty manipulation, it also tells us that the faction of terms that occur **n** times is given by:

$$\frac{1}{n(n+1)}$$

- So, what fraction of the terms occur only once?

Zipf's Law

- With some crafty manipulation, it also tells us that the faction of terms that occur **n** times is given by:

$$\frac{1}{n(n+1)}$$

- About half the terms occur only once!
- About 75% of the terms occur 3 times or less!
- About 83% of the terms occur 5 times or less!
- About 90% of the terms occur 10 times or less!

Zipf's Law

- Note: the fraction of terms that occur **n** times or less is given by:

$$\sum_{i}^{n} \frac{1}{i(i+1)}$$

- That is, we have to add the fraction of terms that appear 1, 2, 3, ... up to **n** times

Verifying Zipf's Law

visualization

Zipf's Law $$f = \frac{k}{r}$$

... still Zipf's Law $$\log(f) = \log(\frac{k}{r})$$

... still Zipf's Law $$\log(f) = \log(k) - \log(r)$$

- So, Zipf's law holds, what would we see if we plotted **log(f)** vs. **log(r)**?

Verifying Zipf's Law
visualization

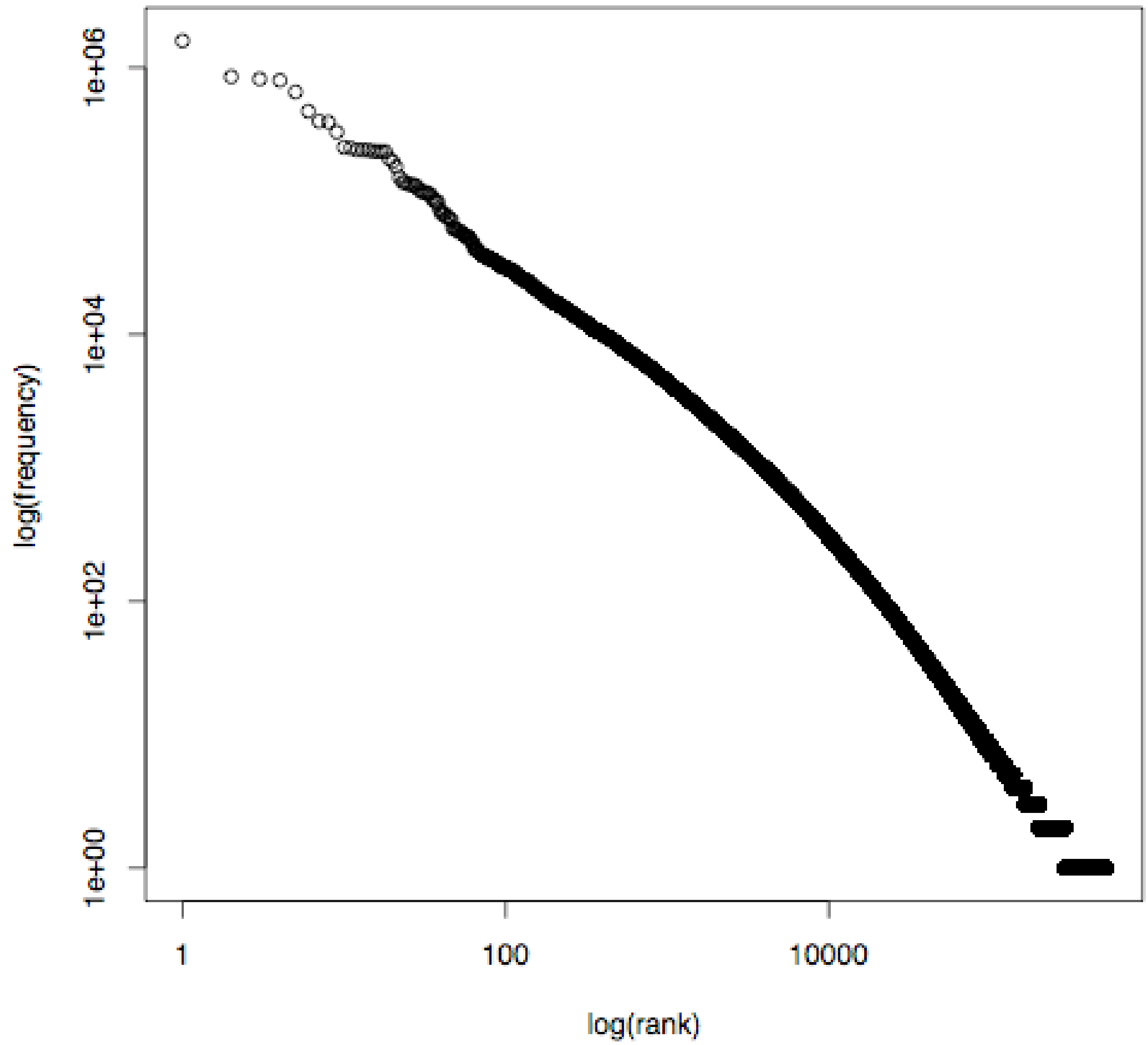
Zipf's Law $$f = \frac{k}{r}$$

... still Zipf's Law $$\log(f) = \log(\frac{k}{r})$$

... still Zipf's Law $$\log(f) = \log(k) - \log(r)$$

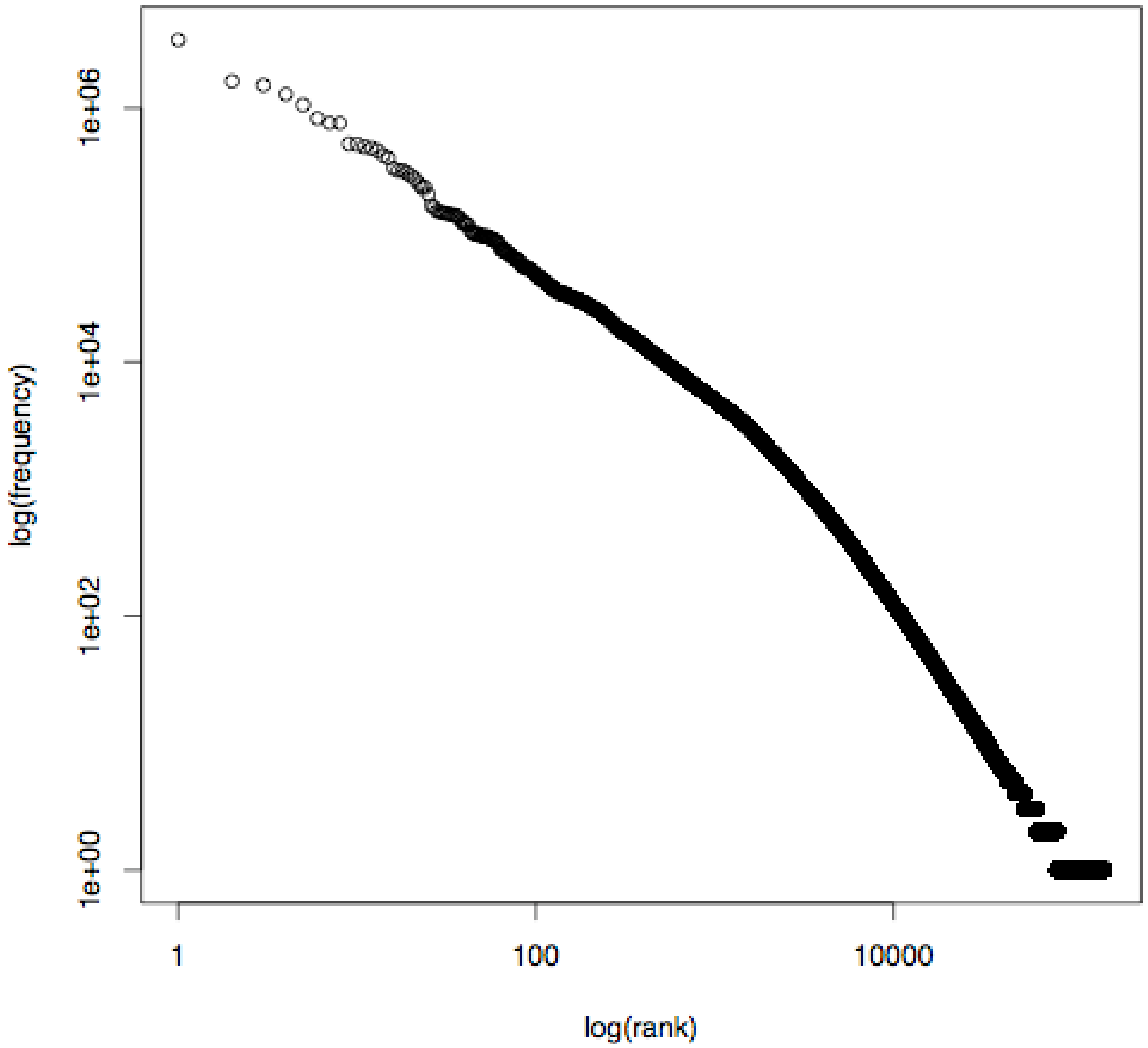
- If Zipf's law holds true, we should be able to plot **log(f)** vs. **log(r)** and see a straight light with a slope of -1

Zipf's Law
IMDB Corpus



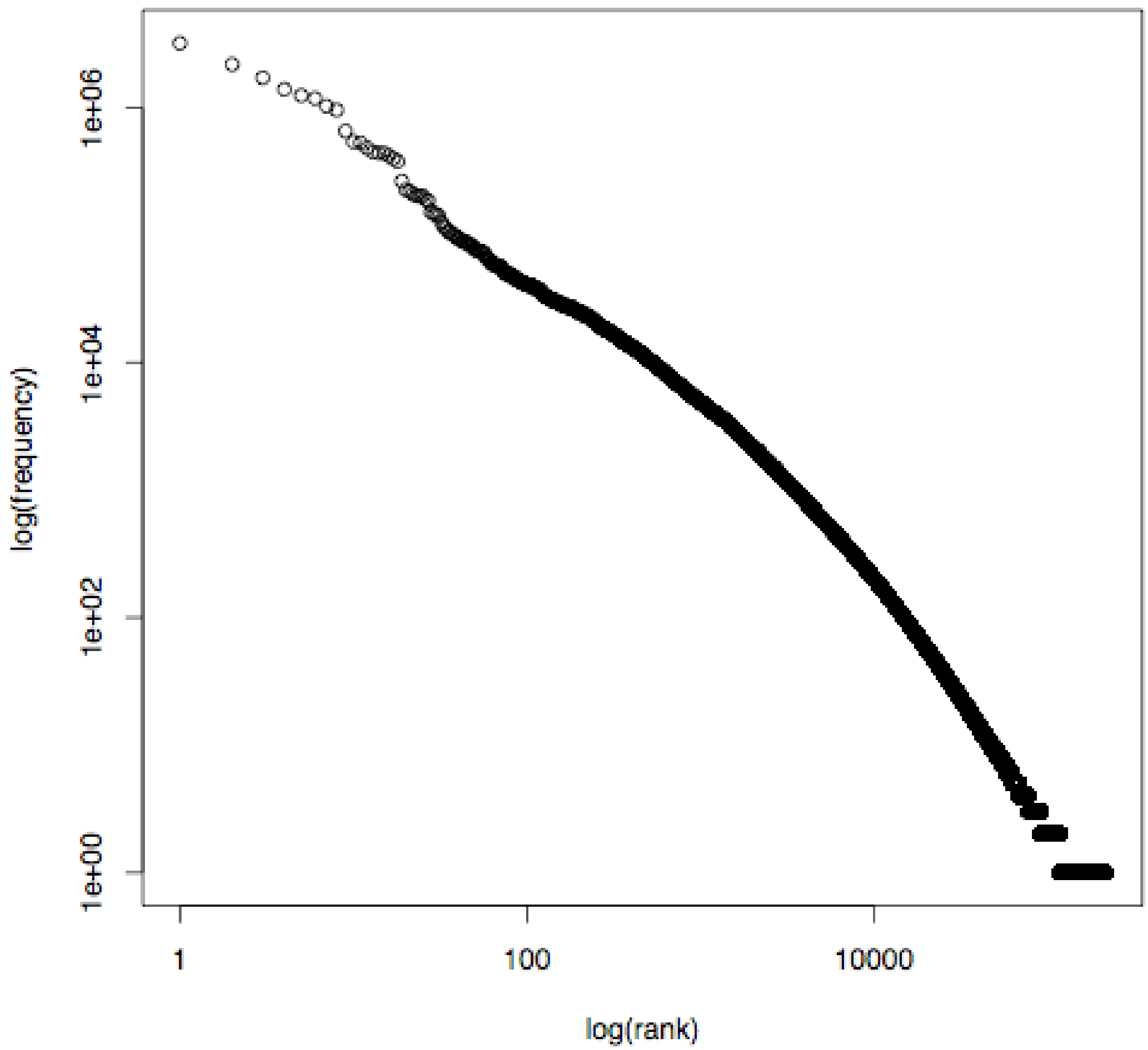
Does Zipf's Law generalize across languages?

Zipf's Law
European Parliament: English

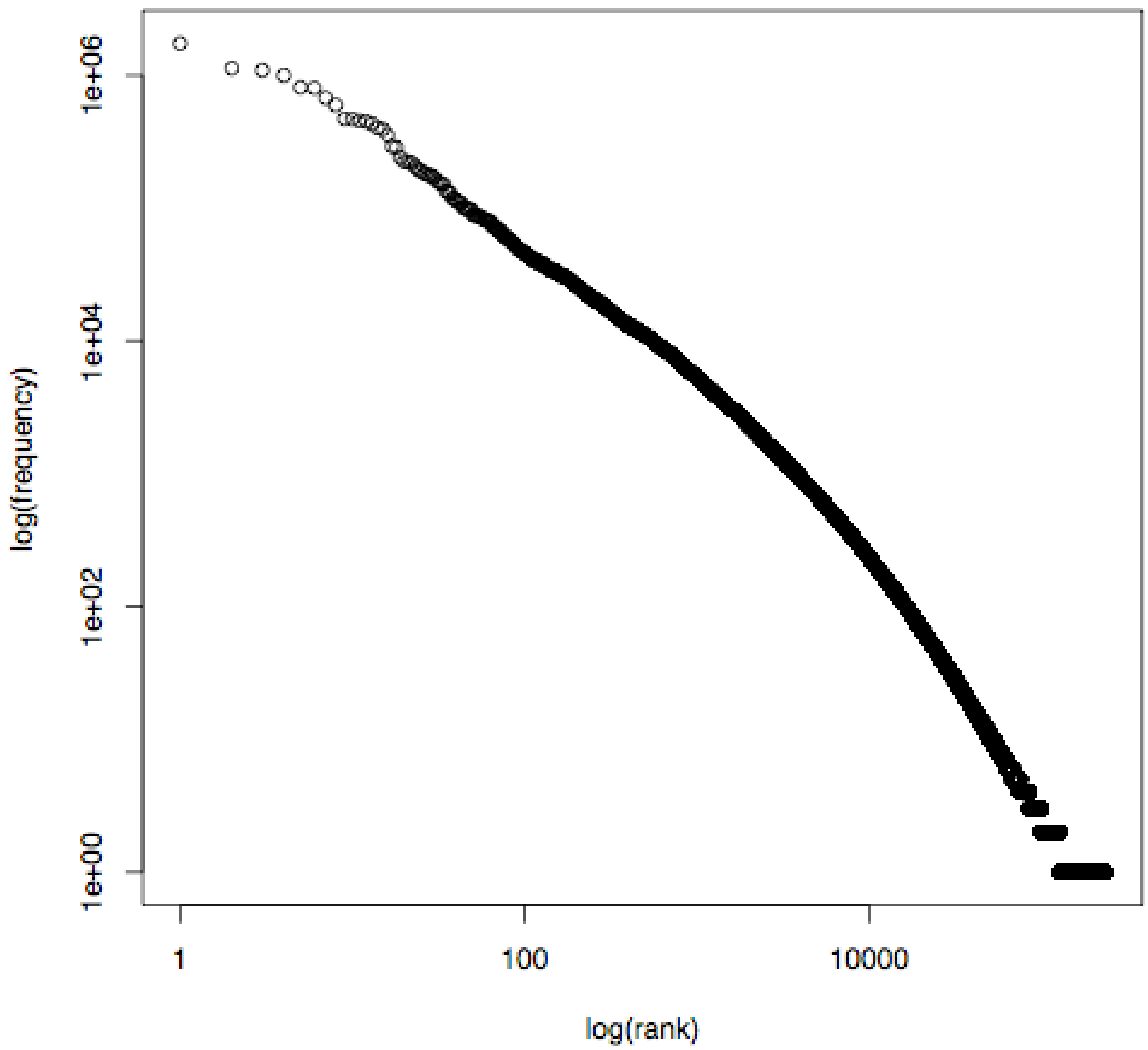


- Transcribed speech from proceedings of the European Parliament **(Koehn '05)**

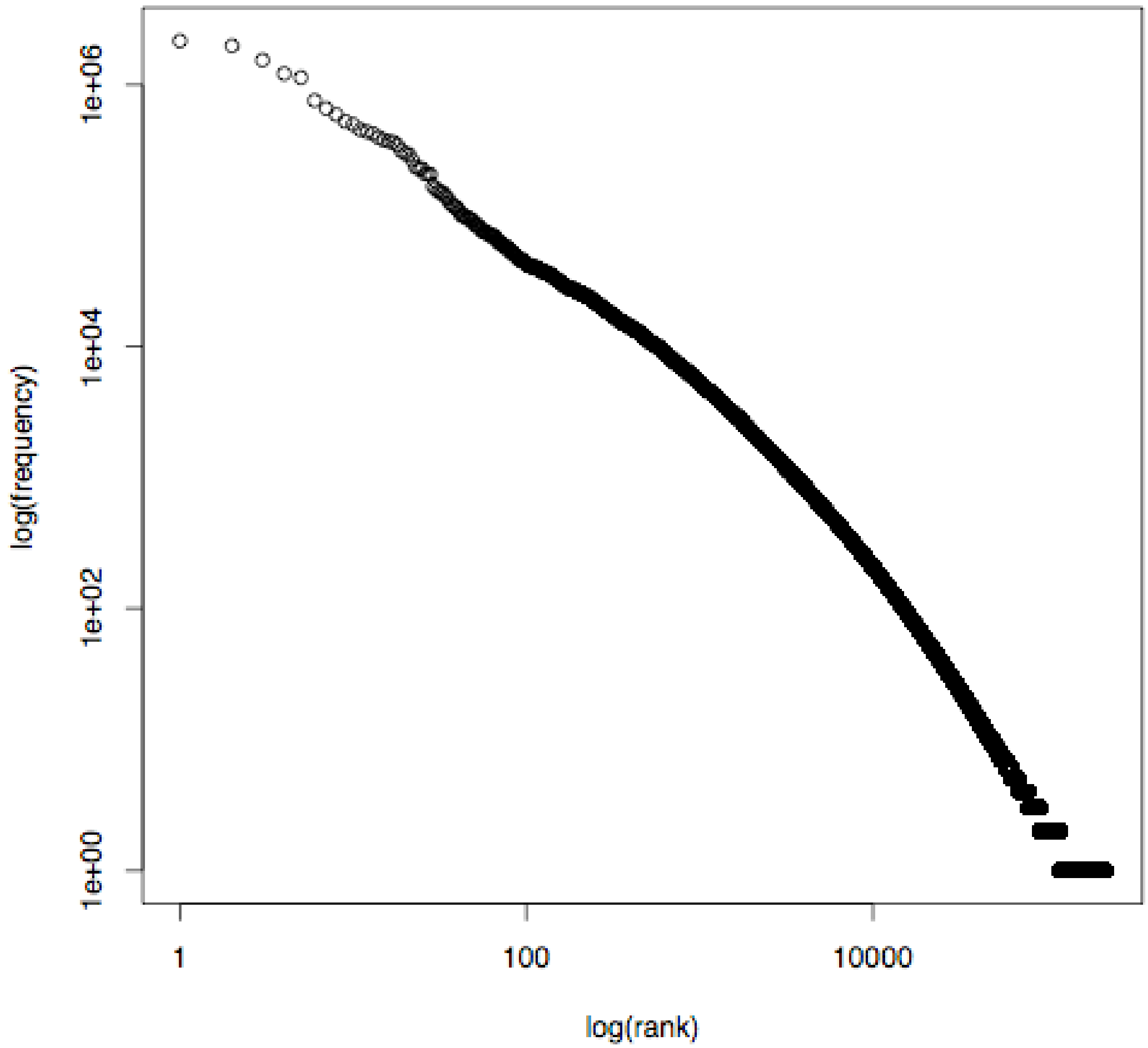
Zipf's Law
European Parliament: Spanish



Zipf's Law
European Parliament: Italian

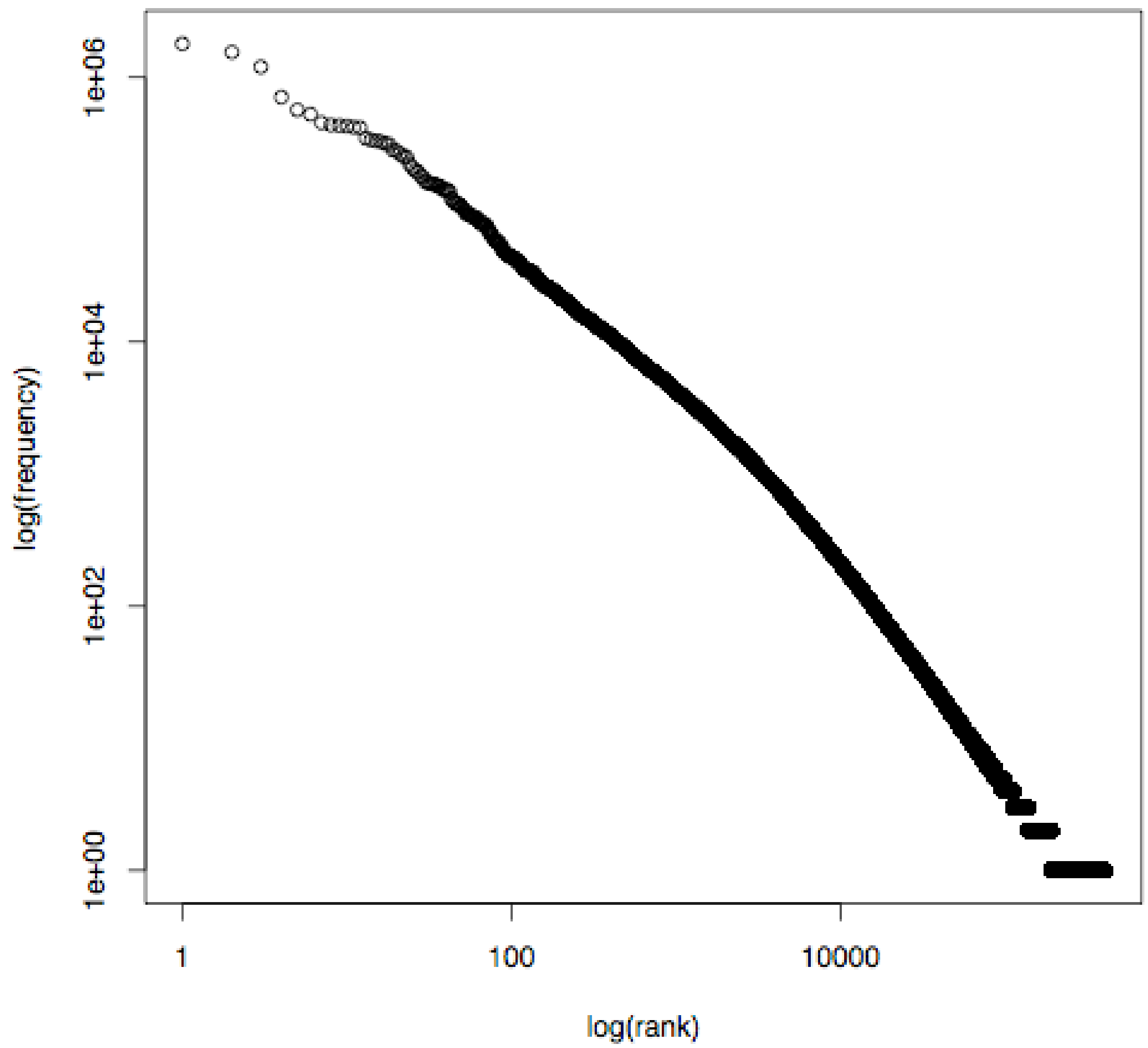


Zipf's Law
European Parliament: Portuguese



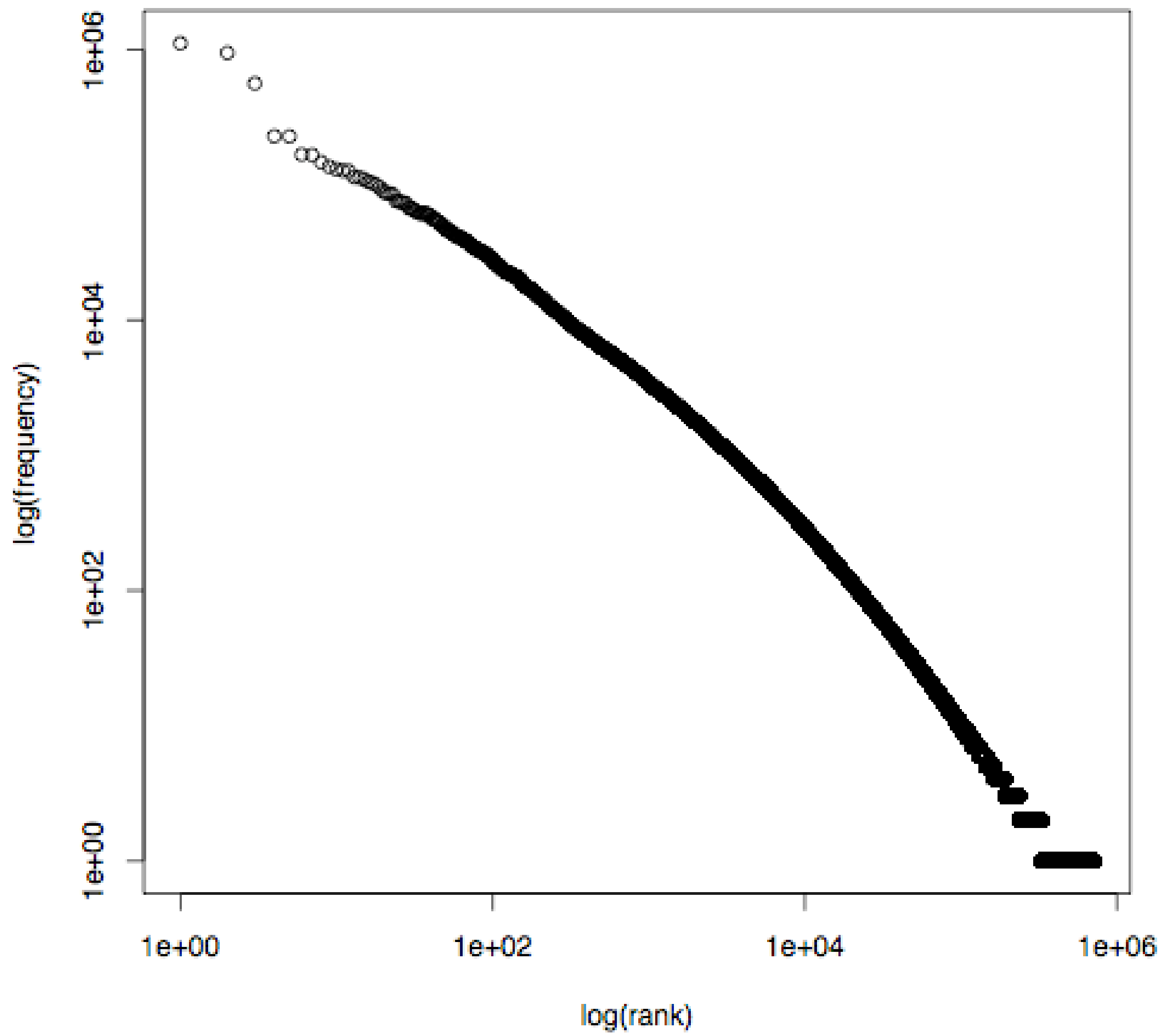
Zipf's Law

European Parliament: German



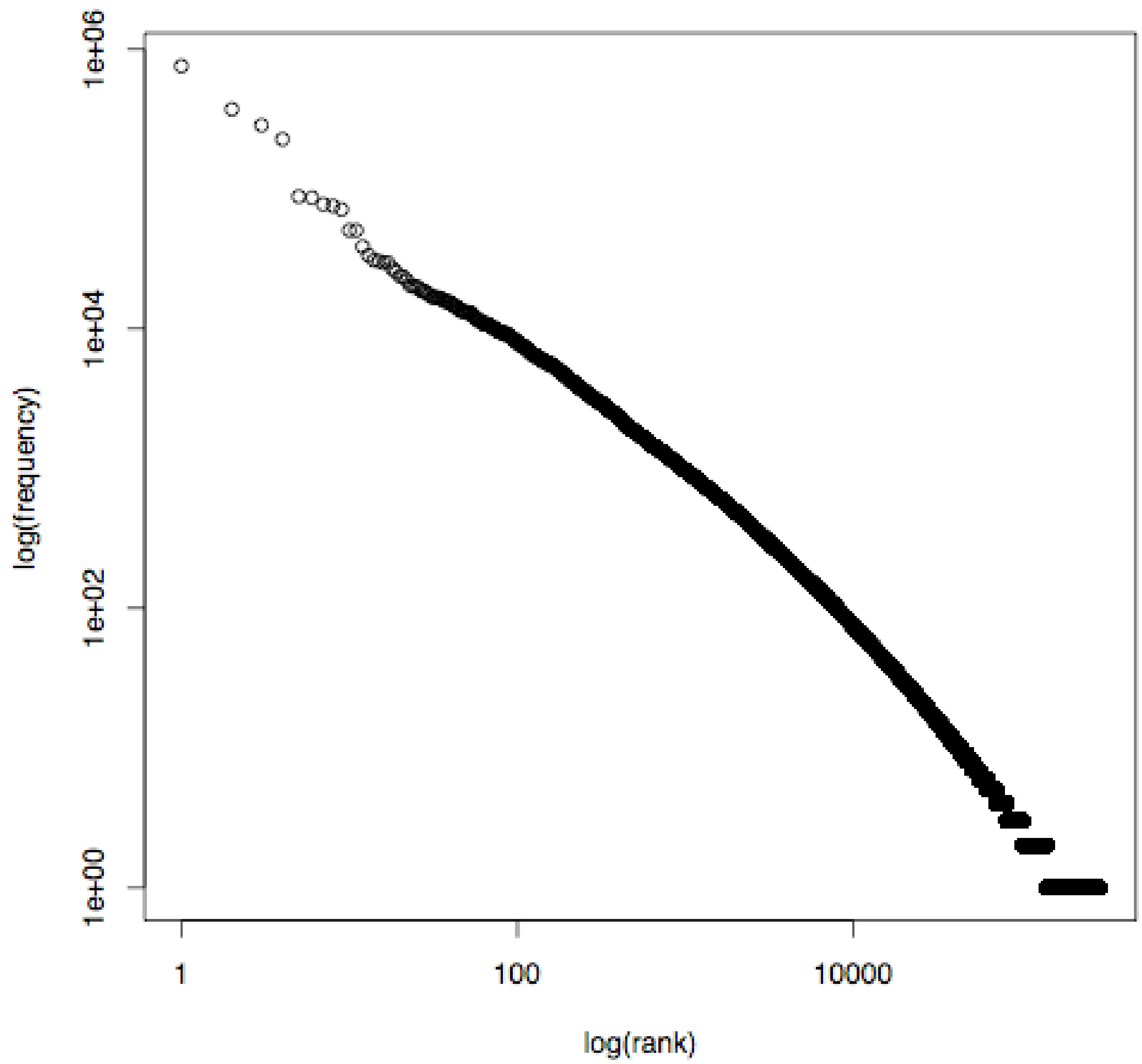
Zipf's Law

European Parliament: Finnish



Zipf's Law

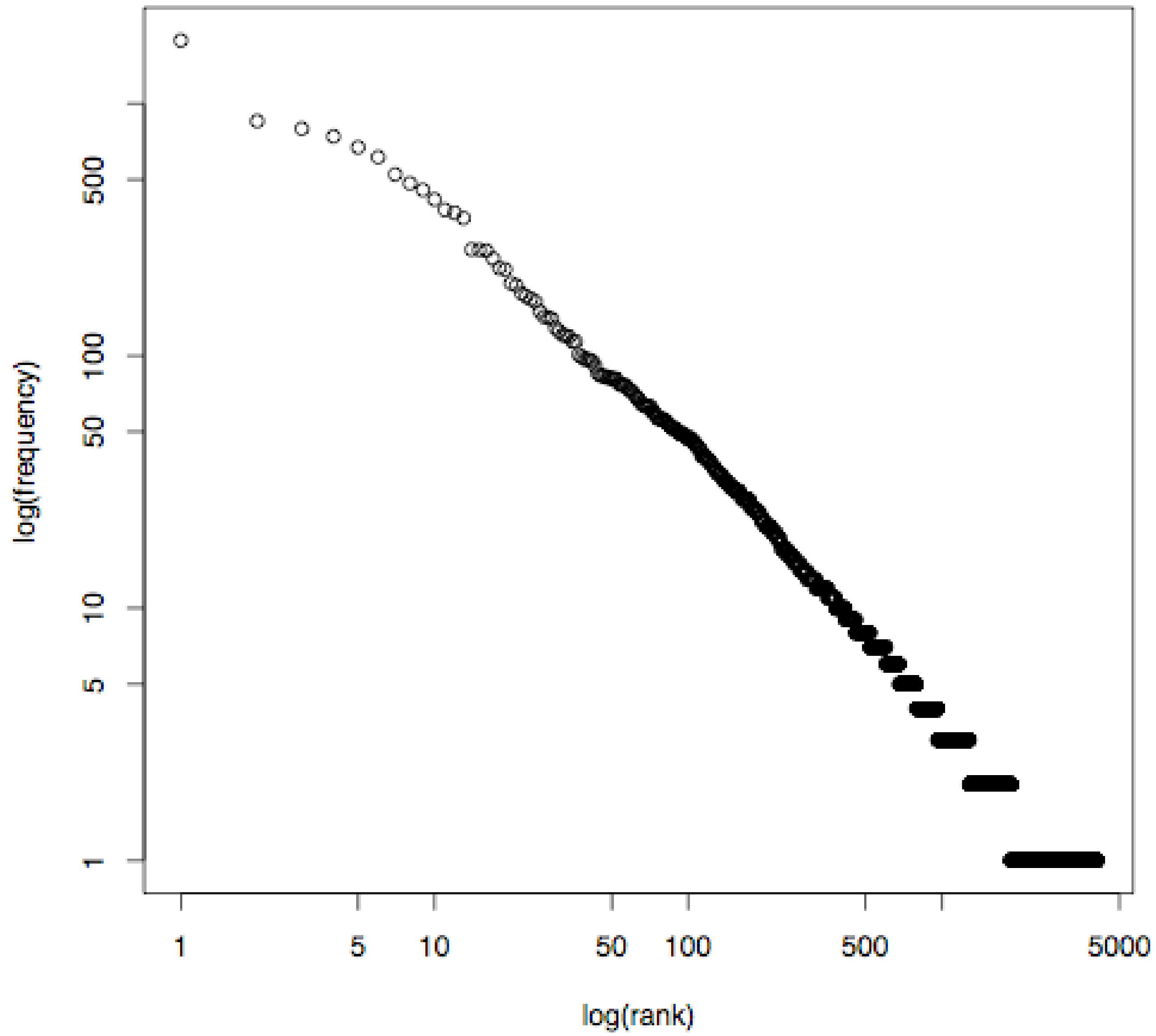
European Parliament: Hungarian



Yes, but these texts are translations of the same content!

What about <u>different</u> texts?
different topics?
different genres?
different sizes?
different complexity?

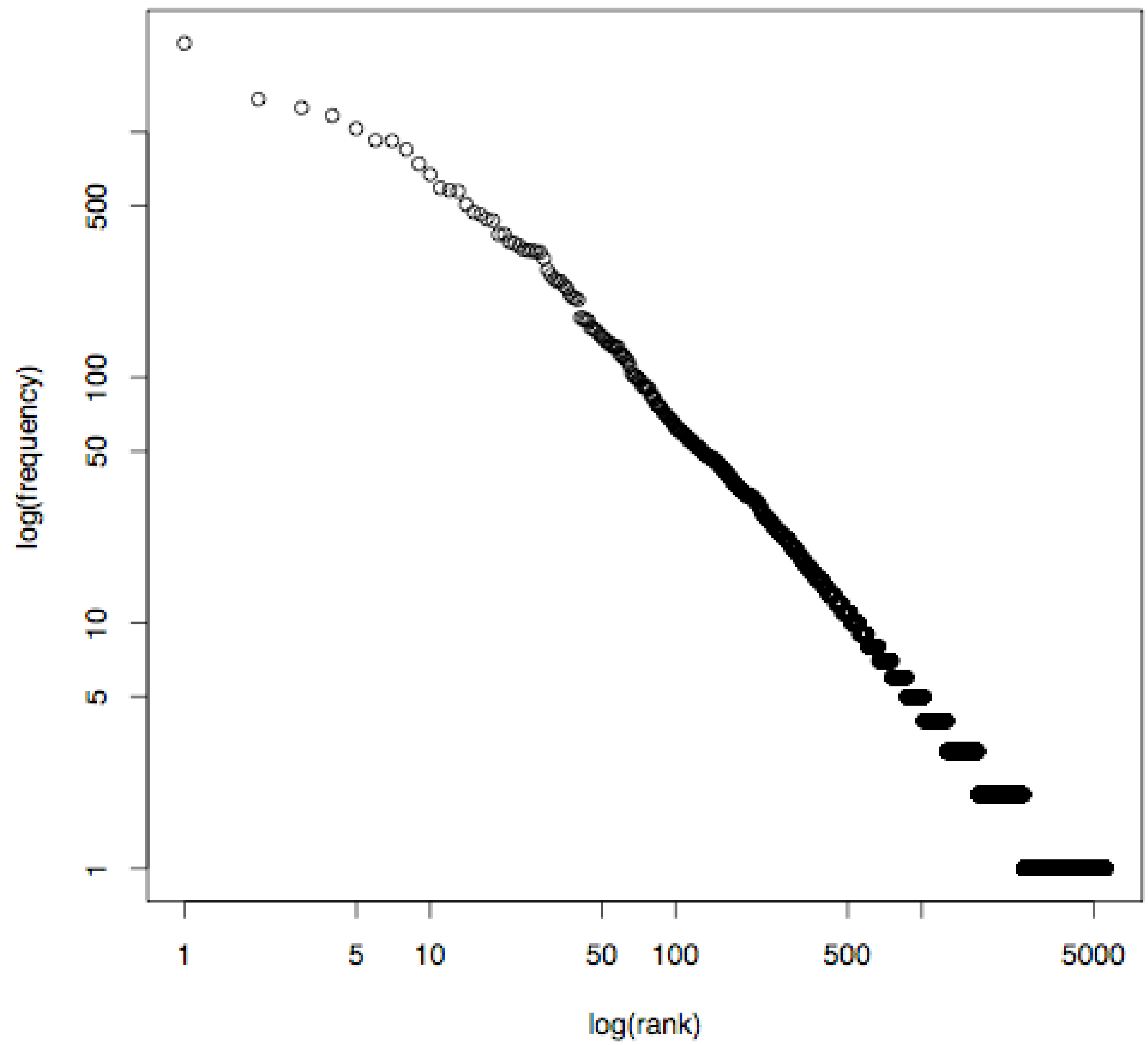
Zipf's Law
Alice in Wonderland



(text courtesy of Project Gutenberg)

Zipf's Law

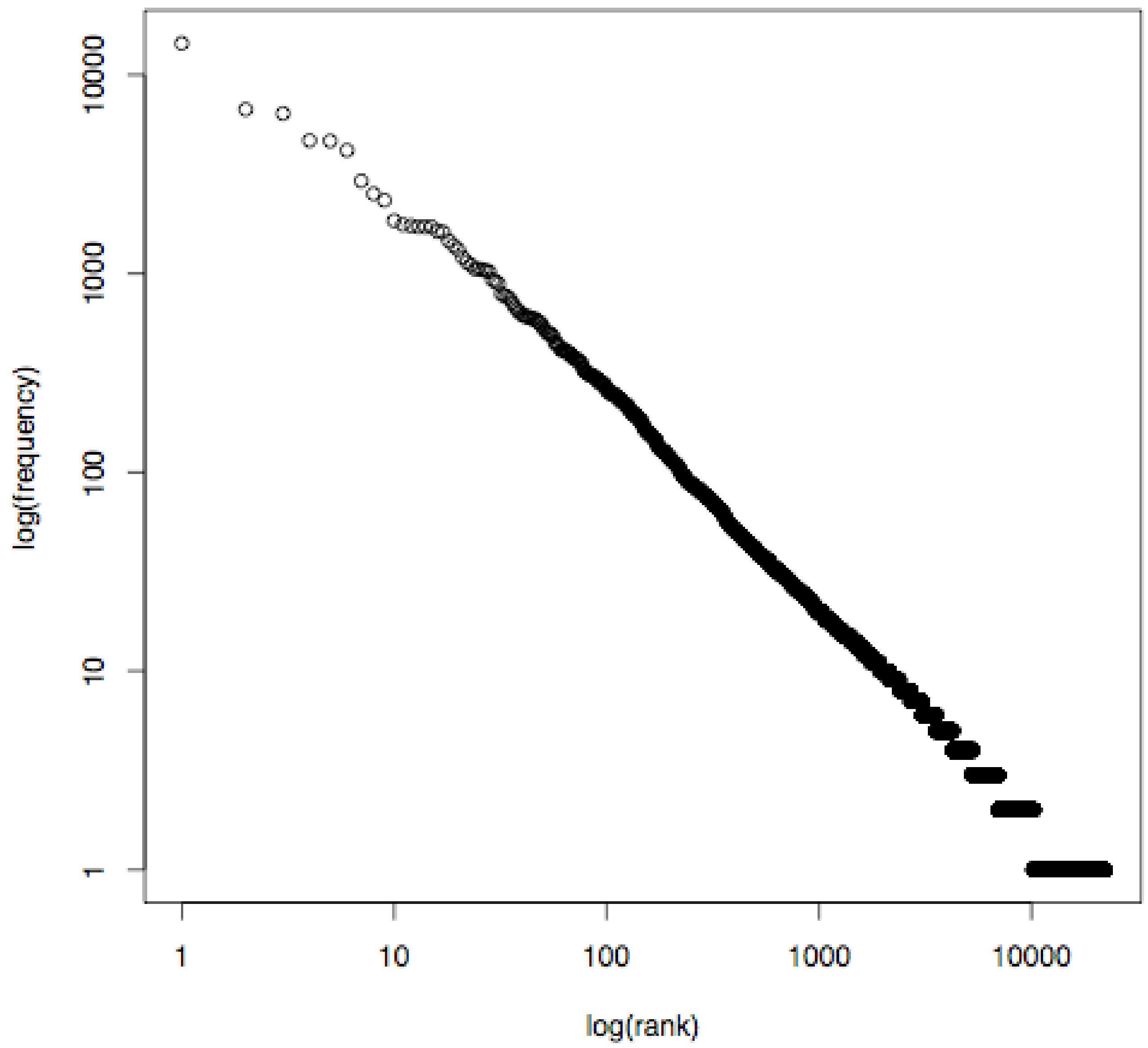
Peter Pan



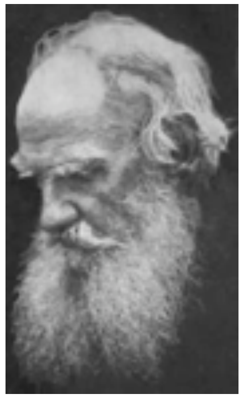
(text courtesy of Project Gutenberg)

Zipf's Law

Moby Dick

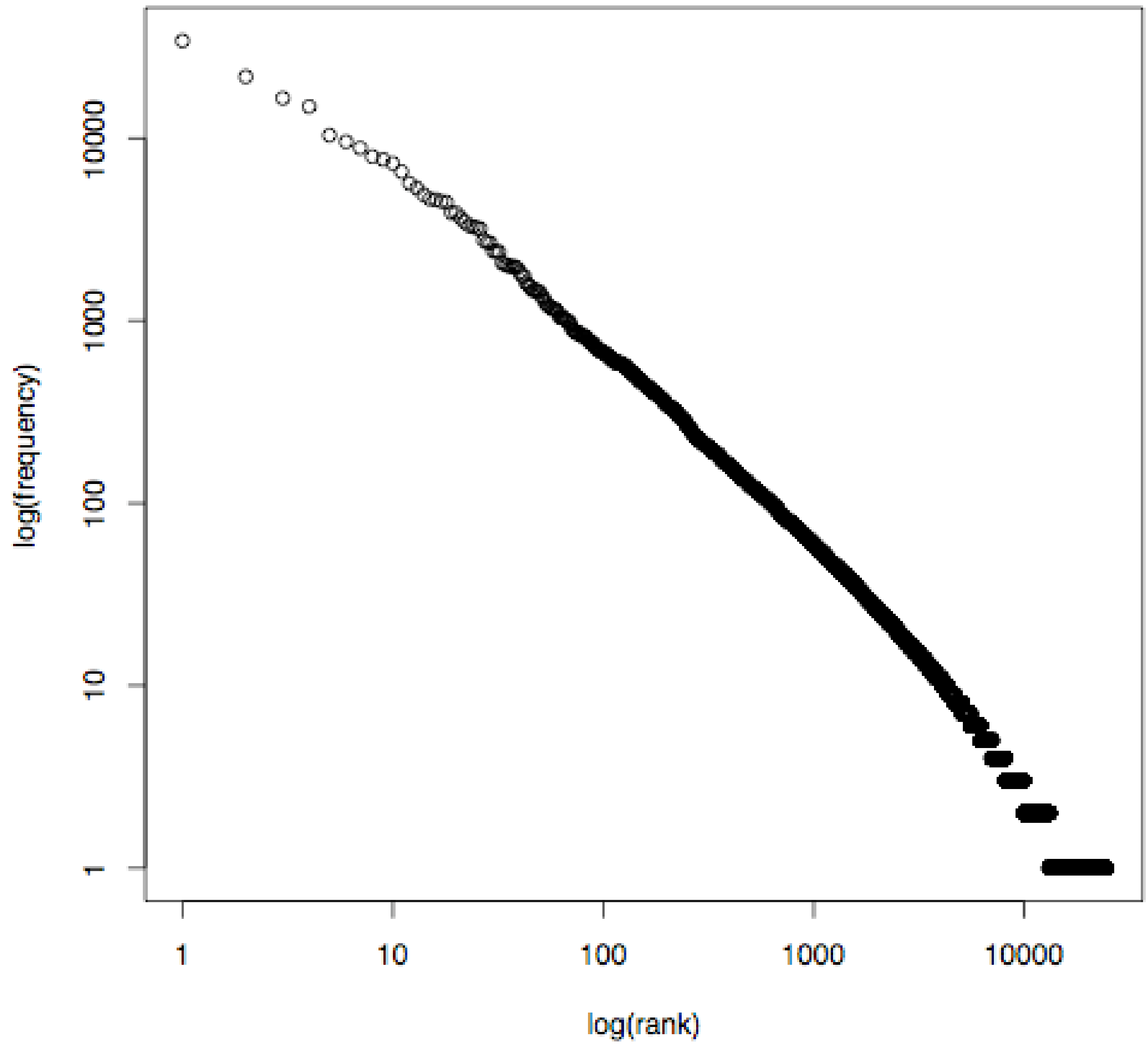


(text courtesy of Project Gutenberg)

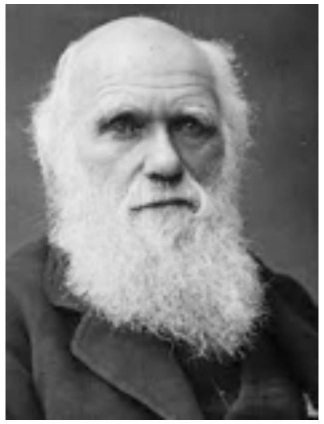


Zipf's Law

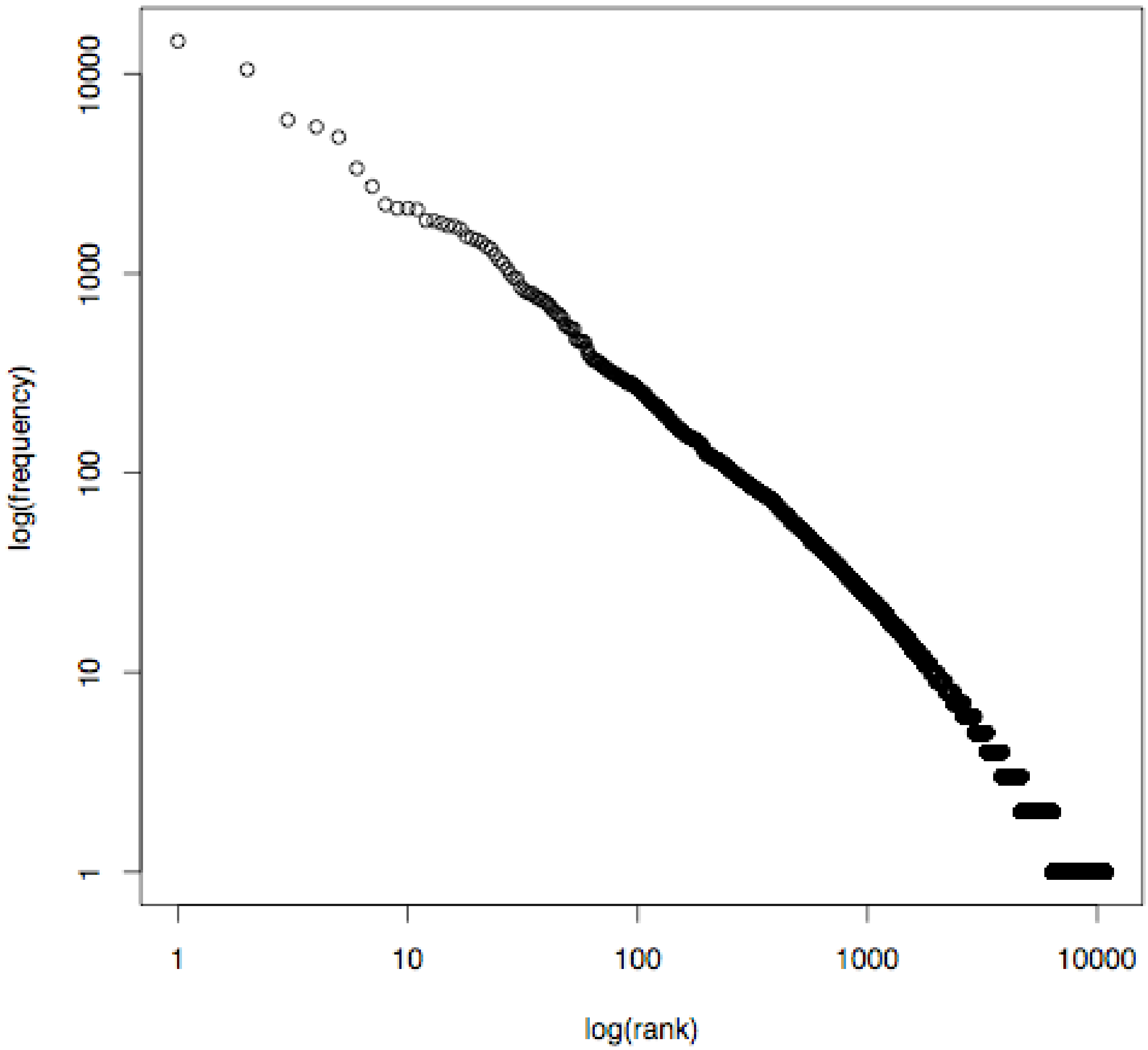
War and Peace



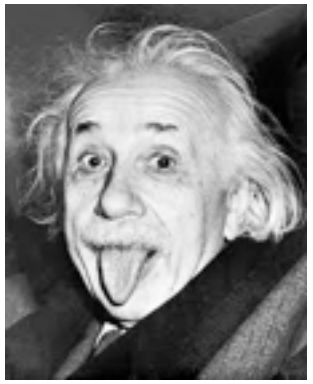
(text courtesy of Project Gutenberg)



Zipf's Law
On the Origin of Species

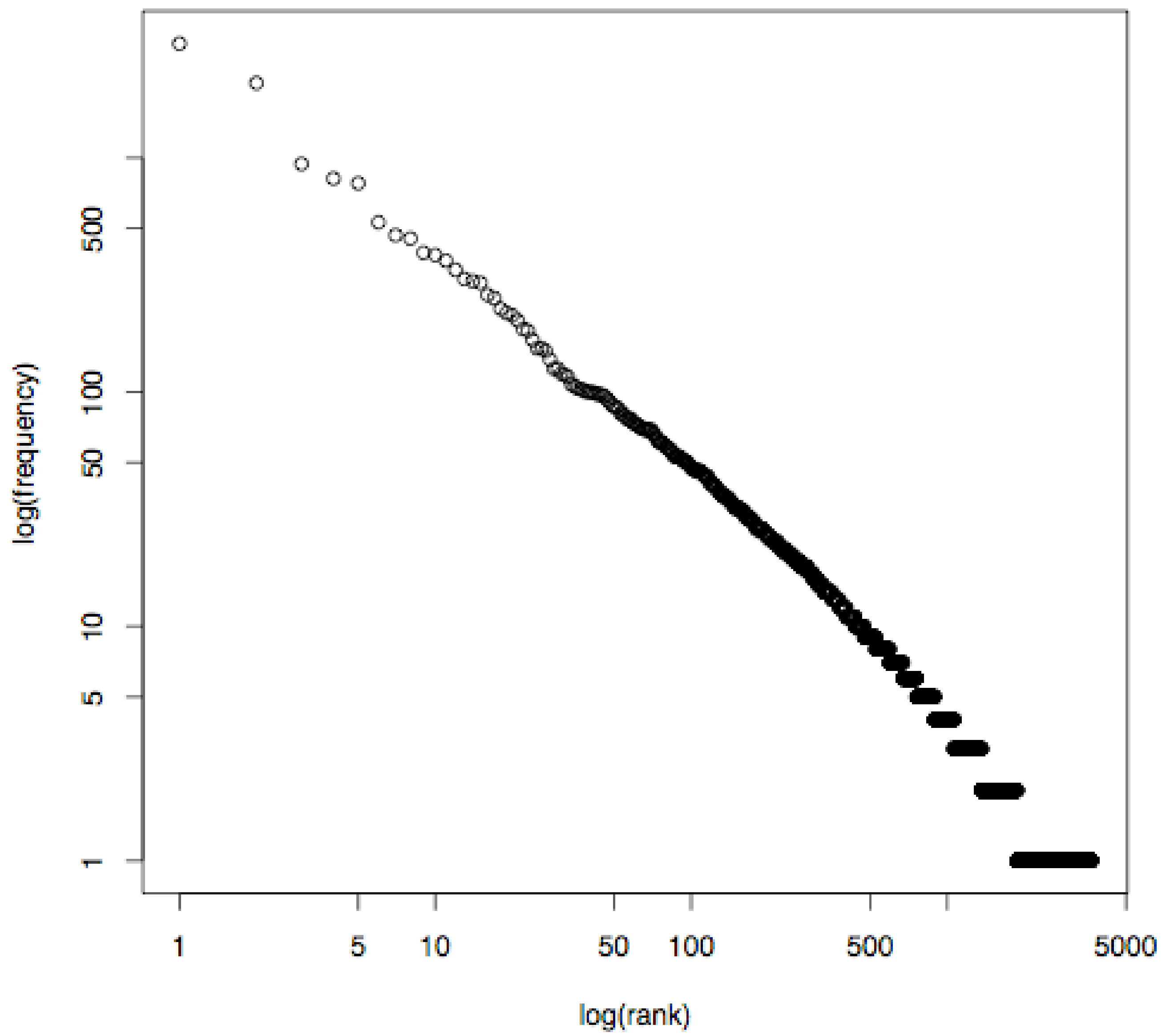


(text courtesy of Project Gutenberg)



Zipf's Law

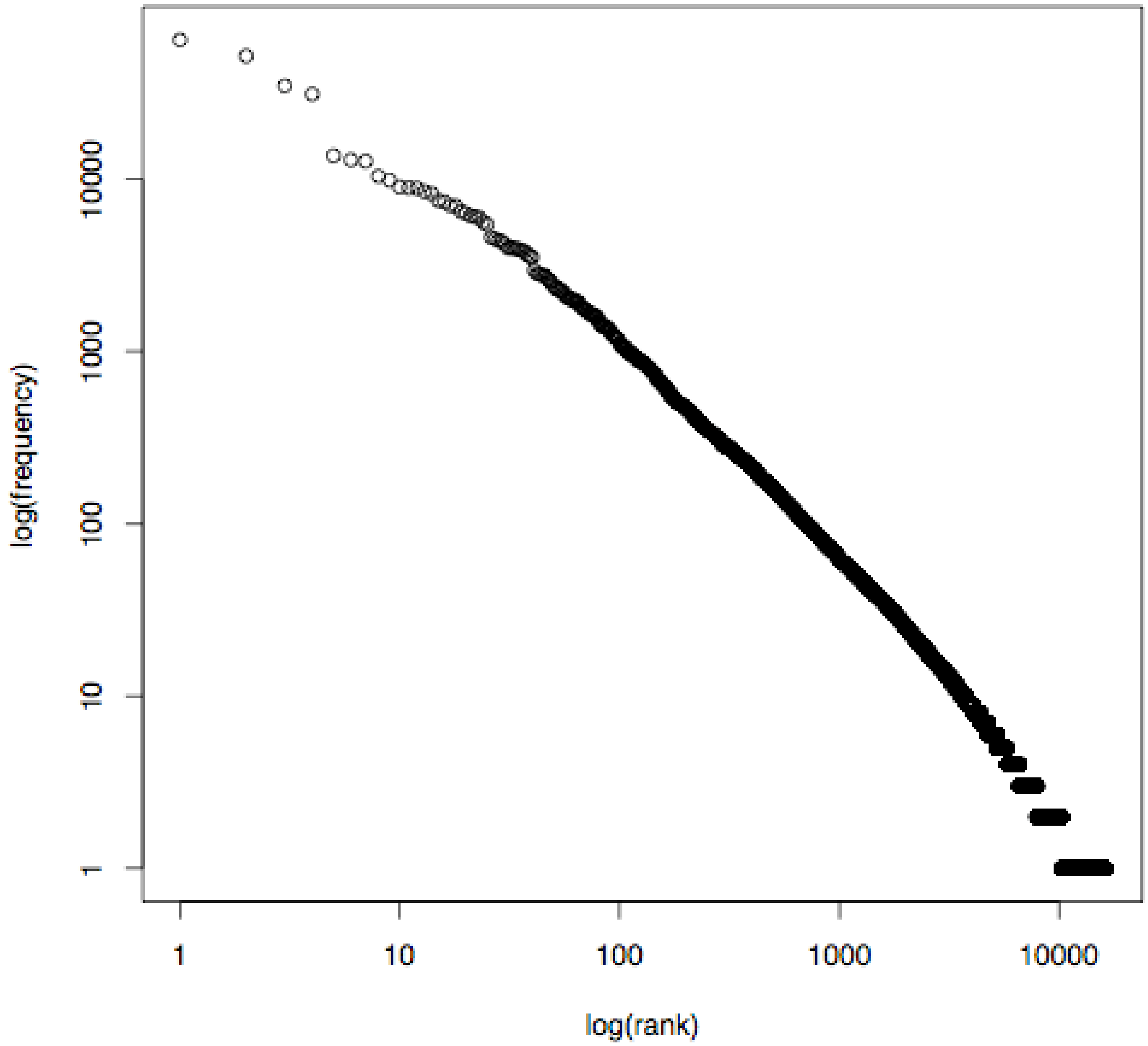
Relativity: The Special and General Theory



(text courtesy of Project Gutenberg)

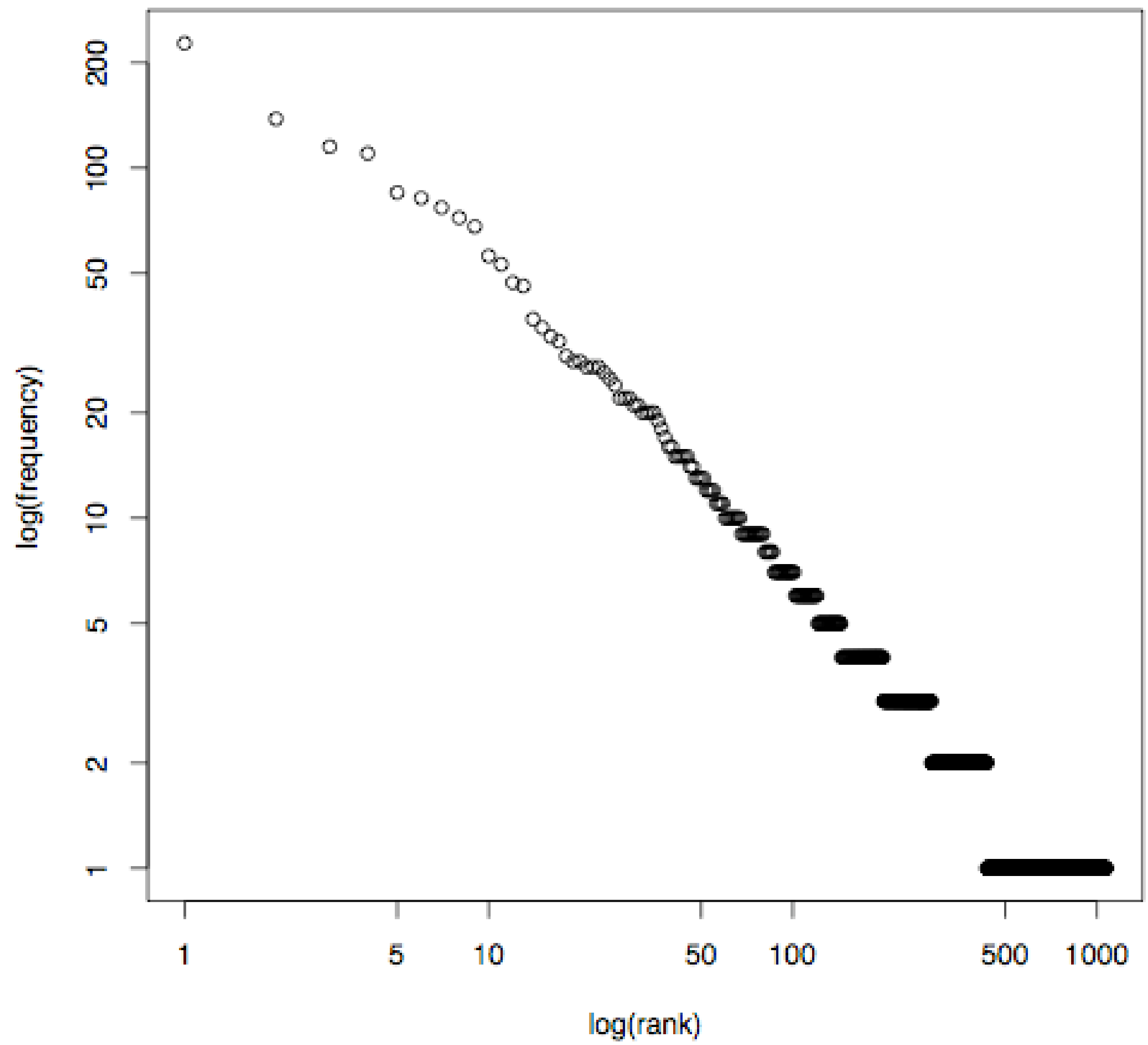
Zipf's Law

The King James Bible



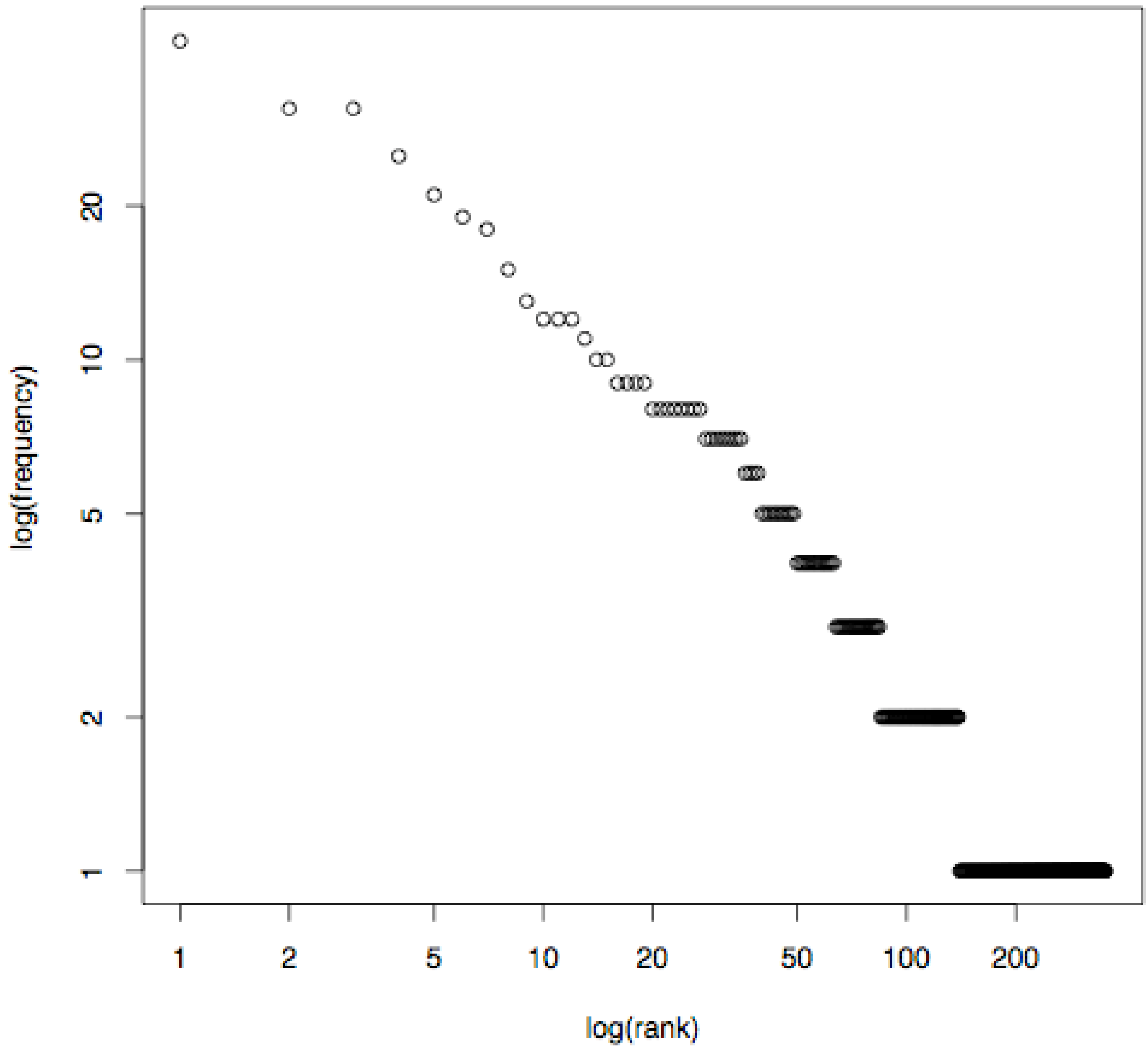
(text courtesy of Project Gutenberg)

Zipf's Law
The Tale of Peter Rabbit



(text courtesy of Project Gutenberg)

Zipf's Law
The Three Bears



(text courtesy of Project Gutenberg)

Zipf's Law

- Zipf's Law holds true for:
  - different languages
  - different sizes of text
  - different genres
  - different topics
  - different complexity of content

Implications of Zipf's Law

gerard salton 8 march 1978 in nuremberg 28 august 1995 also know as gerry salton was professor of computer science at cornell university salton was perhaps the leading computer scientist working in the field of information retrieval during his time his group at cornell developed the smart information retrieval system which he initiated when he was at harvard

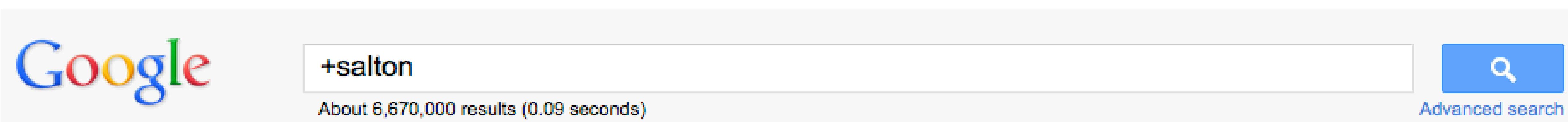
- The most important words are those that are frequent in the document, but not the most frequent in the collection
- Most retrieval models (as we will see) exploit this idea
- Zipf's law allows us to automatically identify these non-descriptive terms and treat them differently
- Example: (gerard OR salton OR at OR cornell)

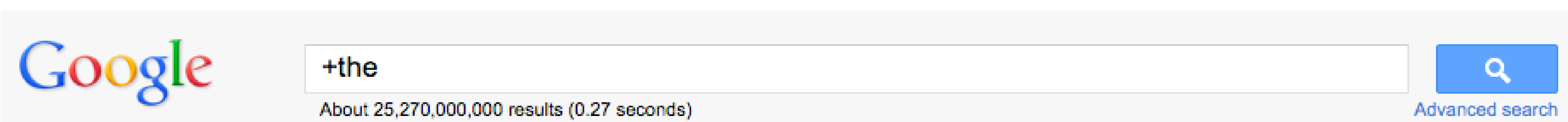
Implications of Zipf's Law

- Ignoring the most frequent terms greatly reduces the size of the index
- The top 50 accounts for about 45% of the collection
- Warning: these words <u>can</u> be important in combination with others (e.g., in proximity operators)
- Example queries: "to be or not to be", "the who", "state of the union", "it had to be you"

Implications of Zipf's Law

- Ignoring the most frequent terms can improve retrieval efficiency (response time)
- The inverted lists associated with the most frequent terms are huge relative to others
- Alternative: leave them in the index and remove them from the query, unless they occur in a proximity operator





Implications of Zipf's Law

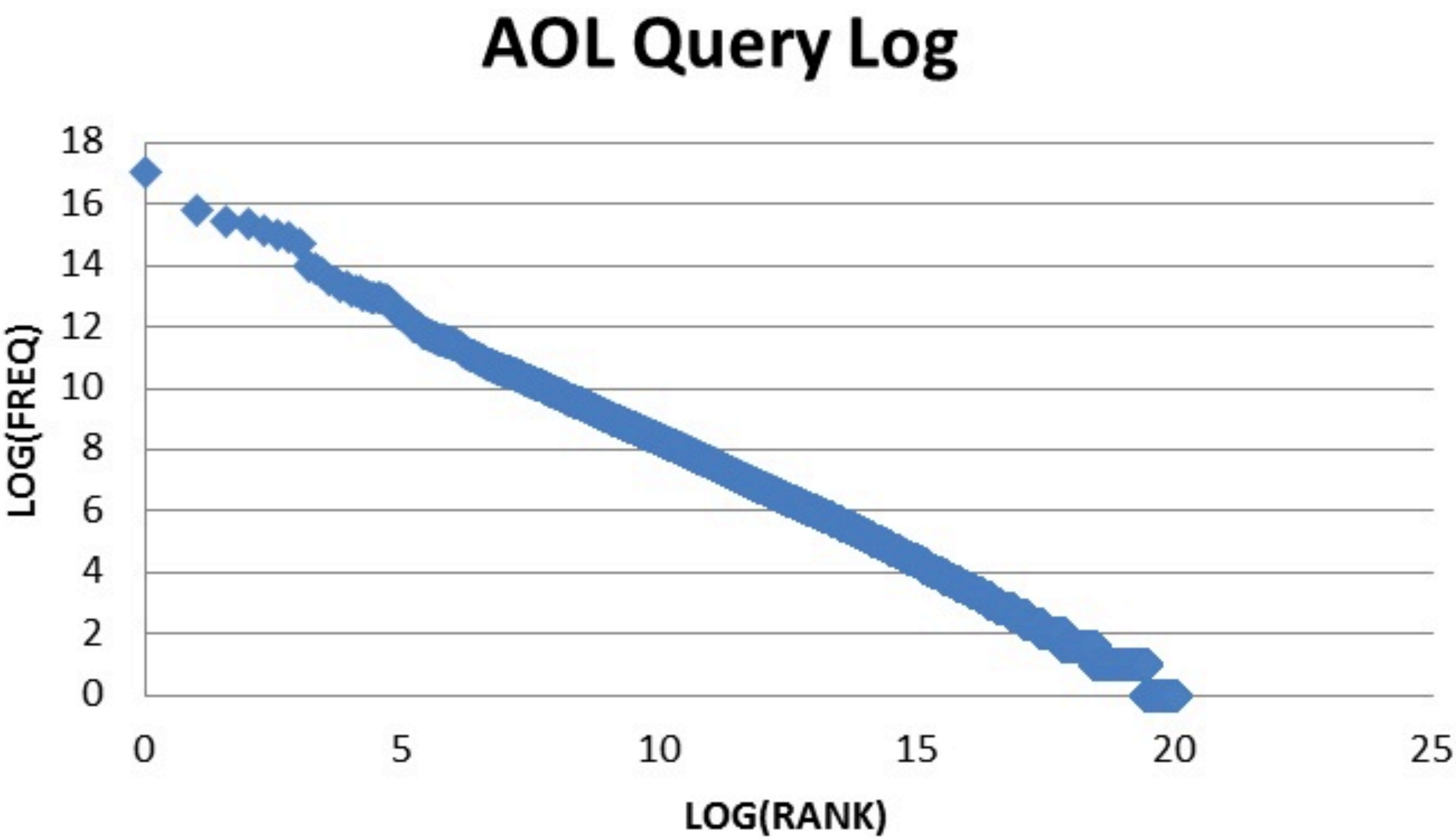
- Ignoring the most frequent terms can improve retrieval effectiveness
- Very frequent terms may not be related to the main content of the doc, but may be a "quirk" of the corpus

rank	term	frequency	rank	term	frequency
1	the	1586358	11	year	250151
2	a	854437	12	he	242508
3	and	822091	13	movie	241551
4	to	804137	14	her	240448
5	of	657059	15	artist	236286
6	in	472059	16	character	234754
7	is	395968	17	cast	234202
8	i	390282	18	plot	234189
9	his	328877	19	for	207319
10	with	253153	20	that	197723

Implications of Zipf's Law

- We've talked about Zipf's Law in the collection
- What about Zipf's Law in queries issued to the search engine?

Implications of Zipf's Law



Implications of Zipf's Law

- Same trend: a few queries occur very frequently, while most occur very infrequently
- Opportunity: the system can be tweaked to do well on those queries it is likely to "see" again and again
- Curse: this is only a partial solution.
- In Web search, about half the queries ever observed are unique
- How does this effect evaluation?

Implications of Zipf's Law

- Given Zipf's Law, as a collection grows, how will the size of the vocabulary grow?

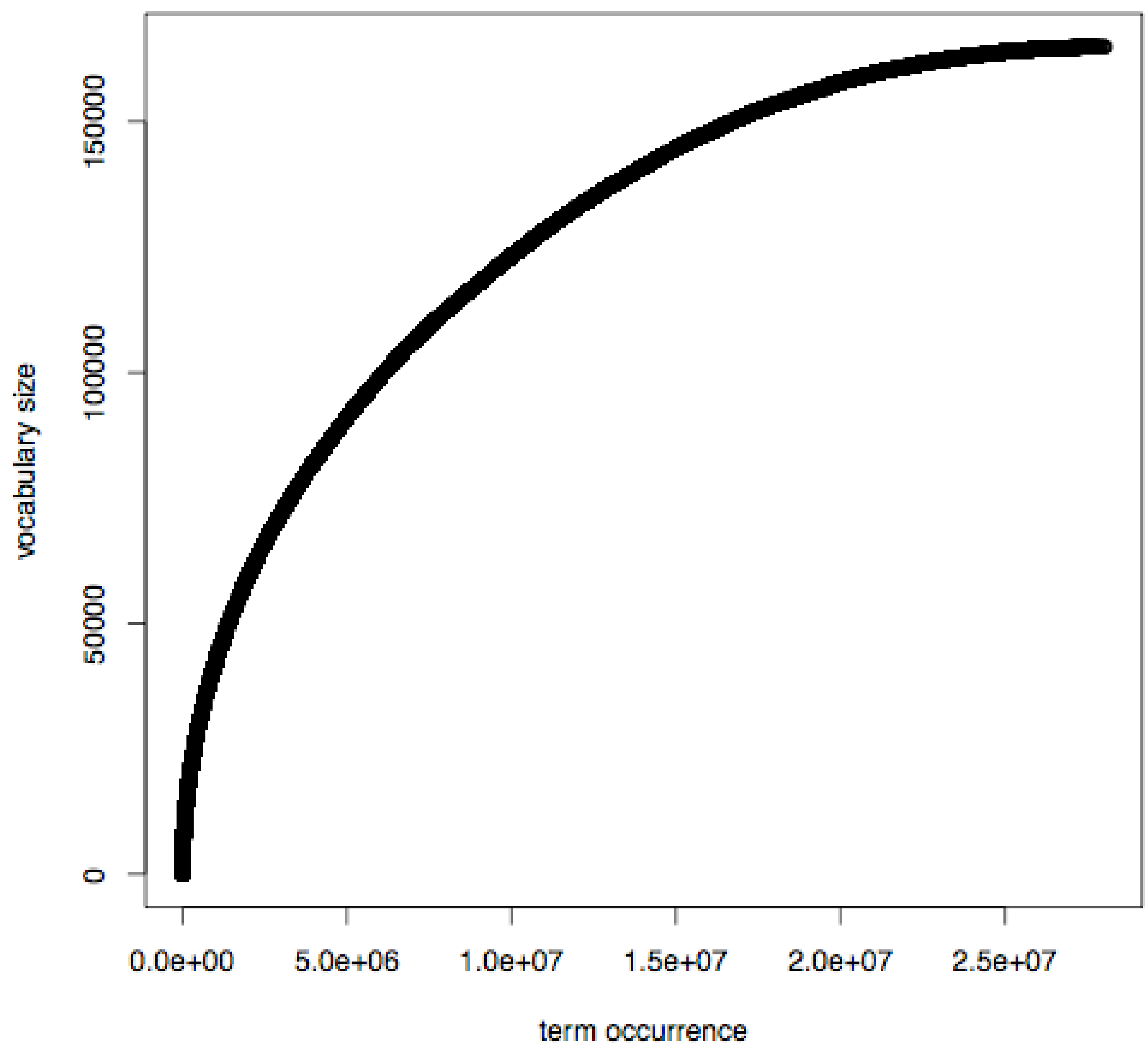
Vocabulary Growth and Heaps' Law

- The number of <u>new</u> words <u>decreases</u> as the size of the corpus <u>increases</u>
- Heaps' Law:

$$v = k \times n^{\beta}$$

- **v** = size of the vocabulary (number of unique words)
- **n** = size of the corpus (number of word-occurrences)
- **k** = constant (10 ≤ **k** ≤ 100)
  - not the same as **k** in Zipf's law
- **B** = constant (**B** ≈ 0.50)

Heaps' Law
IMDB Corpus



Heaps' Law

- As the corpus grows, the number of <u>new</u> terms will increase dramatically at first, but then will increase at a <u>slower rate</u>
- Nevertheless, as the corpus grows, new terms will <u>always</u> be found (even if the corpus becomes huge)
  - there is no end to vocabulary growth
  - invented words, proper nouns (people, products), misspellings, email addresses, etc.

Implications of Heaps' Law

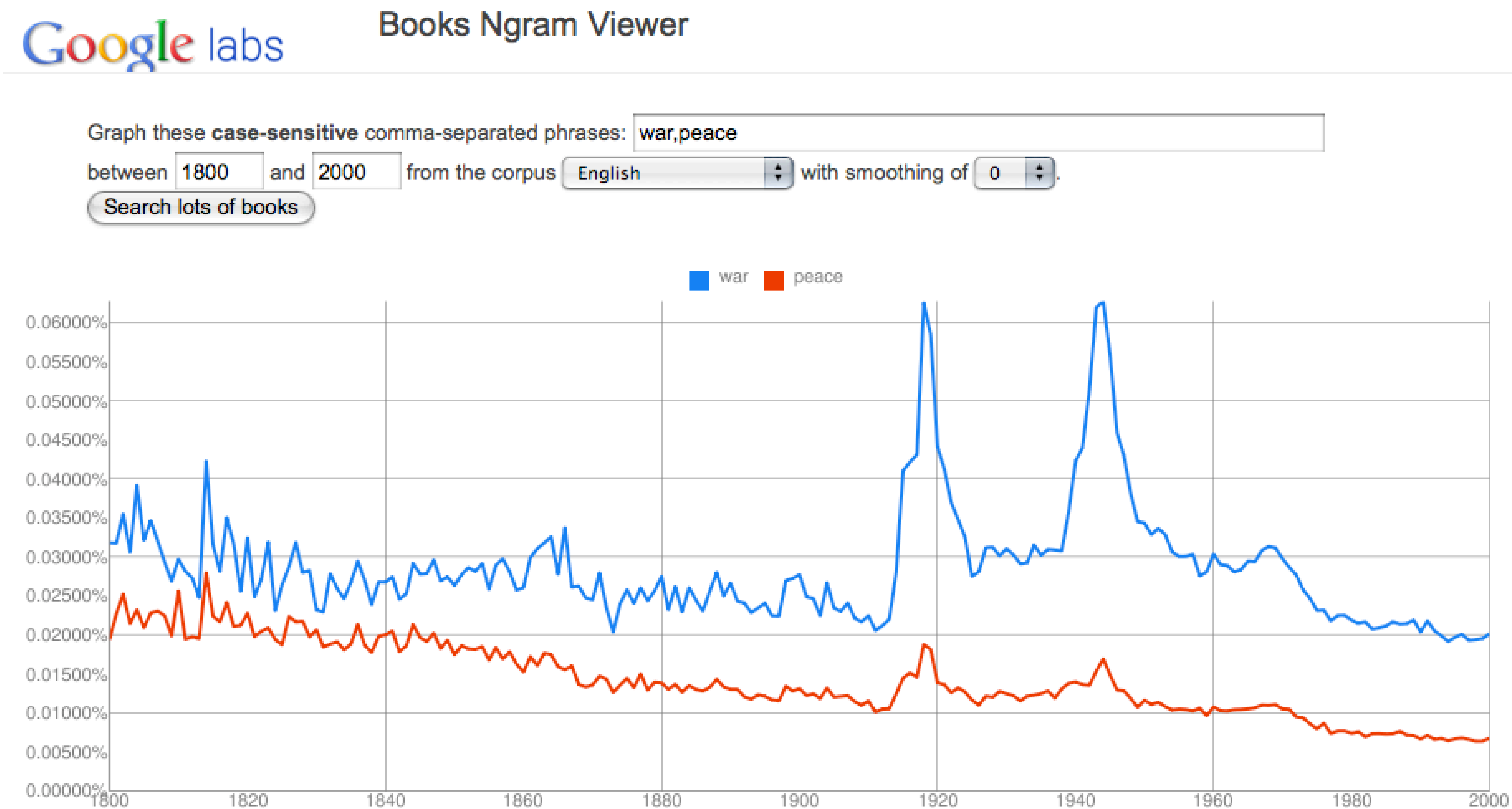
- Given a corpus and a <u>new</u> set of data, the number of new index terms will depend on the size of the corpus
- Given more data, new index terms will always be required
- This may also be true for controlled vocabularies (?)
  - Given a corpus and a new set of data, the requirement for new <u>concepts</u> will depend on the size of the corpus
  - Given more data, new <u>concepts</u> will always be required

Term Co-occurrence

- So far, we've talked about statistics for <u>single</u> terms
- What about statistics for <u>pairs</u> of terms?
- Term co-occurrence considers the extent to which different terms tend to appear <u>together</u> in text
- Does knowledge that one term appears, tell us whether another term is likely to appear?

Term Co-occurrence Example

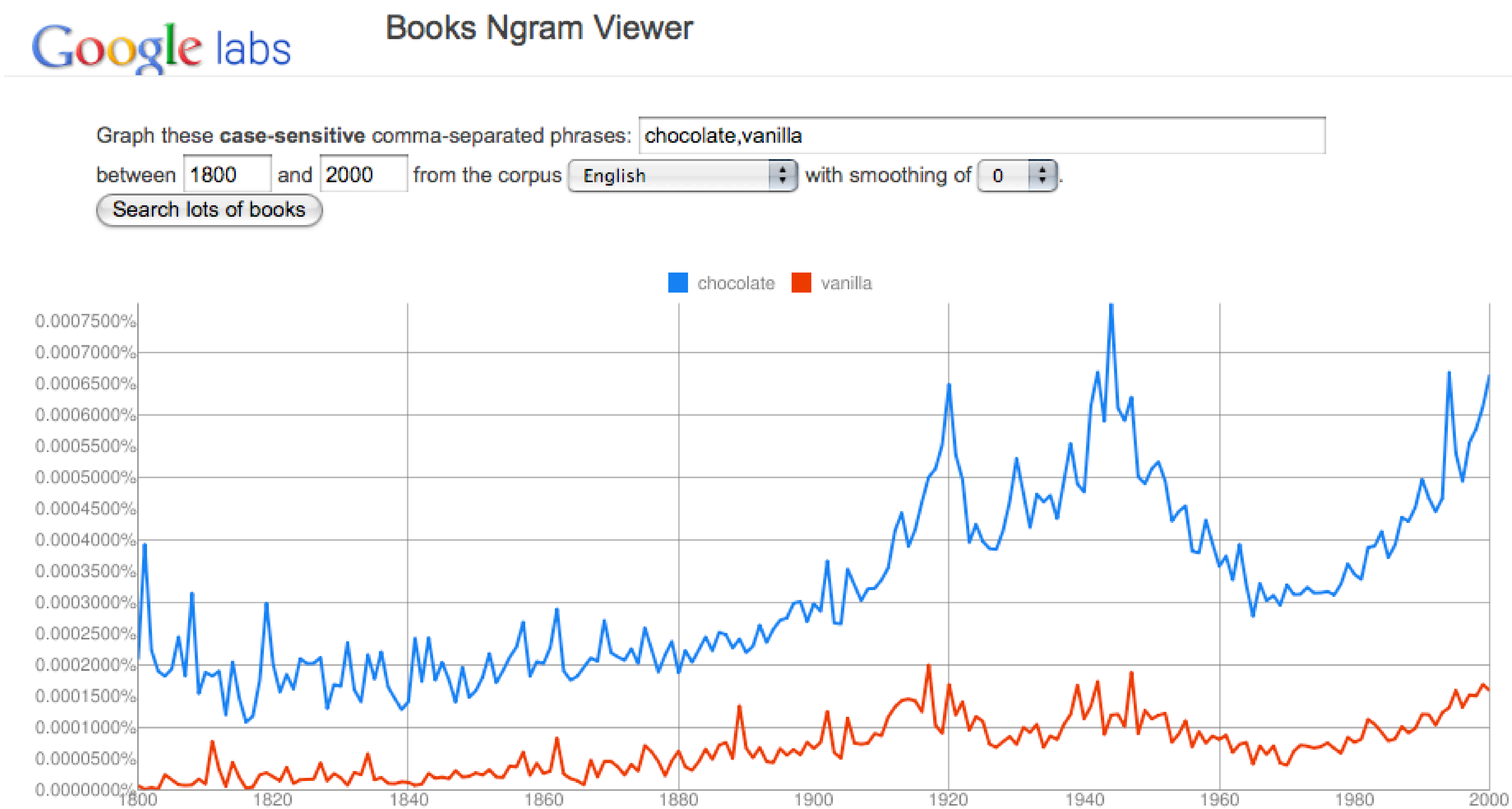
war vs. peace



(The Google Books N-gram Corpus)

Term Co-occurrence Example

chocolate vs. vanilla



(The Google Books N-gram Corpus)

A Few Important Concepts in Probability Theory and Statistics

(Some material courtesy of Andrew Moore: http://www.autonlab.org/tutorials/prob.html)

Discrete Random Variable

- A is a discrete random variable if:
  - A describes an event with a finite number of possible outcomes (discrete vs continuous)
  - A describes and event whose outcome has some degree of uncertainty (random vs. pre-determined)
- A is a boolean-valued random variable if it describes an event with two outcomes: TRUE or FALSE
- Can you name some examples of boolean-valued random variables?

Boolean-Valued Random Variables

Examples

- A = it will rain tomorrow
- A = the outcome of a coin-flip will be heads
- A = the fire alarm will go off sometime this week
- A = The US president in 2023 will be female
- A = you have the flu
- A = the word "retrieval" will occur in a document

Probabilities

- P(A=TRUE): the probability that the outcome is TRUE
  - the probability that it will rain tomorrow
  - the probability that the coin will show "heads"
  - the probability that "retrieval" appears in the doc
- P(A=FALSE): the probability that the outcome is FALSE
  - the probability that it will NOT rain tomorrow
  - the probability that the coin will show "tails"
  - the probability that "retrieval" does NOT appear in the doc

Probabilities

$$0 <= P(A=TRUE) <= 1$$

$$0 <= P(A=FALSE) <= 1$$

$$P(A=TRUE) + P(A=FALSE) = 1$$

Estimating the Probability of an Outcome

- P(heads=TRUE)
- P(rain tomorrow=TRUE)
- P(alarm sound this week=TRUE)
- P(female pres. 2023=TRUE)
- P(you have the flu=TRUE)
- P("retrieval" in a document=TRUE)

Statistical Estimation

- Use data to estimate the probability of an outcome
- Data = observations of previous outcomes of the event
- What is the probability that the coin will show "heads"?
- Statistical Estimation Example:
  - To gather data, you flip the coin 100 times
  - You observe 54 "heads" and 46 "tails"
  - What would be your estimation of P(heads=TRUE)?

Statistical Estimation

- What is the probability that it will rain tomorrow?
- Statistical Estimation Example:
    - To gather data, you keep a log of the past 365 days
    - You observe that it rained on 93 of those days
    - What would be your estimation of P(rain=TRUE)?

Statistical Estimation

- What is the probability that “retrieval” occurs in a document?
- Statistical Estimation Example:
  - To gather data, you take a sample of 1000 documents
  - You observe that “retrieval” occurs in 2 of them.
  - What would be your estimation of P(“retrieval” in a document=TRUE)?
- Usually, the more data, the better the estimation!

Joint and Conditional Probability

- For simplicity, P(A=TRUE) is typically written as P(A)
- P(A,B): the probability that event A and event B both occur together
- P(A|B): the probability of event A occurring given the prior knowledge that event B has occurred

Chain Rule

- $P(A, B) = P(A|B) \times P(B)$
- Example:
  - probability that it will rain today and tomorrow =
  - probability that it will rain today X
  - probability that it will rain tomorrow given that it rained today

Independence

- Events A and B are independent if:

$$P(A,B) = P(A|B) \times P(B) = P(A) \times P(B)$$

Always true! (Chain Rule)

Only true is A and B are independent

- Events A and B are independent if the outcome of A tells us nothing about the outcome of B (and vice-versa)

Independence

- Suppose A = rain tomorrow and B = rain today
  - Are these likely to be independent?
- Suppose A = rain tomorrow and B = fire-alarm today
  - Are these likely to be independent?

Mutual Information

$$MI(w_1, w_2) = \log\left(\frac{P(w_1, w_2)}{P(w_1)P(w_2)}\right)$$

- $P(w_1, w_2)$: probability that words w_1 and w_2 both appear in a text
- $P(w_1)$: probability that word w_1 appears in a text, with or without w_2
- $P(w_2)$: probability that word w_2 appears in a text, with or without w_1
- The definition of "a text" is up to you (e.g., a sentence, a paragraph, a document)

Mutual Information

$$MI(w_1, w_2) = \log\left(\frac{P(w_1, w_2)}{P(w_1)P(w_2)}\right)$$

- If $P(w_1, w_2) = P(w_1)\,P(w_2)$, it means that the words are independent: knowing that one appears conveys no information that the other one appears
- If $P(w_1, w_2) > P(w_1)\,P(w_2)$, it means that the words are not independent: knowing that one appears conveys some information that the other one appears

Mutual Information

estimation (using documents as units of analysis)

	word w_1 appears	word w_1 does not appear
word w_2 appears	a	b
word w_2 does not appear	c	d

every document falls under one of these quadrants

total # of documents
$N = a + b + c + d$

$P(w_1, w_2) = a / N$

$P(w_1) = (a + c) / N$

$P(w_2) = (a + b) / N$

Mutual Information

IMDB Corpus

- Word-pairs with highest mutual information (1-20)

w1	w2	MI	w1	w2	MI
francisco	san	6.619	dollars	million	5.437
angeles	los	6.282	brooke	rick	5.405
prime	minister	5.976	teach	lesson	5.370
united	states	5.765	canada	canadian	5.338
9	11	5.639	un	ma	5.334
winning	award	5.597	nicole	roman	5.255
brooke	taylor	5.518	china	chinese	5.231
con	un	5.514	japan	japanese	5.204
un	la	5.512	belle	roman	5.202
belle	nicole	5.508	border	mexican	5.186

Mutual Information

IMDB Corpus

- Word-pairs with highest mutual information (20-40)

w1	w2	MI	w1	w2	MI
belle	lucas	5.138	brooke	eric	4.941
nick	brooke	5.136	serial	killer	4.927
loved	ones	5.116	christmas	eve	4.911
hours	24	5.112	italy	italian	4.909
magazine	editor	5.103	un	I	4.904
e	fianc	5.088	photo	shoot	4.866
newspaper	editor	5.080	ship	aboard	4.856
donna	brooke	5.064	al	un	4.800
ed	un	5.038	plane	flight	4.792
mexican	mexico	5.025	nicole	victor	4.789

Mutual Information

IMDB Corpus

- Word-pairs with highest mutual information (1-20)

w1	w2	MI	w1	w2	MI
francisco	[illegible]	[illegible]	[illegible]	[illegible]	5.437
angeles	[illegible]	[illegible]	[illegible]	[illegible]	5.405
prime	[illegible]	[illegible]	[illegible]	[illegible]	5.370
united	[illegible]	[illegible]	[illegible]	[illegible]	5.338
9	11	5.639	un	ma	5.334
winning	award	5.597	nicole	roman	5.255
brooke	taylor	5.518	china	chinese	5.231
con	un	5.514	japan	japanese	5.204
un	la	5.512	belle	roman	5.202
belle	nicole	5.508	border	mexican	5.186

Not a perfect metric! Subject to subtleties in the collection (these are pairs of semantically unrelated Spanish words)

Implications of Term Co-occurrence

- Potential to improve search
  - word-variants co-occur: canada, canadian
  - phrases describe important concepts
  - semantically-related terms co-occur
- Multiple paths to improvement
  - document representation: conflating variants, indexing phrases, adding related terms
  - information need representation: conflating variants, proximity operators, adding related terms
  - search assistance and interactions: query suggestions

Implications of Term Co-occurrence

Google

PC repair

Advanced search

[**Computer Repair** | **PC Repair** Directory](www.pcrepairdirectory.com/)
www.**pcrepair**directory.com/ - Cached
Use the PCRepairDirectory to find local **computer repair** business listings and services for **PC repair** in your area. **Laptop repair**, virus removal and other services ...

[**Computer Repair** Directory](www.computerrepairdirectory.com/)
www.**computerrepair**directory.com/ - Cached
COMPUTER REPAIR. Need Help? Find The Best **PC Repair** Shops across the Country. Find a Technician near you Now! More than 2000 **Computer Repair** ...

[**Fix** My **Pc** FREE – Is Your **Computer** Running Slow?](www.fixmypcfree.com/)
www.**fix**mypcfree.com/ - Cached
Fix your **computer** yourself of any problems and situations that can arise. Simple tips and information for anyone to use, retake control of your **computer**.

[Home - Franklin **P. C. Repair** ® **Computer Repair** and Virus ...](www.franklinpcrepair.com/)
www.franklin**pcrepair**.com/ - Cached
Whether it's Home or Business **PC repairs**, installation of new computers, upgrades, advice or Virus Removal, we offer a quality service at competitive prices. ...

[Mobile Computer Wizard- San Diego **Computer Repair**, **PC Repair** ...](mobilecomputerwizard.com/)
mobile**computer**wizard.com/ - Cached
Mobile **Computer** Wizard: Fast, Reliable **Computer** Tech Support for San Diego County, including downtown, Oceanside, La Jolla, El Cajon, Escondido. We **fix** ...

Take-Home Message

- Language use is highly varied
- However, there are statistical properties of language that are highly consistent across domains and languages
- These statistical properties of text make search easier
- Learn them, love them, and use them to your advantage in doing automatic analysis of text