Text Representation

Jaime Arguello
INLS 613: Text Data Mining
jarguell@email.unc.edu

Text Representation

predicting health-related documents

features | concept

instances

w_1	w_2	w_3	...	w_n	label
1	1	0	...	0	health
0	0	0	...	0	other
0	0	0	...	0	other
0	1	0	...	1	other
⋮	⋮	⋮	...	⋮	⋮
1	0	0	...	1	health

Text Representation

predicting positive/negative reviews

features **concept**

instances

w_1	w_2	w_3	...	w_n	label
1	1	0	...	0	positive
0	0	0	...	0	negative
0	0	0	...	0	negative
0	1	0	...	1	negative
⋮	⋮	⋮	...	⋮	⋮
1	0	0	...	1	positive

Text Representation

predicting liberal/conservative bias

features / concept

instances

w_1	w_2	w_3	...	w_n	label
1	1	0	...	0	liberal
0	0	0	...	0	conservative
0	0	0	...	0	conservative
0	1	0	...	1	conservative
⋮	⋮	⋮	...	⋮	⋮
1	0	0	...	1	liberal

Bag of Words Text Representation

- Features correspond to terms in the vocabulary
  - vocabulary: the set of distinct terms appearing in at least one training (positive or negative) instance
  - remember that all (positive and negative) training instances and all test instances must have the same representation!
- Position information and word order is lost
  - dog bites man = man bites dog
- Simple, but often effective

Text Processing

- Down-casing: converting text to lower-case
- Tokenization: splitting text into terms or tokens
  - for example: splitting on one or more non-alpha-numeric characters

Text Processing
in Java

```
public String[] processText(String text){
    text = text.toLowerCase();
    return text.split("[\\W]");
}
```

Text Processing

Steve Carpenter cannot make horror movies. First of all, the casting was very wrong for this movie. The only decent part was the brown haired girl from Buffy the Vampire Slayer. This movies has no gore(usually a key ingredient to a horror movie), no action, no acting, and no suspense(also a key ingredient). Wes Bentley is a good actor but he is so dry and plain in this that it's sad. There were a few parts that were supposed to be funny(continuing the teen horror/comedy movies) and no one laughed in the audience. I thought that this movie was rated R, and I didn't pay attention and realized it had been changed to PG-13. Anyway, see this movie if you liked I Still Know What You Did Last Summer. That's the only type of person who would find this movie even remotely scary. And seriously, this is to you Steve Carpenter, stop making horror movies. This movie makes Scream look like Texas Chainsaw Massacre.

Text Processing

down-casing

steve carpenter cannot make horror movies. first of all, the casting was very wrong for this movie. the only decent part was the brown haired girl from buffy the vampire slayer. this movies has no gore(usually a key ingredient to a horror movie), no action, no acting, and no suspense(also a key ingredient). wes bentley is a good actor but he is so dry and plain in this that it's sad. there were a few parts that were supposed to be funny(continuing the teen horror/comedy movies) and no one laughed in the audience. i thought that this movie was rated r, and i didn't pay attention and realized it had been changed to pg-13. anyway, see this movie if you liked i still know what you did last summer. that's the only type of person who would find this movie even remotely scary. and seriously, this is to you steve carpenter, stop making horror movies. this movie makes scream look like texas chainsaw massacre.

Text Processing

tokenization

steve carpenter cannot make horror movies first of all the casting was very wrong for this movie the only decent part was the brown haired girl from buffy the vampire slayer this movies has no gore usually a key ingredient to a horror movie no action no acting and no suspense also a key ingredient wes bentley is a good actor but he is so dry and plain in this that it s sad there were a few parts that were supposed to be funny continuing the teen horror comedy movies and no one laughed in the audience i thought that this movie was rated r and i didn t pay attention and realized it had been changed to pg 13 anyway see this movie if you liked i still know what you did last summer that s the only type of person who would find this movie even remotely scary and seriously this is to you steve carpenter stop making horror movies this movie makes scream look like texas chainsaw massacre

Bag of Words Text Representation

- Which vocabulary terms should we include as features?
- All of them?
  - why might this be a <u>good</u> idea?
  - why might this be a <u>bad</u> idea?

Bag of Words Text Representation

Steve Carpenter cannot make horror movies. First of all, the casting was very wrong for this movie. The only decent part was the brown haired girl from Buffy the Vampire Slayer. This movies has no gore(usually a key ingredient to a horror movie), no action, no acting, and no suspense(also a key ingredient). Wes Bentley is a good actor but he is so dry and plain in this that it's sad. There were a few parts that were supposed to be funny(continuing the teen horror/comedy movies) and no one laughed in the audience. I thought that this movie was rated R, and I didn't pay attention and realized it had been changed to PG-13. Anyway, see this movie if you liked I Still Know What You Did Last Summer. That's the only type of person who would find this movie even remotely scary. And seriously, this is to you Steve Carpenter, stop making horror movies. This movie makes Scream look like Texas Chainsaw Massacre.

HW1 Training Set

terms that only occurred in negative training set

term	count	term	count	term	count
editor	7	wrestlemania	8	naschy	6
hsien	6	sorvino	7	catastrophe	6
evp	6	boll	19	blah	25
incomprehensible	6	conscience	6	mst3k	9
misery	8	hsiao	6	holmes	6
advise	6	banana	7	physics	10
pc	8	carradine	9	dhoom	7
damme	10	monkey	7	dolph	7
ninja	8	mccabe	11	hess	6
snakes	8	suck	18	transylvania	7
libre	6	stunned	6	wretched	6
streisand	20	tripe	6	moby	6

HW1 Training Set

terms that only occurred in positive training set

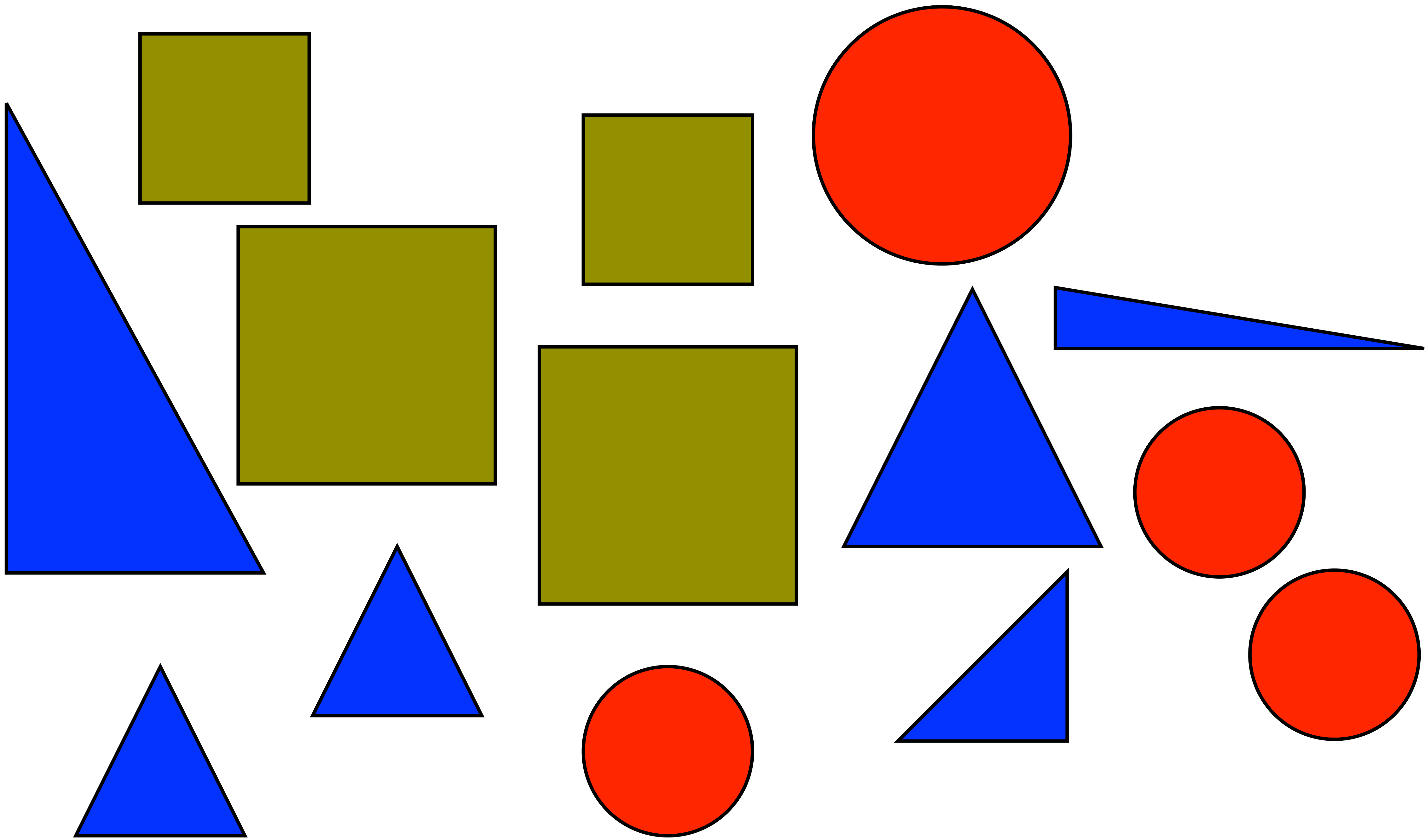
term	count	term	count	term	count
viewings	13	batista	6	captures	16
macy	9	mysteries	11	greene	9
whitaker	6	shemp	8	poison	6
reve	6	brooklyn	8	mum	6
bull	6	bonanza	7	colman	11
shaolin	6	francisco	7	muriel	6
welles	6	palace	8	jesse	9
challenges	6	elvira	11	veronika	13
demonicus	6	hagen	9	soccer	7
scarlett	6	cox	6	ka	6
blake	11	zorak	6	montrose	8
emy	8	bates	6	parsifal	6

Bag of Words Text Representation

- HW1 training set:
  - Number of Instances: 2,000
  - Number of unique terms: 25,637
  - Number of term occurrences: 472,012
- Why should we not include all 25,637 vocabulary terms as features?
- Is there a danger in having 12 times more features than instances?
- We should reduce the feature representation to the most meaningful ones

Training data + Representation

what could possibly go wrong?



Feature Selection

- Objective: reduce the feature set to only the most potentially useful
- Unsupervised Feature Selection
  - does not require training data
  - potentially useful features are selected using term statistics
- Supervised Feature Selection
  - requires training data (e.g., positive/negative labels)
  - potentially useful features are selected using co-occurrence statistics between terms and the target label

Unsupervised Feature Selection

Statistical Properties of Text

- As we all know, language use is highly varied
- There are <u>many</u> ways to convey the same information
- However, there are statistical properties of text that are predictable across domains, and even across languages!
- These can help us determine which terms are less likely to be useful (without requiring training labels)

HW1 Training Set
statistical properties of text

- HW1 training set:
  - Number of Instances: 2,000
  - Number of unique terms: 25,637
  - Number of term occurrences: 472,012

HW1 Training Set

term-frequencies

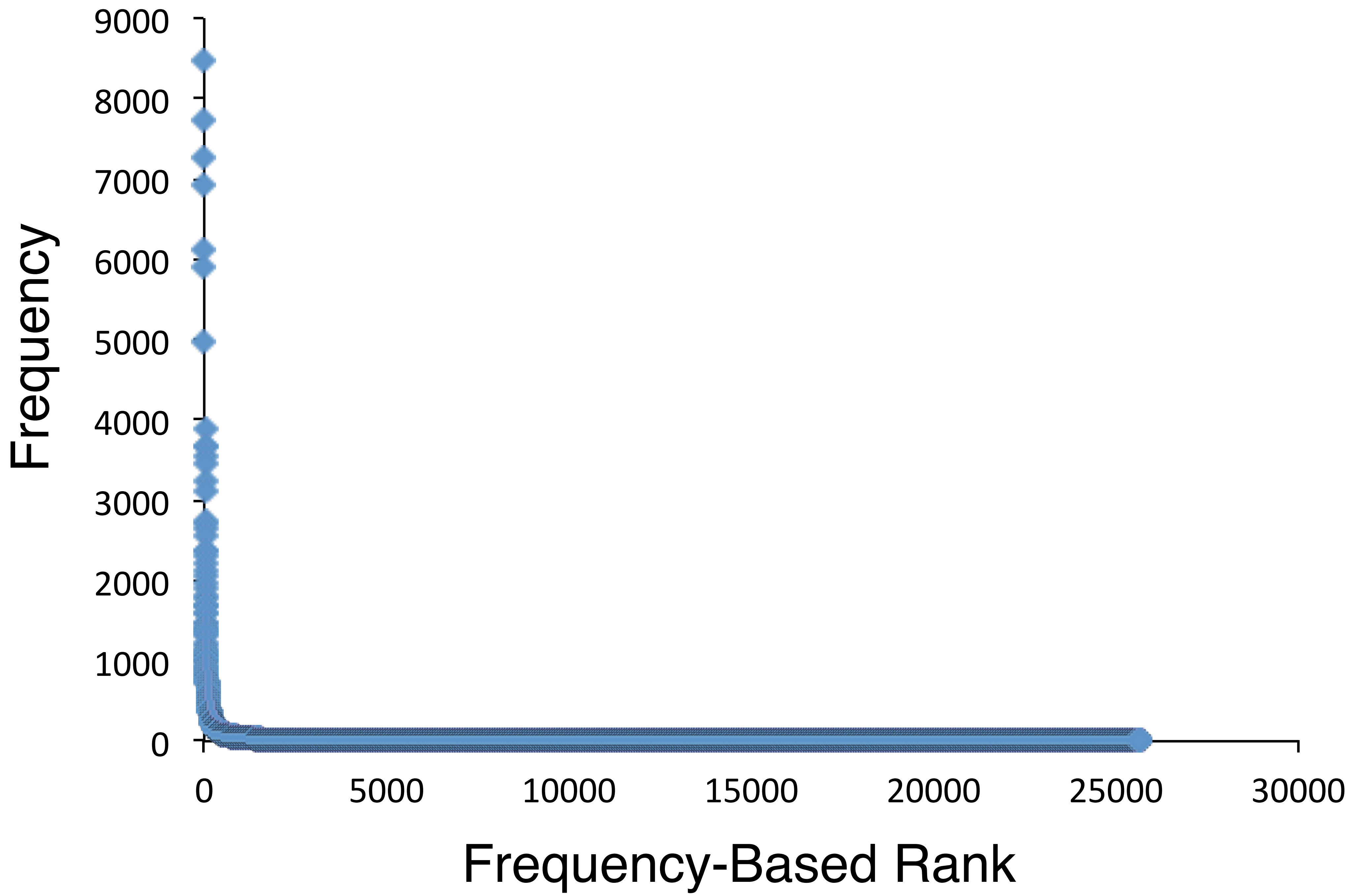
rank	term	frequency	rank	term	frequency
1	the	26638	11	that	5915
2	and	13125	12	s	4975
3	a	12949	13	was	3900
4	of	11715	14	as	3677
5	to	10861	15	movie	3666
6	is	8475	16	for	3540
7	it	7740	17	with	3441
8	in	7259	18	but	3236
9	i	6926	19	film	3124
10	this	6132	20	on	2743

HW1 Training Set

term-frequencies

rank	term	frequency	rank	term	frequency
21	you	2722	31	at	1895
22	t	2660	32	they	1803
23	not	2560	33	by	1793
24	his	2376	34	who	1703
25	he	2366	35	so	1699
26	are	2315	36	an	1681
27	have	2230	37	from	1609
28	be	2133	38	like	1582
29	one	2069	39	there	1483
30	all	1980	40	her	1458

HW1 Training Set
term-frequencies



Zipf's Law

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

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

- f_t = frequency (number of times term **t** occurs)
- r_t = frequency-based rank of term **t**
- **k** = frequency of most frequent term
- 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

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

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 <u>else</u> does Zipf's law tell us?

Zipf's Law

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

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

- So, what <u>fraction</u> 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
HW1 training set

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

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

- About half the terms occur only once! (43.8%)
- About 75% of the terms occur 3 times or less! (67.5%)
- About 83% of the terms occur 5 times or less! (76.7%)
- About 90% of the terms occur 10 times or less! (86.0%)

Zipf's Law

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

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

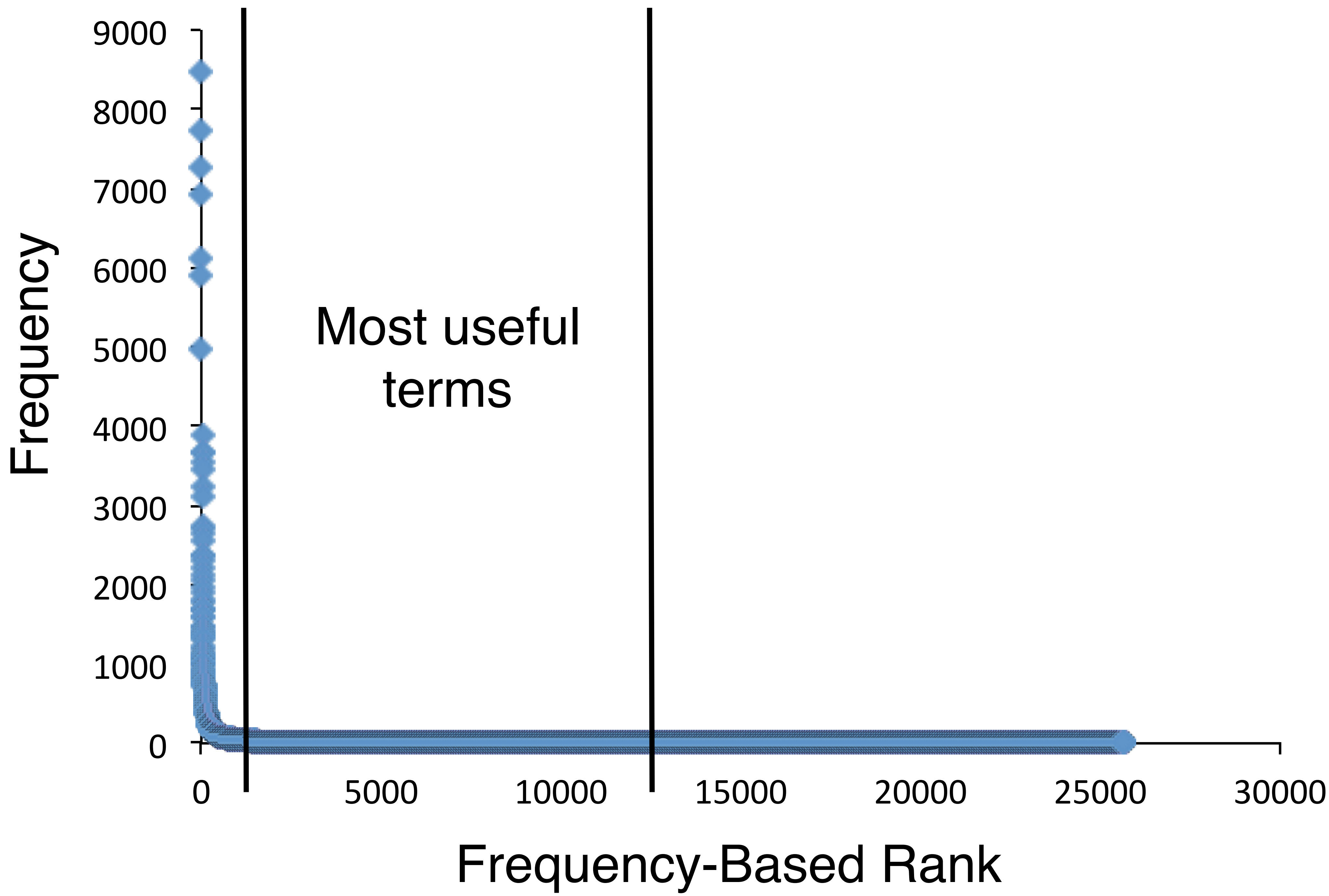
- We have to add the fraction of terms that appear 1, 2, 3, ... up to **n** times

Zipf's Law
Implications for Feature Selection

- The most frequent terms can be ignored
  - assumption: terms that are poor discriminators between instances are likely to be poor discriminators for the target class (e.g., positive/negative sentiment)
- The least frequent terms can be ignored
  - assumption: terms that occur rarely in the training set do not provide enough evidence for learning a model <u>and</u> will occur rarely in the test set

Zipf's Law

Implications for Feature Selection



Zipf's Law

Implications for Feature Selection

- The most frequent terms can be ignored
  - ignore the most frequent 50 terms
  - will account for about 50% of all term occurrences
- The least frequent terms can be ignored
  - ignore terms that occur 5 times or less
  - will account for about 80% of the vocabulary

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 $$\log(f) = \log(k) - \log(r)$$

Verifying Zipf's Law

visualization

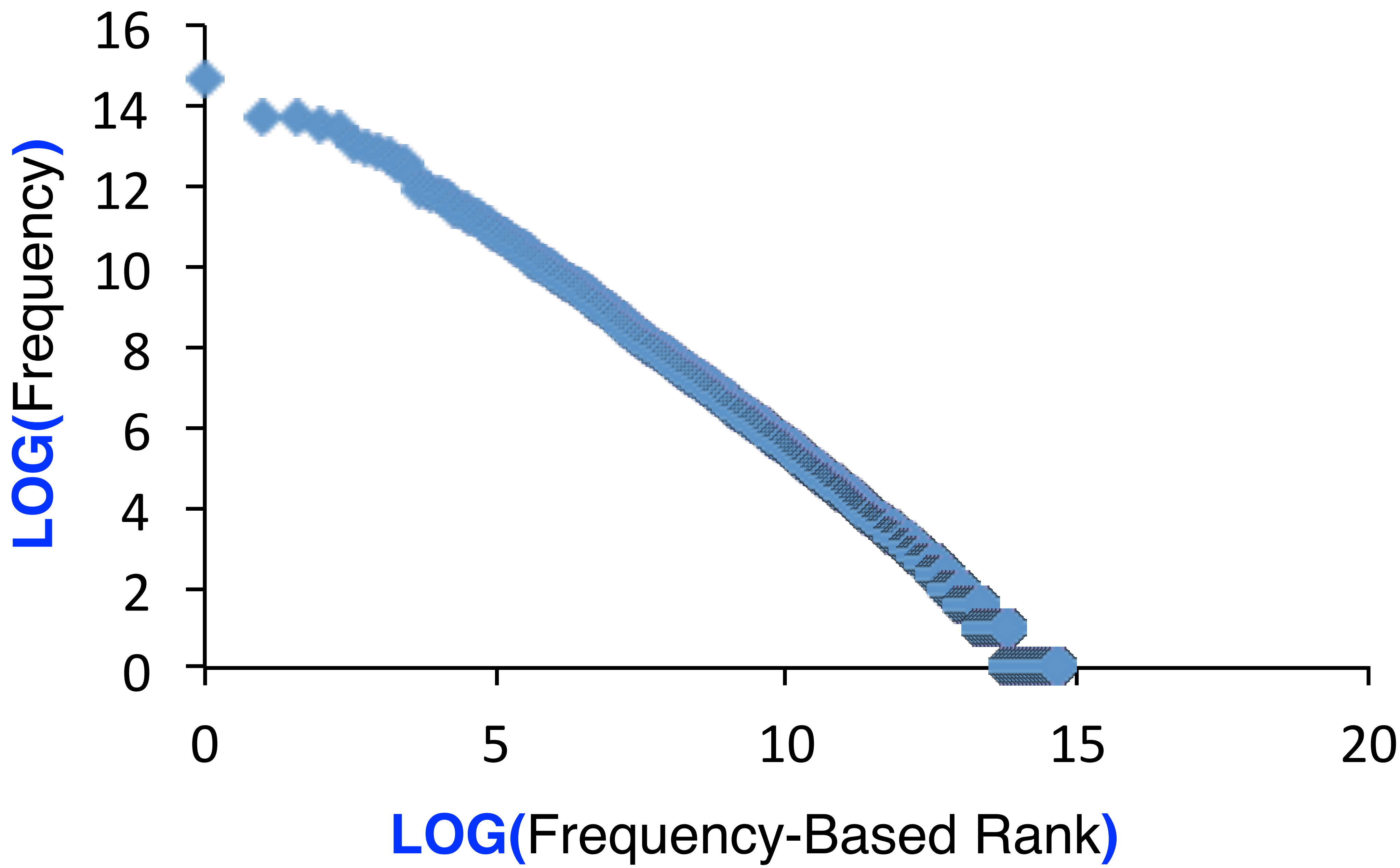
Zipf's Law $$f = \frac{k}{r}$$

... 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 line with a slope of -1

Zipf's Law
HW1 Dataset



Does Zipf's law generalize across collections of different size?

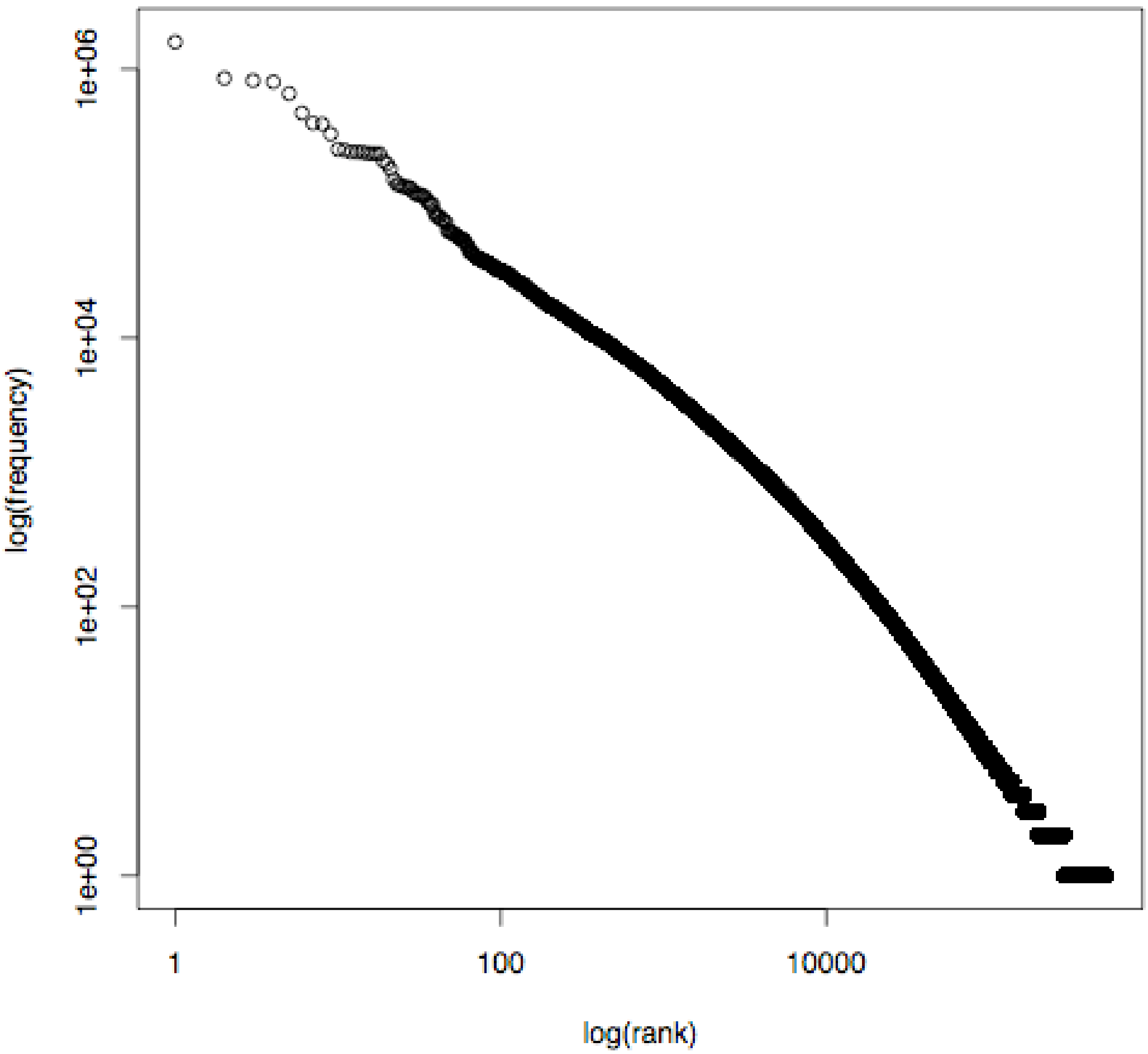
IMDB Corpus

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**

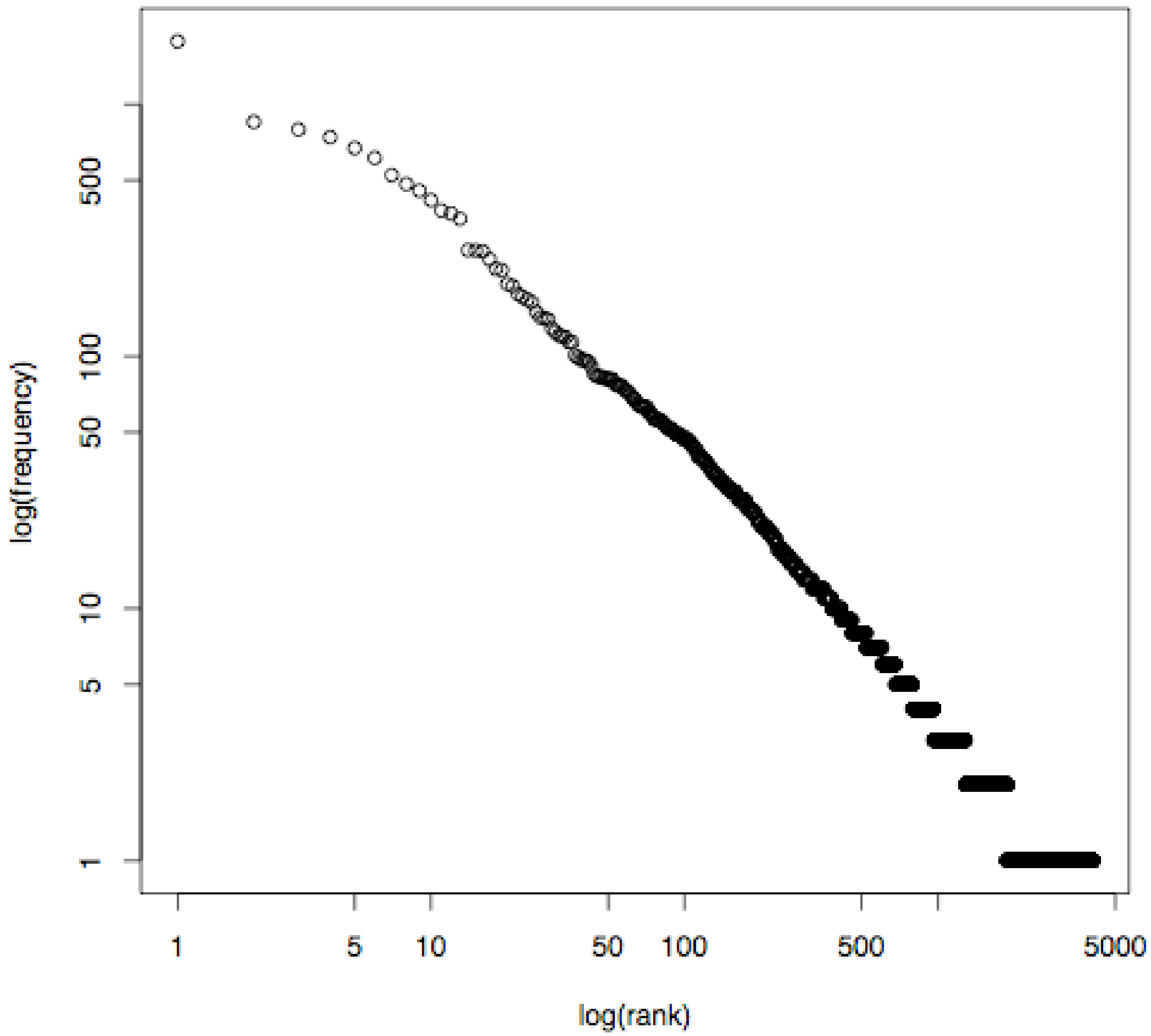
http://www.imdb.com/

Zipf's Law
IMDB Corpus



Does Zipf's law generalize across different domains?

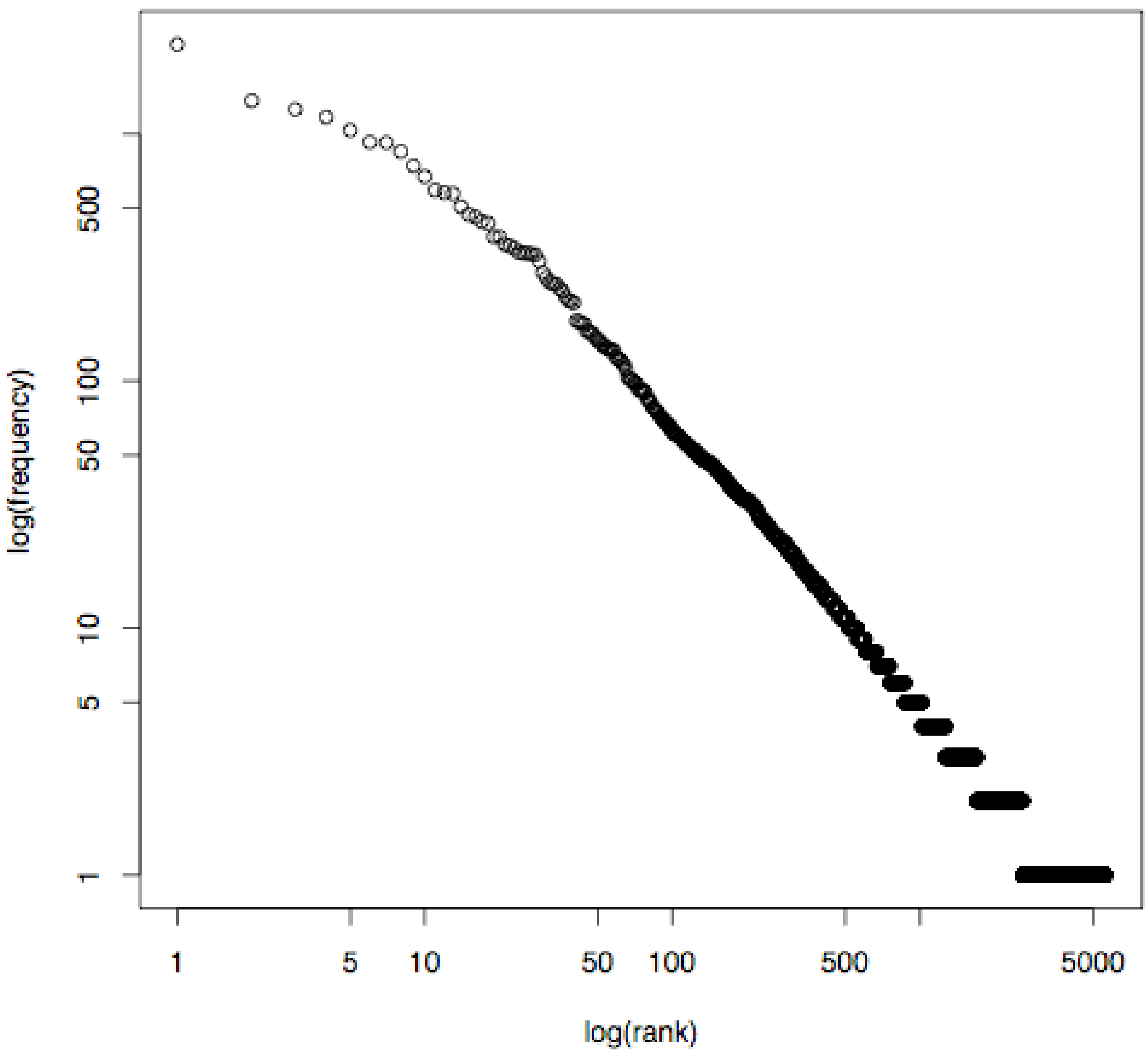
Zipf's Law
Alice in Wonderland



(text courtesy of Project Gutenberg)

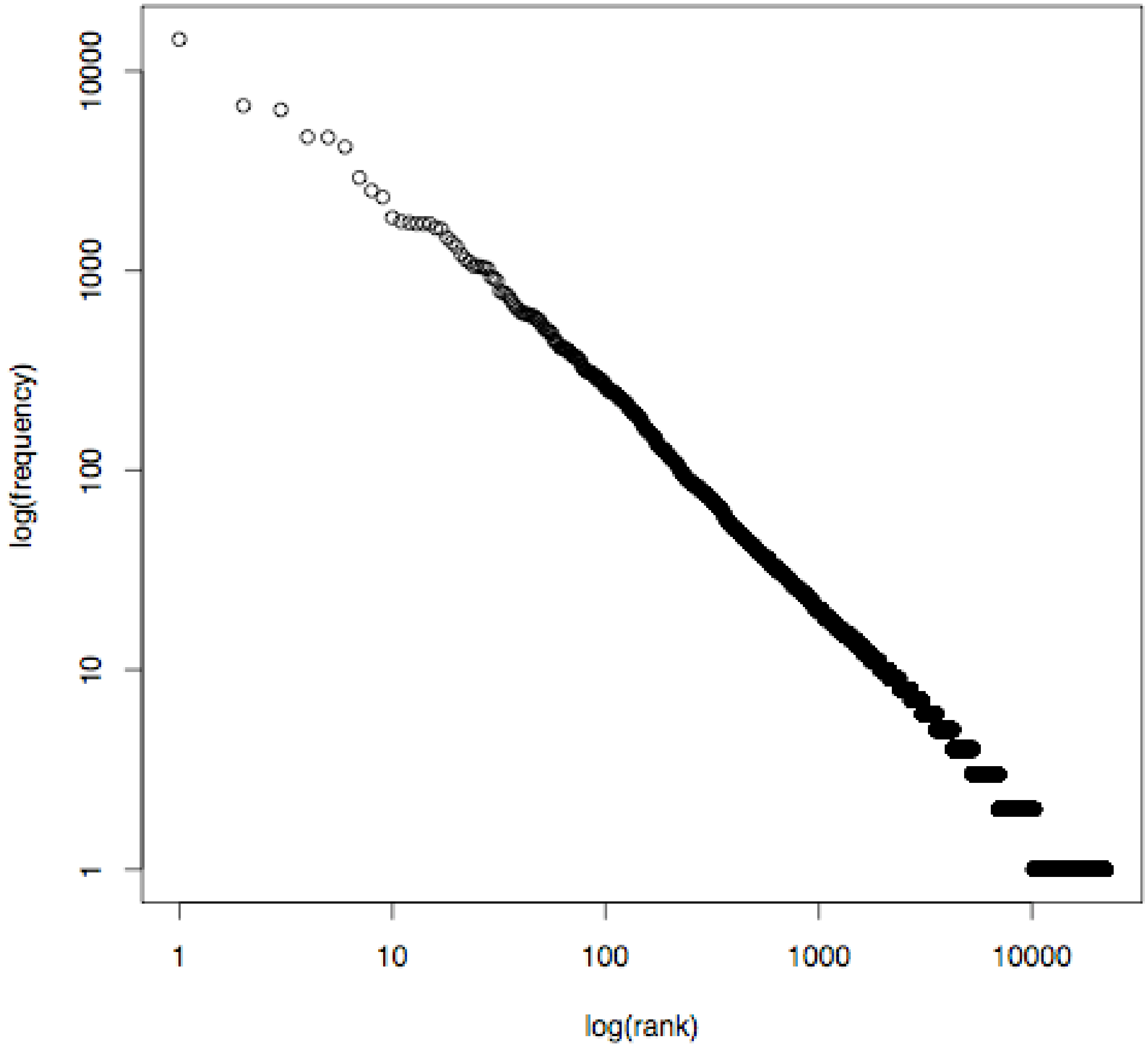
Zipf's Law

Peter Pan

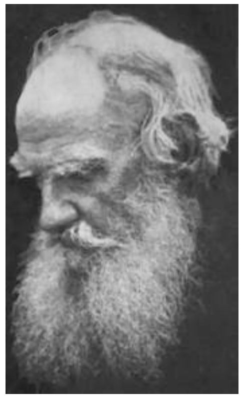


(text courtesy of Project Gutenberg)

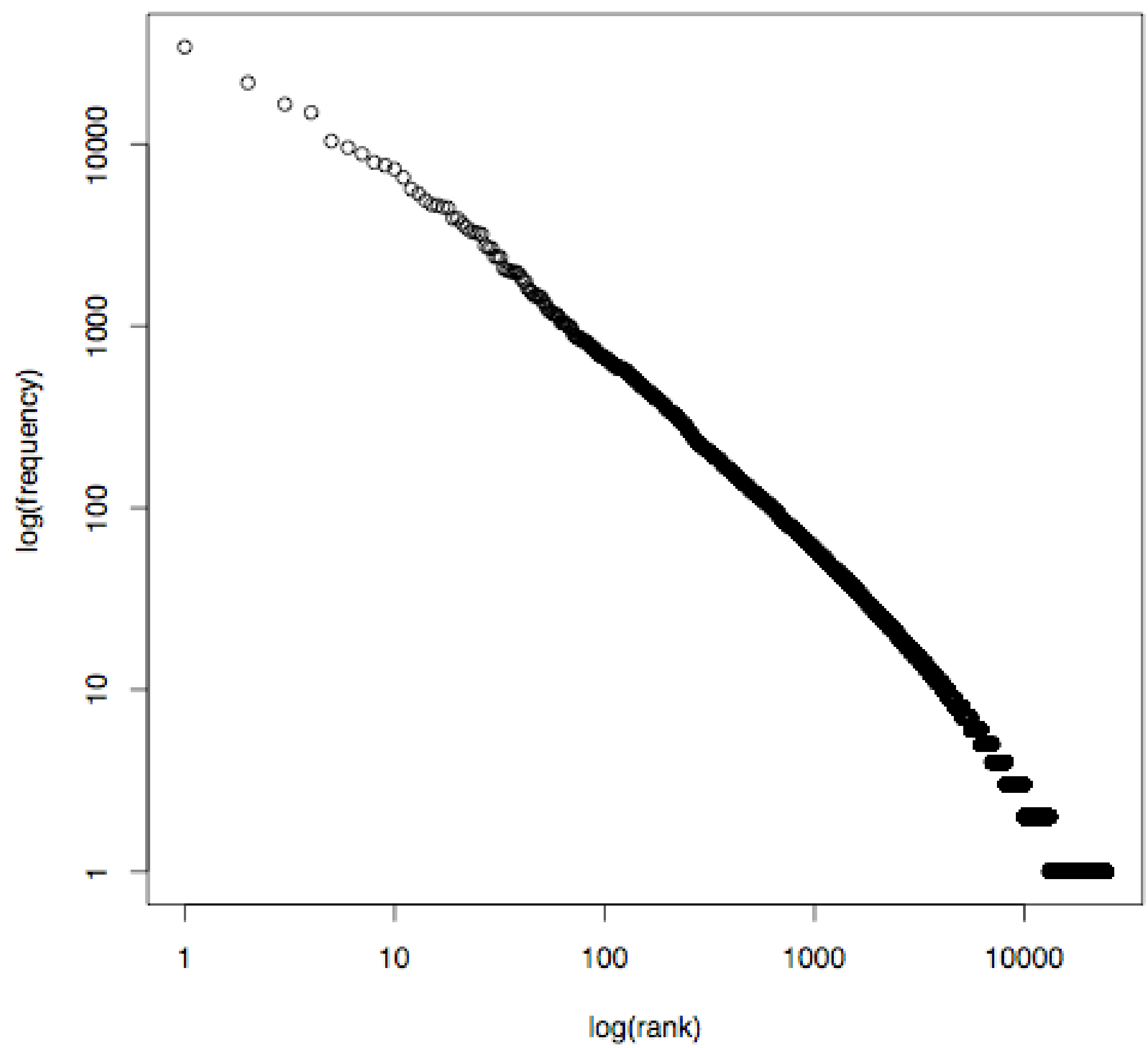
Zipf's Law
Moby Dick



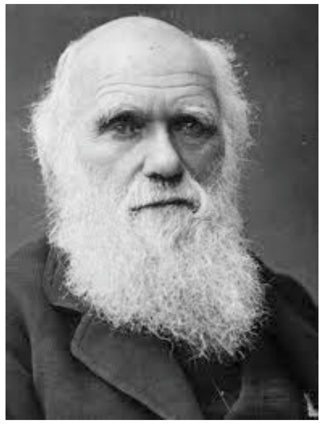
(text courtesy of Project Gutenberg)



Zipf's Law
War and Peace

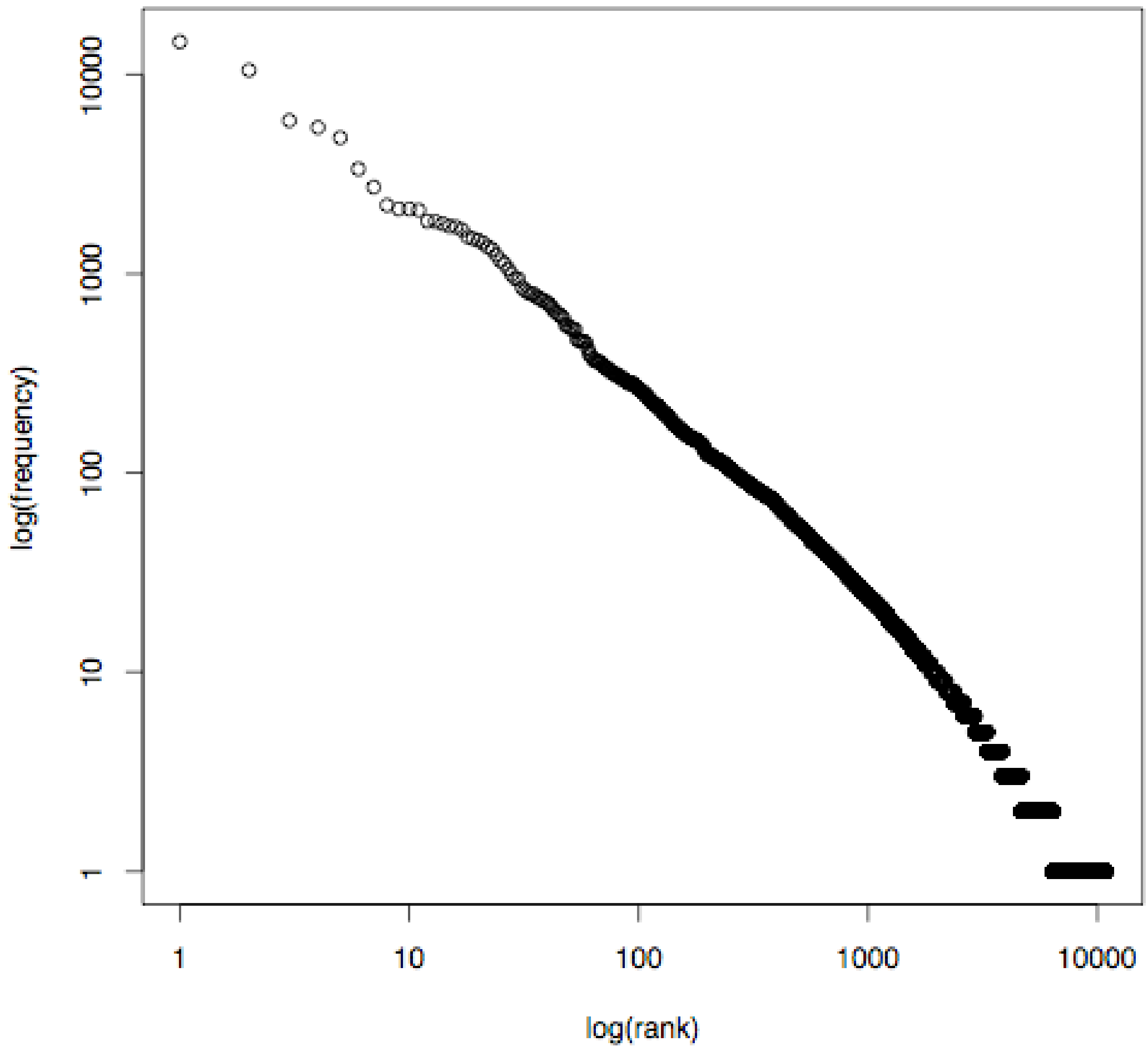


(text courtesy of Project Gutenberg)

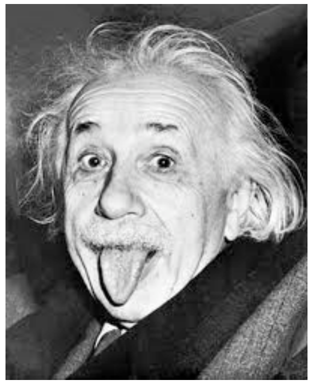


Zipf's Law

On the Origin of Species

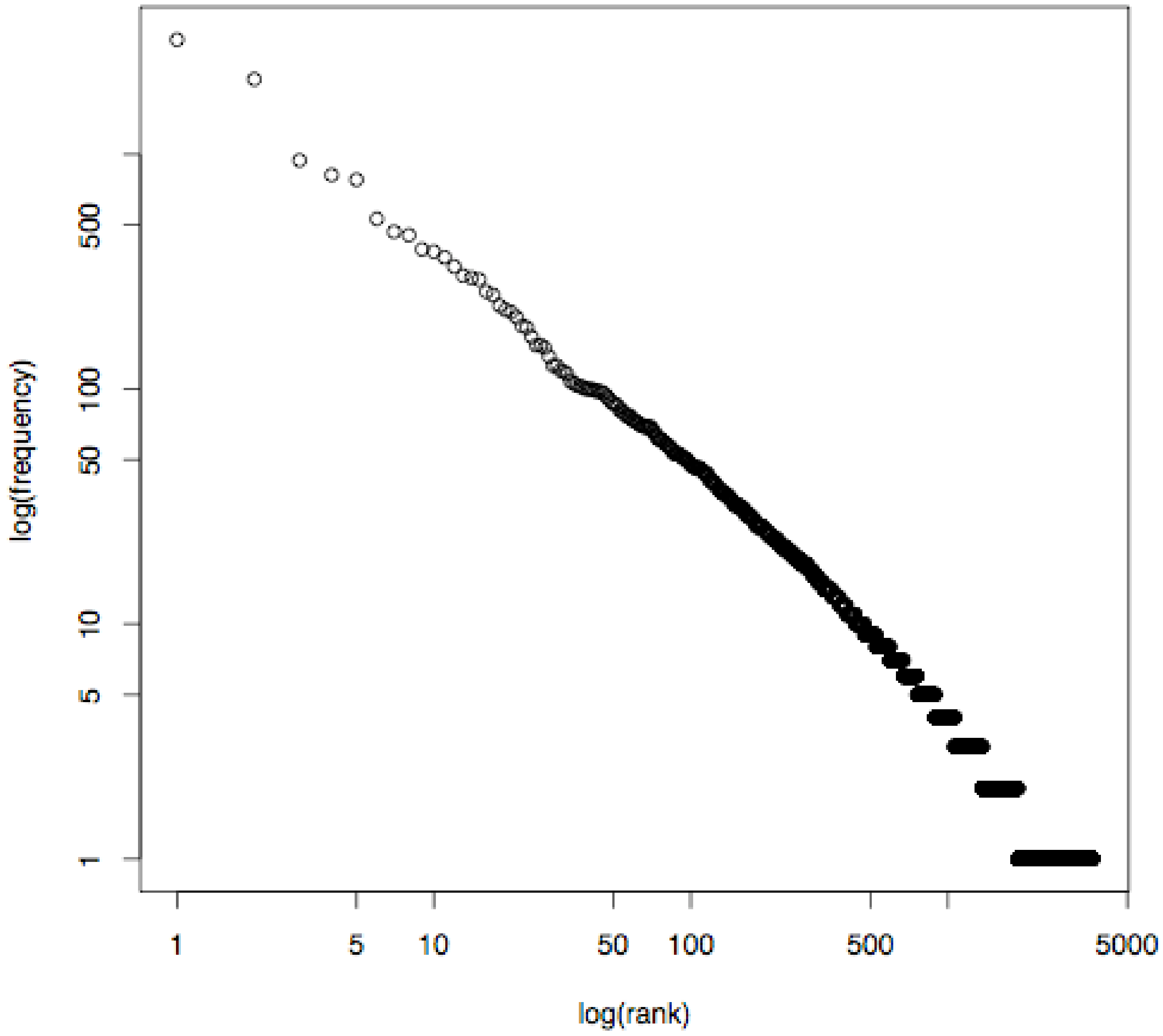


(text courtesy of Project Gutenberg)



Zipf's Law

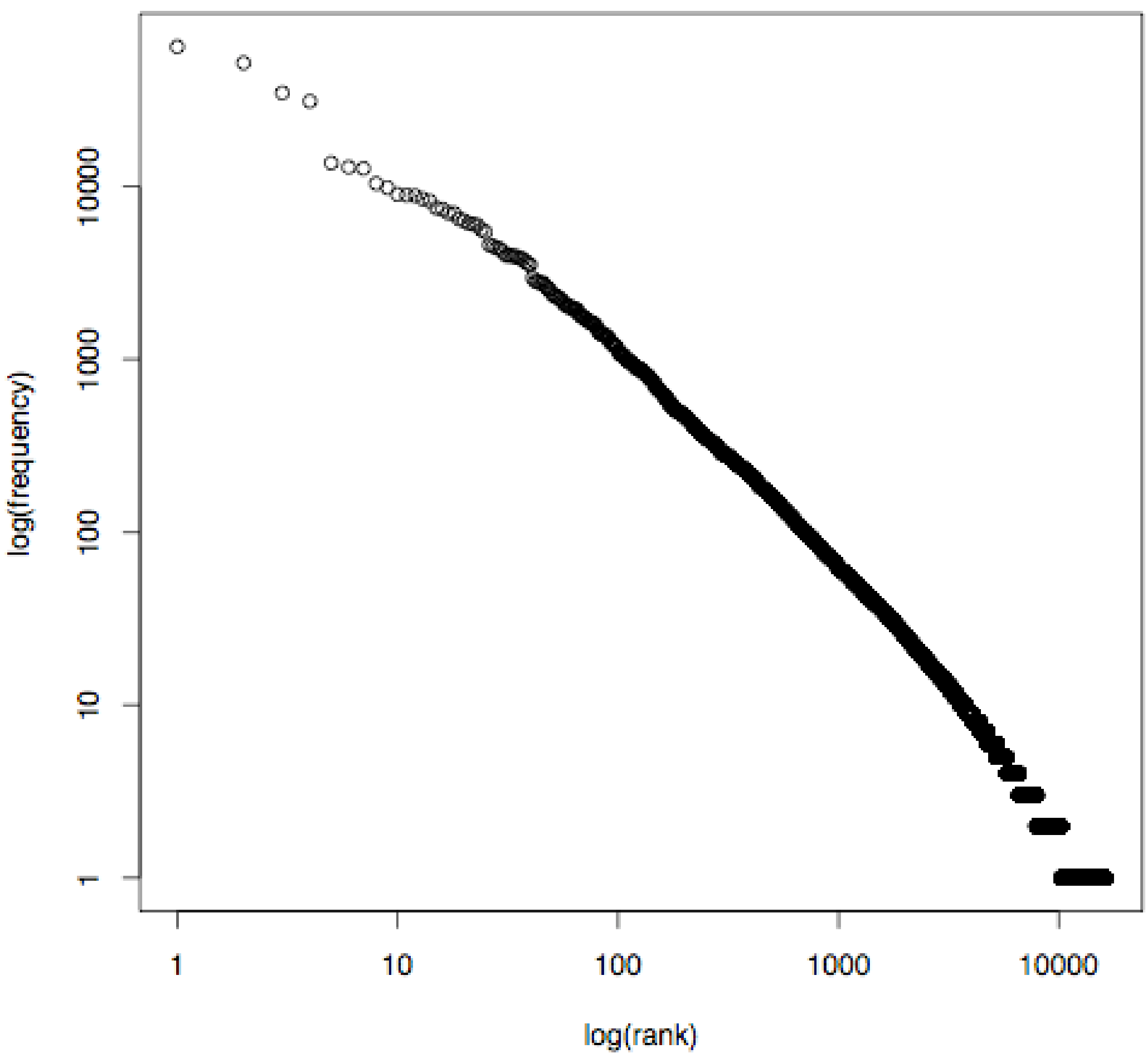
Relativity: The Special and General Theory



(text courtesy of Project Gutenberg)

Zipf's Law

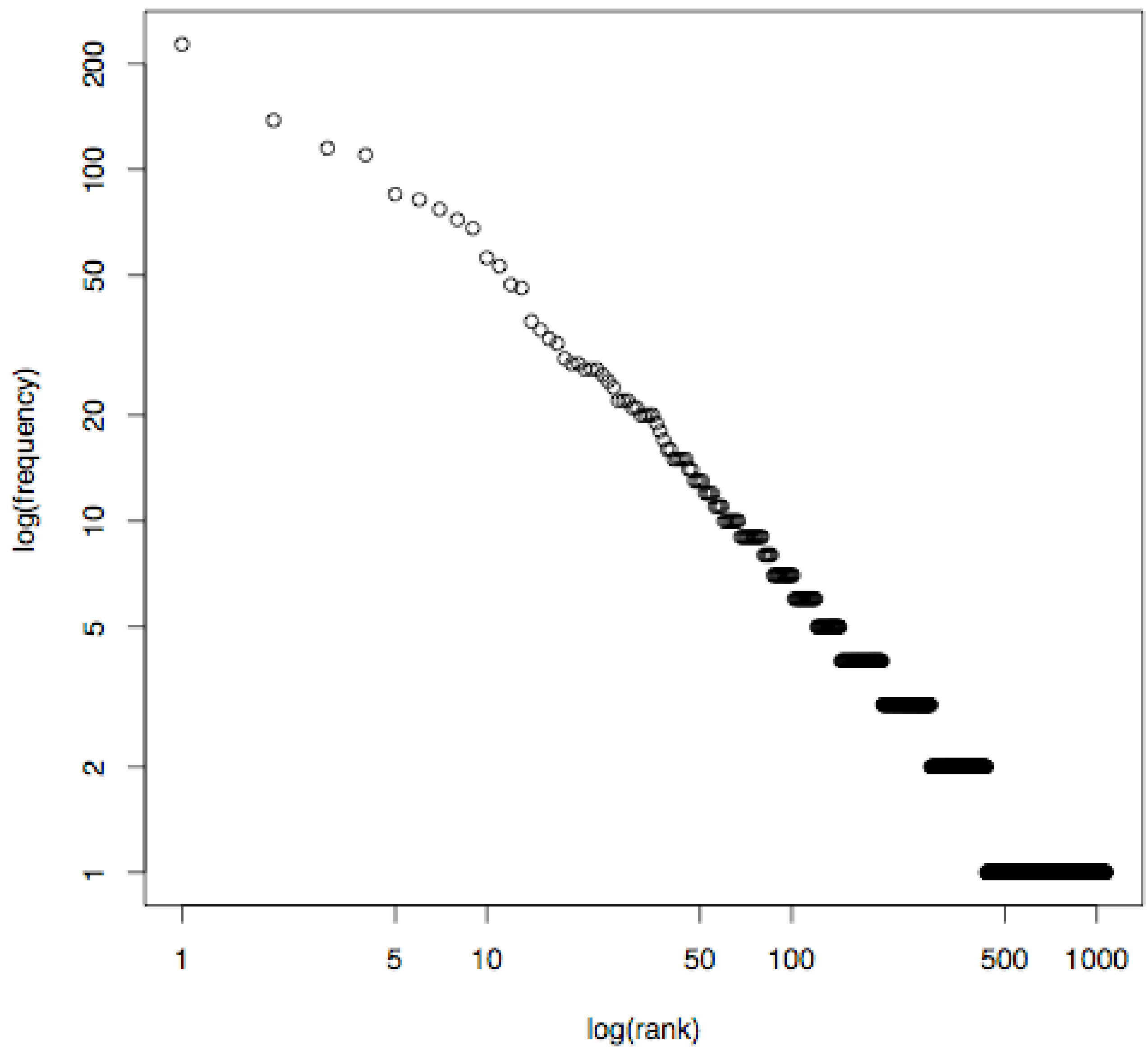
The King James Bible



(text courtesy of Project Gutenberg)

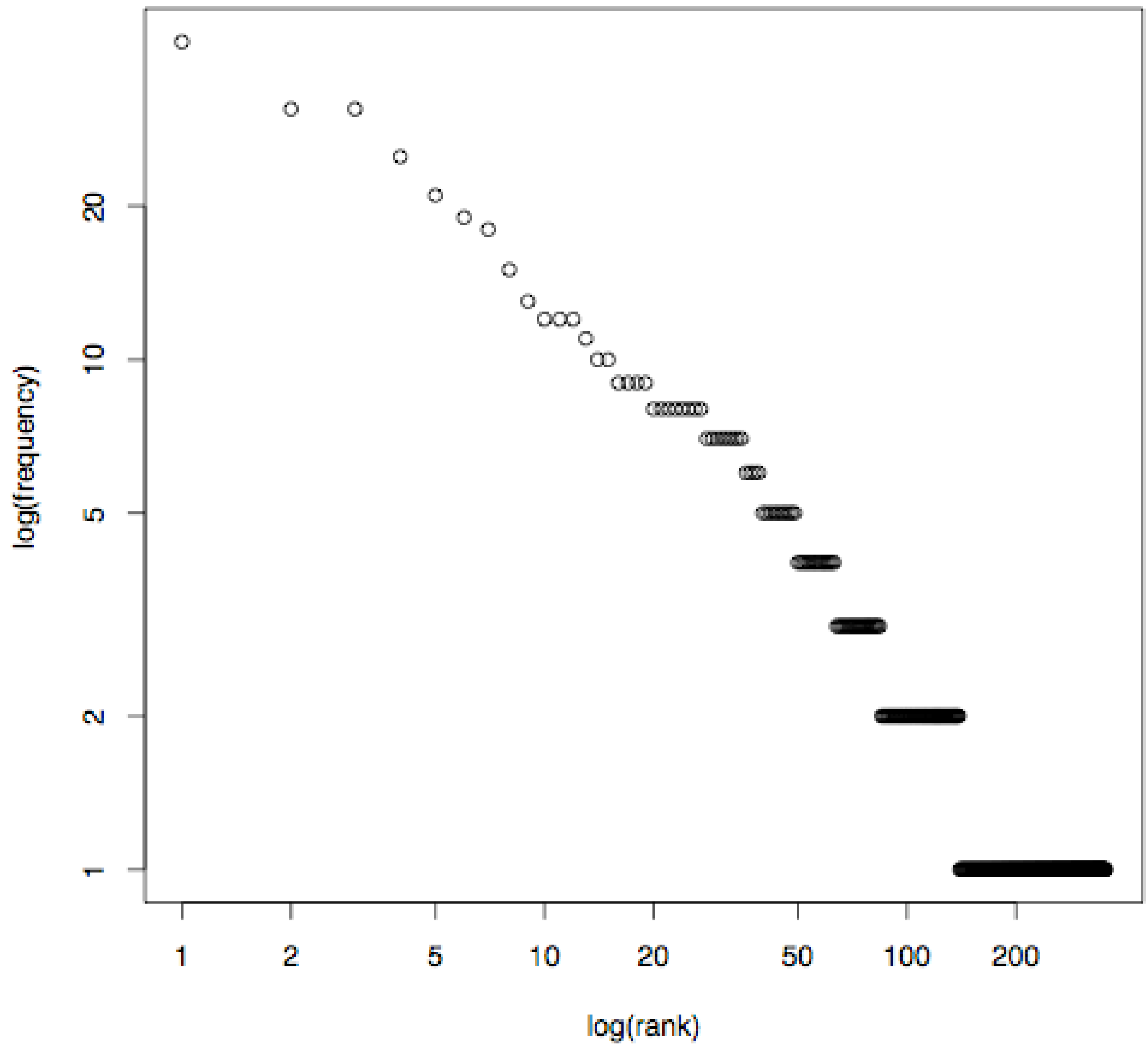
Zipf's Law

The Tale of Peter Rabbit



(text courtesy of Project Gutenberg)

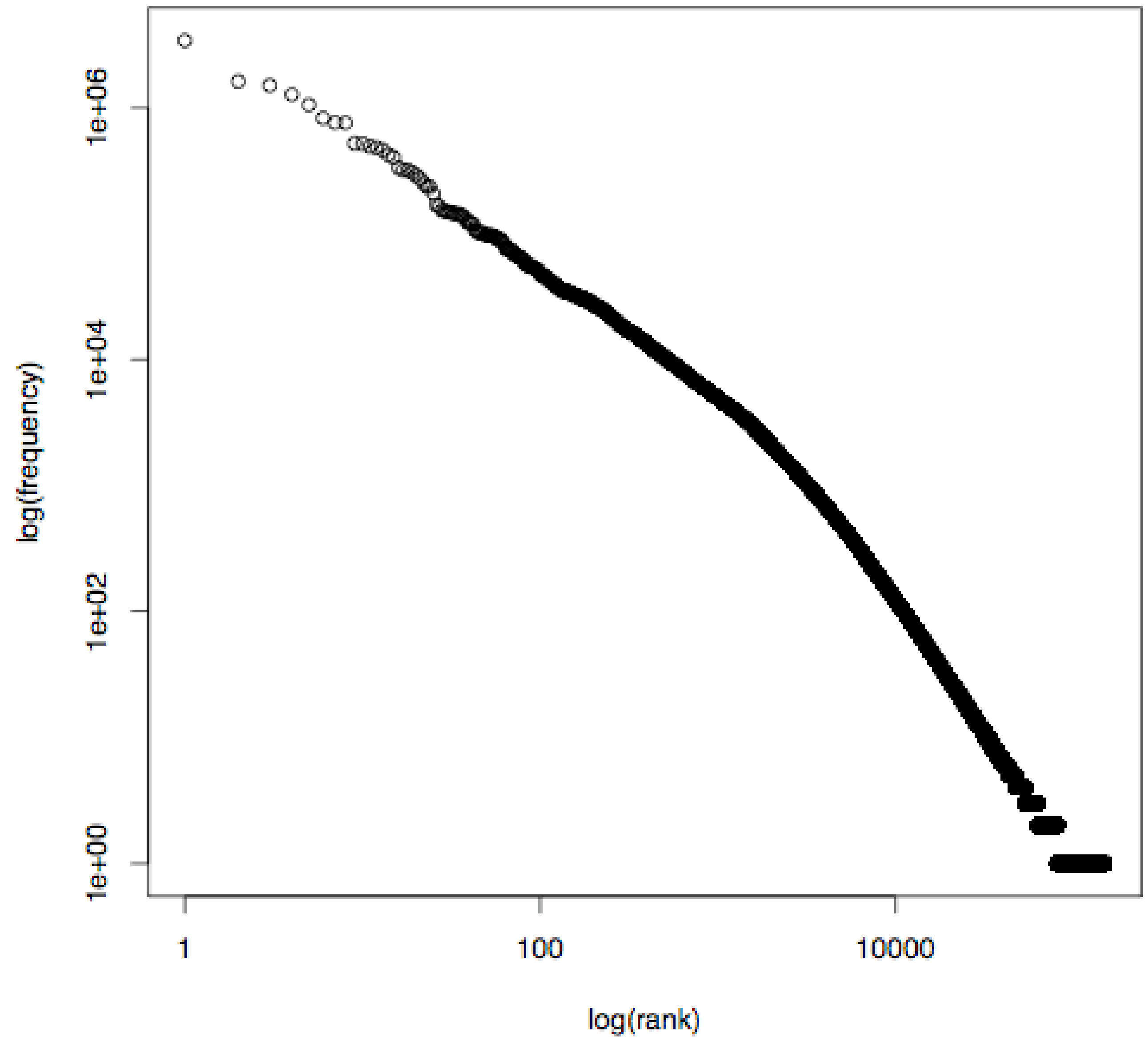
Zipf's Law
The Three Bears



(text courtesy of Project Gutenberg)

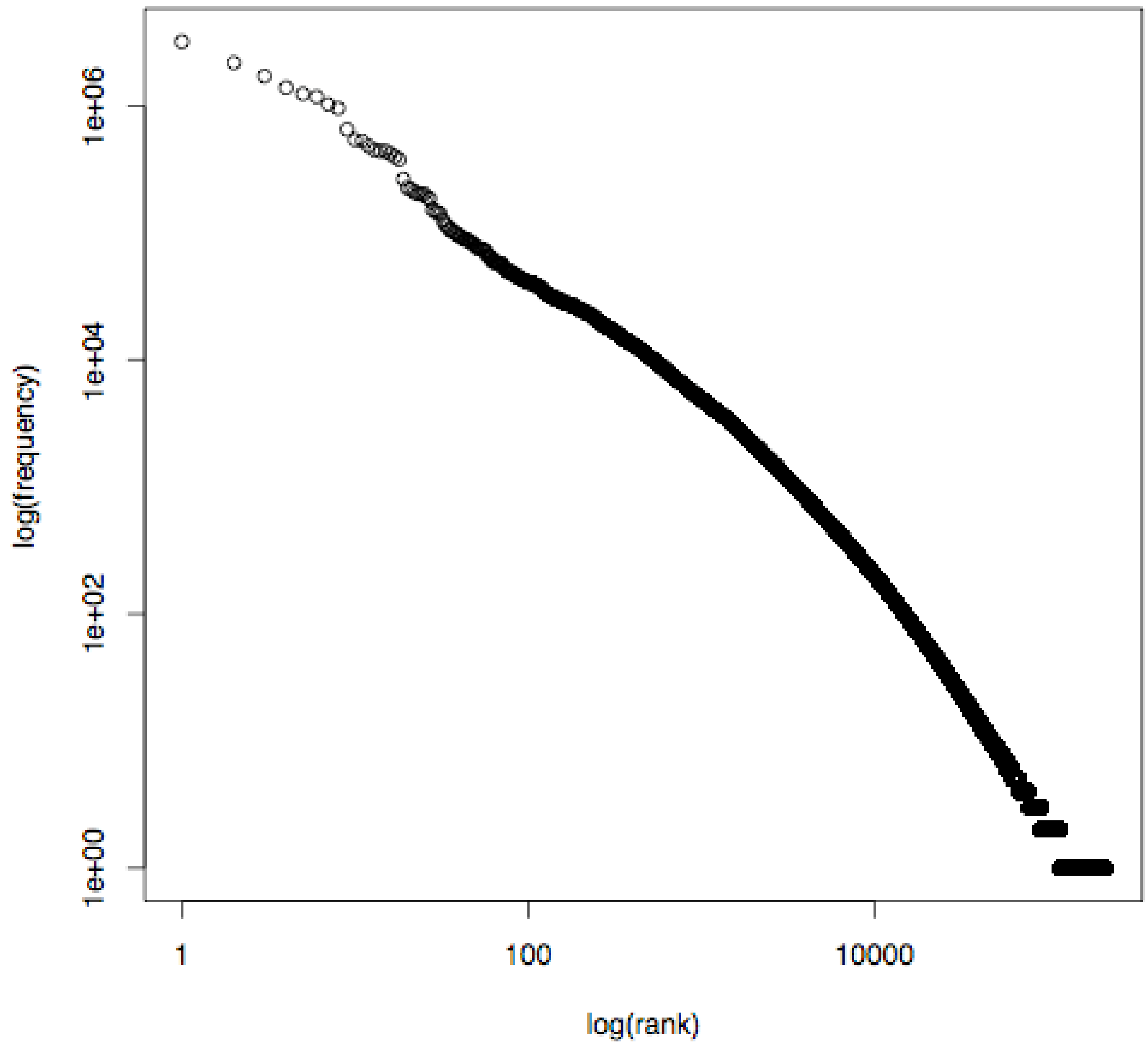
Does Zipf's law generalize across different languages?

Zipf's Law
European Parliament: English



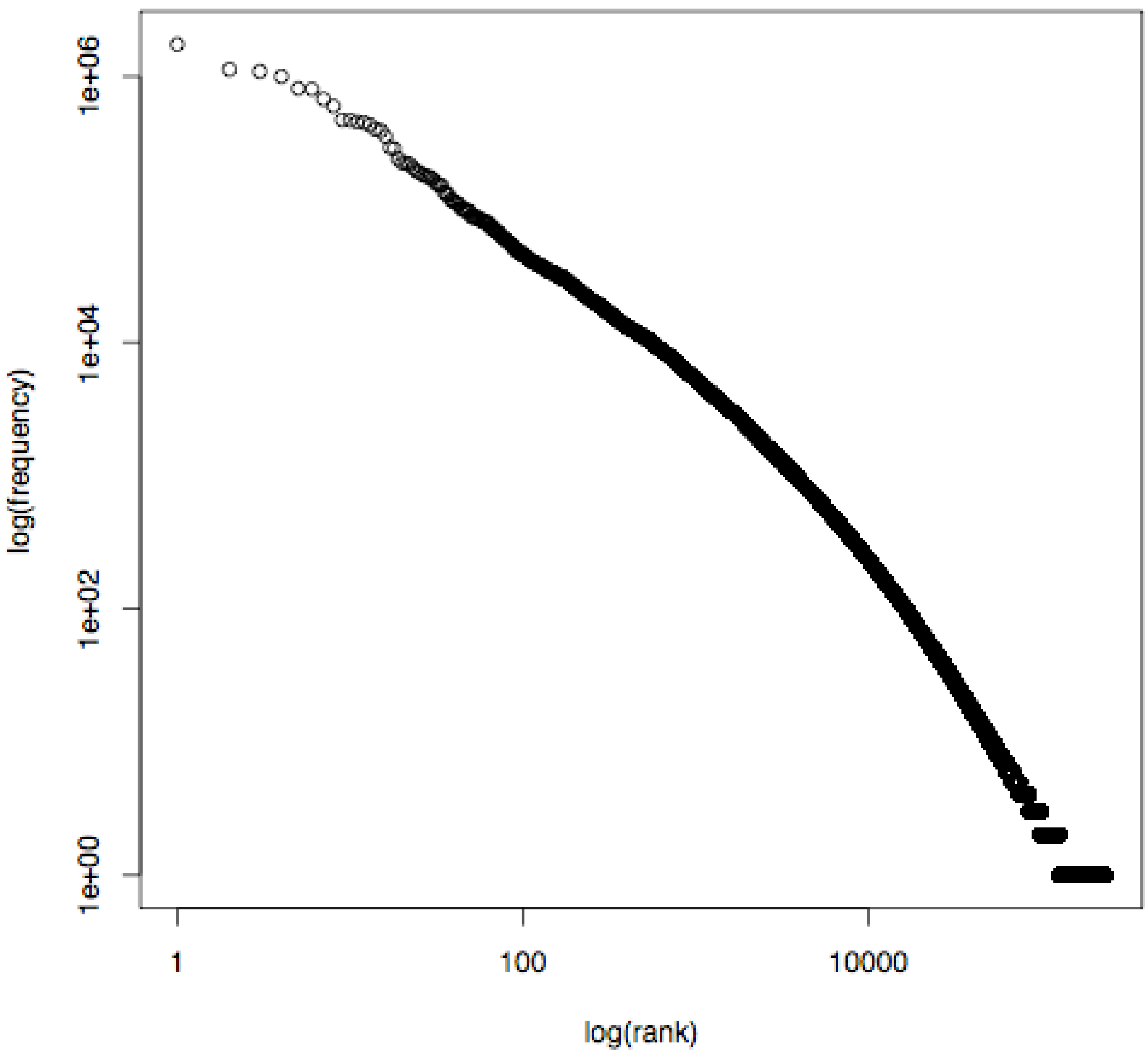
- Transcribed speech from proceedings of the European Parliament **(Koehn '05)**

Zipf's Law
European Parliament: Spanish



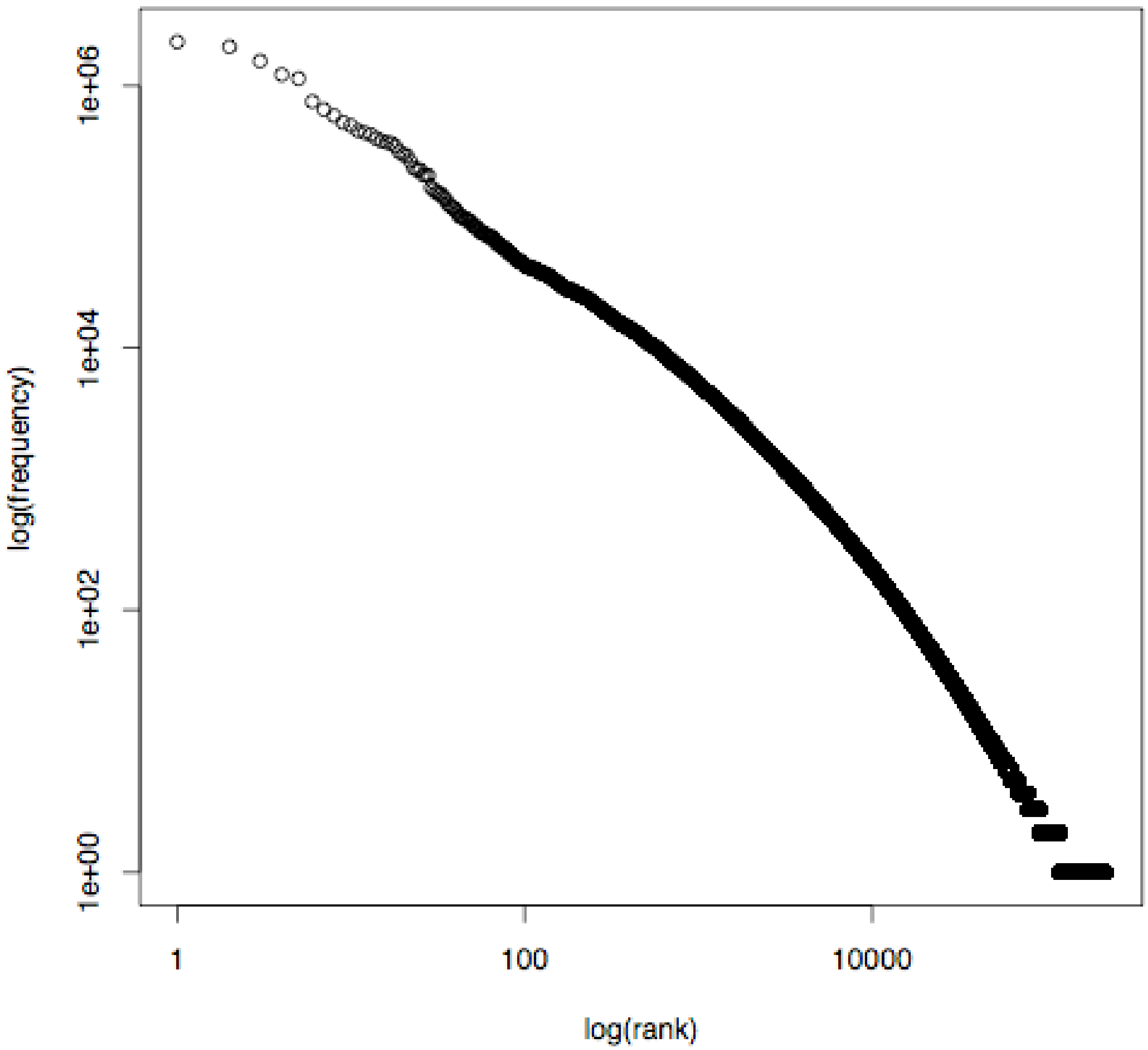
Zipf's Law

European Parliament: Italian

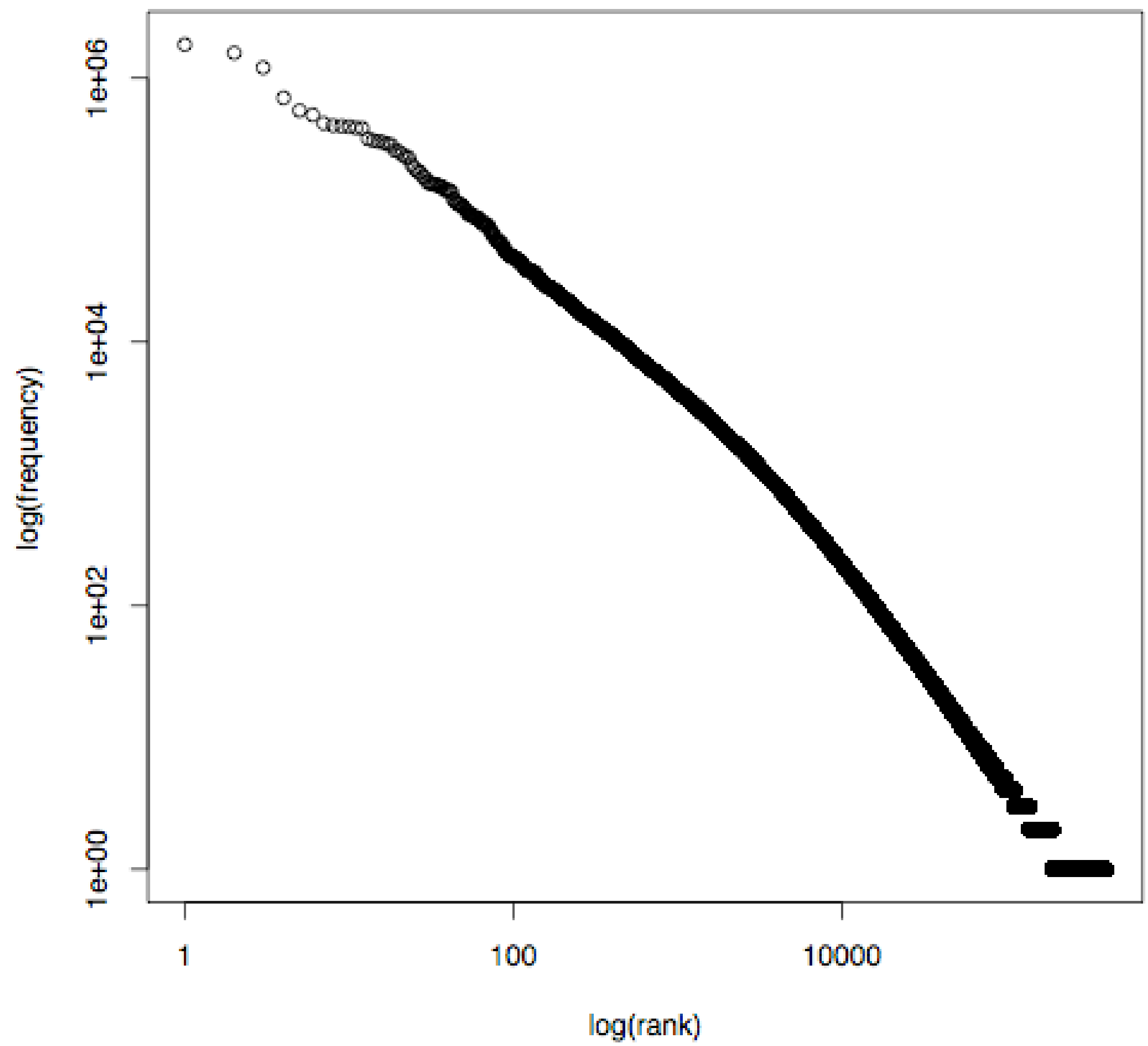


Zipf's Law

European Parliament: Portuguese

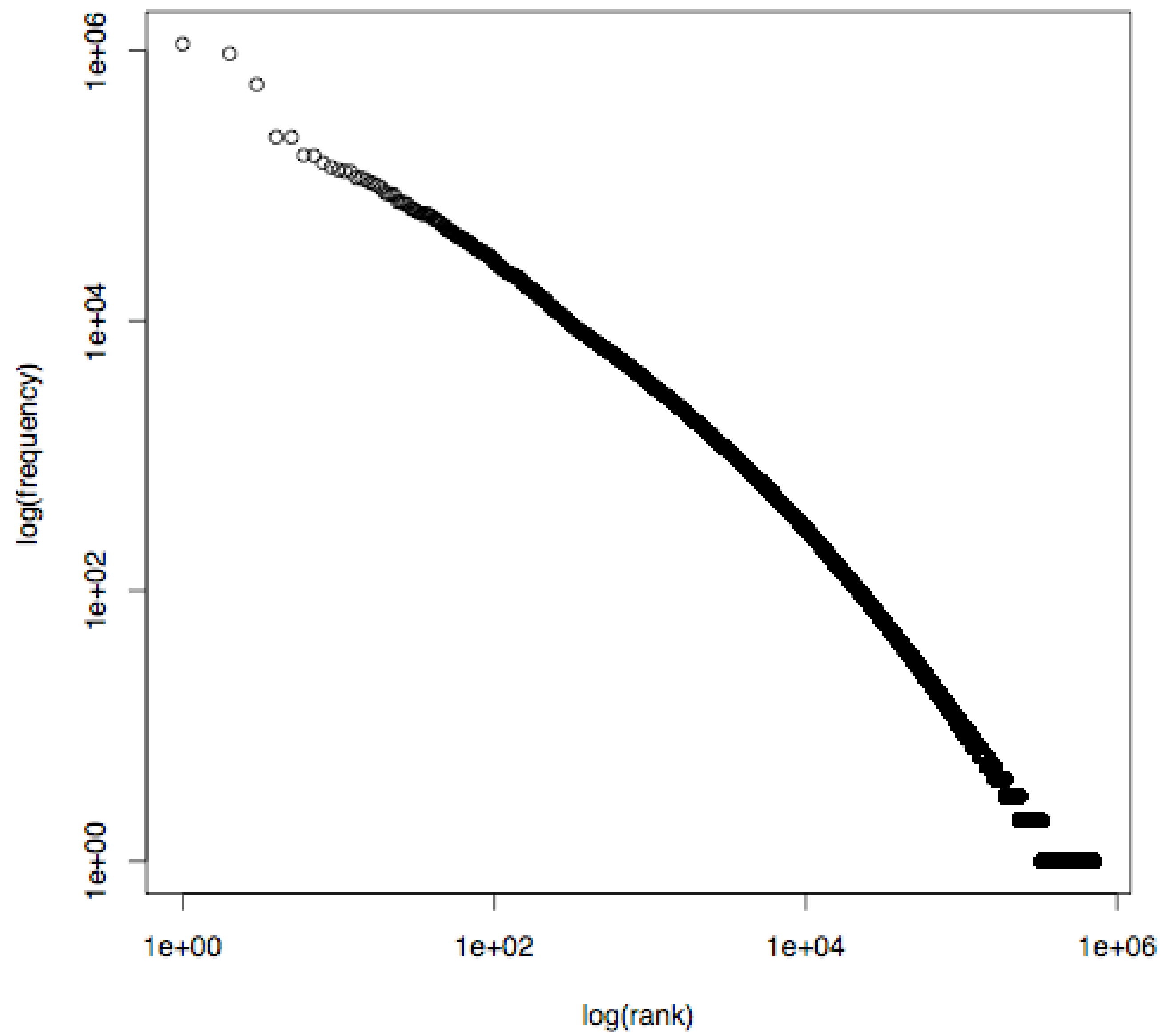


Zipf's Law
European Parliament: German



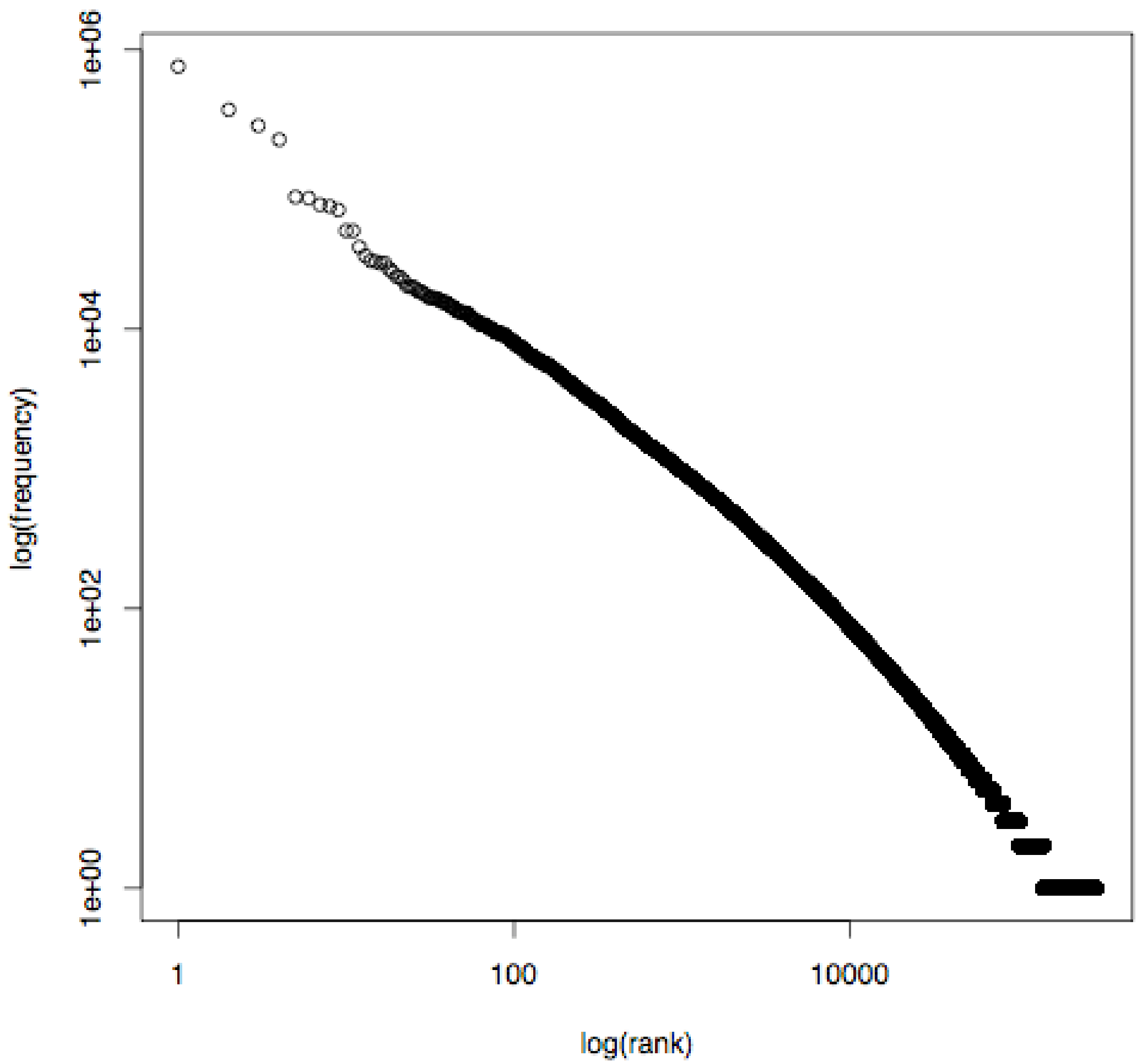
Zipf's Law

European Parliament: Finnish



Zipf's Law

European Parliament: Hungarian



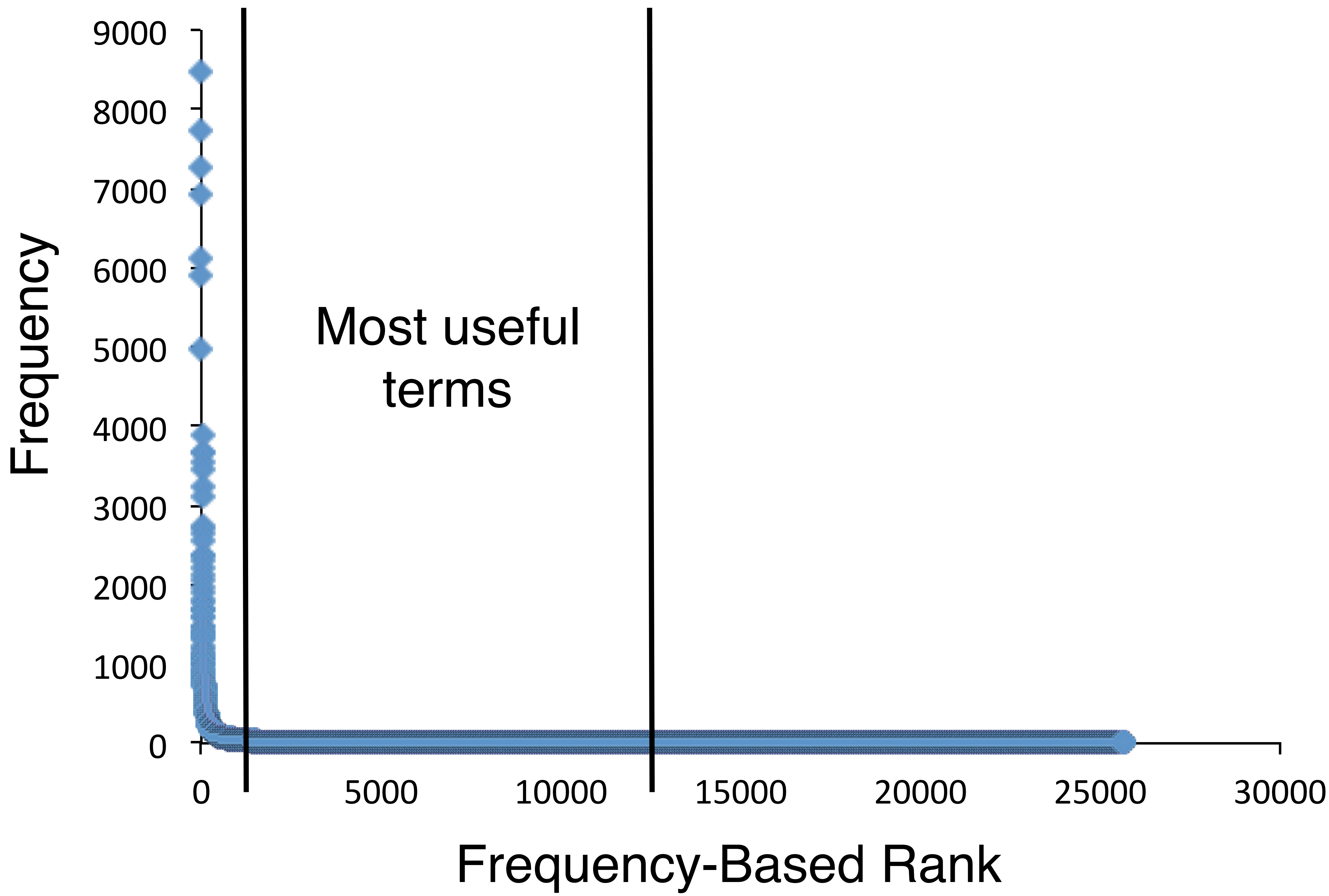
Zipf's Law

- Zipf's Law holds true for:
  - different dataset sizes
  - different domains
  - different languages

Feature Selection

- Unsupervised Feature Selection
  - does not require training data
  - potentially useful features are selected using term and dataset statistics
- Supervised Feature Selection
  - requires training data (e.g., positive/negative labels)
  - potentially useful features are selected using co-occurrence statistics between terms and the target label

Zipf's Law
Implications for Feature Selection



Supervised Feature Selection

Supervised Feature Selection

- What are the terms that tend to co-occur with a particular class value (e.g., positive or negative)?

A Few Important Concepts in Probability 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 possible outcomes: TRUE or FALSE

Boolean-Valued Random Variables

Examples

- A = it will rain tomorrow
- A = the outcome of a coin-flip will be heads
- A = The US president in 2023 will be female
- A = the word “good” appears in a positive review

Probabilities

- P(A=TRUE): the probability that the outcome is TRUE
  - the probability that it will rain tomorrow
  - the probability that a coin-flip will be "heads"
  - the probability that "good" appears in a positive review
- P(A=FALSE): the probability that the outcome is FALSE
  - the probability that it will NOT rain tomorrow
  - the probability that a coin-flip will NOT be "heads"
  - the probability that "good" does NOT appears in a positive review

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(coin-flip is “heads”=TRUE)
- P(rain tomorrow=TRUE)
- P(female president in 2023=TRUE)
- P(“good” in a positive review=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 "good" appears in a positive review?
- Statistical Estimation Example:
  - To gather data, you take a sample of 1000 positive reviews
  - You observe that "good" appears in 550 of them
  - What would be your estimation of P("good" in positive review=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 A <u>and</u> B both true
- P(A|B): the probability that A is true give prior knowledge that B is also true

Chain Rule

- $P(A, B) = P(A|B) \times P(B)$
- Example:
  - probability that it will rain today <u>and</u> tomorrow
  - probability that it will rain today
  - probability that it will rain tomorrow given prior knowledge that it rains 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 = coin flip is “heads”
  - Are these likely to be independent?

Mutual Information

$$\mathrm{MI}(w,c) = \log\left(\frac{P(w,c)}{P(w)P(c)}\right)$$

- **P(w,c)**: the probability that word **w** and class value **c** occur together
- **P(w)**: the probability that word **w** occurs (with or without class value **c**)
- **P(c)**: probability that class value **c** occurs (with or without word **w**)

Mutual Information

$$\mathrm{MI}(w,c) = \log\left(\frac{P(w,c)}{P(w)P(c)}\right)$$

- If **P(w,c) = P(w) P(c)**, it means that the word **w** is <u>independent</u> of class value **c**
- If **P(w,c) > P(w) P(c)**, it means that the word **w** is <u>dependent</u> of class value **c**

Mutual Information

- Every instance falls under one of these quadrants

	class value **c** occurs	class value **c** does not occur
word **w** occurs	a	b
word **w** does not occur	c	d

total # of instances N = a + b + c + d

P(w, c) = ?

P(c) = ?

P(w) = ?

$$\mathrm{MI}(w, c) = \log\left(\frac{P(w, c)}{P(w)P(c)}\right)$$

Mutual Information

- Every instance falls under one of these quadrants

	class value **c** occurs	class value **c** does not occur
word **w** occurs	a	b
word **w** does not occur	c	d

total # of instances N = a + b + c + d

P(w, c) = a / N

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

P(w) = (a + b) / N

$$\mathrm{MI}(w,c) = \log\left(\frac{P(w,c)}{P(w)P(c)}\right)$$

HW1 Training Set

terms correlated with positive class

term	MI	term	MI	term	MI
captures	0.69315	urban	0.60614	fellow	0.58192
viewings	0.69315	overlooked	0.59784	masterpiece	0.57808
extraordinary	0.62415	breathtaking	0.59784	legend	0.57536
allows	0.62415	biography	0.59784	awards	0.55962
delight	0.61904	intensity	0.59784	donald	0.55962
wayne	0.61904	represent	0.59784	journey	0.555
unforgettable	0.61904	elegant	0.59784	traditional	0.55005
sentimental	0.61904	emma	0.59784	seasons	0.55005
touching	0.61619	deliberate	0.59784	mass	0.539
essence	0.6131	friendship	0.59784	court	0.539
superb	0.6131	splendid	0.59784	princess	0.539
underrated	0.6131	desires	0.59784	refreshing	0.539
devoted	0.60614	terrific	0.59784	drunken	0.539
frightening	0.60614	delightful	0.59306	adapted	0.539
perfection	0.60614	gorgeous	0.59306	stewart	0.539

HW1 Training Set

terms correlated with negative class

term	MI	term	MI	term	MI
atrocious	0.693147181	gross	0.613104473	existent	0.575364145
blatant	0.693147181	appalling	0.606135804	dumb	0.572519193
miserably	0.693147181	unintentional	0.606135804	zero	0.571786324
unfunny	0.693147181	drivel	0.606135804	!@#$	0.568849464
unconvincing	0.693147181	pointless	0.60077386	amateurish	0.567984038
stupidity	0.693147181	unbelievably	0.597837001	garbage	0.559615788
blah	0.693147181	blockbuster	0.597837001	dreadful	0.559615788
suck	0.693147181	stinker	0.597837001	horribly	0.559615788
sounded	0.693147181	renting	0.597837001	tedious	0.550046337
redeeming	0.660357358	idiotic	0.597837001	uninteresting	0.550046337
laughable	0.652325186	awful	0.596154915	wasted	0.550046337
downright	0.624154309	lame	0.585516516	insult	0.550046337
irritating	0.619039208	worst	0.58129888	horrible	0.547193268
waste	0.613810438	brain	0.579818495	pretentious	0.546543706
horrid	0.613104473	sucks	0.575364145	offensive	0.546543706

Co-occurrence Statistics

- Mutual Information
- Chi-squared
- Term strength
- Information Gain
- For a nice review, see:
  - Yang and Pedersen. A Comparative Study of Feature Selection for Text Categorization. 1997

Chi Squared

- Every instance falls under one of these quadrants

	class value **c** occurs	class value **c** does not occur
word **w** occurs	a	b
word **w** does not occur	c	d

$$\chi^2(w,c) = \frac{N \times (ad - cb)^2}{(a+c) \times (b+d) \times (a+b) \times (c+d)}$$

HW1 Training Set

chi-squared term statistics

term	chi-squared	term	chi-squared	term	chi-squared
bad	160.9971465	best	42.61226642	guy	30.21744225
worst	129.7245814	love	40.85783977	highly	30.18018867
great	114.4167082	even	39.61387169	very	29.04056204
waste	90.05925899	don	38.87461084	masterpiece	28.83716791
awful	84.06935342	superb	38.22460907	amazing	28.79058228
nothing	49.63235294	excellent	36.35817308	fantastic	28.42431877
boring	48.08302214	only	35.37872166	i	28.07171446
!@#$	47.01798462	minutes	34.16970651	redeeming	27.55615262
stupid	47.01038257	worse	33.43003177	dumb	26.86372932
terrible	46.87740534	no	33.13496711	ridiculous	26.73027231
t	46.72237358	poor	32.66596825	any	25.86206897
acting	46.36780576	lame	31.82041653	like	25.69031789
horrible	44.78927425	annoying	31.32494449	mess	25.58837466
supposed	44.48292448	brilliant	30.89314779	poorly	25.58837466
wonderful	43.24661832	make	30.61995968	not	25.47840442

HW1 Training Set

chi-squared term statistics

term	chi-squared	term	chi-squared	term	chi-squared
avoid	24.64813529	cheap	22.26804124	gore	19.46385538
plot	24.32739264	favorite	22.21941826	this	19.3814528
loved	24.13368514	always	21.72980415	perfect	19.28060105
oh	24.10901468	laughable	21.4278481	so	19.26007925
lives	23.93399462	family	21.40903284	beautiful	19.25267715
m	23.85882353	better	21.35884719	role	19.14529915
pointless	23.45760278	zero	21.19956379	classic	19.13622759
garbage	22.95918367	unless	20.938872	anything	19.02801032
they	22.8954747	1	20.88669951	unfortunately	18.9261532
or	22.68259489	there	20.4478906	also	18.48036413
script	22.60364052	half	20.23467433	8	18.18641071
terrific	22.46152424	unfunny	20.2020202	suck	18.16347124
performance	22.42822967	low	19.89567408	brain	17.53115039
money	22.34443913	touching	19.86071221	guess	17.52876709
movie	22.34161803	attempt	19.75051975	were	17.49633958

Conclusions

- Bag-of-words feature representation: describing textual instances using individual terms
- Feature selection: reducing the number of features to only the most meaningful/predictive ones
- Unsupervised feature selection: omitting terms that are very frequent and very infrequent
- Supervised features selection: omitting terms that do not have a strong co-occurrence with each target-class value