Text Representation

Sandeep Avula
asandeep@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
⋮	⋮	⋮	…	0	⋮
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
⋮	⋮	⋮	...	0	⋮
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
⋮	⋮	⋮	…	0	⋮
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
- Stemming: transforming terms into some root-form representation
  - compute, computed, computing, computer, computers, compontational ➔ comput
  - a form of conflation

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

Text Processing
in Java

```
public String[] processText(String text){

    text = text.toLowerCase();

    return text.split(“[\\W]”);

}
```

Text Processing
in Python

```
import nltk

Import codecs

…

file = codecs.open(”file_name.txt", "r",
encoding='utf-8')

    lines = file.readlines()

    for text in lines:

        lower_text = text.lower()

        temp_tokens = nltk.word_tokenize(lower_text)
```

Slide borrowed from Heejun Kim

Bag of Words Text Representation

- Which vocabulary terms should we include as features?
- All of them?
  - why might this be a good idea?
  - why might this be a bad 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 (neg)	term	count (neg)	term	count (neg)
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 (neg)	term	count (neg)	term	count (neg)
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

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 many 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

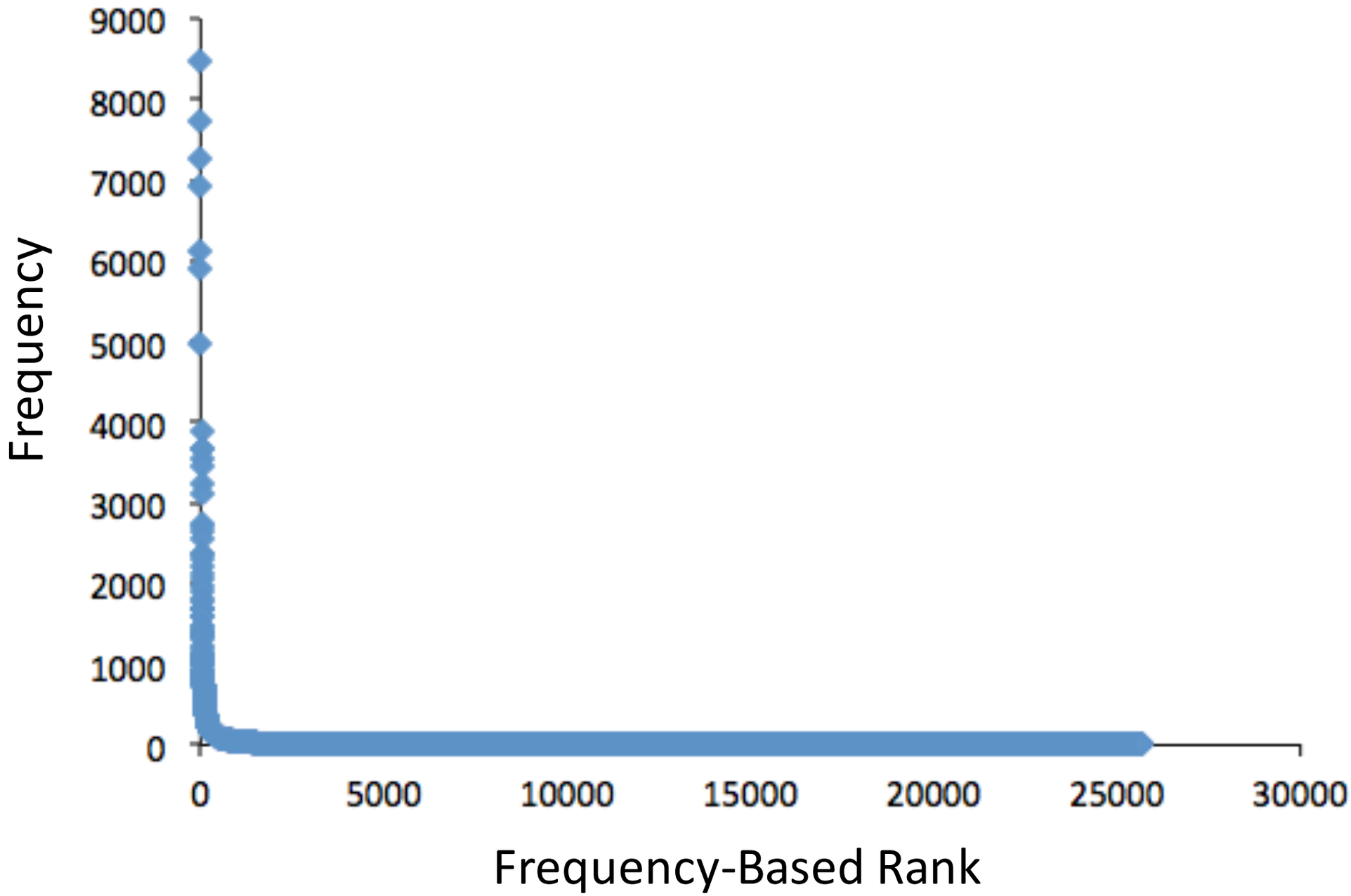
rank	term	frequency	rank	term	frequency
21	you	2722	31	at	1895
22	t	2660	32	they	1803
23	not	2560	33	by	1793
24	his	2376	34	who	1703
25	he	2366	35	so	1699
26	are	2315	36	an	1681
27	have	2230	37	from	1609
28	be	2133	38	like	1582
29	one	2069	39	there	1483
30	all	1980	40	her	1458

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

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** = constant (specific to the collection of text)
- 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

$$P_t = \frac{c}{r_t} \qquad \mathbf{c} = 0.1$$

- 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 algebraic manipulation, it also tells us that the fraction 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 fraction 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 faction 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:

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

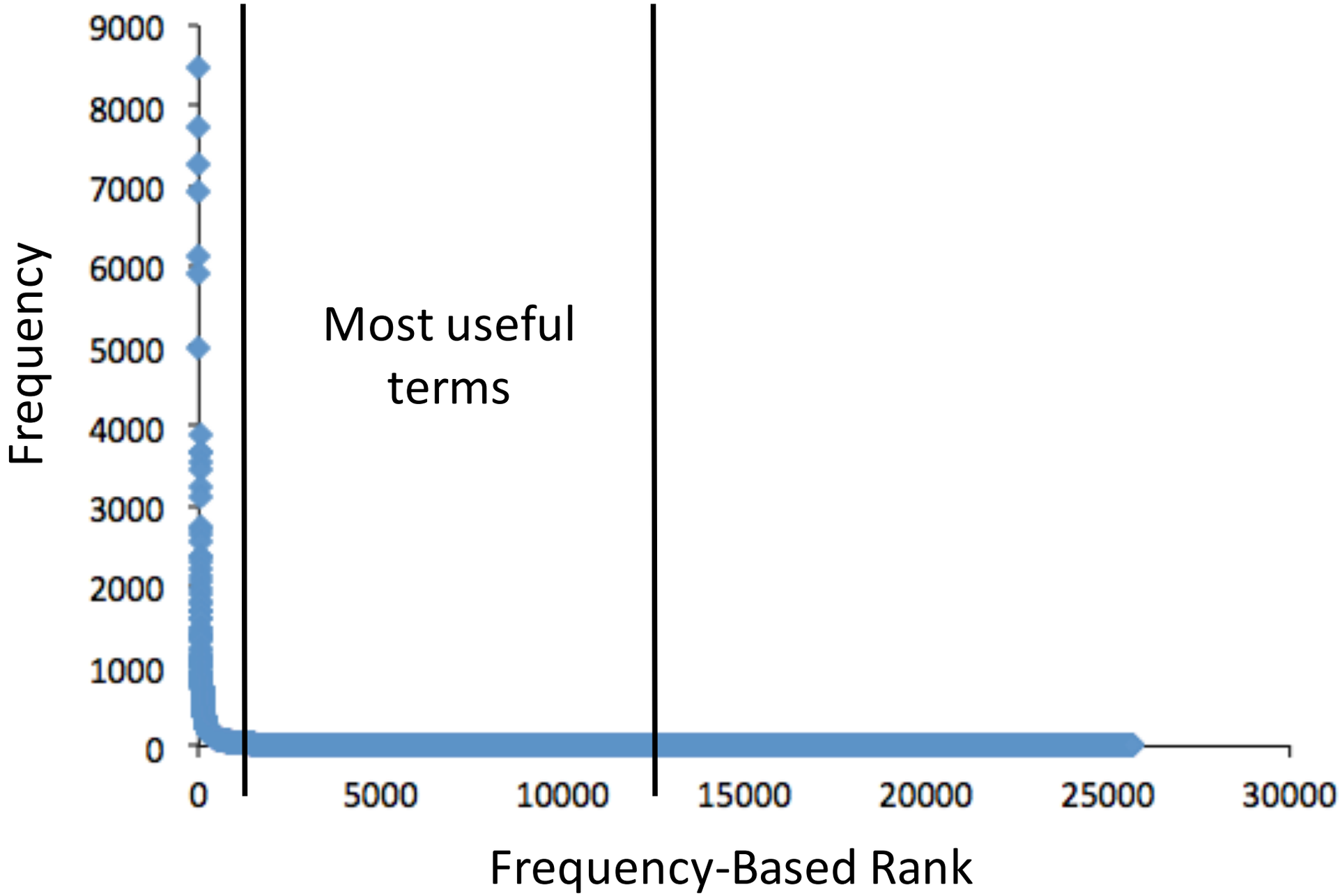
- That is, we have to add the fraction of terms that appear 1, 2, 3, ... up to **n** times

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 and 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 $$f = \frac{k}{r}$$

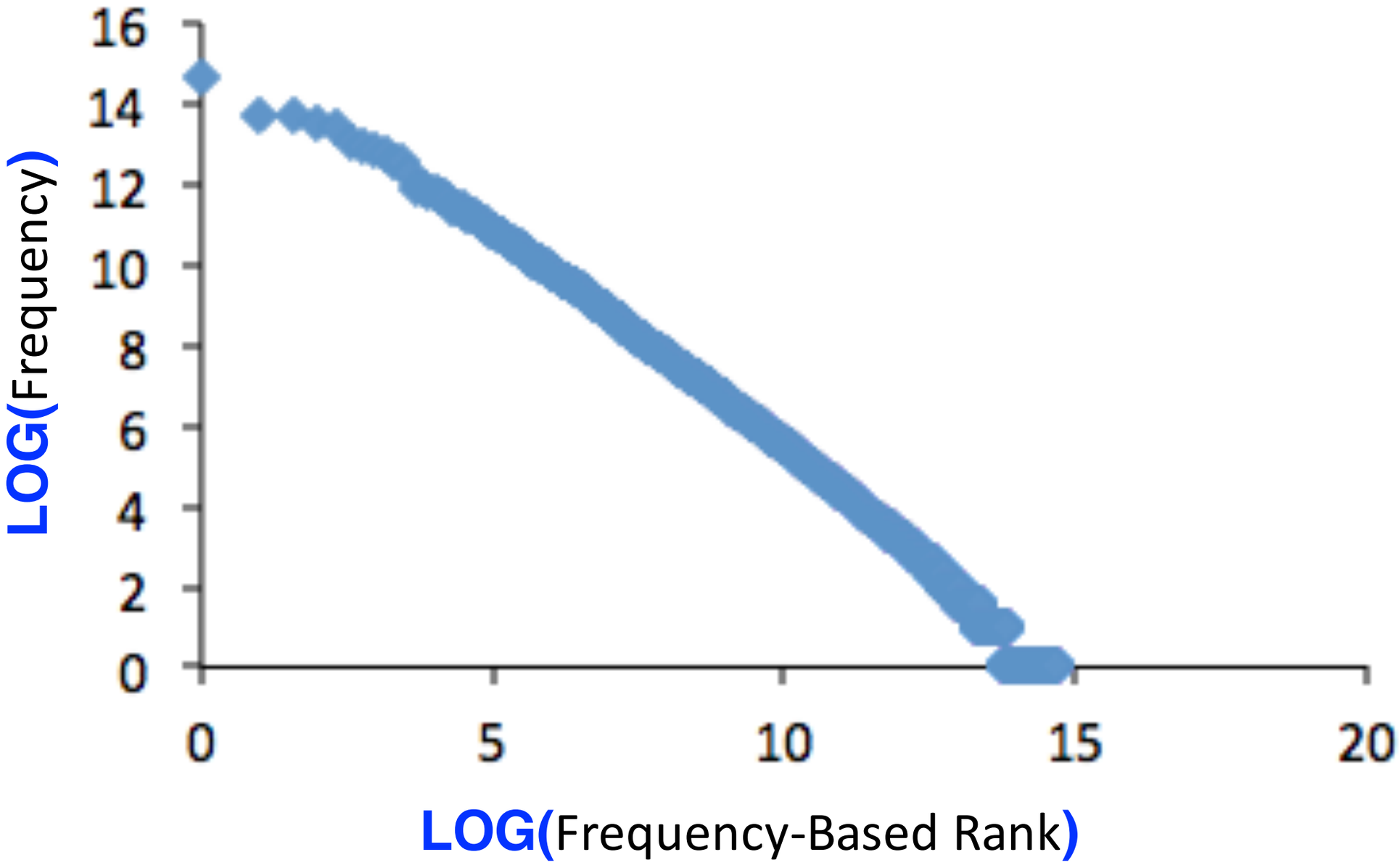
... still Zipf's Law $$\log(f) = \log(\frac{k}{r})$$

... 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 light with a slope of -1

Zipf's Law

HW1 Dataset



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

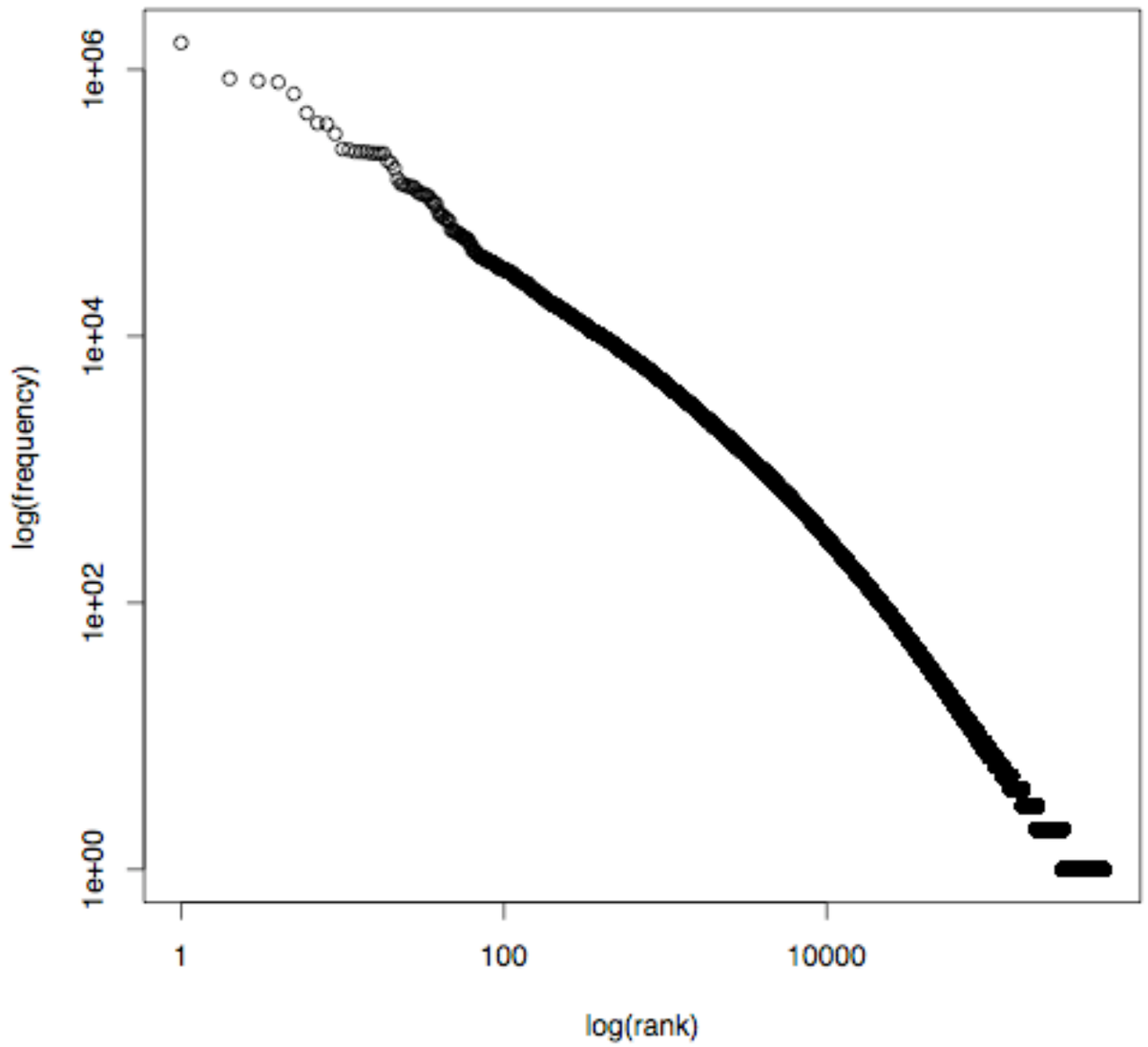
IMDB Corpus
internet movie database

- 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

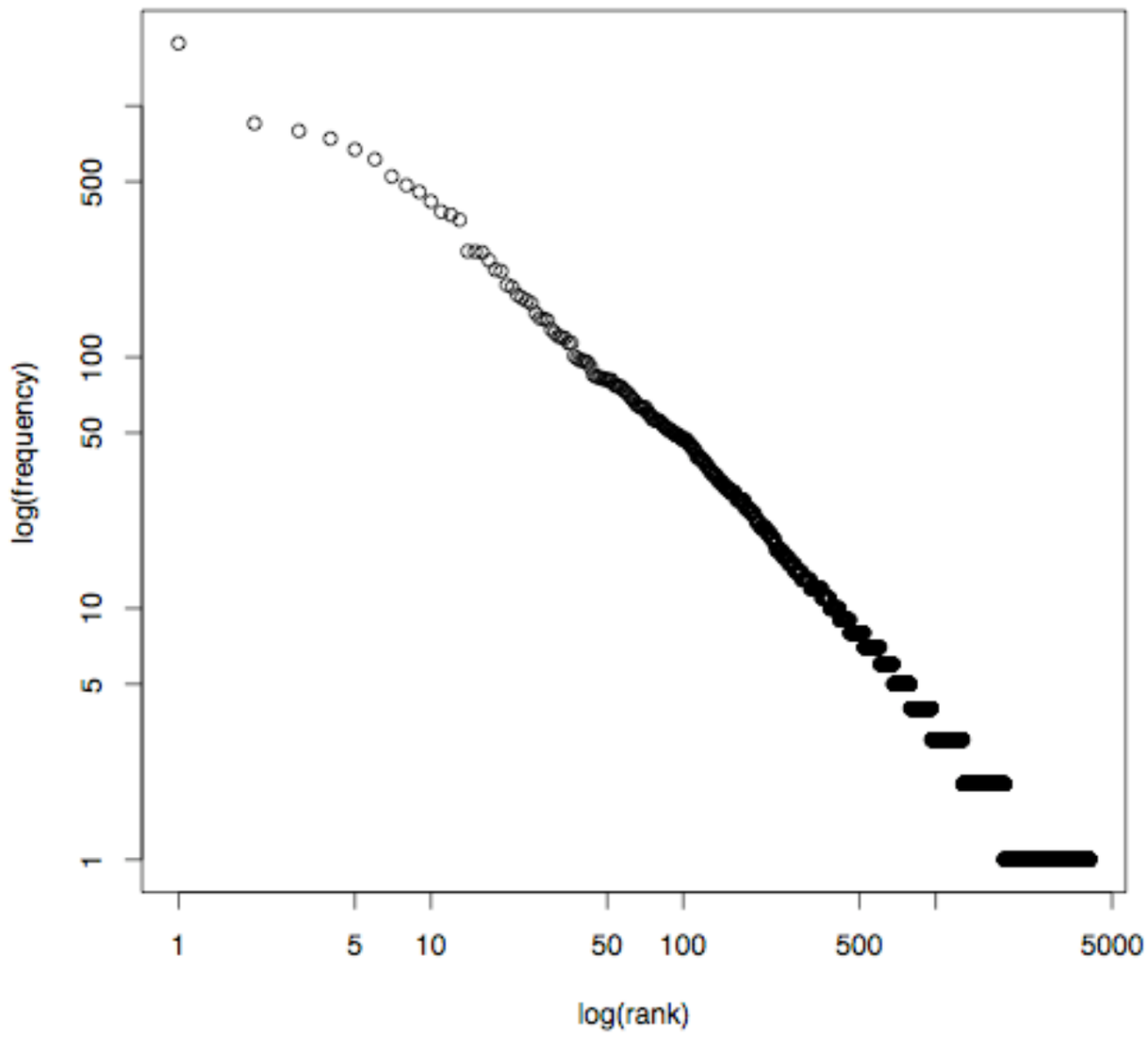
IMDB Corpus



Does Zipf's law generalize across different domains?

Zipf's Law

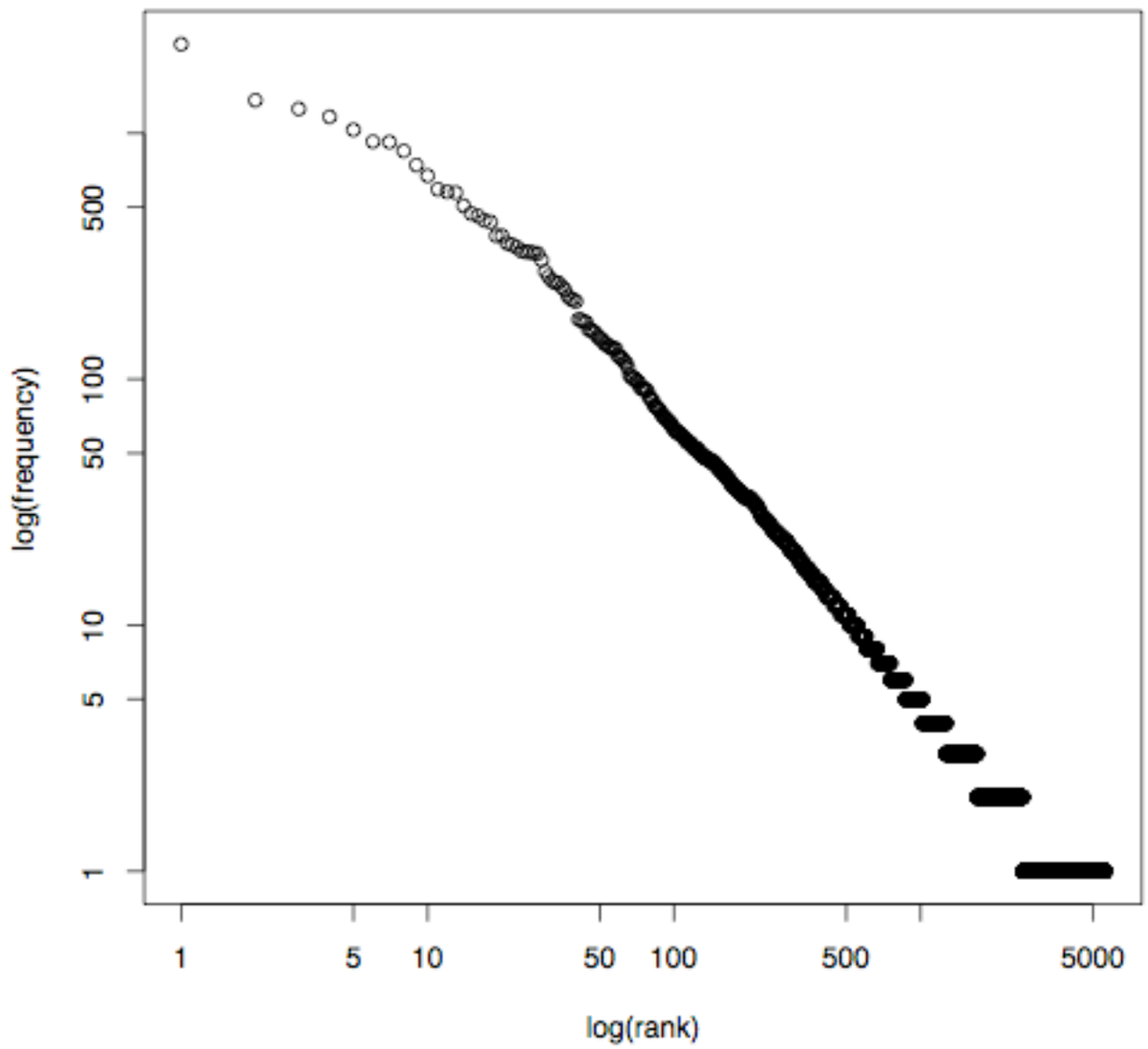
Alice in Wonderland



(text courtesy of Project Gutenberg)

Zipf's Law

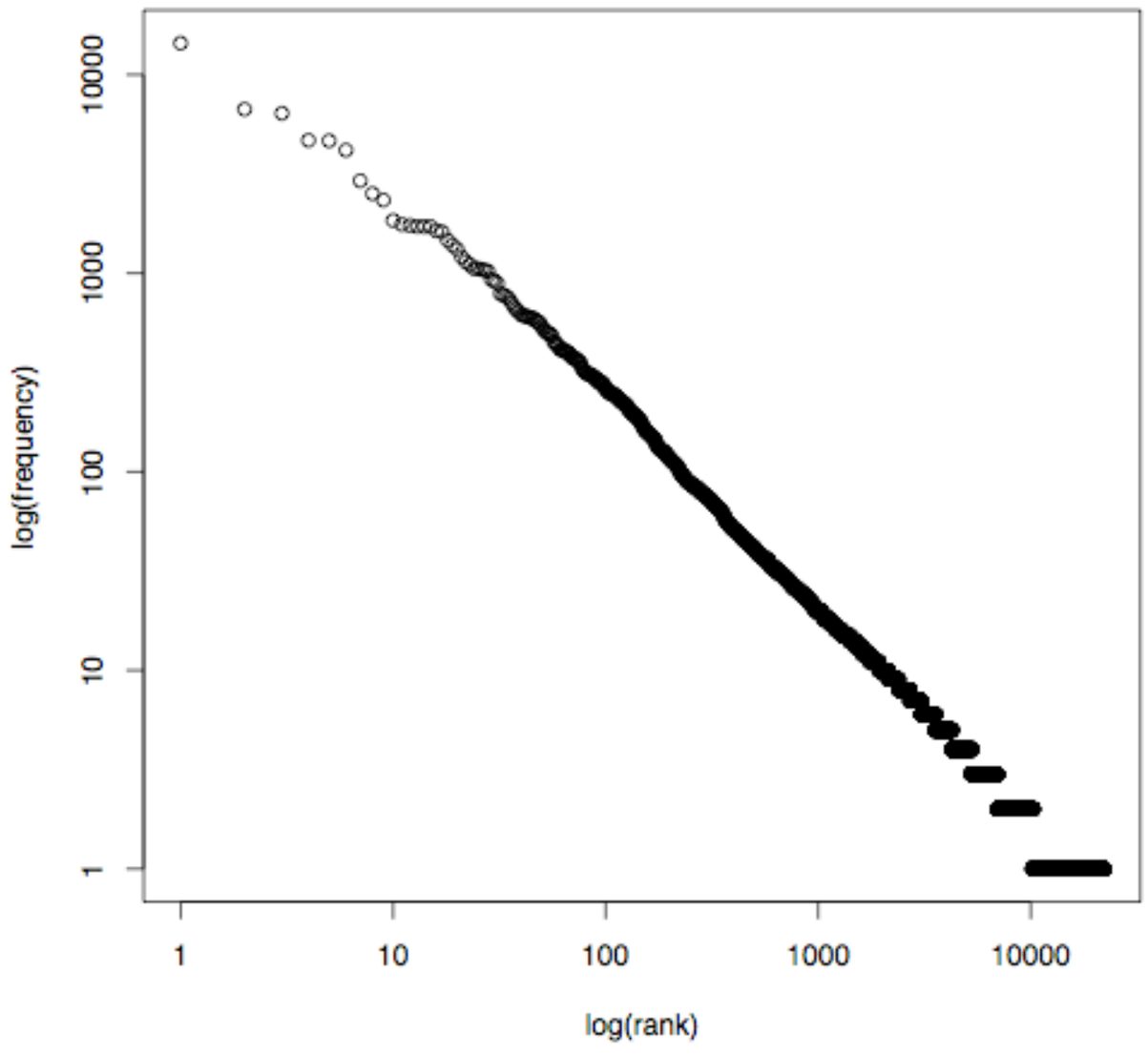
Peter Pan



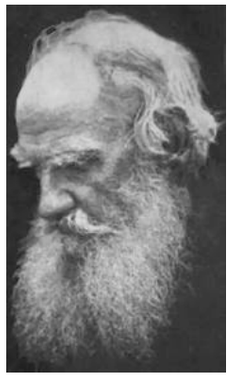
(text courtesy of Project Gutenberg)

Zipf's Law

Moby Dick

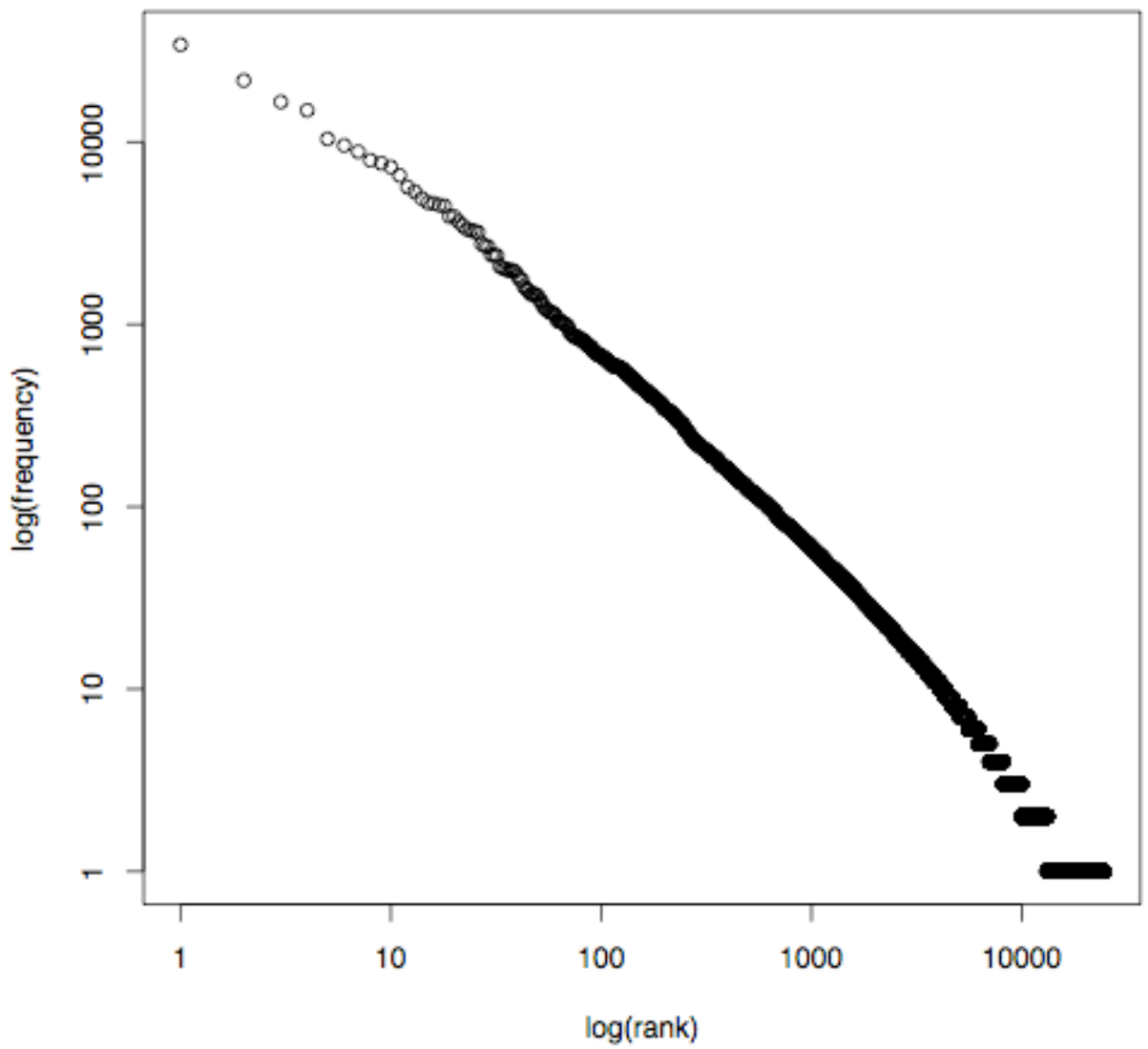


(text courtesy of Project Gutenberg)

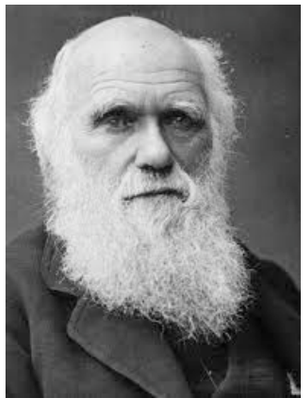


Zipf's Law

War and Peace

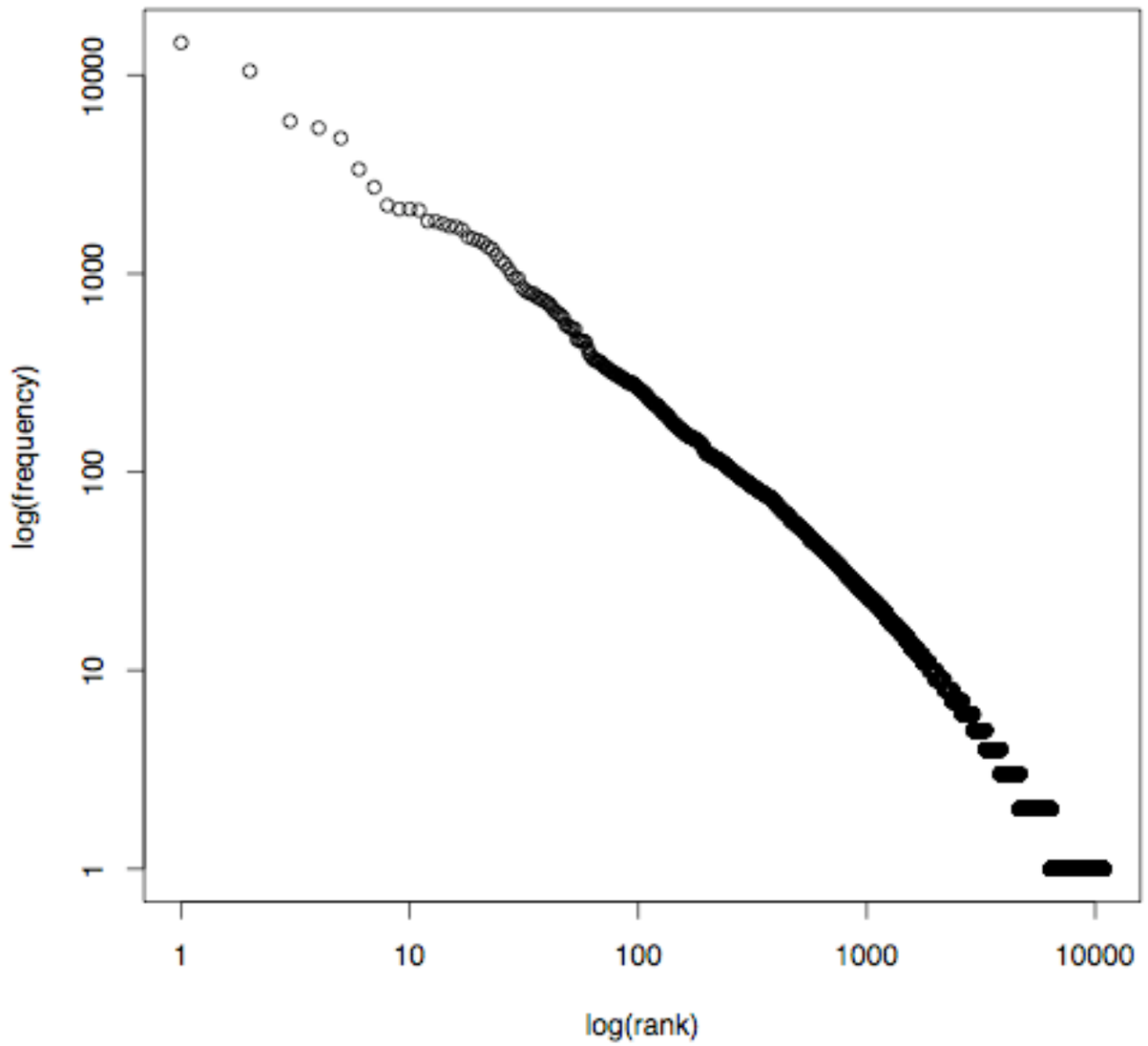


(text courtesy of Project Gutenberg)

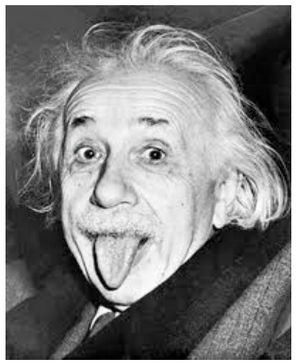


Zipf's Law

On the Origin of Species

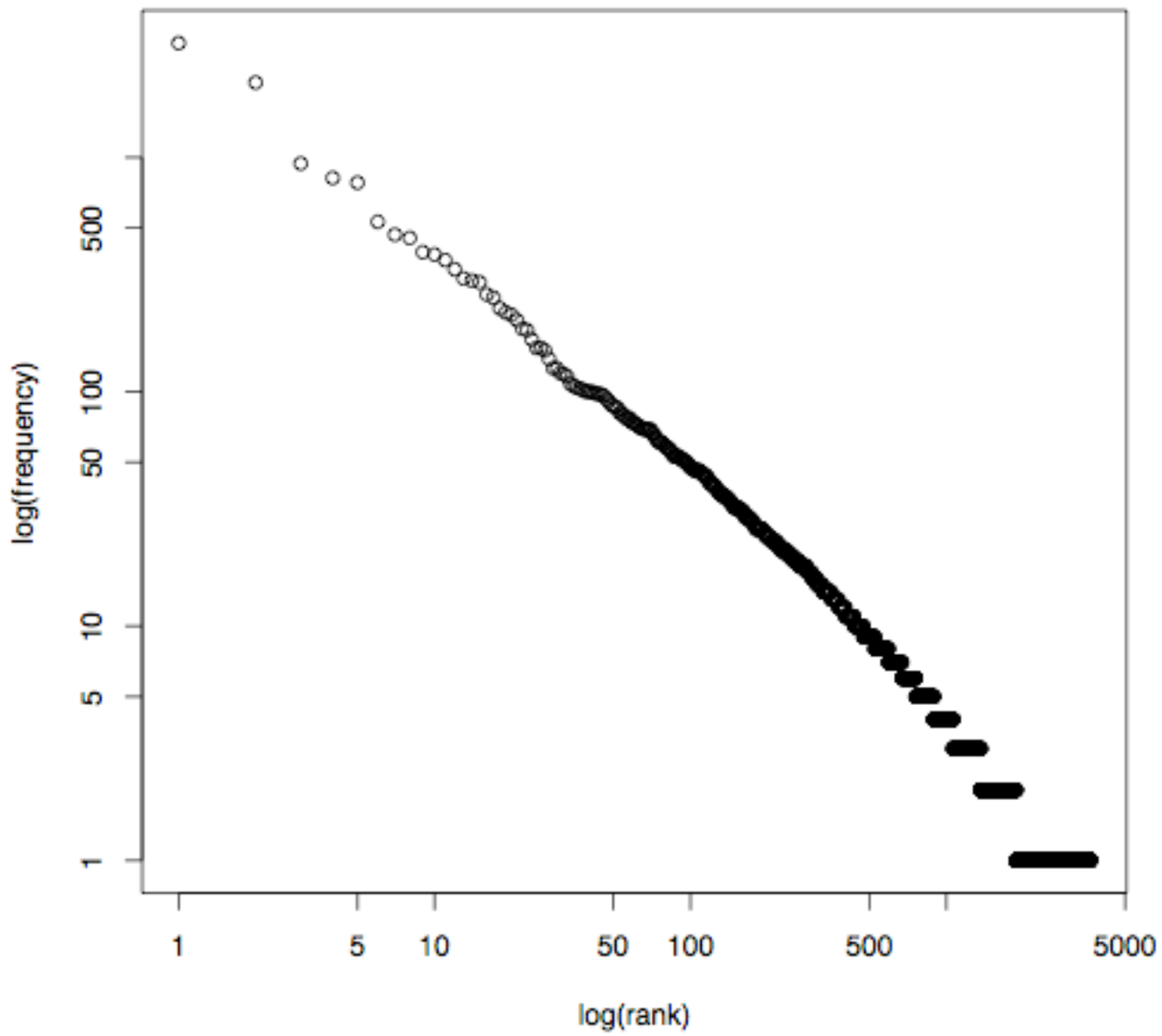


(text courtesy of Project Gutenberg)



Zipf's Law

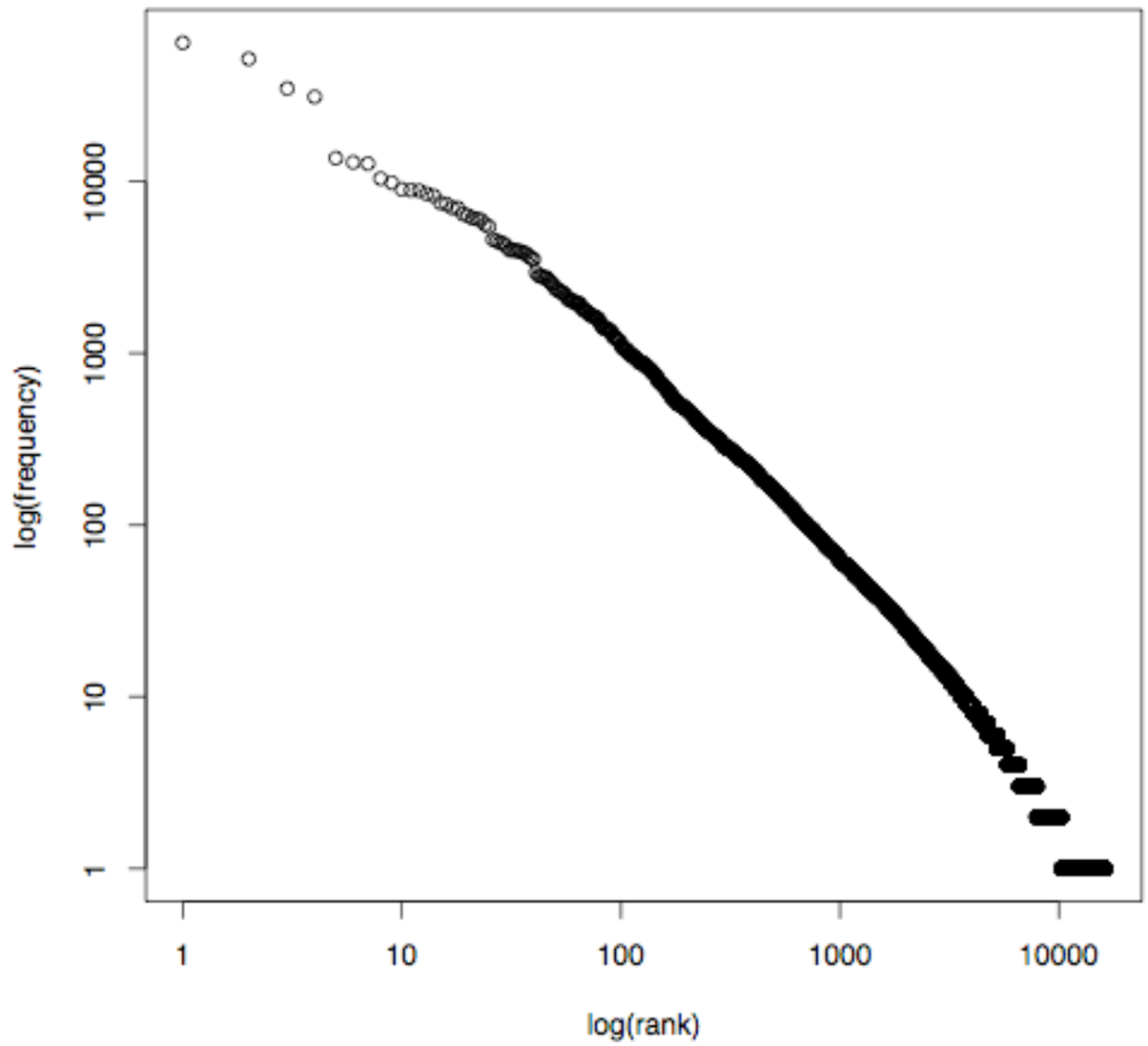
Relativity: The Special and General Theory



(text courtesy of Project Gutenberg)

Zipf's Law

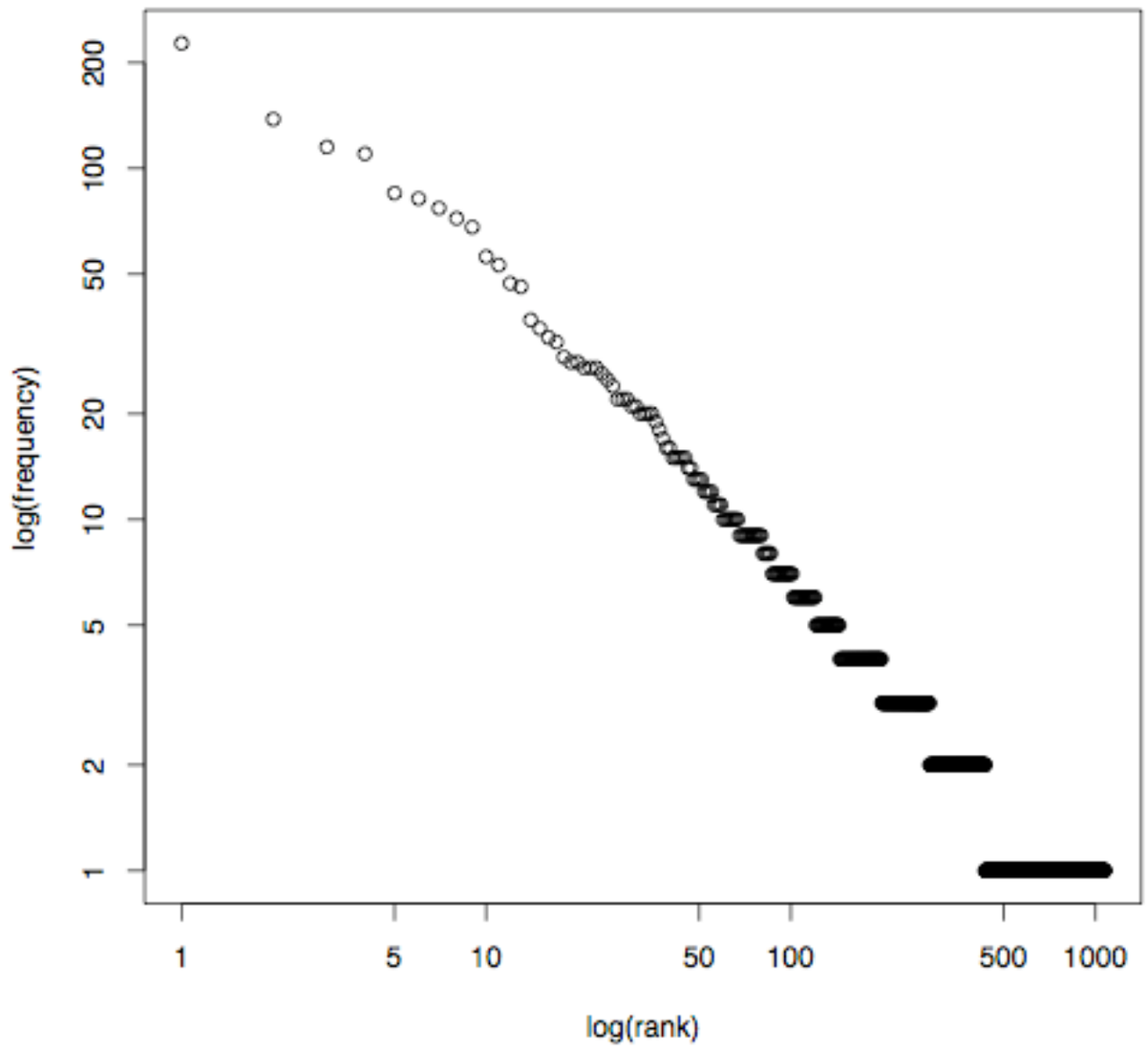
The King James Bible



(text courtesy of Project Gutenberg)

Zipf's Law

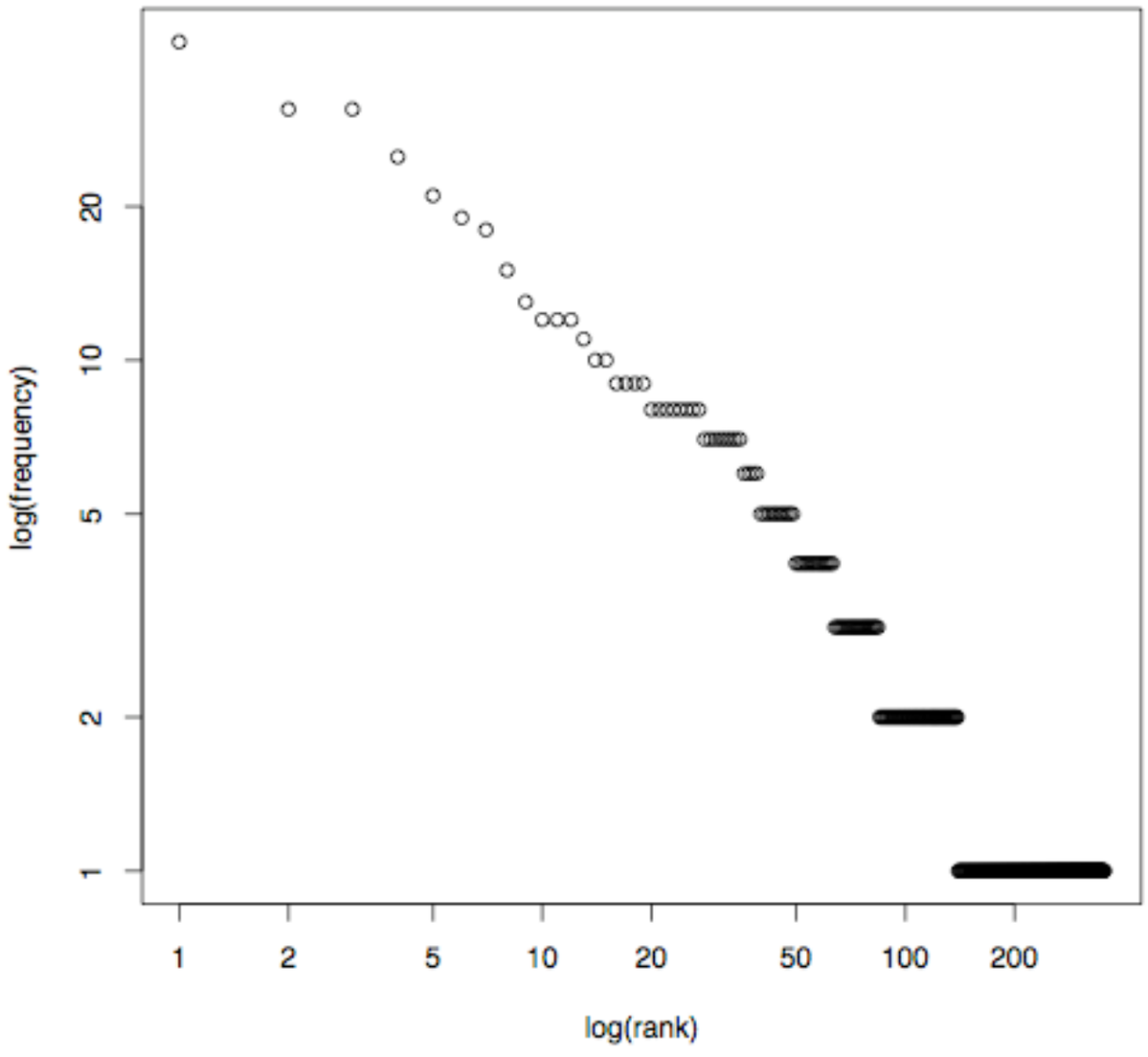
The Tale of Peter Rabbit



(text courtesy of Project Gutenberg)

Zipf's Law

The Three Bears

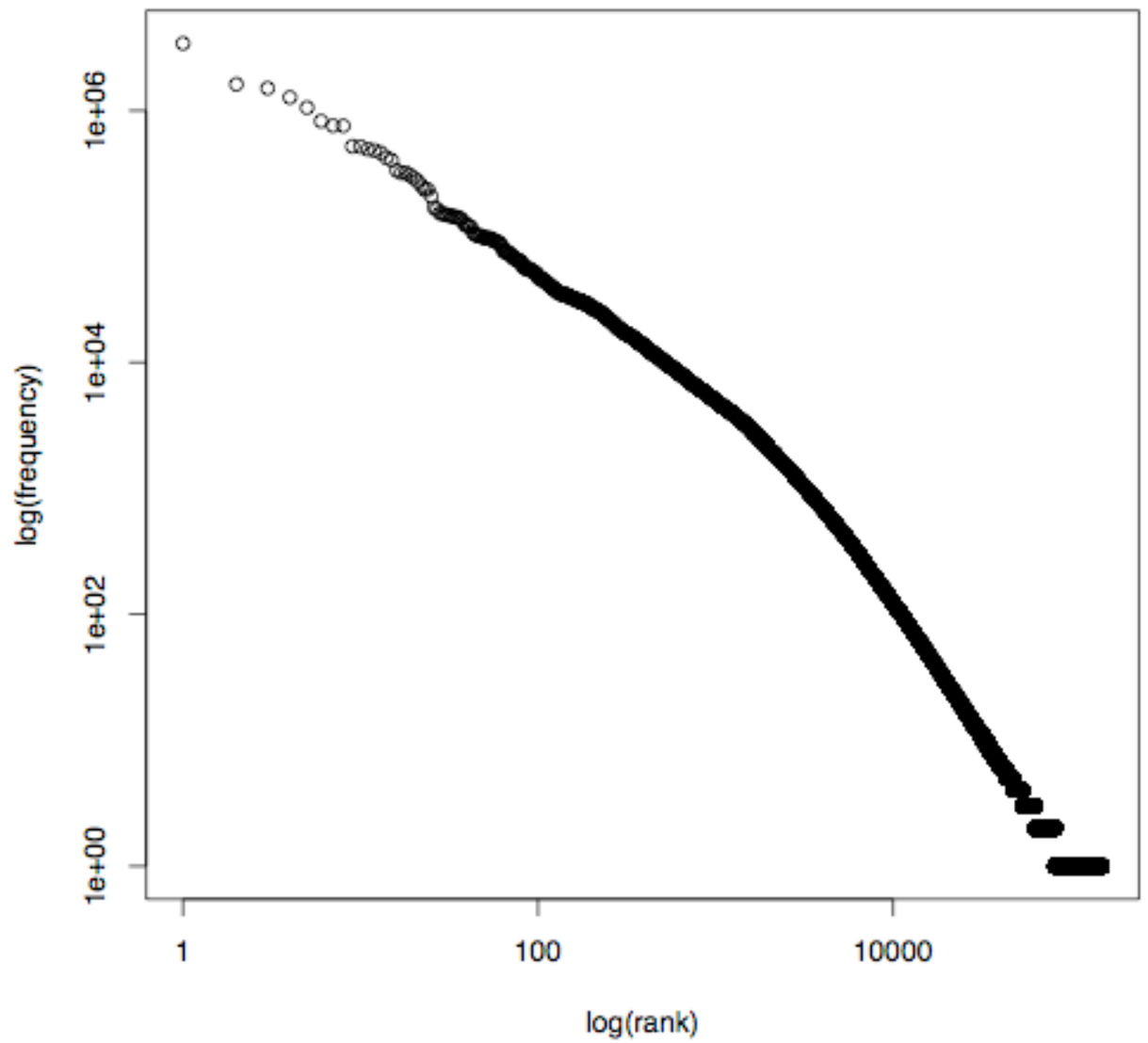


(text courtesy of Project Gutenberg)

Does Zipf's law generalize across different languages?

Zipf's Law

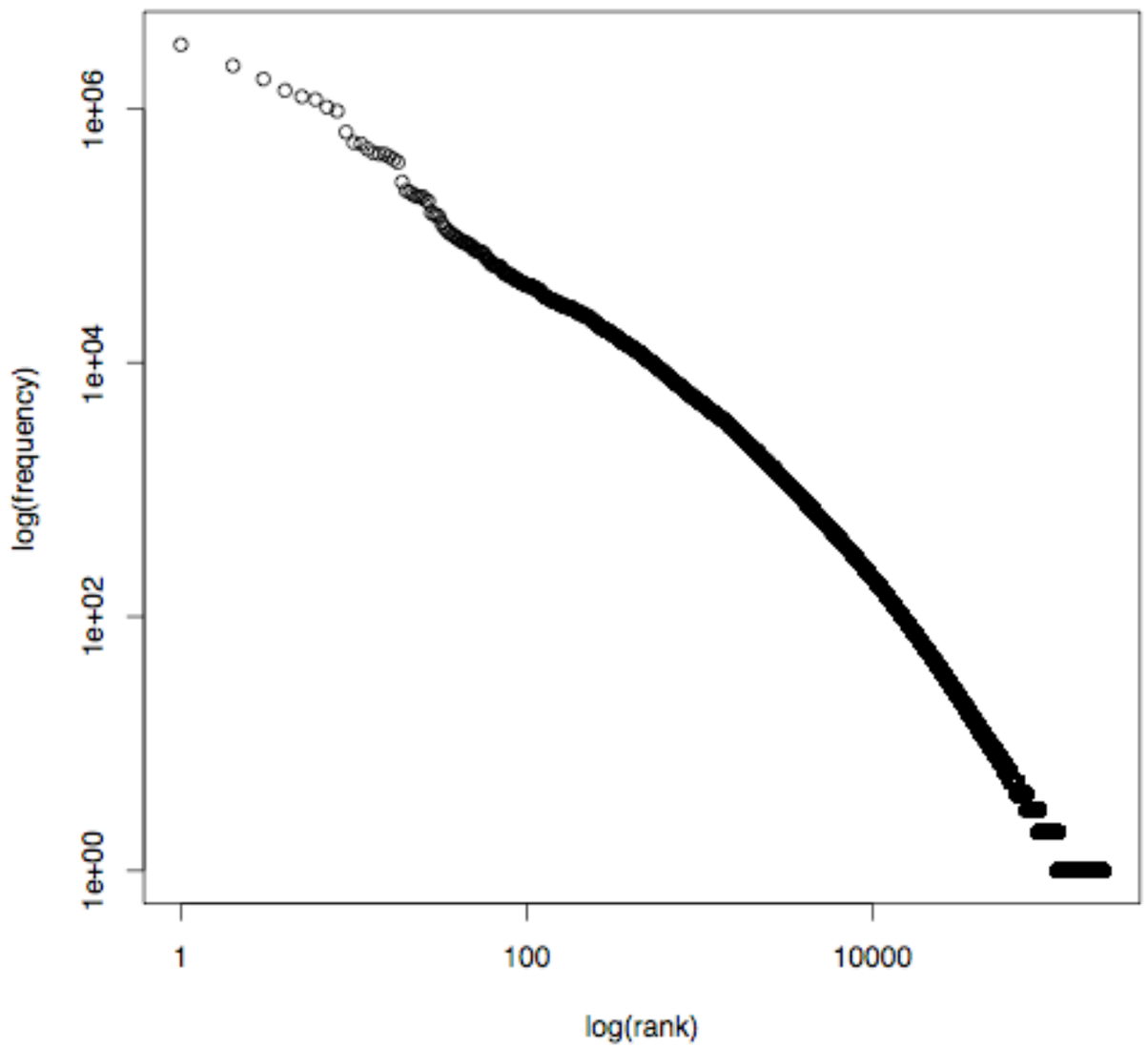
European Parliament: English



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

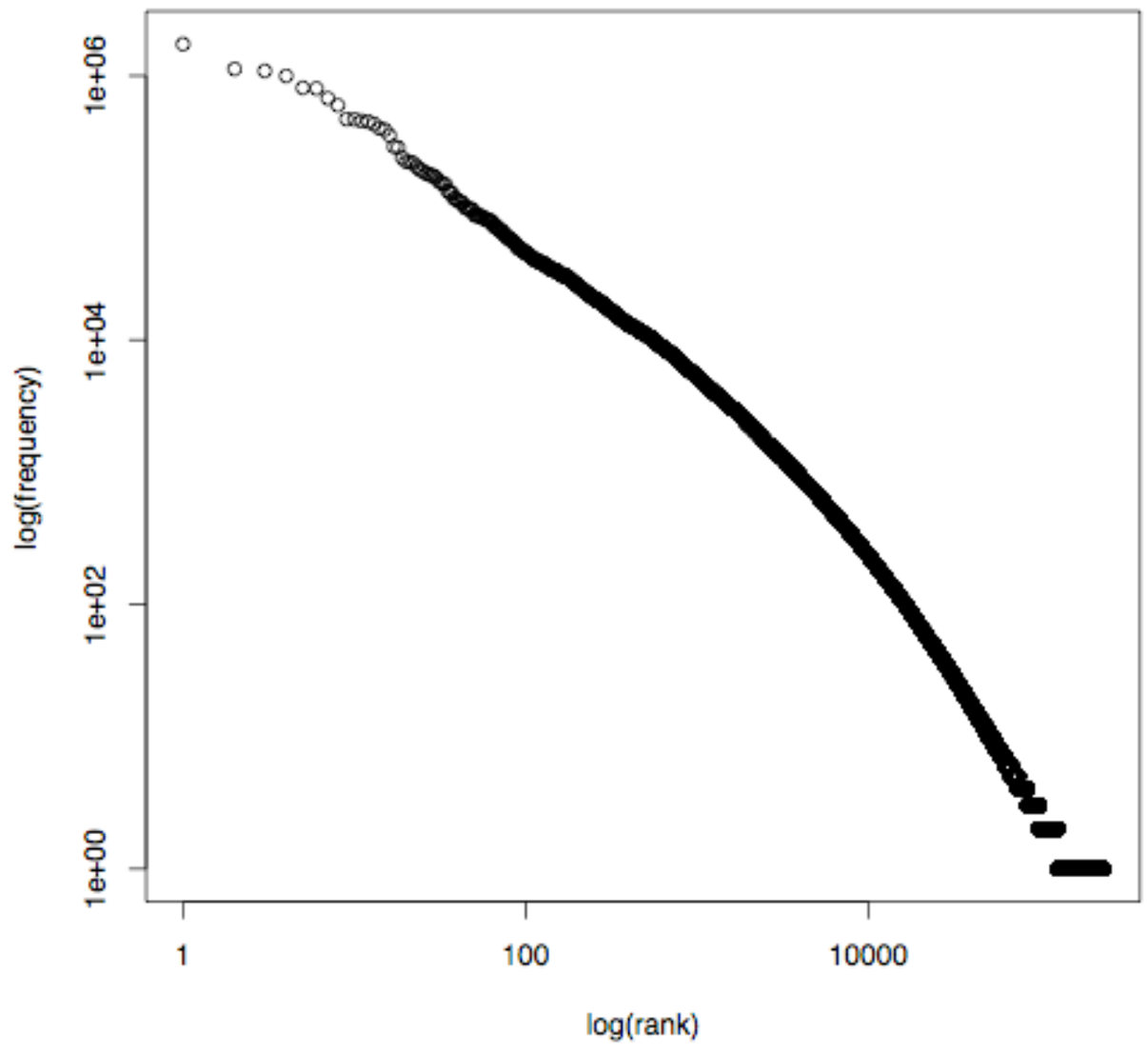
Zipf's Law

European Parliament: Spanish



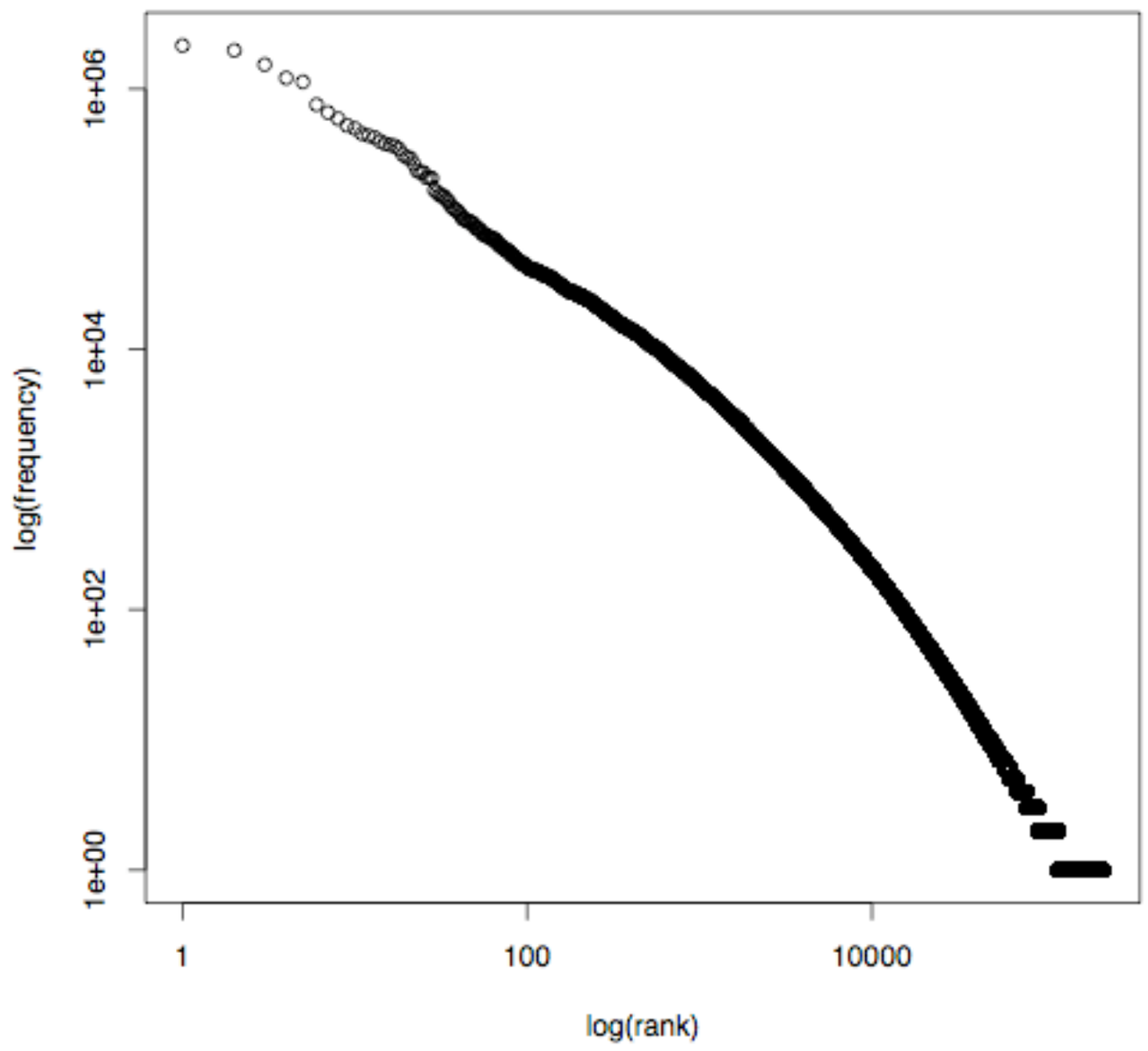
Zipf's Law

European Parliament: Italian



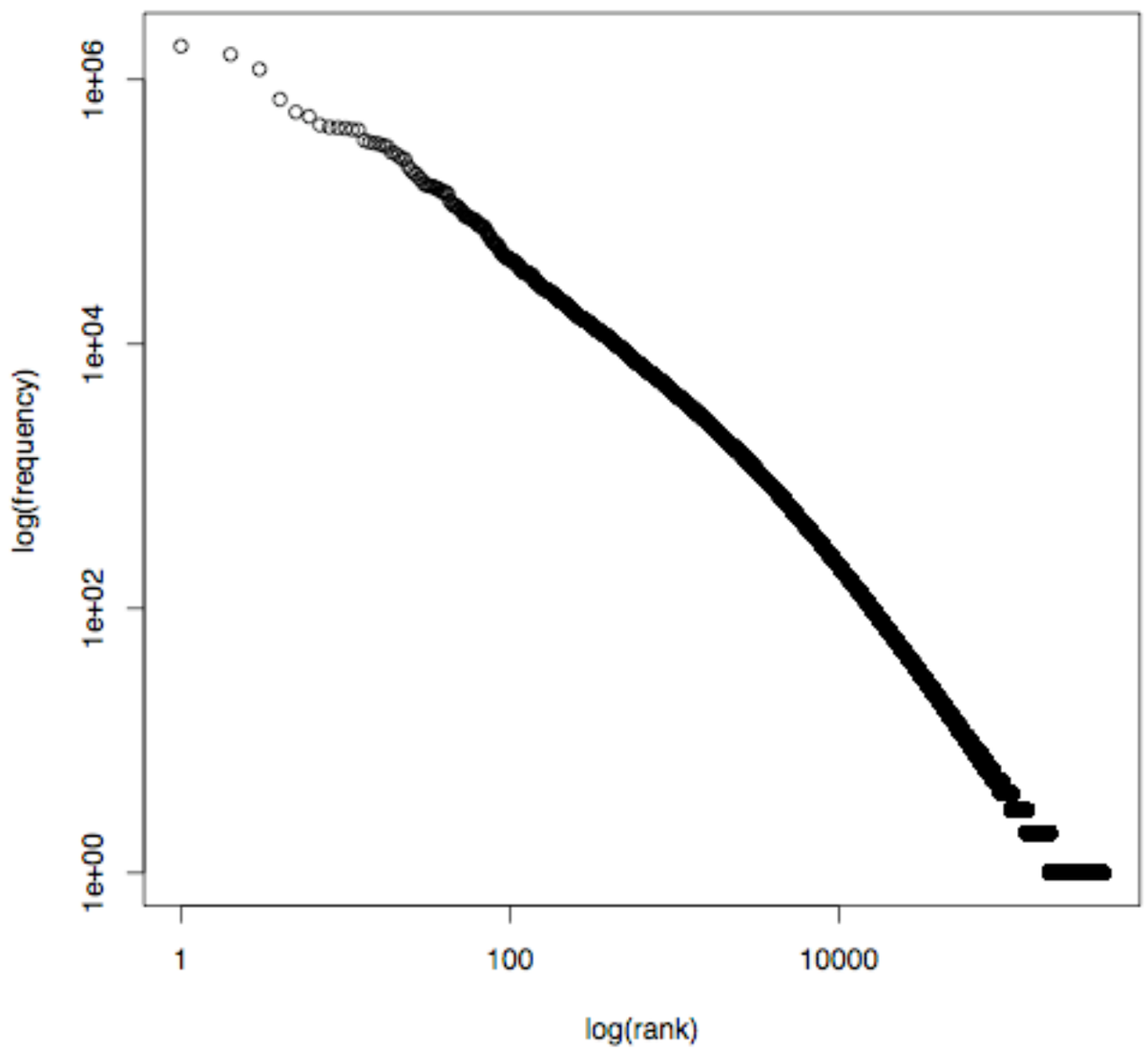
Zipf's Law

European Parliament: Portuguese



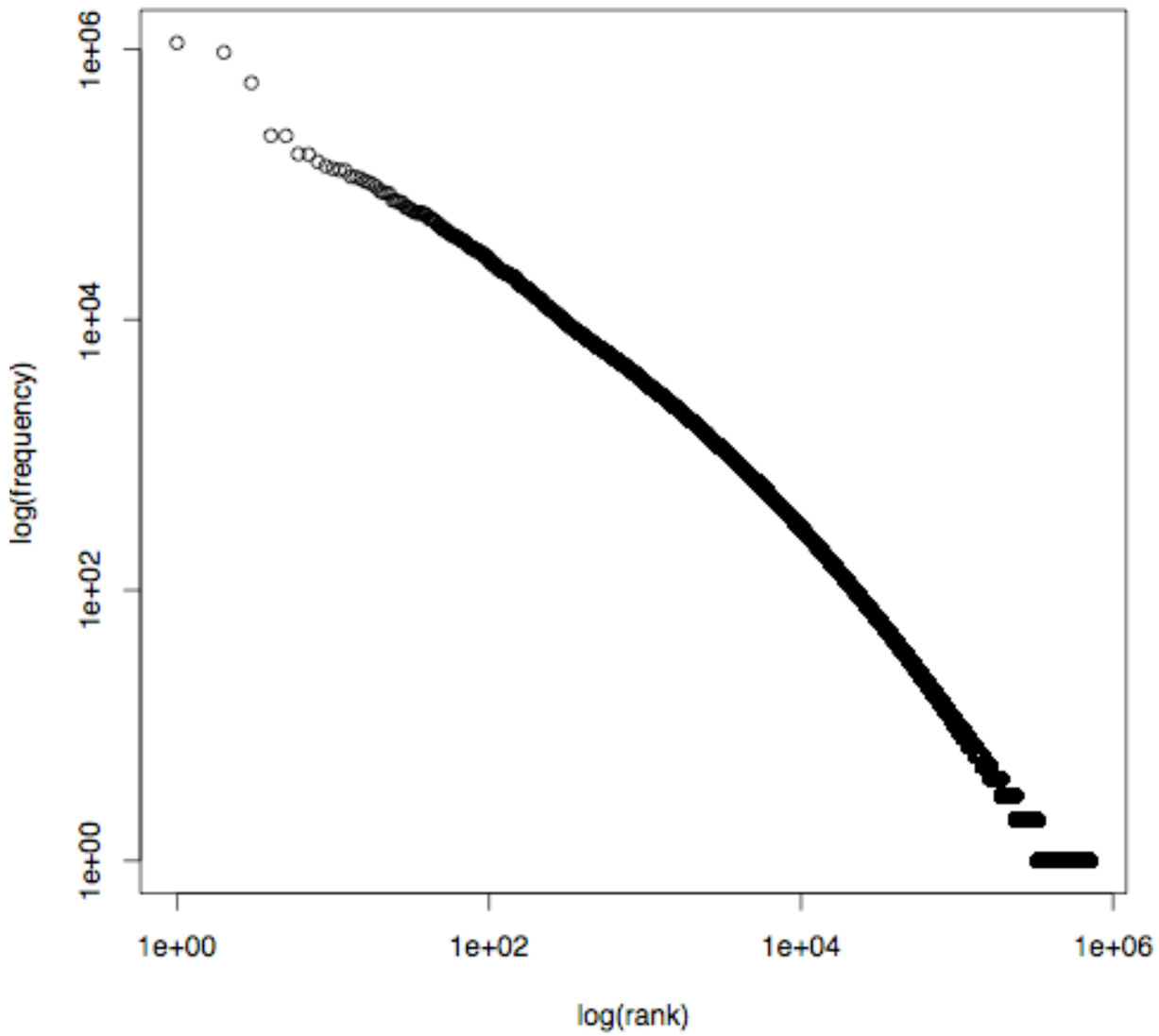
Zipf's Law

European Parliament: German



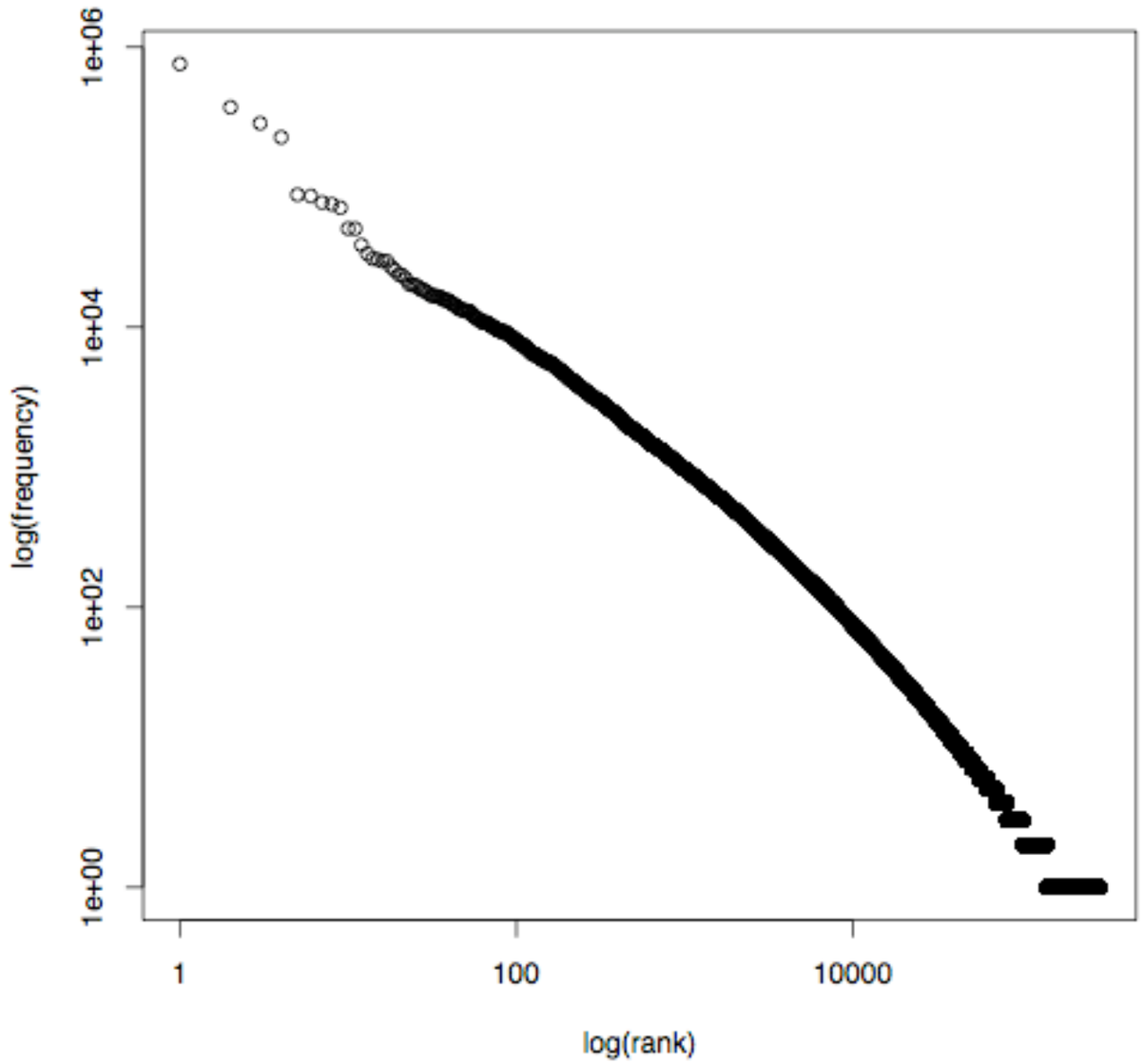
Zipf's Law

European Parliament: Finnish



Zipf's Law

European Parliament: Hungarian



Zipf's Law

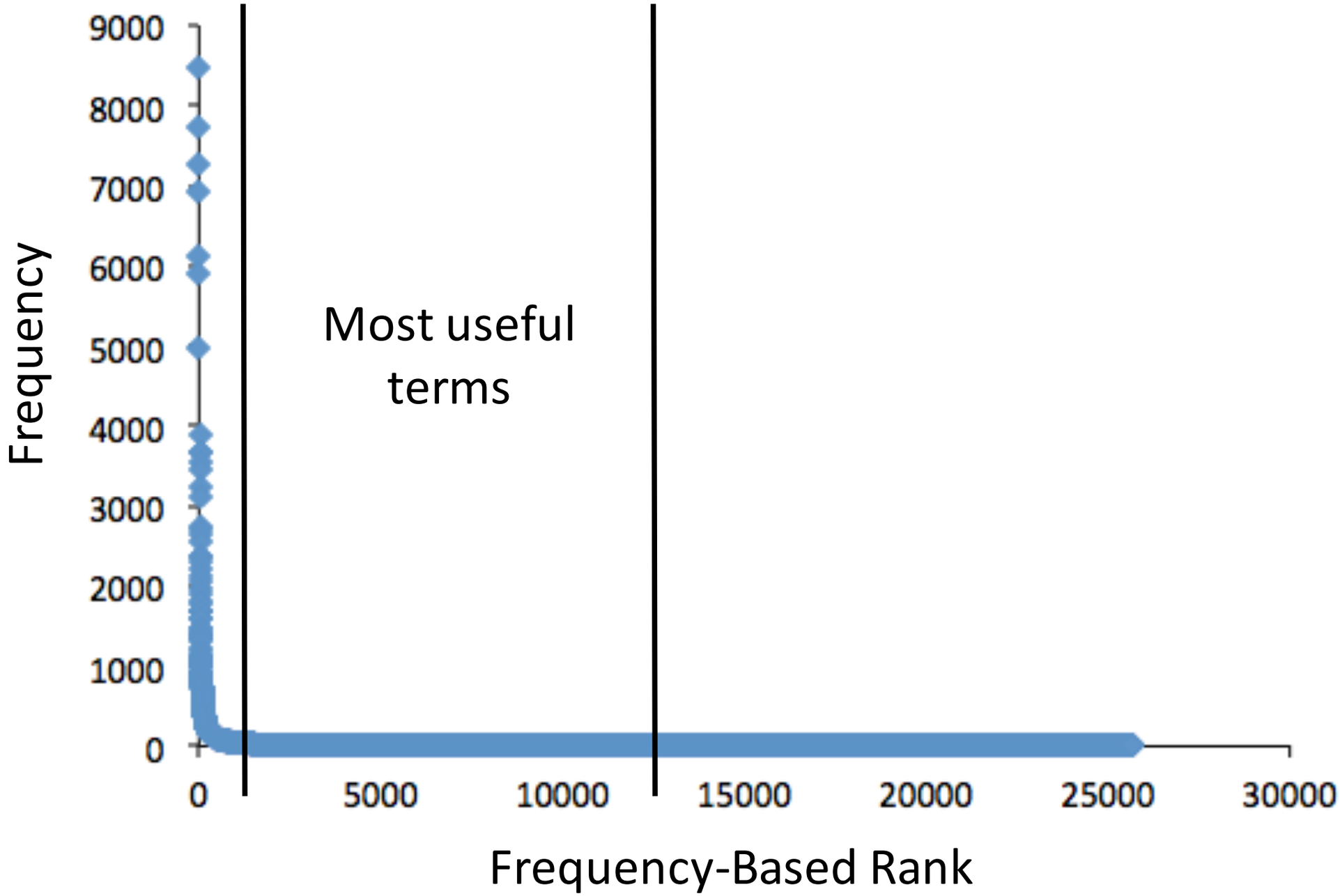
- Zipf's Law holds true for:
  - different dataset sizes
  - different domains
  - different languages

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 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 an 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

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 that event B has occurred

Chain Rule

- P(A, B) = P(A|B) x 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

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

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

Final Project Topics

- Fake News detection
- Gender bias on Twitter.
- News to predict stock price.
- Categorize posts on reddit forums.
- Topic categorization for emails.
- Chapel Hill restaurant recommender.
- Detect cyber-bullying on social media.
- Tweet sentiment.
- Personality trait recognition.
- Search engine auto-fill queries.
- Clustering companies by social responsibilities based on public reports.
- Validity of reviews on Amazon.
- Observable machine learning – Patient analysis.
- Gender bias on Twitter.