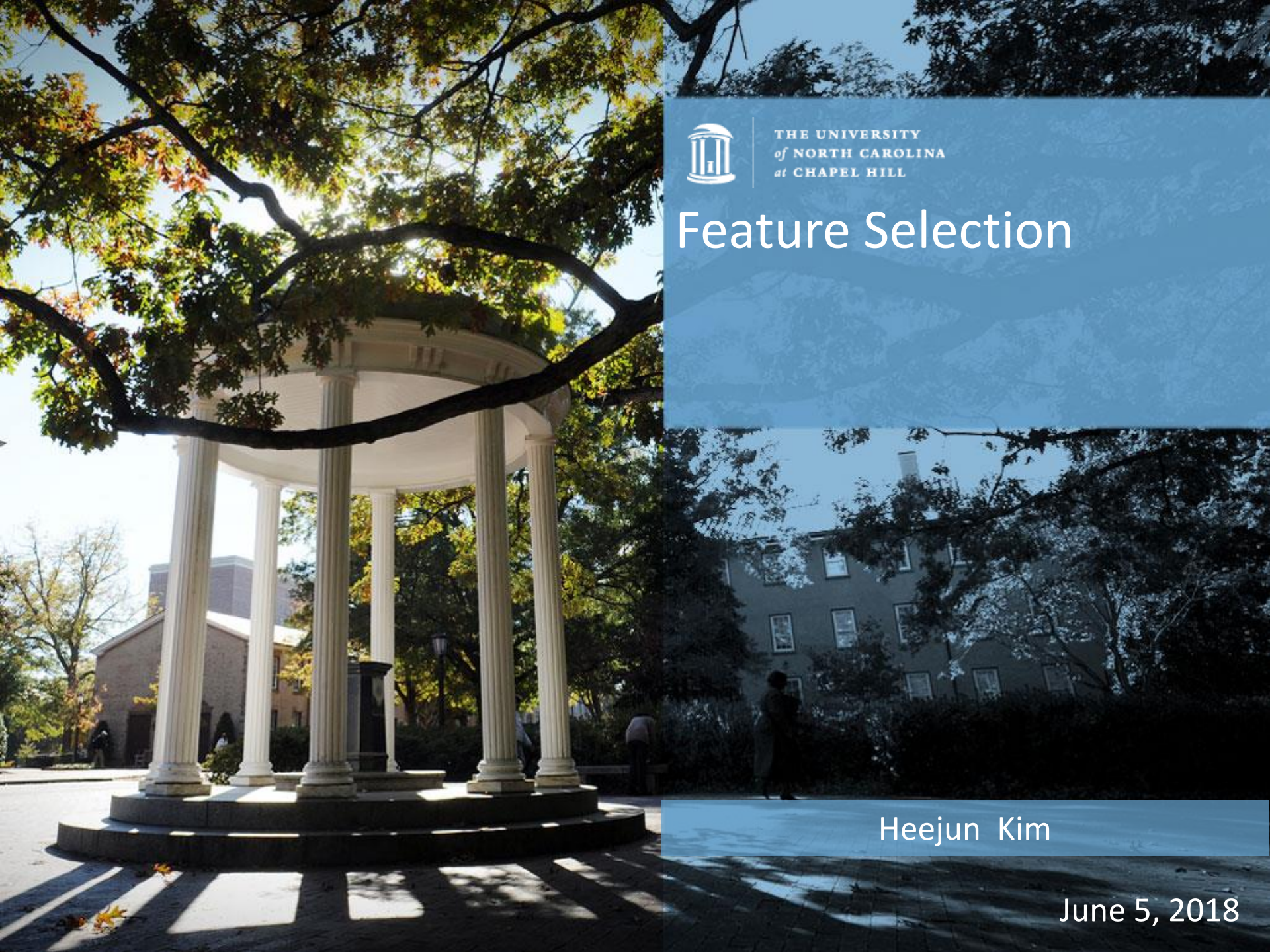


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Feature Selection

Heejun Kim

June 5, 2018

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?

Hands-on Exercise 2 Training Set

terms that only occurred in Alzheimer's diseases class

term	count	term	count	term	count
dementia	100	filaments	3	halting	1
amyloid	61	vessel	3	j2	1
memory	52	benefits	3	prostaglandin	1
mild	46	fibrils	3	degradation-implications	1
tau	45	angiopathy	3	'omics	1
amyloid-??????	31	homeostasis	3	expansion	1
diseases	23	members	2	ventricular	1
decline	22	chitin	2	d-galactose	1
a??????	22	provide	2	spinosa	1
neurodegenerative	19	preserved	2	capparis	1
deficits	19	ltp	2	retirement	1
cognition	19	legal	2	day	1

Hands-on Exercise 2 Training Set

terms that only occurred in BRCA class

term	count	term	count	term	count
breast	818	lymphatic	3	bis	1
chemotherapy	47	mutant	3	mammalian	1
metastatic	37	survivor	3	her2-amplified	1
metastasis	33	anhydrase	3	kinases	1
reconstruction	27	carbonic	3	ibrutinib	1
survival	27	tyrosine	3	absolute	1
invasive	24	bolus	2	projecting	1
negative	24	balance	2	portraying	1
ductal	22	fine-needle	2	tnfaip8	1
biopsy	20	th1	2	differing	1
adjuvant	20	stroma	2	homing	1
survivors	19	oxygen	2	title	1

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

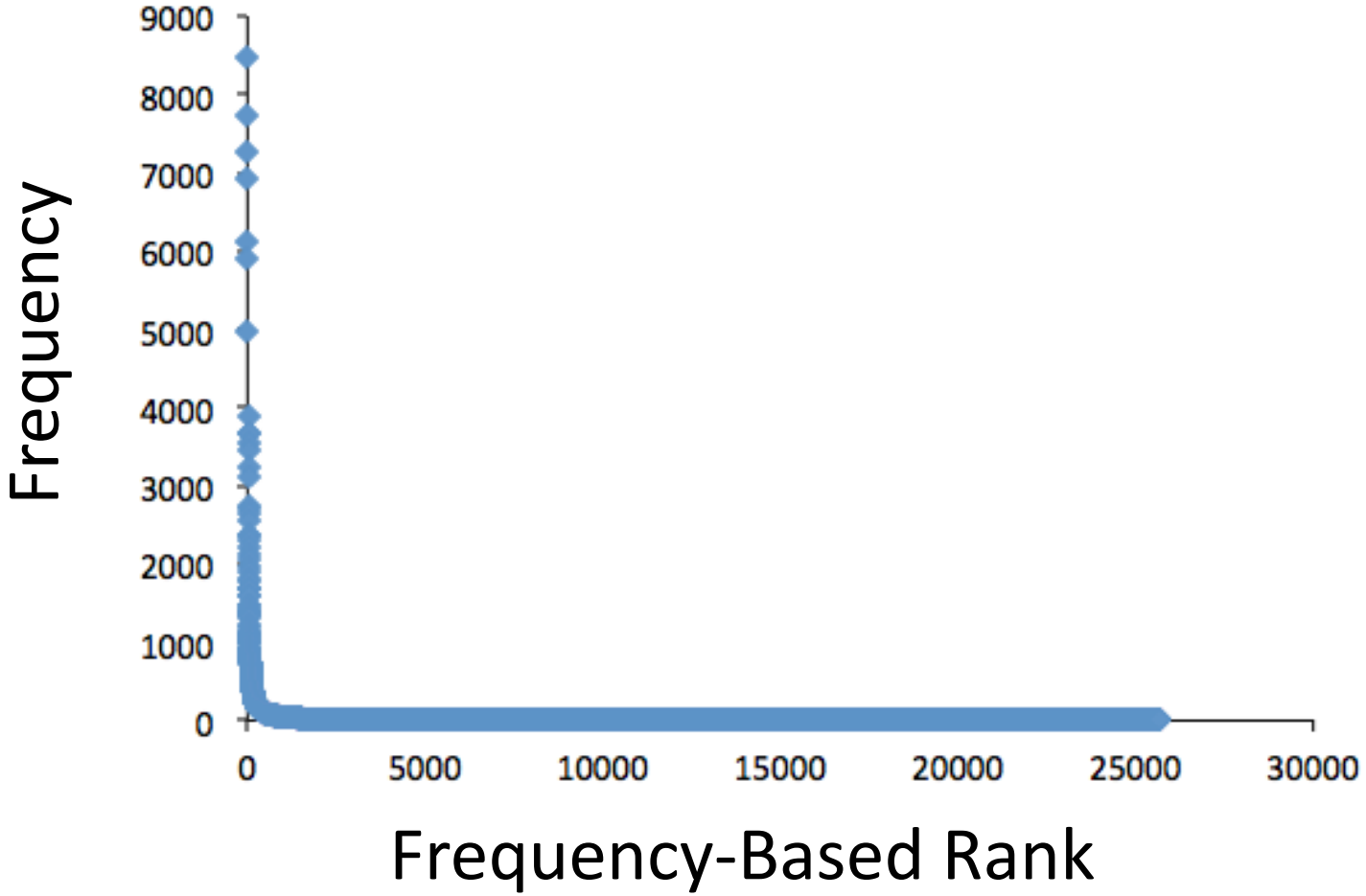
Statistical Properties of Text

- Statistical properties of text can be predictable across domains, and even across languages!
- These can help us determine which terms are less likely to be useful (without requiring training labels)

Hands-on Exercise 2 Training Set: statistical properties of text

- Number of Instances: 2,000
- Number of unique terms: 5,949
- Number of term occurrences: 25,409
- Why should we not include all 5,949 vocabulary terms as features?
- Is there a danger in having 3 times more features than instances?
- We should reduce the feature representation to the most meaningful ones

Hands-on Exercise 2 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
- 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 term t corresponding to the collection
- c = constant
- r_t = frequency-based rank of term t
- For English c = 0.1 (more or less)
- What does this mean?

Zipf's Law

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

- 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 terms account for about 30%
- Together, the top 20 terms account for about 36%
- Together, the top 50 terms 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:

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

- So, what is fraction of the terms which 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

- 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:

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

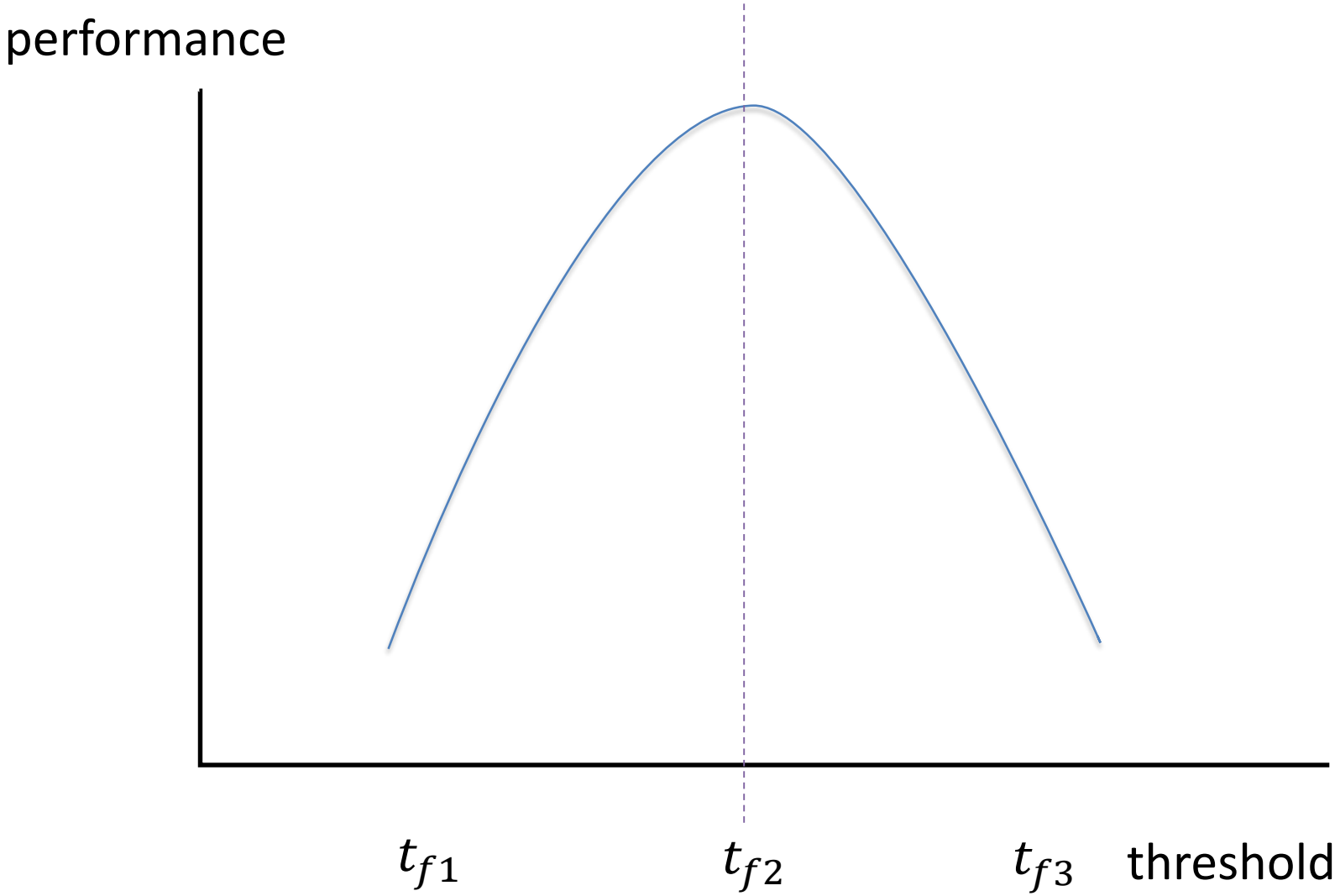
- That is, we have to add the fraction of terms that appear 1, 2, 3, ... up to n times

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

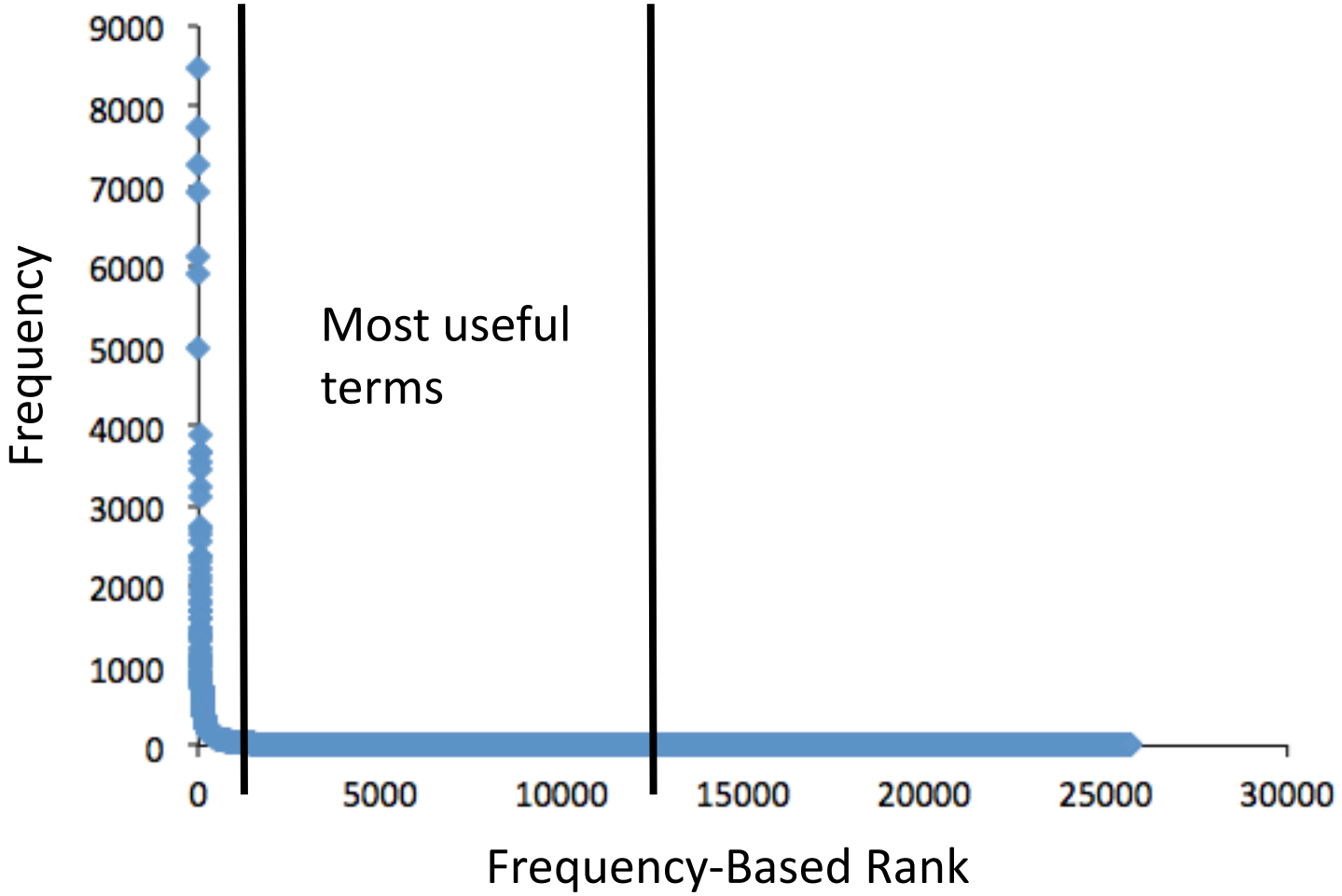
Zipf's Law

Implications for Feature Selection



Zipf's Law

Implications for Feature Selection



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

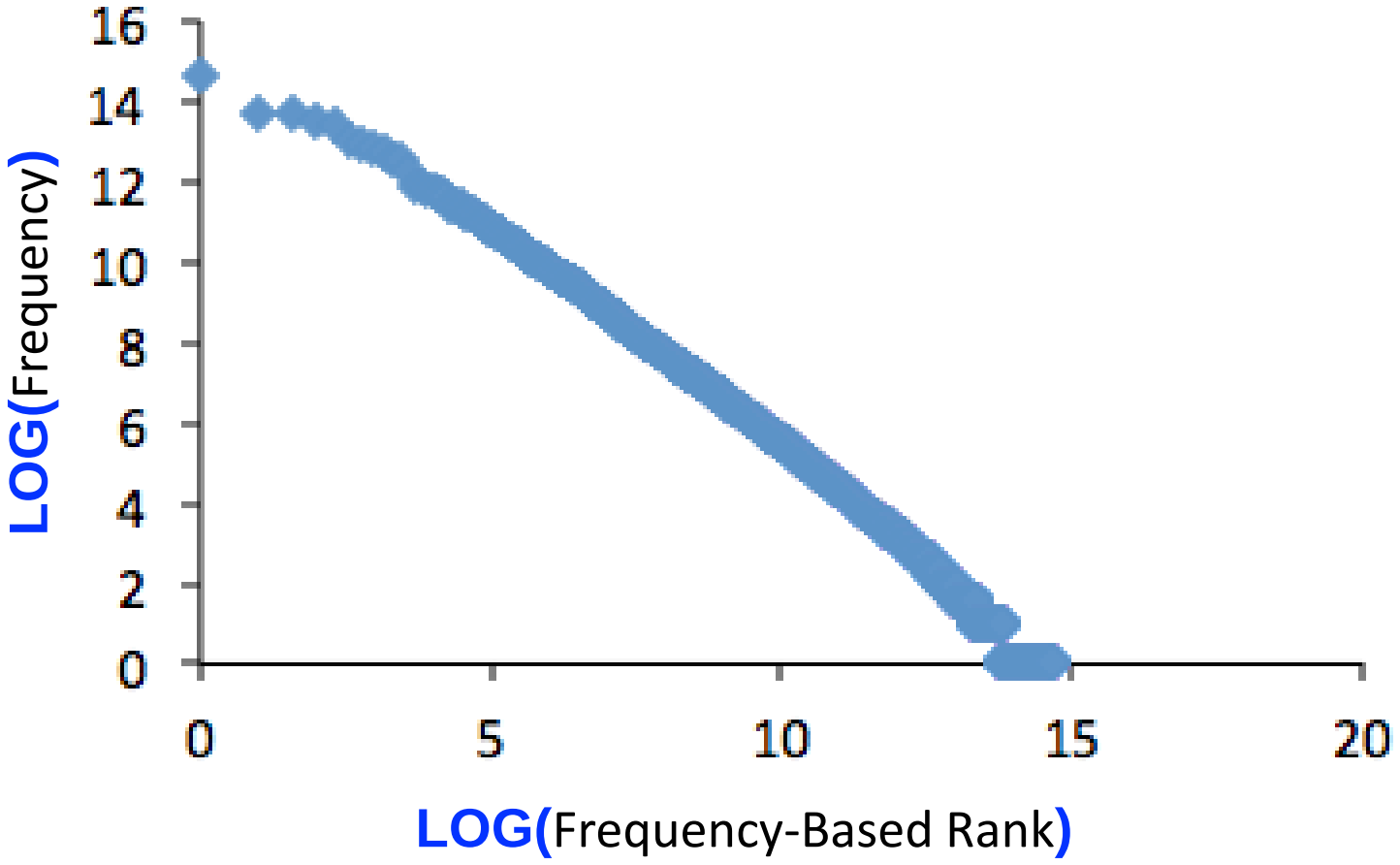
... still Zipf's Law $$\log(f) = \log(\frac{k}{r})$$

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

- 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

Hands-on Exercise 2 Dataset



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

IMDB Corpus

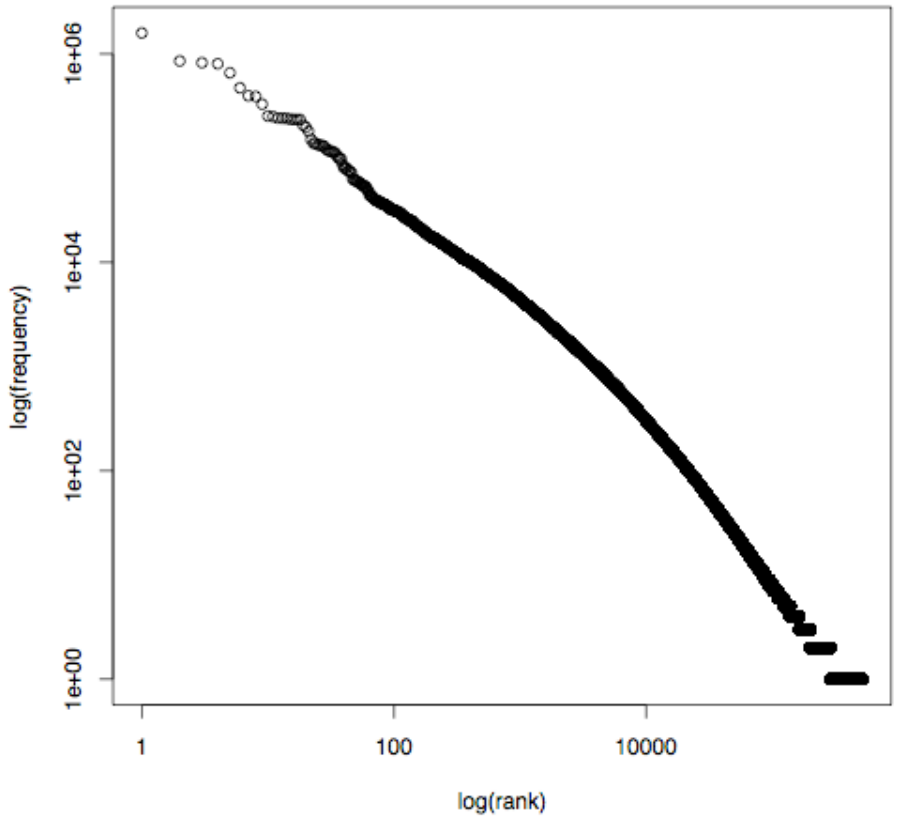
internet movie database

- 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/

IMDB Corpus

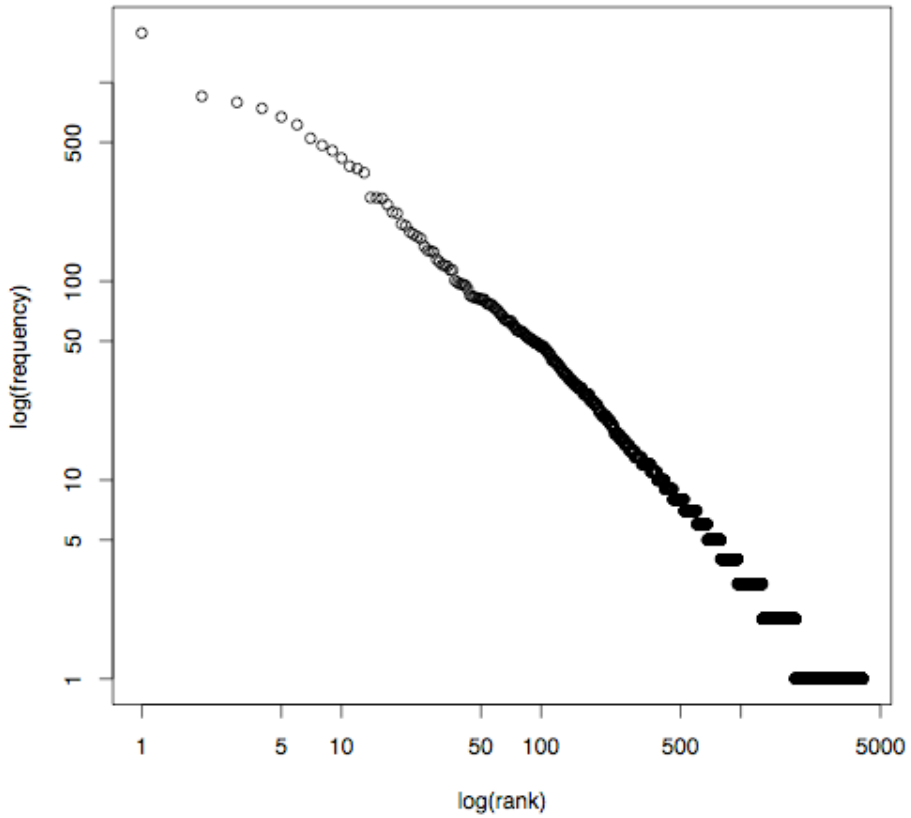
internet movie database



Does Zipf's law generalize across different domains?

Zipf's Law

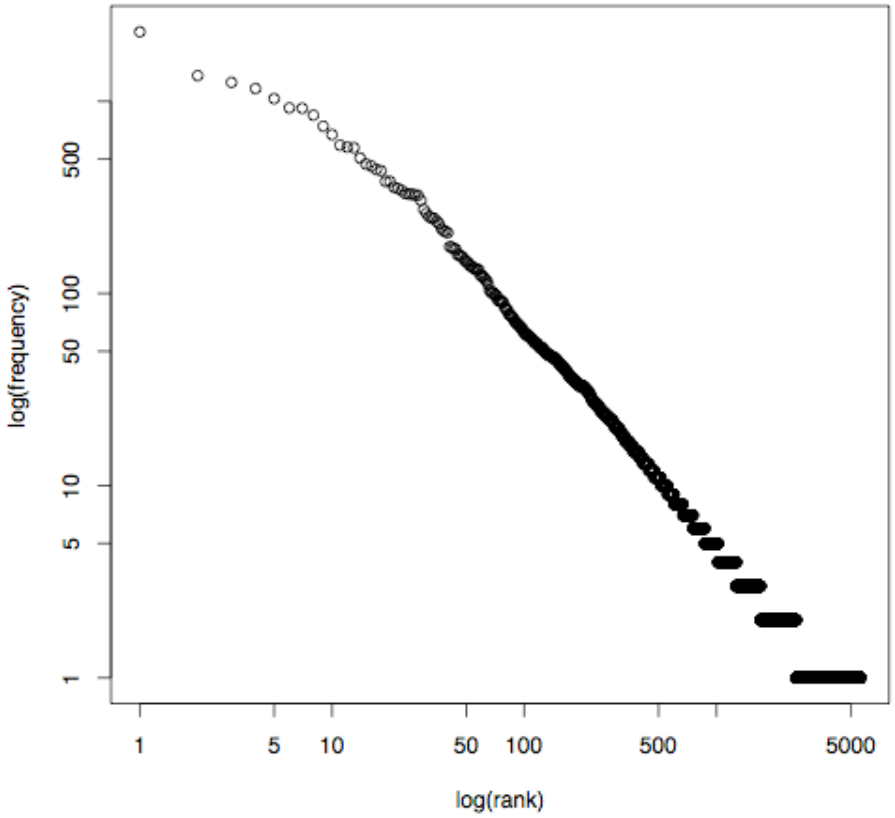
Alice in Wonderland



(text courtesy of Project Gutenberg)

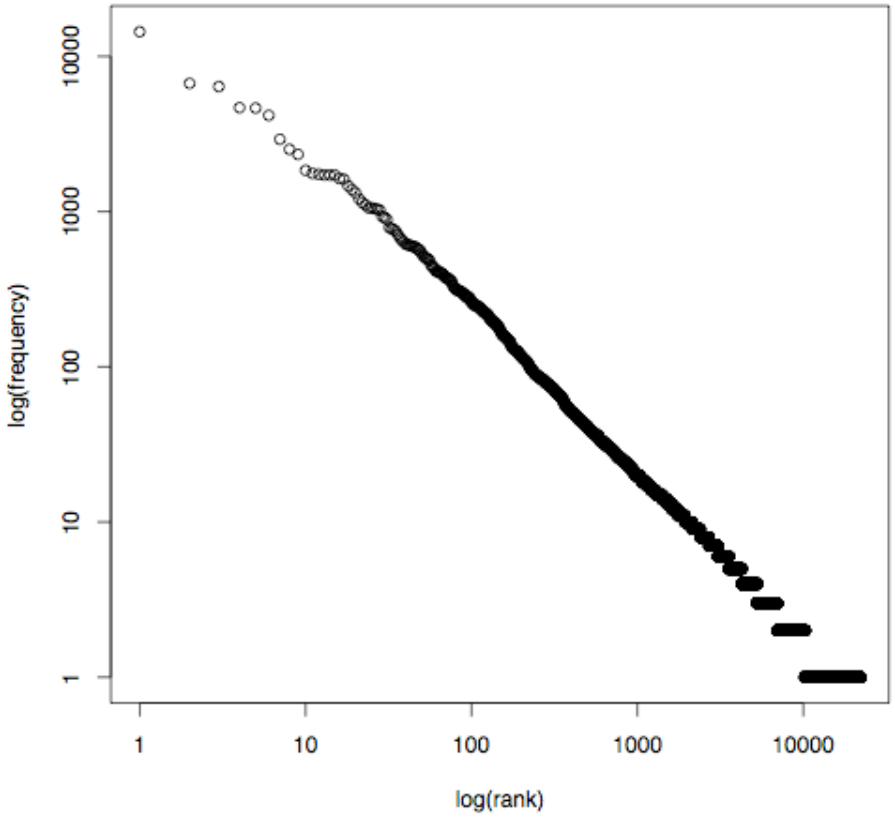
Zipf's Law

Peter Pan



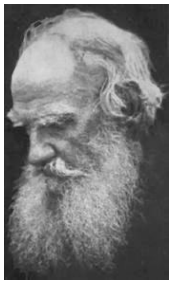
(text courtesy of Project Gutenberg)

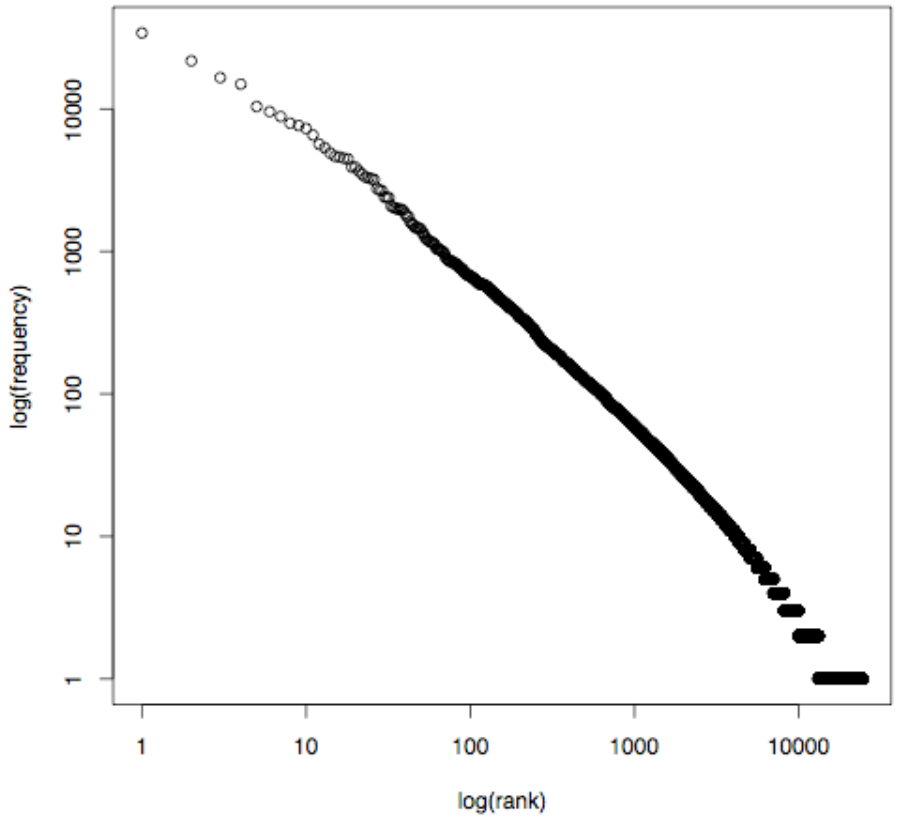
Zipf's Law
Moby Dick



(text courtesy of Project Gutenberg)

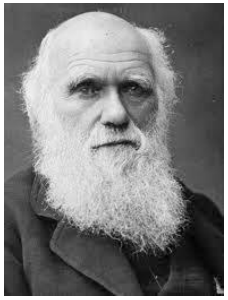
Zipf's Law
War and Peace

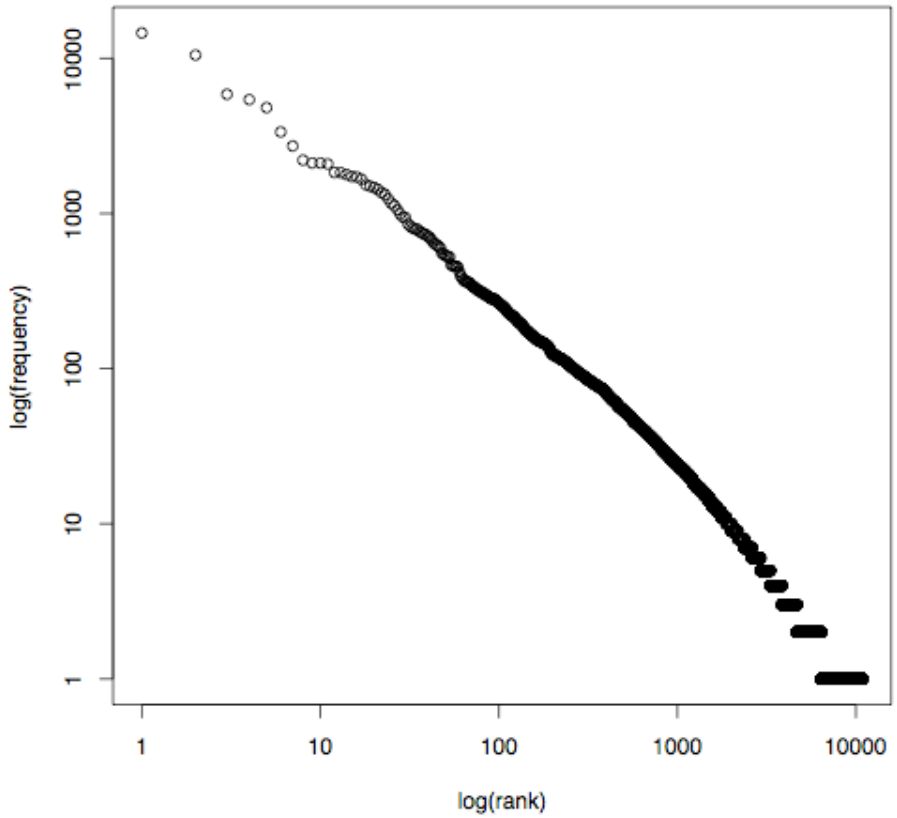




(text courtesy of Project Gutenberg)

Zipf's Law
On the Origin of Species

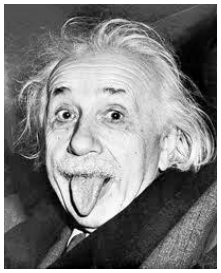


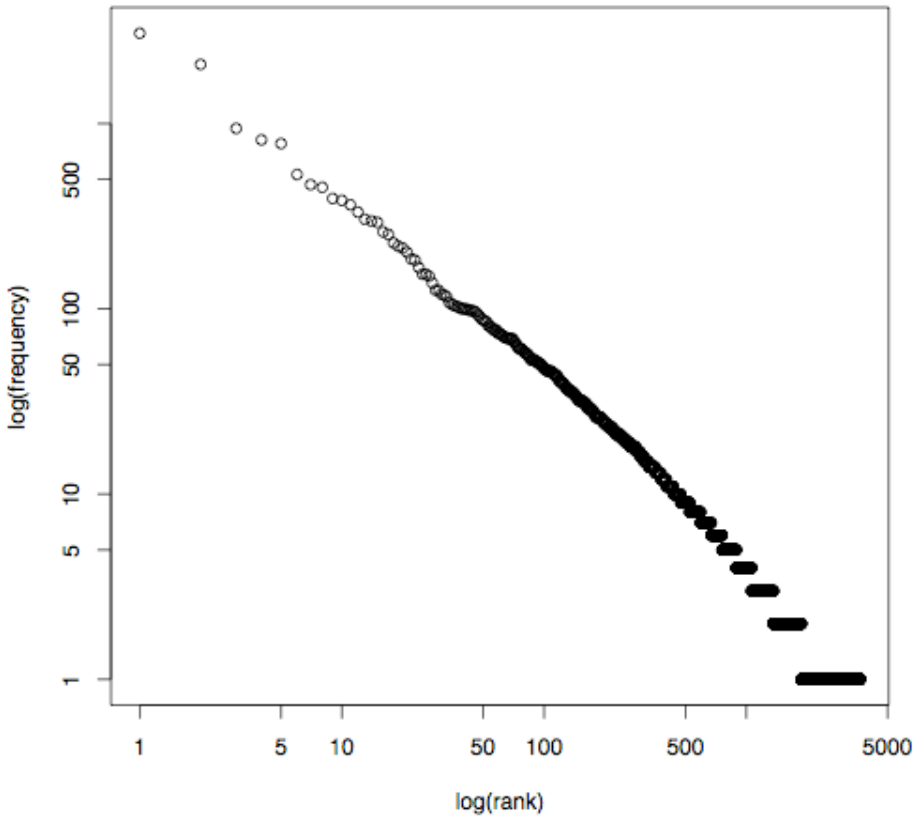


(text courtesy of Project Gutenberg)

Zipf's Law

Relativity: The Special and General Theory

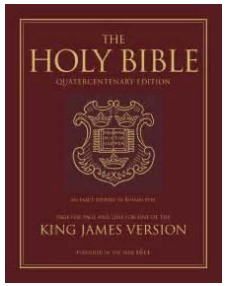


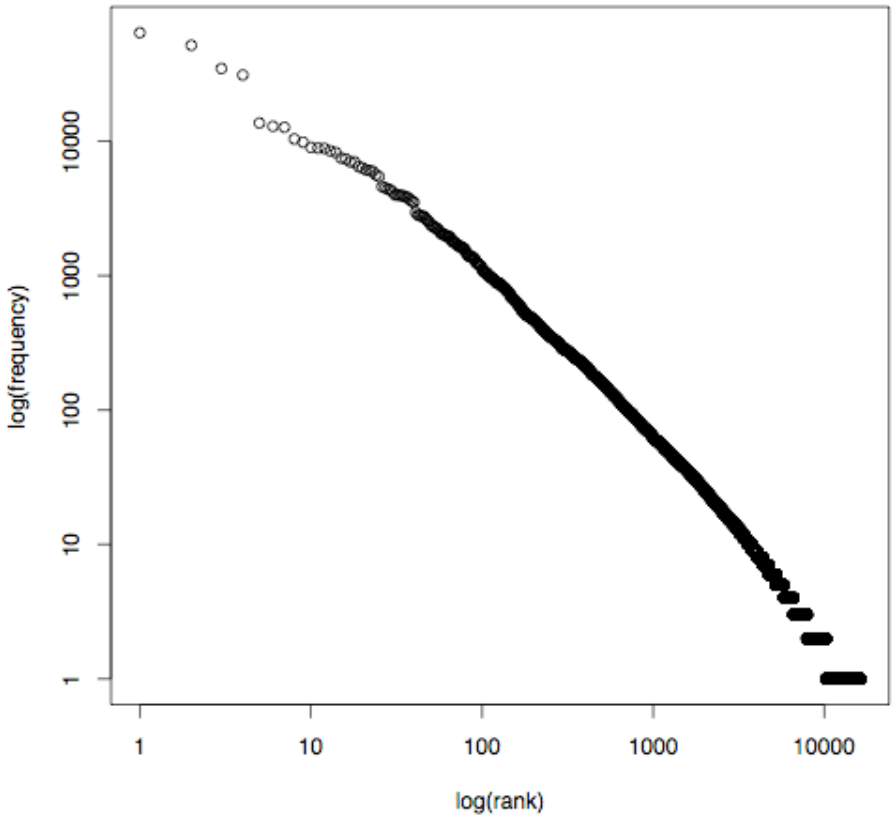


(text courtesy of Project Gutenberg)

Zipf's Law

The King James Bible

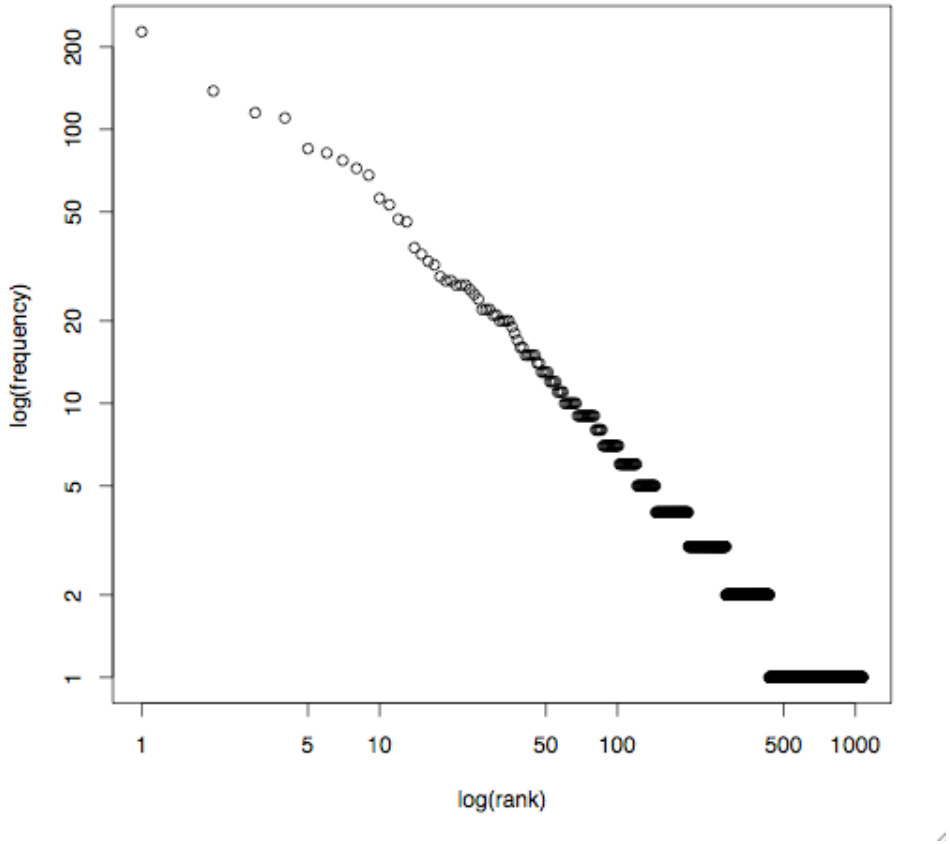




(text courtesy of Project Gutenberg)

Zipf's Law

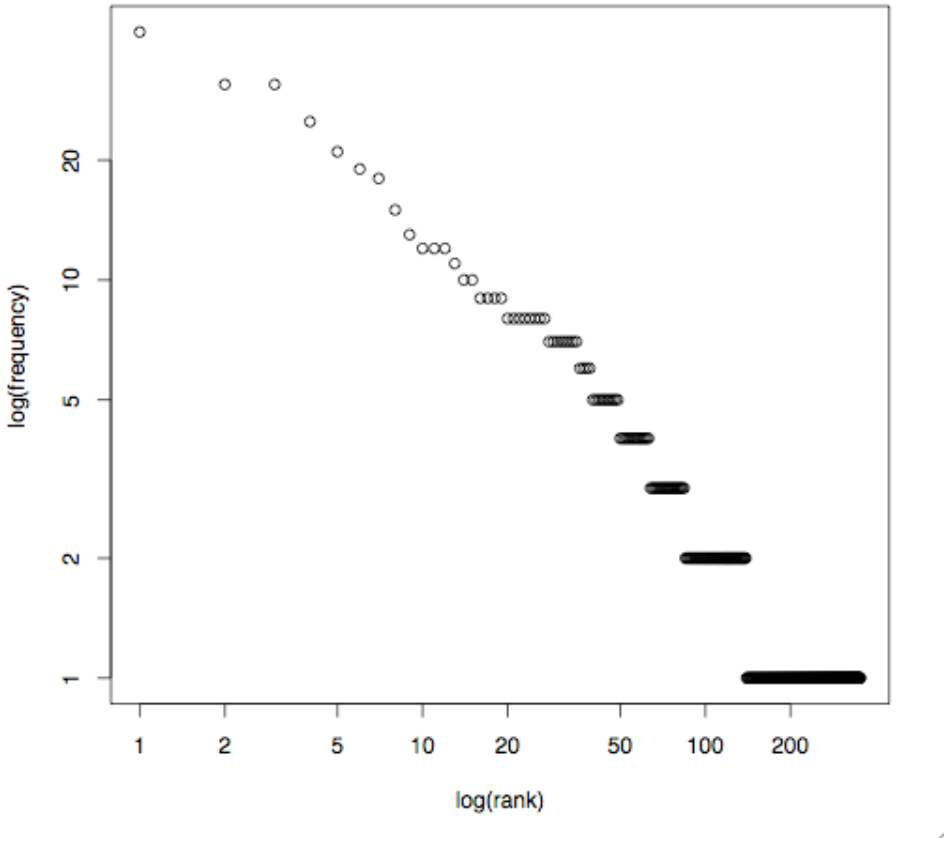
The Tale of Peter Rabbit



(text courtesy of Project Gutenberg)

Zipf's Law

The Three Bears

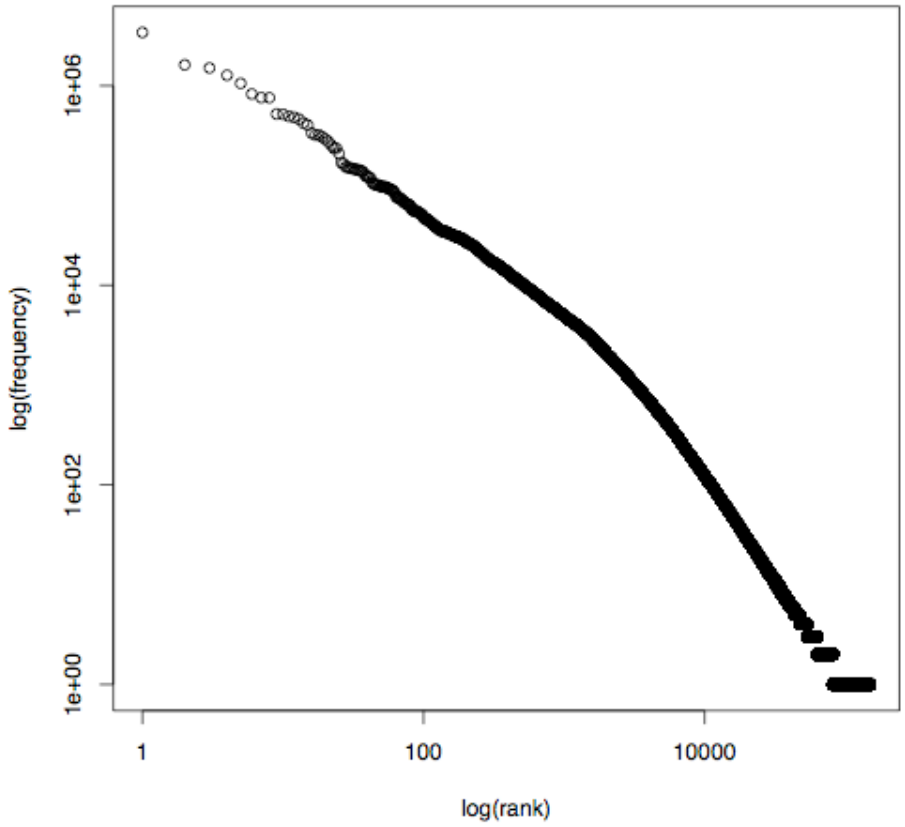


(text courtesy of Project Gutenberg)

Does Zipf's law generalize across different languages?

Zipf's Law

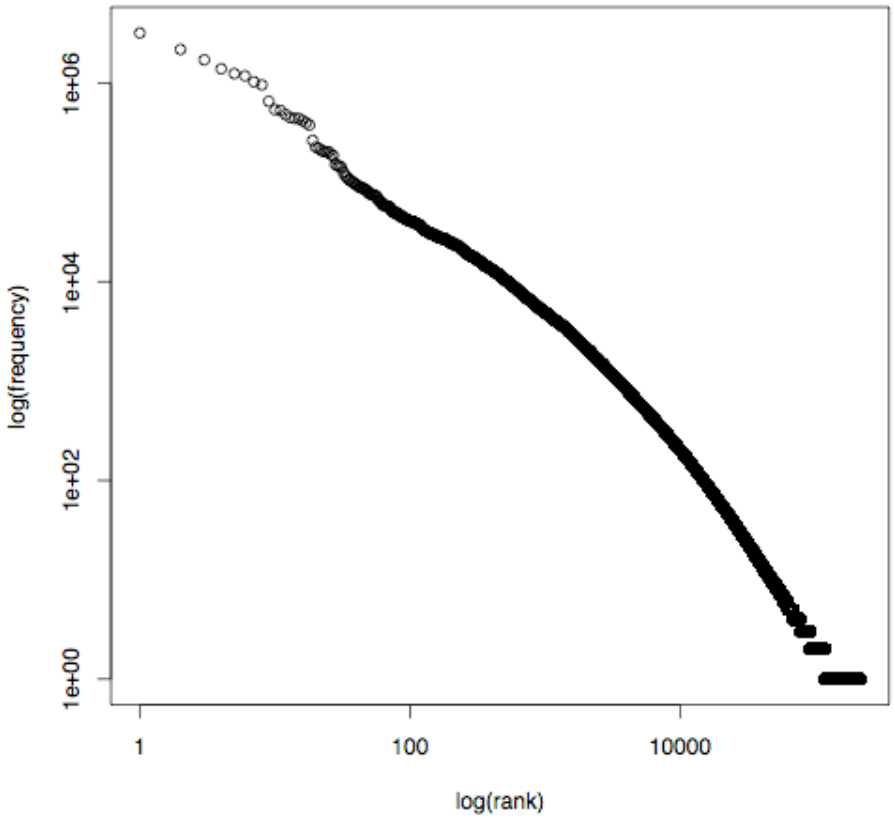
European Parliament: English



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

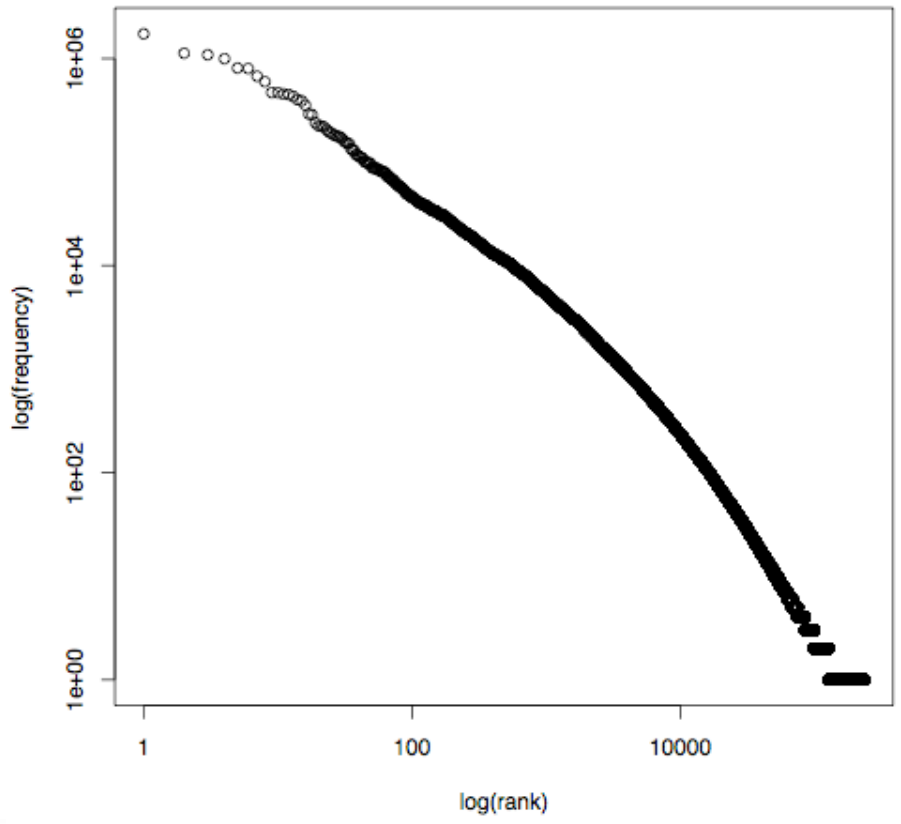
Zipf's Law

European Parliament: Spanish



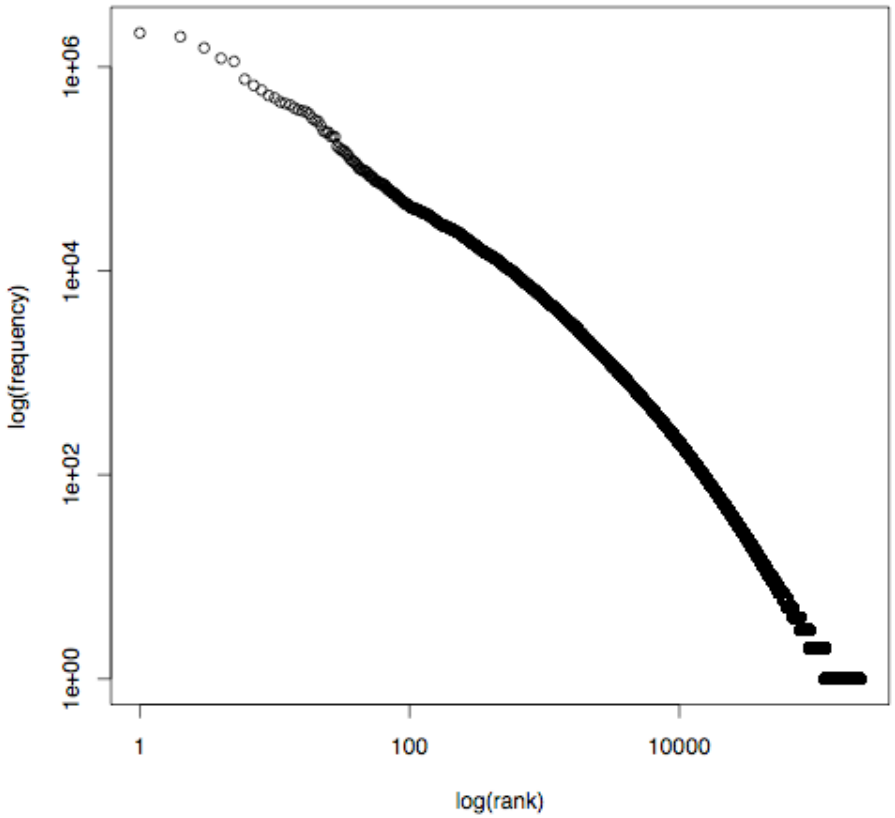
Zipf's Law

European Parliament: Italian



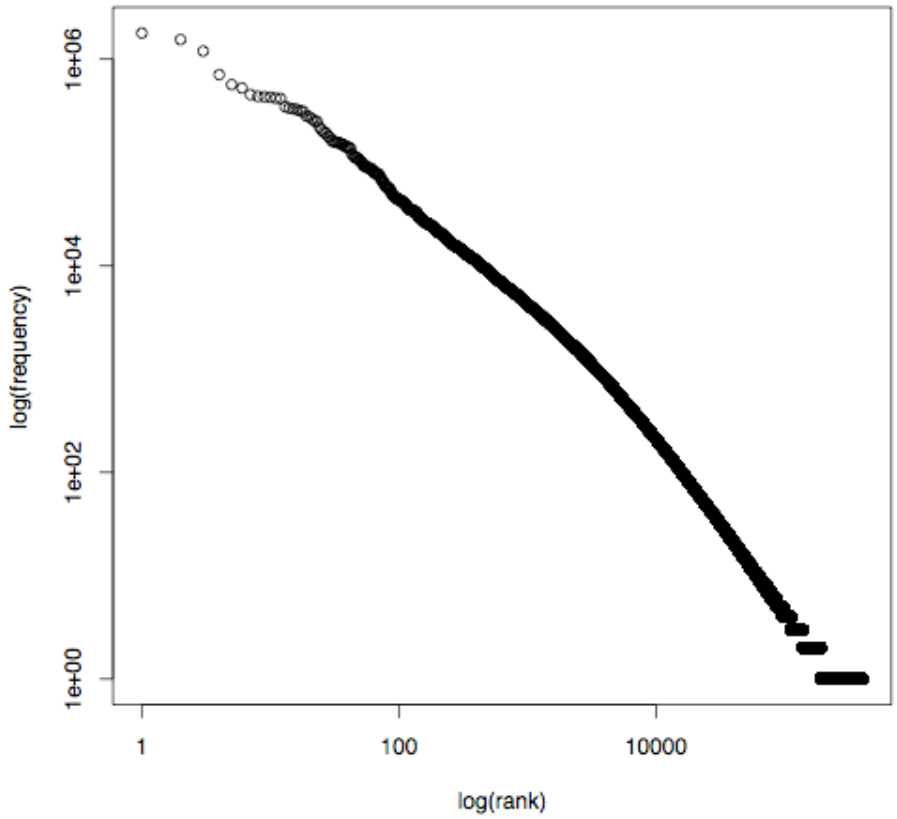
Zipf's Law

European Parliament: Portuguese



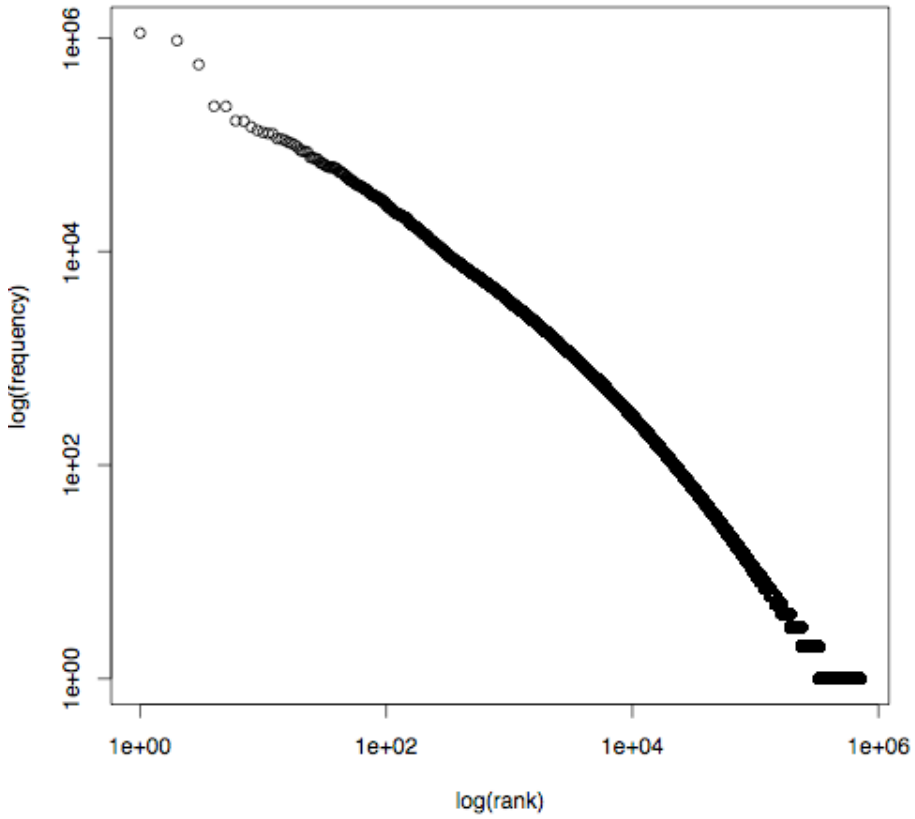
Zipf's Law

European Parliament: German



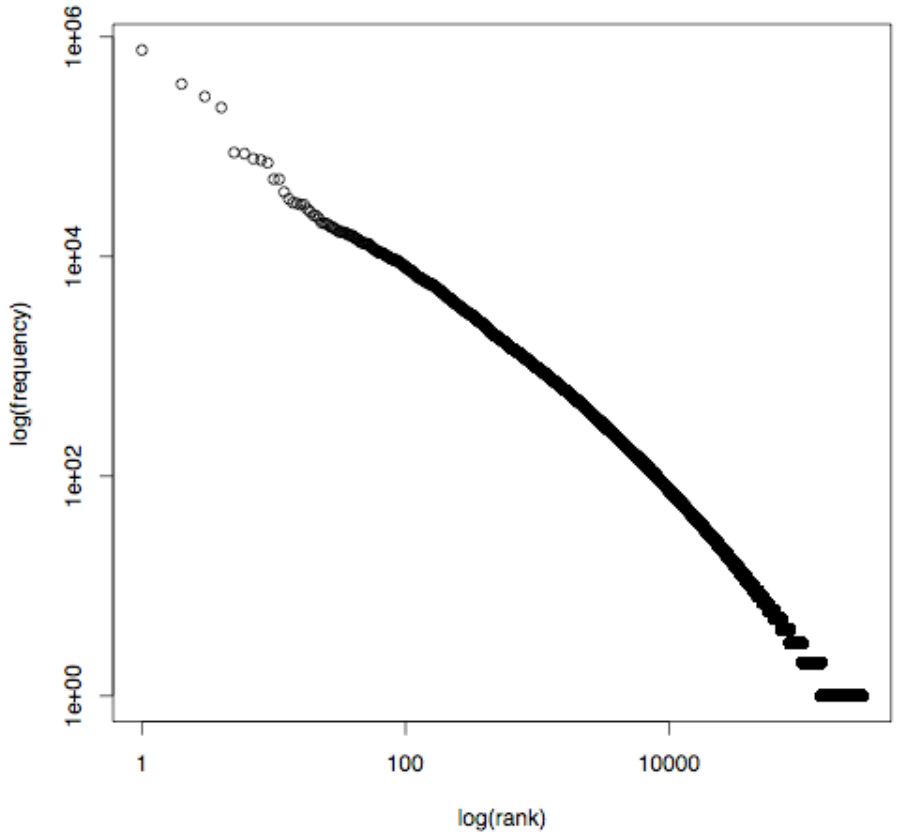
Zipf's Law

European Parliament: Finnish



Zipf's Law

European Parliament: Hungarian

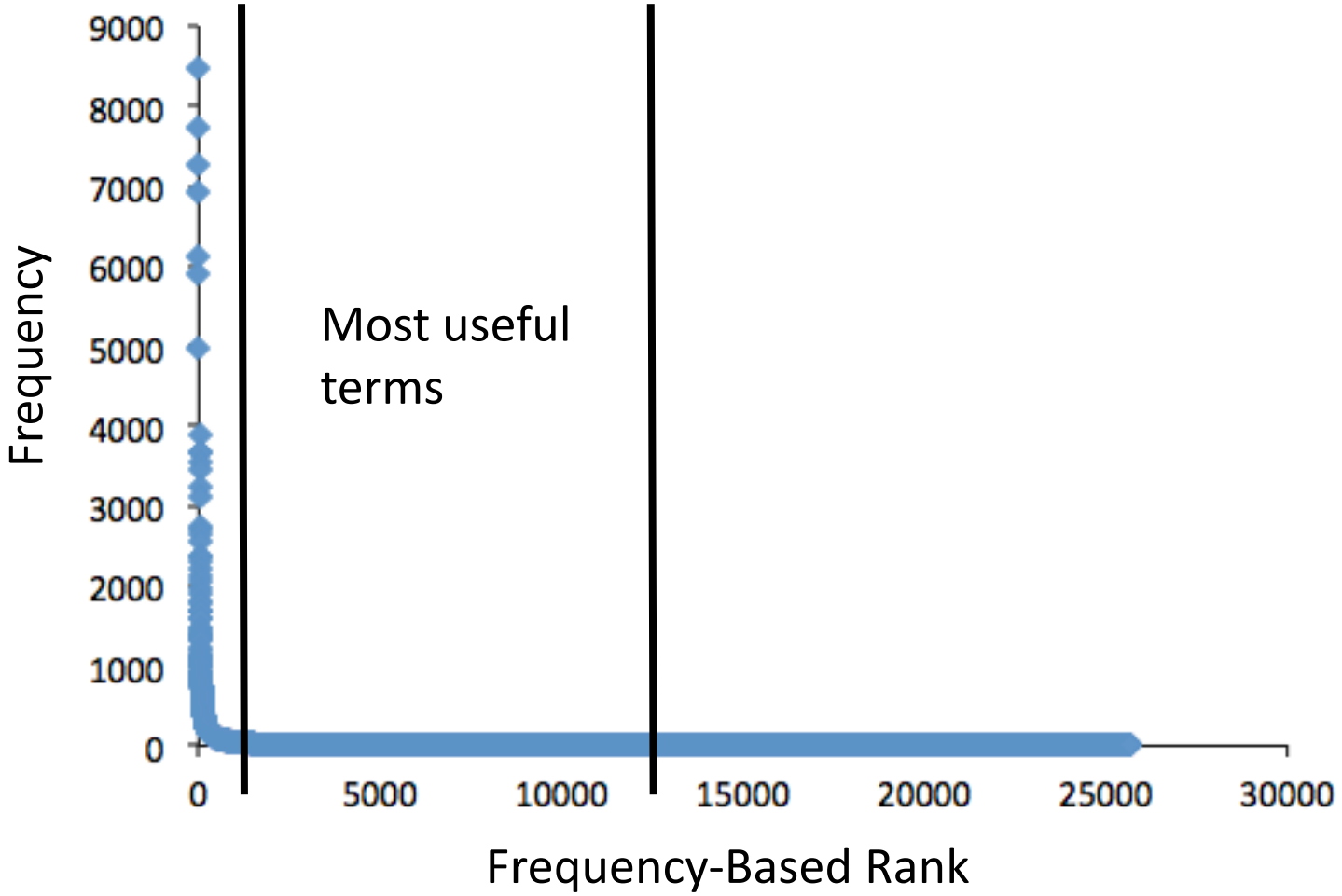


Zipf's Law

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

Zipf's Law

Implications for Feature Selection



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

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)?
- Usually, the more data, the better the estimation!

Statistical Estimation

Link

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

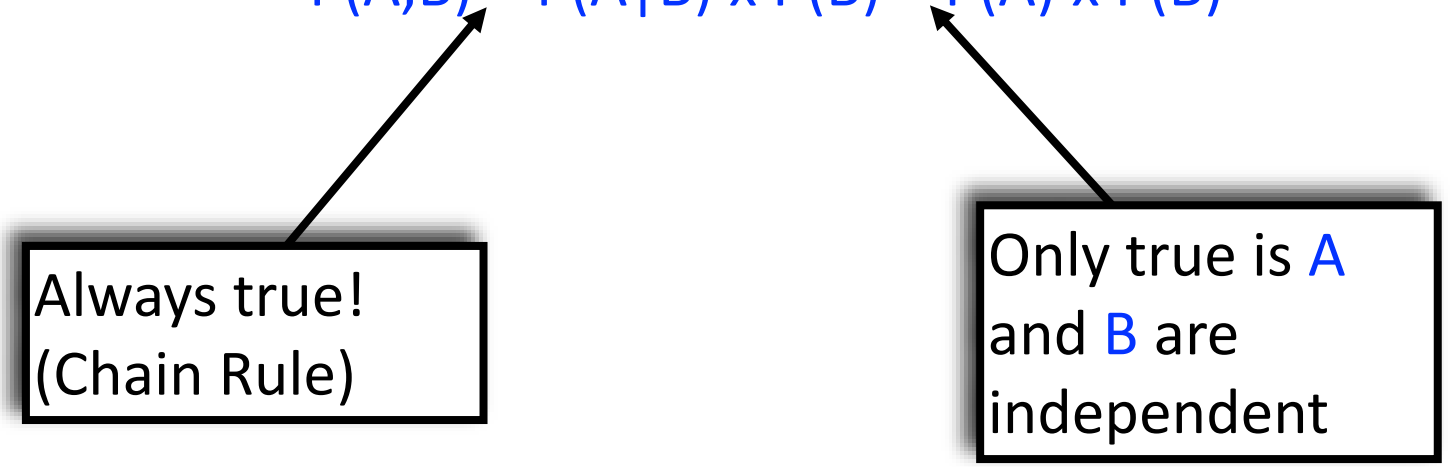
Chain Rule

- 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:

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



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

Mutual Information

- **Mutual Information:** a measure of the mutual dependence between two variables.

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

- **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**

$$P(W,C) = P(W|C) \times P(C) = P(W) \times P(C)$$

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

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

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

$P(w, c^-) = b / N$

$P(c^-) = (b + d) / N$

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

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

Hands-on Exercise 2 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

Mutual Information

be cautious!

- Very infrequent words might get very high mutual information.

	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 = 100

P(w, c) = 1/ 100

P(c) = 1 / 100

P(w) = 1 / 100

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

Hands-on Exercise 2 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

- The basic idea is to examine whether the number of observations (counts) within the categories are different from some expected counts for the corresponding categories.
- Chi-squared distribution is the sum of independent squared z-scores.

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

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

Hands-on Exercise 2 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

Hands-on Exercise 2 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

Feature Selection Methods

- Forward selection
  - Start with *null* model
  - Add one feature to the model at a time
  - For each step, each feature that is yet included in the model is tested for inclusion in the model
  - The most significant of these features is added to the model

Feature Selection Methods

- Backward selection
  - Start with *full* model
  - Remove one feature to the model at a time
  - For each step, each feature that is already included in the model is tested for exclusion
  - The least significant of these features is removed from the model
  - Repeat this until all remaining variable are statistically significant

Feature Selection Methods

- Step-wise selection
  - In this method, we can move either forward or backward by adding or removing features at various steps.

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
- Unsupervised feature selection: filtering terms that are very frequent and very infrequent
- Supervised features selection: focusing on the terms with the highest co-occurrence with each target-class value

Any Questions?

Machine Learning Algorithm

Next Class