3: Statistical Properties of Language

Machine Learning and Real-world Data

Simone Teufel and Ann Copestake

Computer Laboratory
University of Cambridge

Lent 2017

Last session: Naive Bayes Classifier

- You built a smoothed and an unsmoothed NB classifier.
- You evaluated them in terms of accuracy.
- The unsmoothed classifier mostly produced equal probabilites = 0.
- In the smoothed version, this problem has been alleviated.
- Why are there so many zero frequencies, and why does smoothing work?

Statistical Properties of Language I

- How many frequent vs. infrequent terms should we expect in a collection?
- Zipf's law states that there is a direct inverse relationship between a word's frequency rank and the absolute value of that frequency.
- This is an instance of a Power Law.
- The law is astonishingly simple . . .
- . . . given how complex the sentence-internal constraints between some of the words concerned are.

Statistical Properties of Language II

- Heaps' law concerns the relationship between all items of a language and unique items of a language.
- There is an exponential relationship between the two.
- This is also surprising because one might expect saturation.
- Surely at some point all words of a language have been "used up"?

Frequencies of words

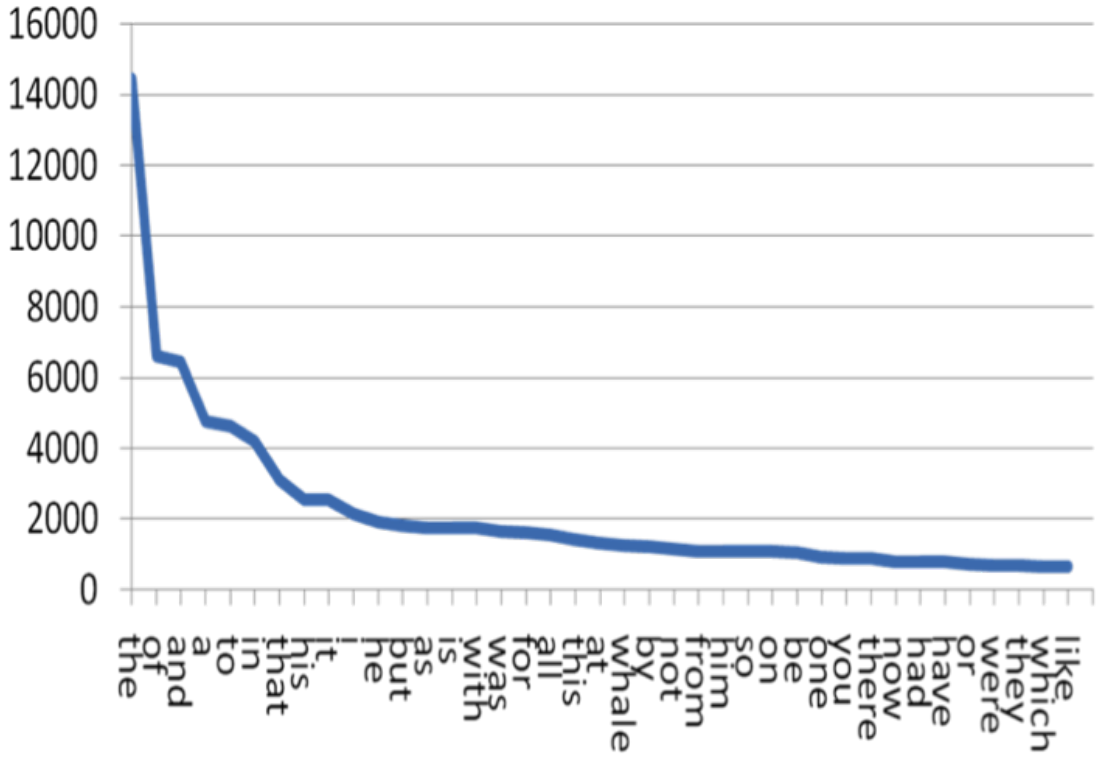
Zipf's law:

There is a direct inverse relationship between a word's frequency rank and the absolute value of that frequency.

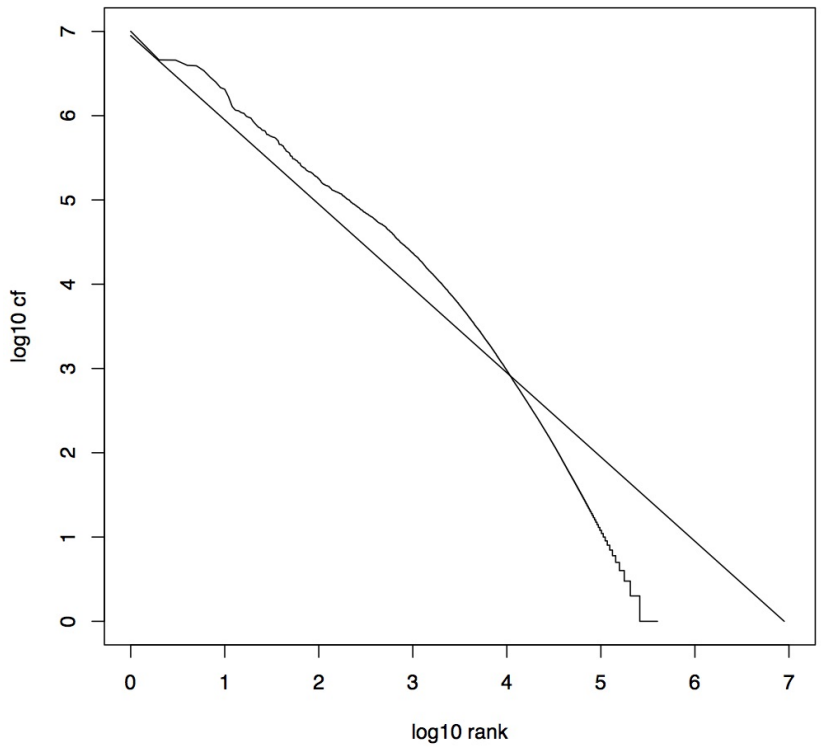
$$f_w \approx \frac{k}{{r_w}^{\alpha}}$$

- f_w: frequency of word w
- r_w: frequency rank of word w
- α, k: constants (language-dependent)
  - α around 1 for English, 1.3 for German
- Zipf's Law means that in language, there are a few very frequent terms and very many very rare terms.

Zipf's Law


16000
14000
12000
10000
8000
6000
4000
2000
0
the
of
and
a
to
in
that
it
i
he
but
as
is
with
was
for
all
this
at
by
whale
not
from
him
so
on
be
one
you
there
now
had
have
or
were
they
which
like

Zipf's Law in log-log space



(Reuters dataset)

Zipf's Law: Examples from 5 Languages

Top 10 most frequent words in some large language samples:

Zipf's Law: Examples from 5 Languages

Top 10 most frequent words in some large language samples:

English

1	the	61,847
2	of	29,391
3	and	26,817
4	a	21,626
5	in	18,214
6	to	16,284
7	it	10,875
8	is	9,982
9	to	9,343
10	was	9,236

BNC,
100Mw

Zipf's Law: Examples from 5 Languages

Top 10 most frequent words in some large language samples:

	English		German	
1	the	61,847	der	7,377,879
2	of	29,391	die	7,036,092
3	and	26,817	und	4,813,169
4	a	21,626	in	3,768,565
5	in	18,214	den	2,717,150
6	to	16,284	von	2,250,642
7	it	10,875	zu	1,992,268
8	is	9,982	das	1,983,589
9	to	9,343	mit	1,878,243
10	was	9,236	sich	1,680,106

BNC, 100Mw

"Deutscher Wortschatz", 500Mw

Zipf's Law: Examples from 5 Languages

Top 10 most frequent words in some large language samples:

	English		German		Spanish	
1	the	61,847	der	7,377,879	que	32,894
2	of	29,391	die	7,036,092	de	32,116
3	and	26,817	und	4,813,169	no	29,897
4	a	21,626	in	3,768,565	a	22,313
5	in	18,214	den	2,717,150	la	21,127
6	to	16,284	von	2,250,642	el	18,112
7	it	10,875	zu	1,992,268	es	16,620
8	is	9,982	das	1,983,589	y	15,743
9	to	9,343	mit	1,878,243	en	15,303
10	was	9,236	sich	1,680,106	lo	14,010

BNC, 100Mw

"Deutscher Wortschatz", 500Mw

subtitles, 27.4Mw

Zipf's Law: Examples from 5 Languages

Top 10 most frequent words in some large language samples:

	English		German		Spanish		Italian	
1	the	61,847	der	7,377,879	que	32,894	non	25,757
2	of	29,391	die	7,036,092	de	32,116	di	22,868
3	and	26,817	und	4,813,169	no	29,897	che	22,738
4	a	21,626	in	3,768,565	a	22,313	è	18,624
5	in	18,214	den	2,717,150	la	21,127	e	17,600
6	to	16,284	von	2,250,642	el	18,112	la	16,404
7	it	10,875	zu	1,992,268	es	16,620	il	14,765
8	is	9,982	das	1,983,589	y	15,743	un	14,460
9	to	9,343	mit	1,878,243	en	15,303	a	13,915
10	was	9,236	sich	1,680,106	lo	14,010	per	10,501

BNC, 100Mw

"Deutscher Wortschatz", 500Mw

subtitles, 27.4Mw

subtitles, 5.6Mw

Zipf's Law: Examples from 5 Languages

Top 10 most frequent words in some large language samples:

	English		German		Spanish		Italian		Dutch	
1	the	61,847	der	7,377,879	que	32,894	non	25,757	de	4,770
2	of	29,391	die	7,036,092	de	32,116	di	22,868	en	2,709
3	and	26,817	und	4,813,169	no	29,897	che	22,738	het/'t	2,469
4	a	21,626	in	3,768,565	a	22,313	è	18,624	van	2,259
5	in	18,214	den	2,717,150	la	21,127	e	17,600	ik	1,999
6	to	16,284	von	2,250,642	el	18,112	la	16,404	te	1,935
7	it	10,875	zu	1,992,268	es	16,620	il	14,765	dat	1,875
8	is	9,982	das	1,983,589	y	15,743	un	14,460	die	1,807
9	to	9,343	mit	1,878,243	en	15,303	a	13,915	in	1,639
10	was	9,236	sich	1,680,106	lo	14,010	per	10,501	een	1,637

BNC, 100Mw

"Deutscher Wortschatz", 500Mw

subtitles, 27.4Mw

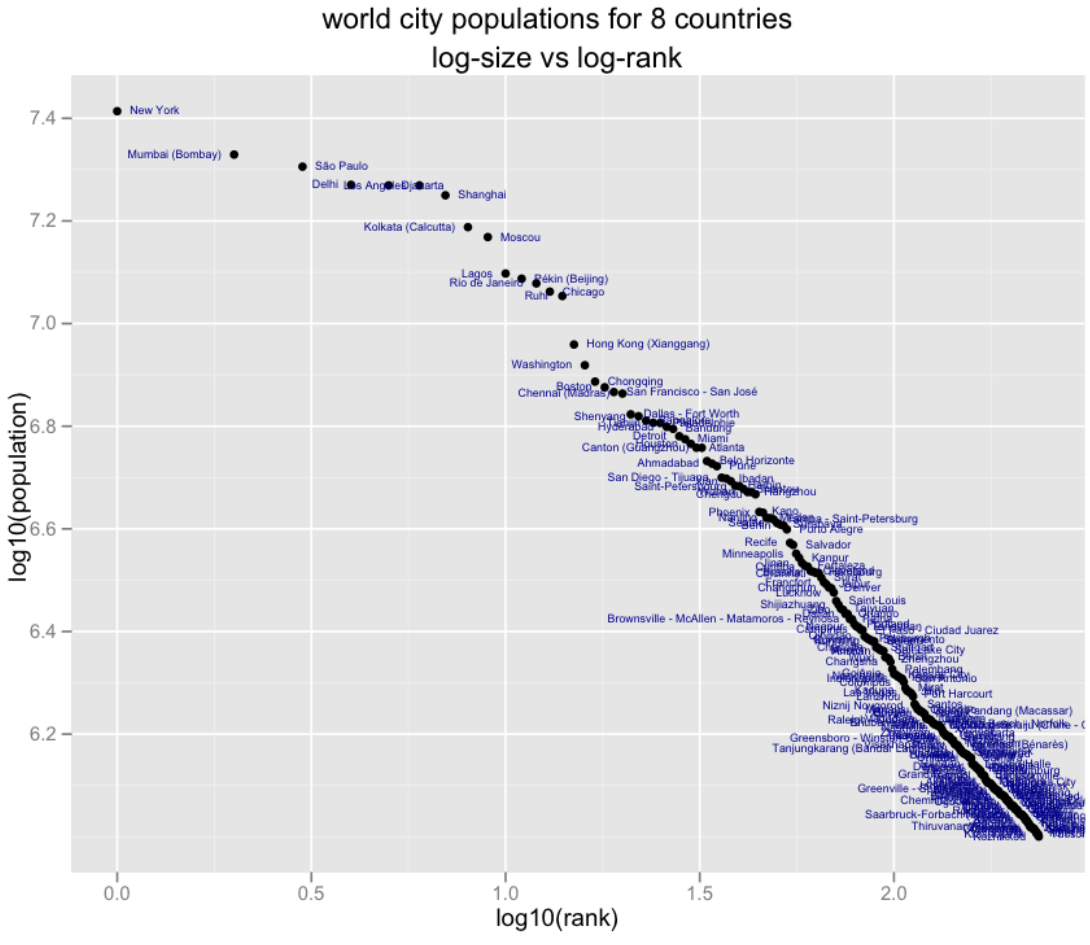
subtitles, 5.6Mw

subtitles, 800Kw

Other collections (allegedly) obeying power laws

- Sizes of settlements
- Frequency of access to web pages
- Income distributions amongst top earning 3% individuals
- Korean family names
- Size of earth quakes
- Word senses per word
- Notes in musical performances
- ...

World city populations



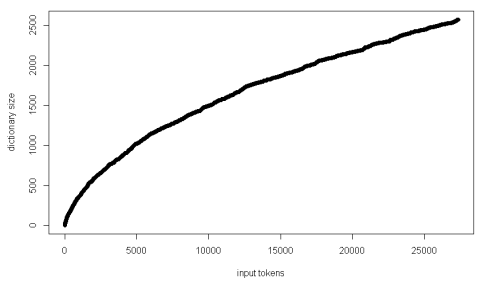
Vocabulary size

Heaps' Law:

The following relationship exists between the size of a vocabulary and the size of text that gave rise to it:

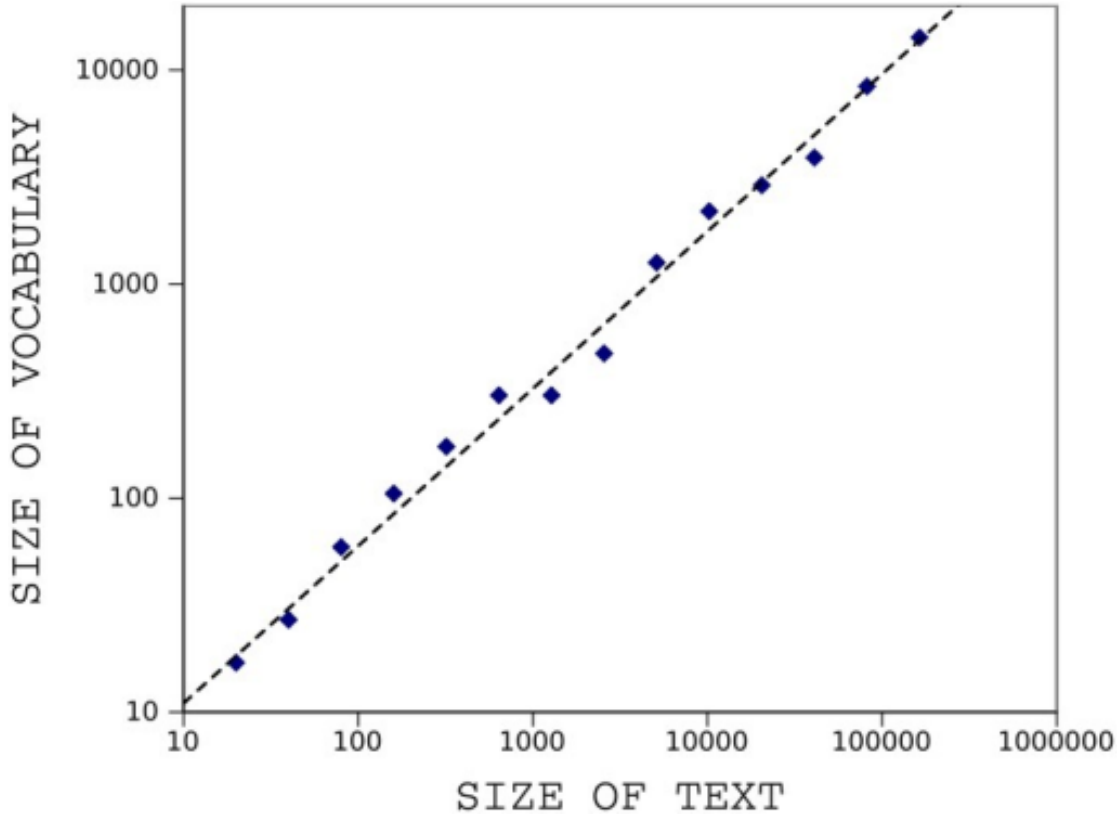
$$u_n = kn^{\beta}$$

- u_n: number of types (unique items); vocabulary size
- n: number of tokens; text size
- β, k: constants (language-dependent)
  - β normally around $\frac{1}{2}$
  - $30 \leq k \leq 100$



Heaps' Law

- In log-log space:

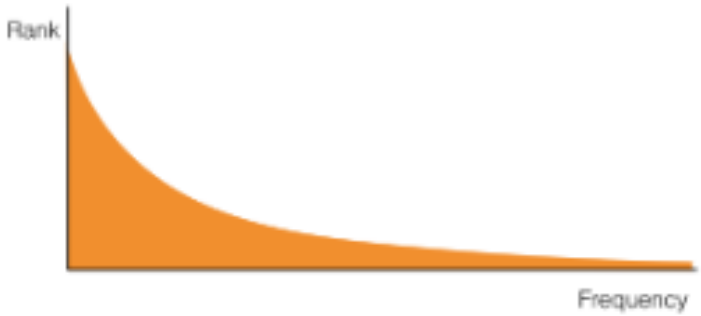


- Reasons for infinite vocabulary growth?

Consequences for our experiment

- Zipf's law and Heaps' law taken together explain why smoothing is necessary and effective:
  - MLE overestimates the likelihood for seen words.
  - Smoothing redistributes some of this probability mass.

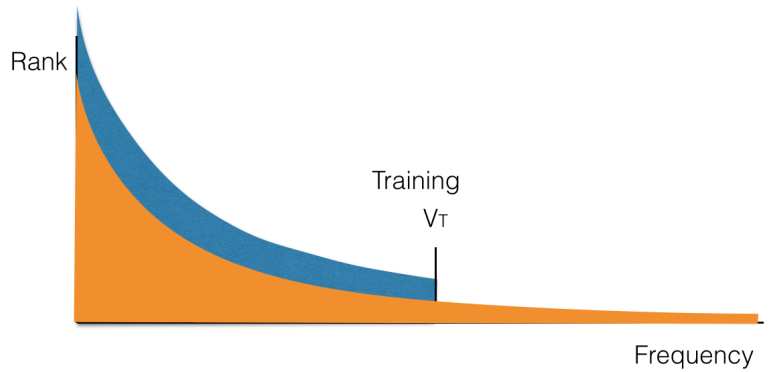
The real situation



- Most of the probability mass is in the long tail.

The situation according to MLE

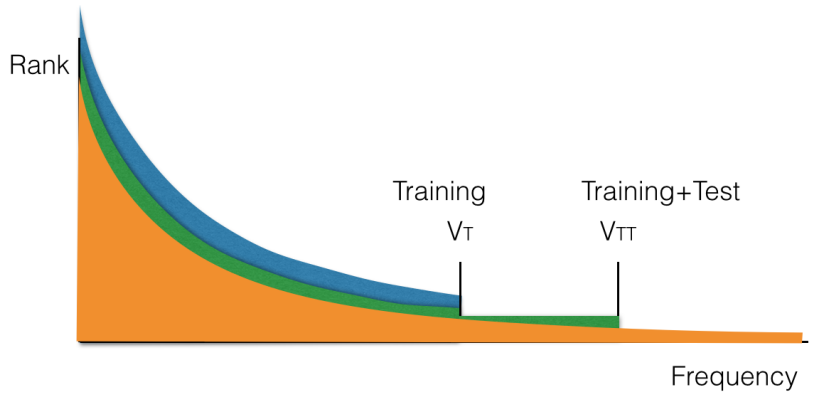
$$\hat{P}_{MLE}(w_i|c) = \frac{count(w_i, c)}{\sum_{w \in V_T} count(w, c)}$$



- With MLE, only seen words can get a frequency estimate.
- Probability mass is still 1.
- Therefore, the probability of seen words is a (big) overestimate.

What smoothing does

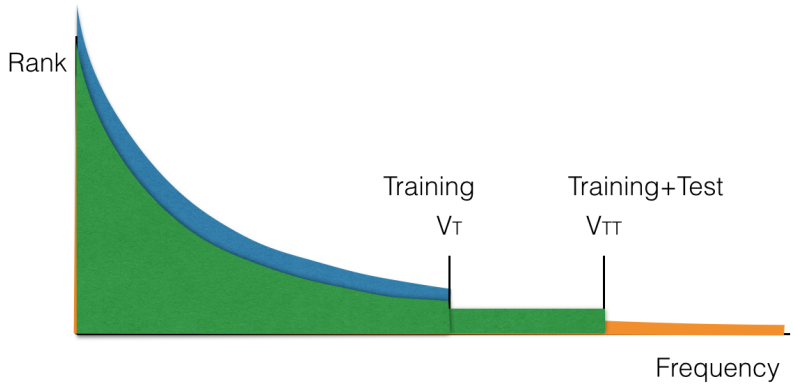
$$\hat{P}_S(w_i|c) = \frac{count(w_i, c) + 1}{(\sum_{w \in V_{TT}} count(w, c)) + |V_{TT}|}$$



- Smoothing redistributes the probability mass towards the real situation.
- It takes some portion away from the MLE overestimate for seen words.
- It redistributes this portion to a certain, finite number of unseen words (in our case, as a uniform distribution).

What smoothing does

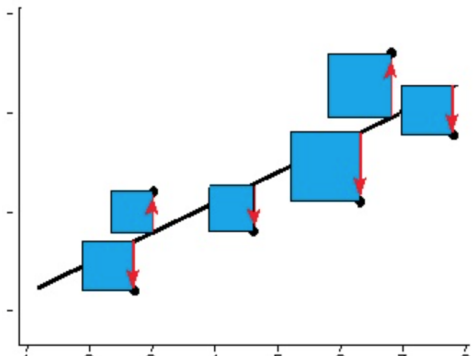
$$\hat{P}_S(w_i|c) = \frac{count(w_i, c) + 1}{(\sum_{w \in V_{TT}} count(w, c)) + |V_{TT}|}$$



- Smoothing takes some portion away from the MLE overestimate for seen words.
- It redistributes this portion to a certain, finite number of unseen words (in our case, as a uniform distribution).
- As a result, the real situation is approximated more closely.

Your first task today

- Plot frequeny vs frequency rank for larger dataset (i.e., visually verify Zipf's Law)
- Estimate parameters k, α for Zipf's Law
- Use least-squares algorithm for doing so.
- Is α really 1 for English?
  - There is much scientific discussion of this question.



Your second task today

- Plot type/token ratio for IMDB dataset (verify Heaps' Law)

Ticking today

- Task 2 – NB Classifier

Literature

- Introduction to Information Retrieval, Christopher C. Manning, Prabhakar Raghavan, Hinrich Schutze, Cambridge University Press, 2008. Section 5.1, pages 79-82.
- (Please note that $\alpha = 1$ is assumed in the Zipf Law formula on page 82.)