Natural Language Processing: Revision of Zipf's law

Katja Markert

School of Computing
University of Leeds
markert@comp.leeds.ac.uk

October 9, 2013

Previously

We talked about

- NLP questions
- NLP applications
- Ambiguity in Language

An example for counting words

> ...A certain selection and discretion must be used in producing a realistic effect," remarked Holmes. ...

350	the		....
212	and	58	but
191	to	47	have
167	of	47	which
165	a	46	me
160	i	46	holmes
134	that	20	windibank

An example for counting words

> …A certain selection and discretion must be used in producing a realistic effect,” remarked Holmes. …

350	the		....
212	and	58	but
191	to	47	have
167	of	47	which
165	a	46	me
160	i	46	holmes
134	that	20	windibank

Frequencies of Frequencies

Frequencies of frequencies in “A Case of Identity”

Word Frequency	Frequency of Frequency
1	993
2	248
3	93
4	70
5	40
10	8
50	2
>100	11

7105	word tokens
1625	word types

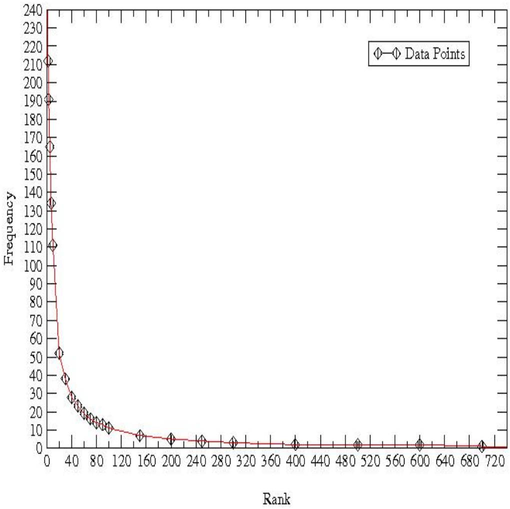
Zipf's law I

word	Freq. (f)	Rank (r)	$f \cdot r$
the	350	1	350
and	212	2	424
to	191	3	573
was	111	10	1110
her	52	20	1040
had	38	30	1140
very	28	40	1120
what	23	50	1150
father	19	60	1140
come	16	70	1120

Zipf's law I

word	Freq. (f)	Rank (r)	$f \cdot r$
out	14	80	1120
can	13	90	1170
street	11	100	1100
time	7	150	1050
leadenhall	5	200	1000
went	3	300	900
violent	2	400	800
mean	2	500	1000
certain	2	600	1200
unprof	1	700	700
pleasant	1	1000	1000

Frequency and Rank



Zipf’s law II

Captures the relationship between **frequency** and **rank**. (observation made by Harvard linguist George Kingsley Zipf).

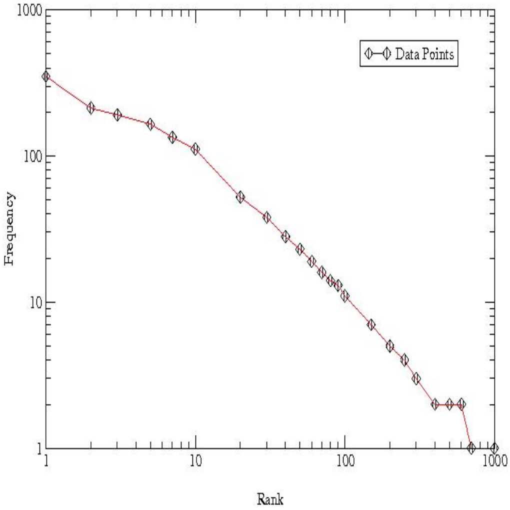
There is a constant *k* such that:

$$f \cdot r = k$$

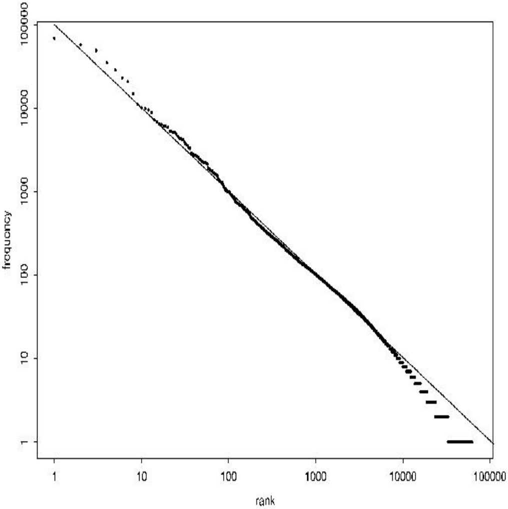
Or *f* is a **power-law function of** r:

$$f \propto \frac{1}{r}$$

Logarithmic Scales



Zipf’s law for a 1m word corpus



Zipf's law in layman's terms

- There is a very small number of very common words
- There is a small-medium number of middle frequency words
- There is a very large number of words that are infrequent
- The relationship between frequency and rank can be approximated by a line (in logarithmic scales)
- Different from bell curve (Gaussian/normal distribution).

More Zipf

Question: Which other phenomena are governed by Zipf's law?

- Now frequently invoked for the web too!
  (See `www.hpl.hp.com/research/idl/papers/ranking/adamicglottometrics.pdf`)
- income distribution amongst individuals
- size of earthquakes

Data Sparseness

Data Sparseness: For most words we will have very few or no examples, which can lead to unreliable counts

You can try out corpus counts at `http://sara.natcorp.ox.ac.uk/lookup.html` The BNC is the British National Corpus with 100 million tokens.

Discussion Topics

Questions:

1. How does Zipf's law relate to Rationalist criticism of empiricsm?
2. What can we do with unseen events? Should they all be treated the same?
3. Other mathematical laws in language?

Summary

- Estimation in NLP starts with sampling textual data and is dependent on the sampled data
- Estimation can suffer from adverse distributions
- As exemplified by Zipf's law
- Word frequencies can be predictors of other properties of words