Word frequencies

HAP/LAP. Corpus Linguistics.

Word frequencies

- How many times do words occur in a corpus?
- How is the frequency of different words distributed?
  - Are all words about equally frequent?
- How fast does vocabulary size grow with the size of a corpus?
- Word frequencies give insights of human language.
- Let's count words!
- Also, let's obtain the histogram of word freqs.

Using R for displaying frequencies

```
> v <- c(1,1,1,1,1,1,2,2,2,3,4,5,5,5,6)
% histogram
> hist(v)
> hist(v,breaks=10) % 10 bins
% histogram of 1000 number sample following uniform distribution
> hist(runif(1000))
% histogram of 2000 number sample following normal distribution (mean 4, sd 1)
> hist(rnorm(2000, 4, 1))
% load vector from file
> source("simple.R")
```

Word frequencies and Zipf's law

- Sort words according to their frequency (decreasing order)
- Exploit relationship between a word's frequency (f) count and its position in the list (rank, r).
- Zipf law: $f \propto \frac{1}{r}$
  - more exactly: $f \propto \frac{1}{r^{\alpha}}$ and $\alpha \propto 1$
- alternatively: there exist a constant a st. $f \cdot r = a$
- Zipf's law says: 50^{th} most common word should occur three times the frequency of the 150^{th} most common word.

Word frequencies and Zipf's law

Word	f	r	$f \cdot r$	Word	f	r	$f \cdot r$
the	133582	1	133582	with	17598	15	263970
and	95366	2	190732	is	16501	16	264016
of	71229	3	213687	you	16417	17	279089
to	47719	4	190876	be	16096	18	289728
a	33822	5	169110	as	14527	19	276013
in	33494	6	200964	but	13946	20	278920
i	29234	7	204638	all	13687	21	287427
that	28810	8	230480	they	13107	22	288354
he	25861	9	232749	him	13025	23	299575
it	22304	10	223040	shall	11684	24	280416
his	21411	11	235521	her	11574	25	289350
for	19522	12	234264	my	10515	26	273390
was	18779	13	244127	had	10351	27	279477
not	18199	14	254786	them	10249	28	286972

Counting words

Counting words (again)

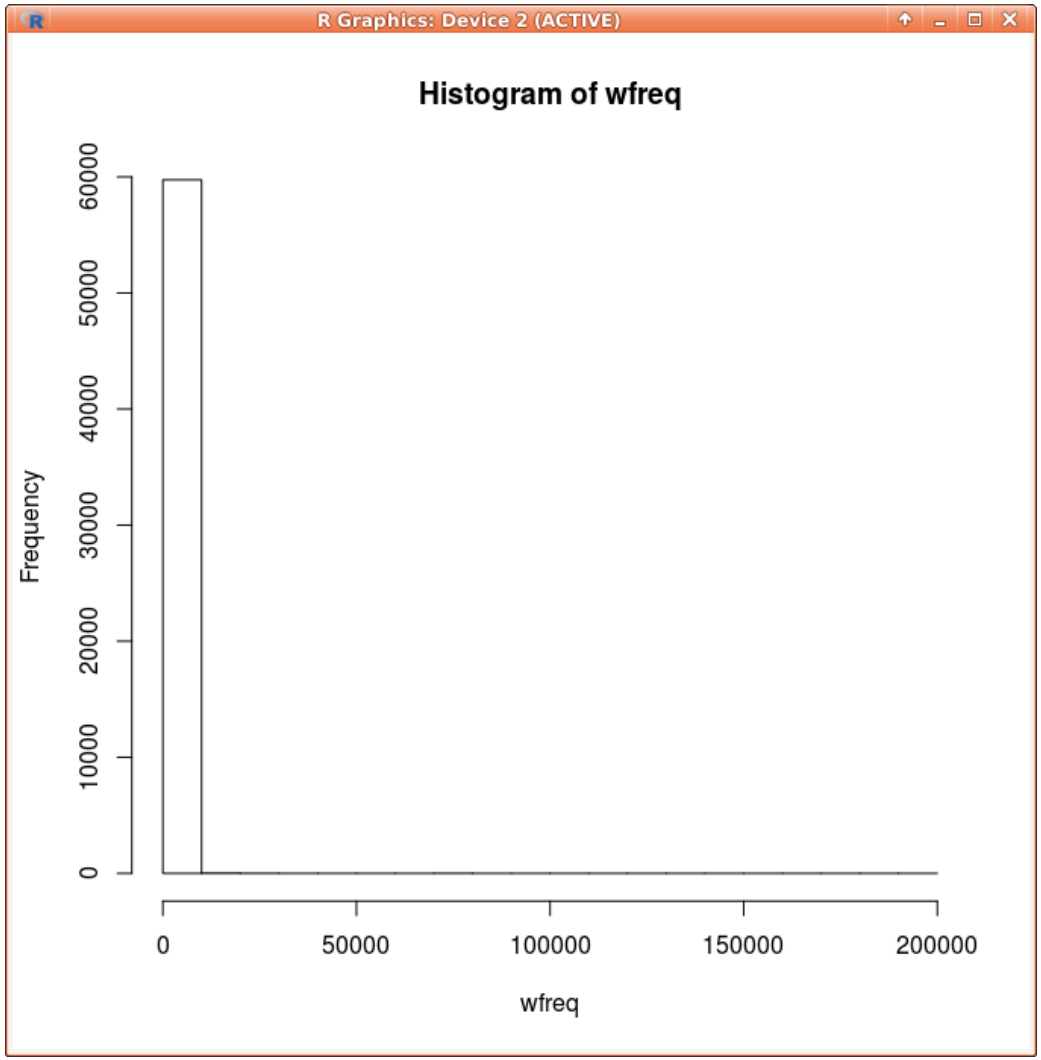
Implement a script which outputs word frequencies

- `freq TAB word`

Store the output in a file called `wfreq.list`

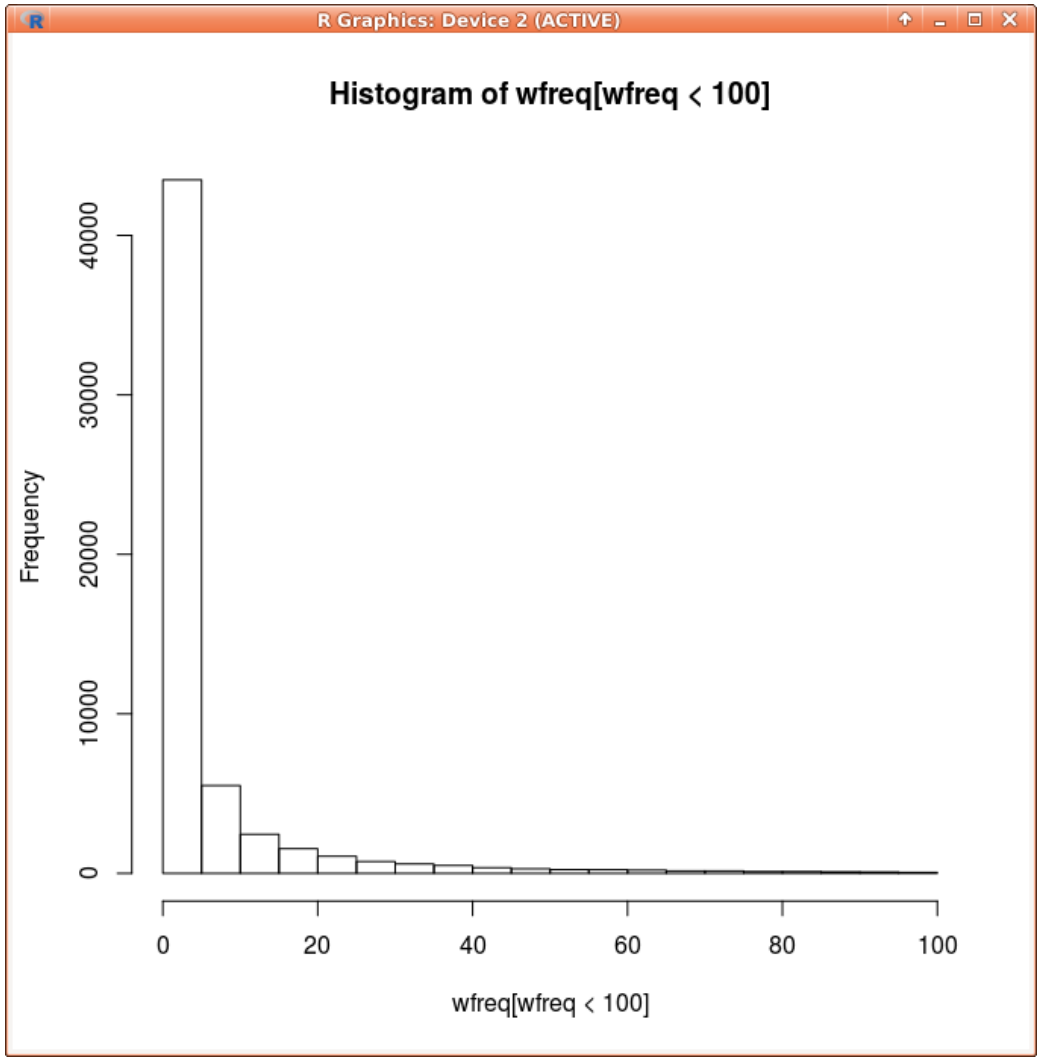
Histogram of word frequencies

```
% python3 createR.py wfreq.list > wfreq.R
% R
> source("wfreq.R")
> hist(wfreq)
```



Histogram using R

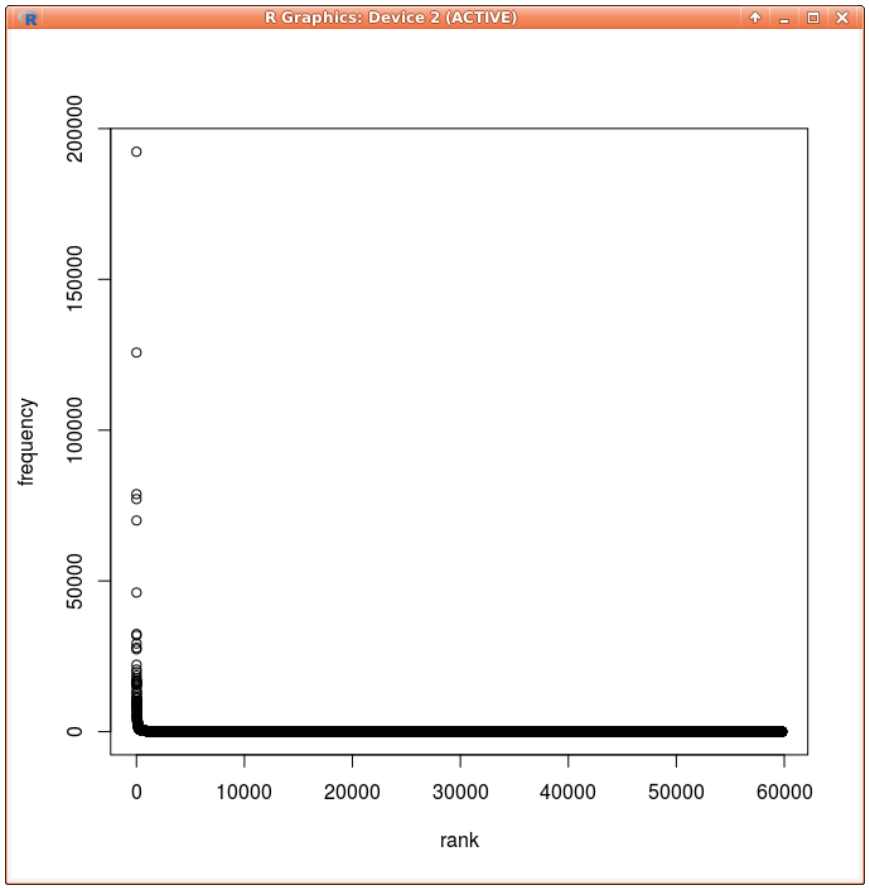
```
% python3 createR.py wfreq.list > wfreq.R
% R
> source("wfreq.R")
> hist(wfreq[wfreq < 100])
```



Plotting word frequencies

- Another option is to plot the frequencies:
- Need `wfreq` to be sorted according to freq.

```
% python3 createR.py wfreq.list > wfreq.R
% R
> source("wfreq.R")
> plot(wfreq)
```

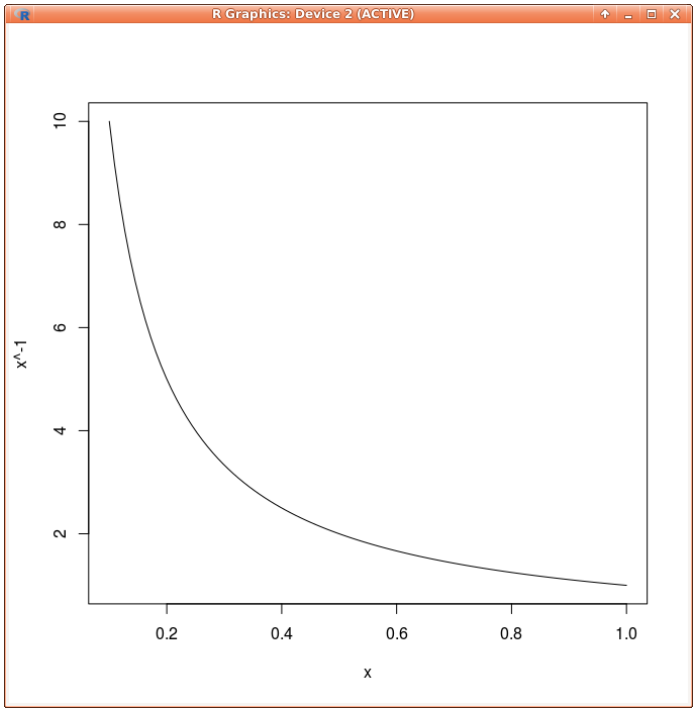


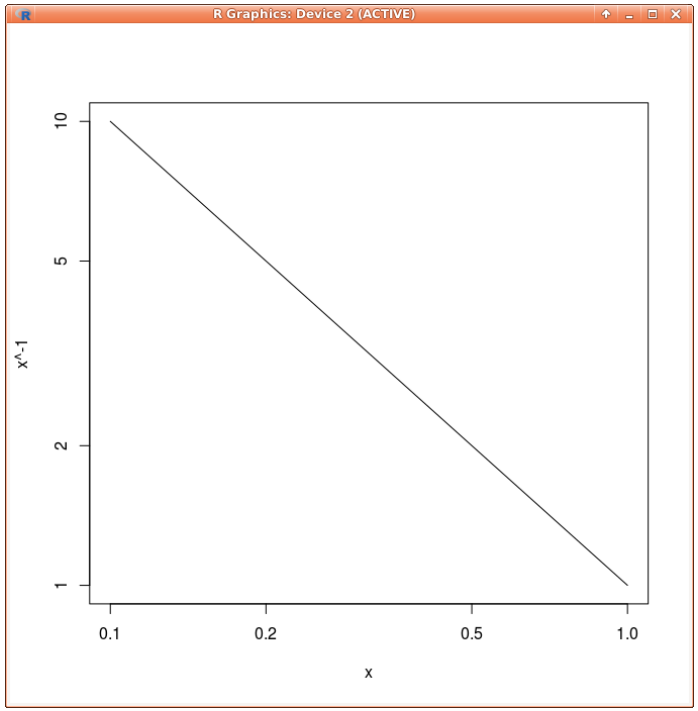
Power law

- Zipf law is a special type of power law function

  $y = ax^k$, where $a = 1, k = -1$

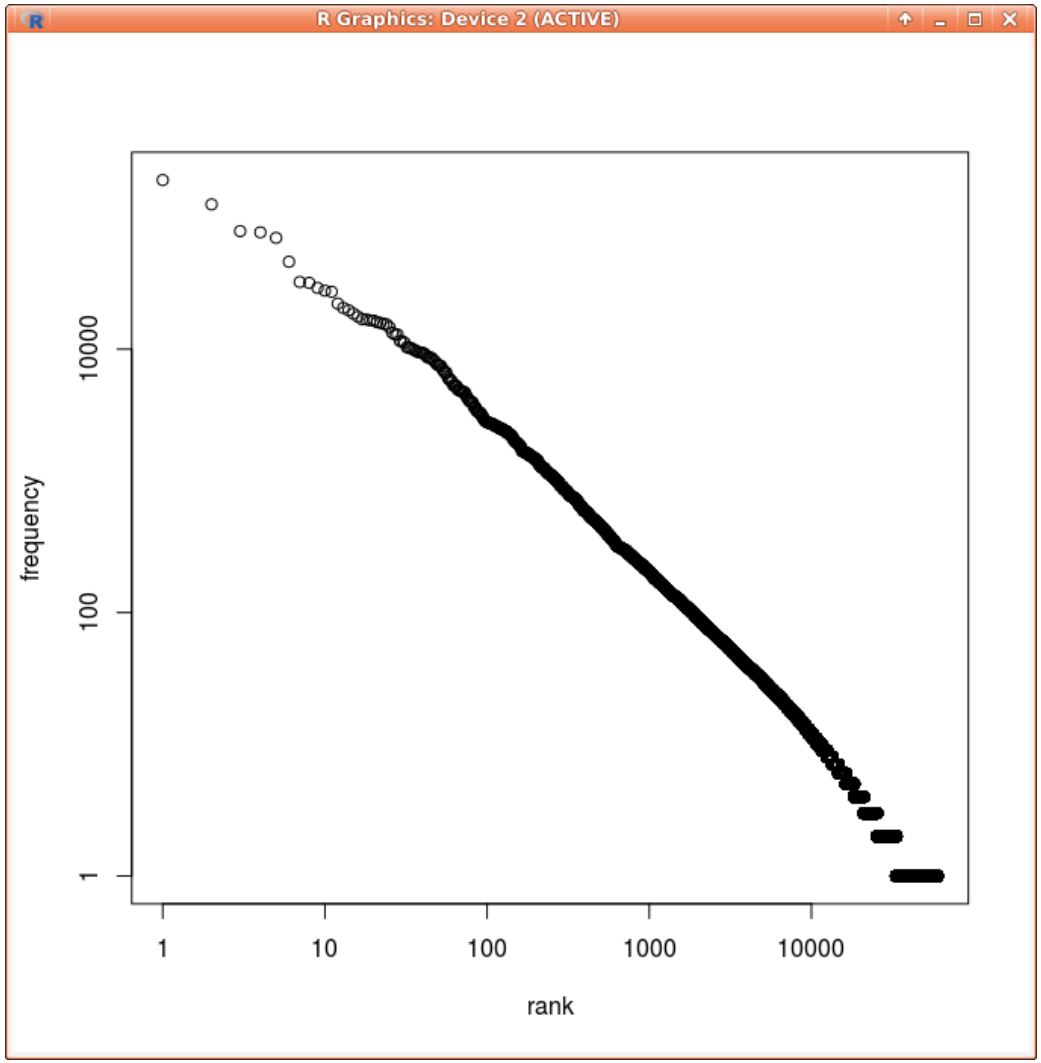
- These functions are displayed as lines when using a log-log plot





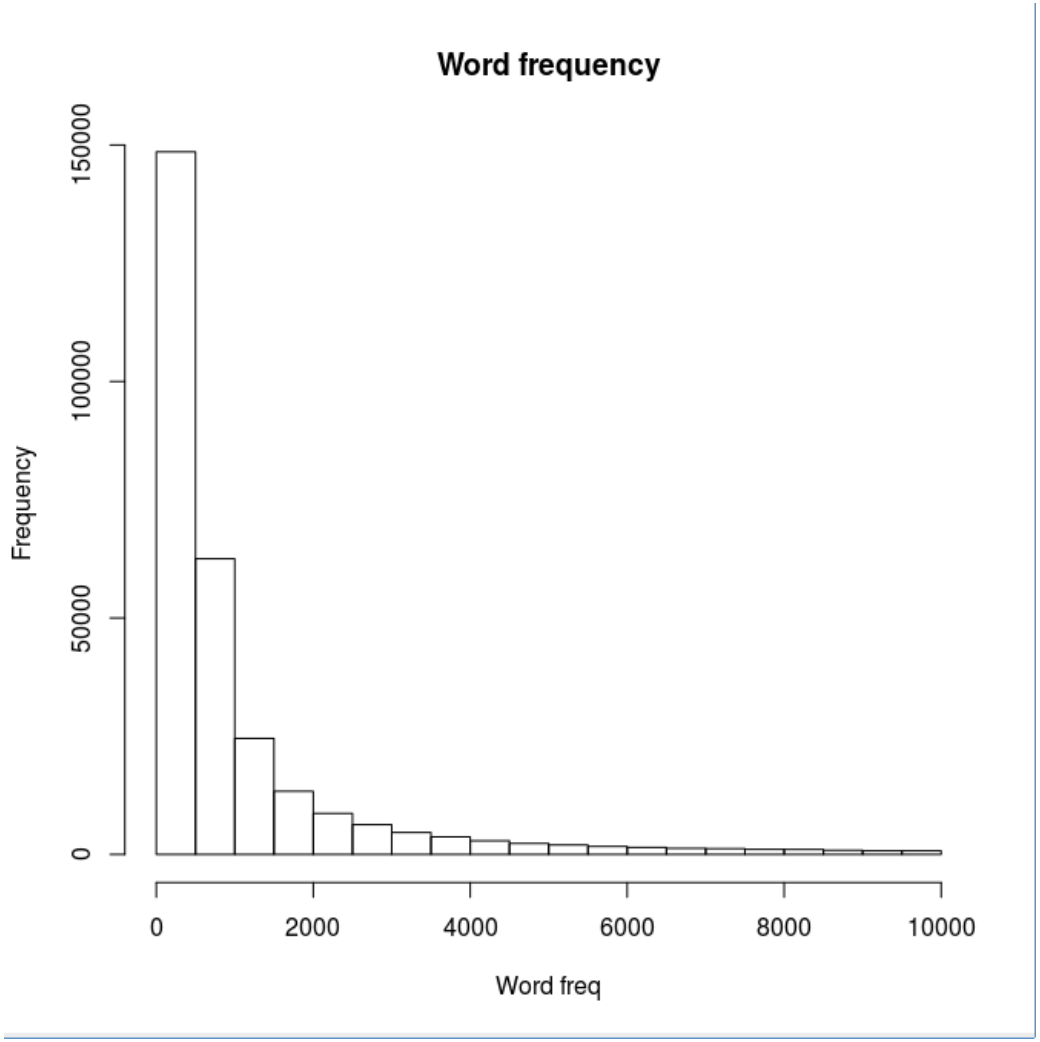
Plotting word frequencies

```
% python3 createR.py wfreq.list > wfreq.R
% R
> source("wfreq.R")
> plot(wfreq, log="xy")
```



Wikipedia word frequencies

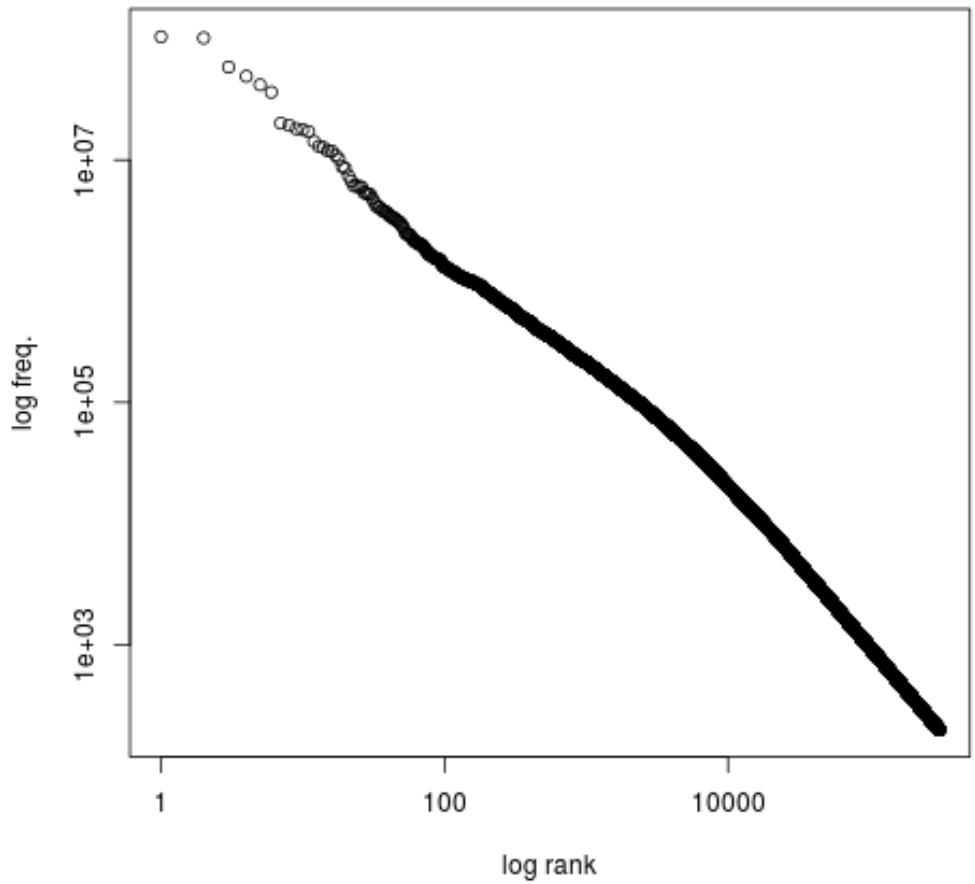
- Histogram of most frequent Wikipedia words.



Wikipedia word frequencies and Zipf's law

- Log-log plot of word frequencies

```
> load("wikipedia_cut200.RR")
> plot(wfreq, log="xy")
```



Zipf's law consequences

- Word distribution
  - few very common.
  - middling number of medium frequency.
  - many low frequency.
- Speaker and hearer effort compromise:
  - speaker effort to choose a word
    - extreme case: one word with all the meanings.
  - hearer effort in choosing meaning from word
    - extreme case: each meaning different word.
- Maximal economical compromise between these competing needs leads to Zipf law.
- Sparseness in Corpora:
  - lot of examples for few words.
  - very few examples for the majority of words.

Zipf law everywhere

Zipf law holds in many areas:

- City populations
- Corporation sizes
- Income ratings
- Social relations
  - 6 degrees of separation