THE UNIVERSITY of EDINBURGH

Text Technologies for Data Science

INFR11145

Laws of Text

Instructor:
Walid Magdy

26-Sep-2017

Reminder: Skills to be gained

- Working with large text collections
- Few shell commands
- Some Perl programming (regex)
- IR tools: Lemur / Indri / Solr
- Crawling: Web / Tweets
- TEAM WORK

Lecture Objectives

- Learn about some text laws
  - Zipf's law
  - Benford's law
  - Heap's law
  - Clumping/contagion
- Index size estimation

Try with me …

- Shell commands: cat, sort, uniq, grep
- Perl
- Excel (or alternative)
- Download the following:
  - Bible: http://www.gutenberg.org/cache/epub/10/pg10.txt
  - Unix commands for windows https://sourceforge.net/projects/unxutils

- Piazza (PLEASE) https://piazza.com/class/j766gisdu46m

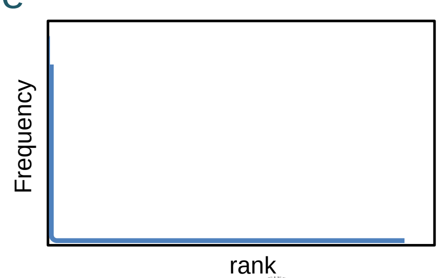
Words' nature

- Word → basic unit to represent text
- Certain characteristics are observed for the words we use!
- These characteristics are very consistent, that we can apply laws for them
- These laws apply for:
  - Different languages
  - Different domains of text

THE UNIVERSITY of EDINBURGH

Frequency of words

- Some words are very frequent
  e.g. "the", "of", "to"
- Many words are not that frequent
  e.g. "schizophrenia", "covfefe"
- ~50% terms appears once
- Frequency of words has hard exponential decay



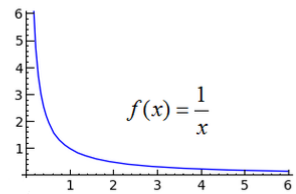
THE UNIVERSITY of EDINBURGH

Zipf's Law:

- For a given collection of text, ranking unique terms according to their frequency, then:

$$r \times P_r \cong const$$

- r, rank of term according to frequency
- P_r, probability of appearance of term
- $P_r \cong \frac{const}{r} \rightarrow f(x) \cong \frac{1}{x}$



Zipf's Law:

Wikipedia abstracts

→ 3.5M En abstracts

$r \times P_r \cong const \rightarrow$

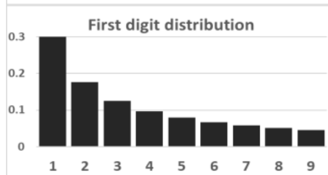
$r \times freq_r \cong const$

Term	Rank	Frequency	r x freq
the	1	5,134,790	5,134,790
of	2	3,102,474	6,204,948
in	3	2,607,875	7,823,625
a	4	2,492,328	9,969,312
is	5	2,181,502	10,907,510
and	6	1,962,326	11,773,956
was	7	1,159,088	8,113,616
to	8	1,088,396	8,707,168
by	9	766,656	6,899,904
an	10	566,970	5,669,700
it	11	557,492	6,132,412
for	13	493,374	5,970,456
as	14	480,277	6,413,862
on	15	471,544	6,723,878
from	16	412,785	7,073,160

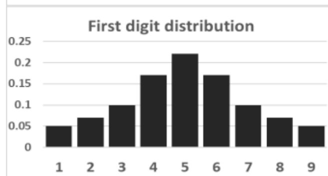
Distribution of first digit in frequencies?

1) Uniform →

2) Exp decay →



3) Normal →



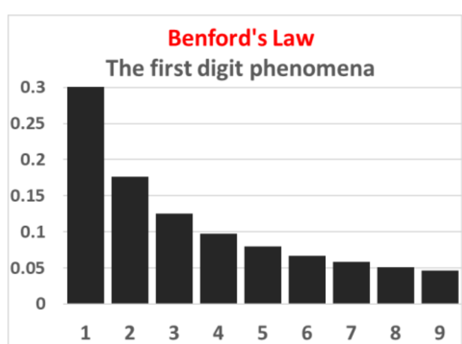
Term	Rank	Frequency
the	1	5,134,790
of	2	3,102,474
in	3	2,607,875
a	4	2,492,328
is	5	2,181,502
and	6	1,962,326
was	7	1,159,088
to	8	1,088,396
by	9	766,656
an	10	566,970
it	11	557,492
for	13	493,374
as	14	480,277
on	15	471,544
from	16	412,785

THE UNIVERSITY of EDINBURGH

Benford's Law:

- First digit of a number follows a Zipf's like law!
  - Terms frequencies
  - Physical constants
  - Energy bills
  - Population numbers
- Beford's law:

$$P(d) = \log(1 + \frac{1}{d})$$



THE UNIVERSITY of EDINBURGH

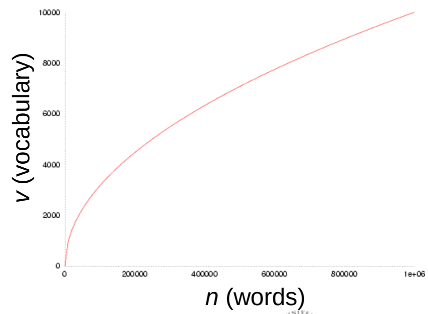
Heap's Law:

- While going through documents, the number of new terms noticed will reduce over time
- For a book/collection, while reading through, record:
  - n: number of words read
  - v: number of news words (unique words)
- Vocabulary growth:

$$v(n) = k \times n^b$$

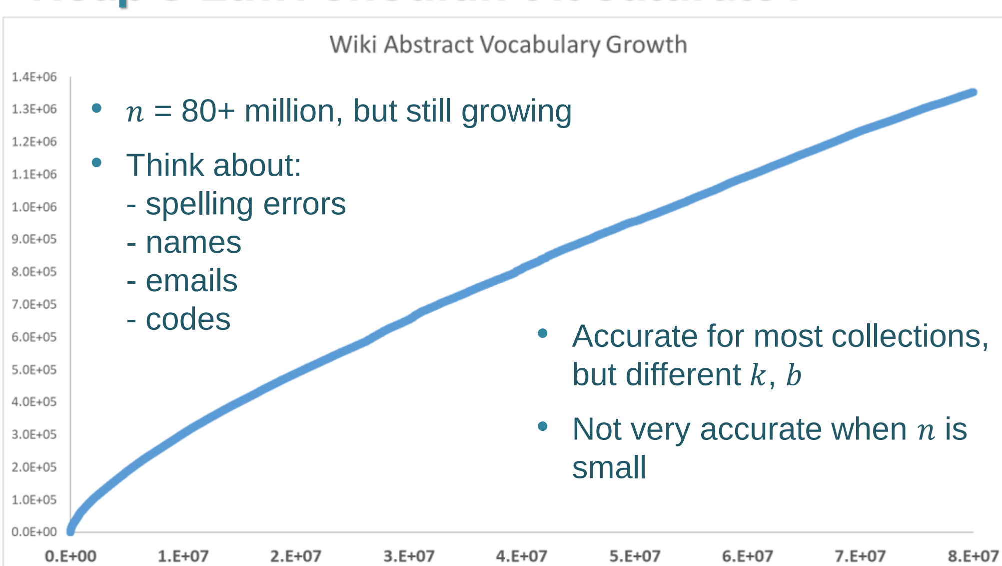
where, $b < 1$

typically, $0.4 < b < 0.7$



THE UNIVERSITY of EDINBURGH

Heap's Law: shouldn't it saturate?



- n = 80+ million, but still growing
- Think about:
  - spelling errors
  - names
  - emails
  - codes
- Accurate for most collections, but different k, b
- Not very accurate when n is small

THE UNIVERSITY of EDINBURGH

Clumping/Contagion in text

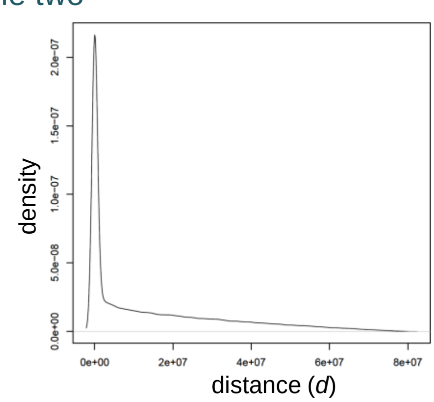
- From Zipf's law, we notice:
  - Most words do not appear that much!
  - Once you a word once → expect to see again!
  - Words are like:
    - Rare contagious disease
    - Not, rare independent lightening

- Words are rare events, but they are contagious

Clumping/Contagion in text

- Wiki abstract collection
  - Identify terms appeared only twice
  - Measure distance between the two occurrences of the terms:
    $d = n_{occurence2} - n_{occurence1}$
  - Plot density function of d

- Majority of terms appearing only twice appear close to each other.



Applying the laws

- Given a collection of 20 billion terms,
- What is the number of unique terms?

Heap's law: $v(n) = k \times n^b$, assume $k = 0.25, b = 0.5$
➔ $v(n) = 0.25 \times (20B)^{0.5} \cong 35M$

- What is the number of terms appearing once?

Zipf's law ➔ ~17M appeared only once

Estimating Index size

- How many pages Google have in index?
- Assume two independent words: t_1, t_2
- Search for $\{t_1\}$, $\{t_2\}$, $\{t_1,t_2\}$, and report number of results n_1, n_2, $n_{1,2}$
- $P(t_1) = \frac{n_1}{N}, P(t_2) = \frac{n_2}{N}, P(t_1, t_2) = \frac{n_{1,2}}{N}$

  but, t_1, t_2 independent → $P(t_1, t_2) = P(t_1).P(t_2)$

  → $\frac{n_{1,2}}{N} = \frac{n_1}{N}.\frac{n_2}{N}$ ➔ $N = \frac{n_1 n_2}{n_{1,2}}$
- Repeat for different t_1 and t_2, and estimate N

* It worth noting that observed n's are estimated as well

Testing on Google

t_1	t_2	n_1	n_2	$n_{1,2}$	N
yellow	water	7.26B	4.37B	628M	50.5B
John	green	3.13B	14.44B	801M	56.4B
purple	politics	4.41B	799M	66.6M	52.9B
Irma	car	233M	5.09B	17.6M	67.3B
falafel	pencil	23.8M	480M	319K	35.9B

- Index size → 40-60 billion

* Google index size is over 60 trillion web pages

Summary

- Text Laws:
  - Zipf
  - Benford
  - Heab
  - Index size

Recourses

- Text book:
  - Search engines: IR in practice → chapter 4
- Videos:
  - Zipf's law, Vsouce: https://www.youtube.com/watch?v=fCn8zs912OE
  - Benford's law, Numberphile: https://www.youtube.com/watch?v=XXjlR2OK1kM

Next Lecture

- Getting ready for indexing?
- Pre-processing steps before the indexing process

Walid Magdy, TTDS 2017/2018

THE UNIVERSITY of EDINBURGH