THE UNIVERSITY
of EDINBURGH

Text Technologies for Data Science

INFR11145

Laws of Text

Instructor:
Youssef Al Hariri

27-Sep-2023

Pre-Lecture

- Lab 0: How did it go?
- Lab 1: this week, important to everyone
  - Try to implement directly after the lectures
  - Ask questions / Share results over Piazza
  - Only attend in-person lab next Tuesday if needed
- Join Piazza

Reminder: Skills to be gained

- Working with large text collections
- Few shell commands
- Python/Perl regex
- TEAM WORK

Lecture Objectives

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

- This lecture is practical

You can try with me …

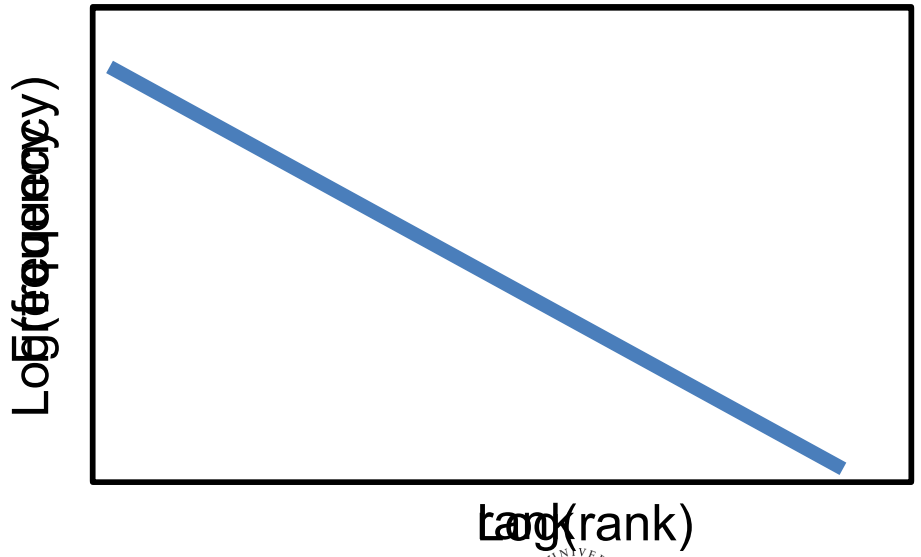
- Shell commands: cat, sort, uniq, grep
- Python (or alternative)
- Excel (or alternative)
- Download the following:
  - Bible: http://www.gutenberg.org/cache/epub/10/pg10.txt

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

Frequency of words

- Some words are very frequent
  e.g. “the”, “of”, “to”
- Many words are less frequent
  e.g. “schizophrenia”, “bazinga”
- ~50% terms appears once
- Frequency of words has hard exponential decay



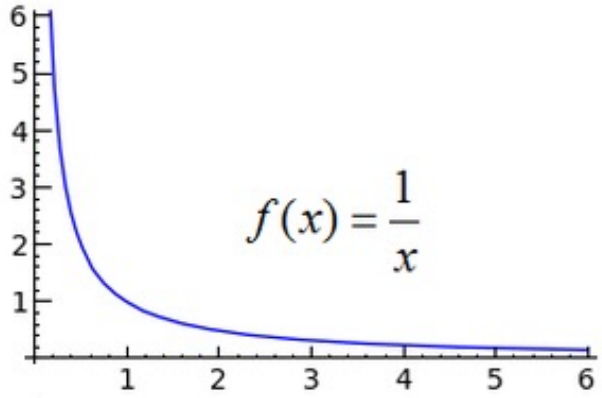
Zipf's Law:

- 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

$r \times P_r \cong const \rightarrow$

$r \times freq_r \cong const$

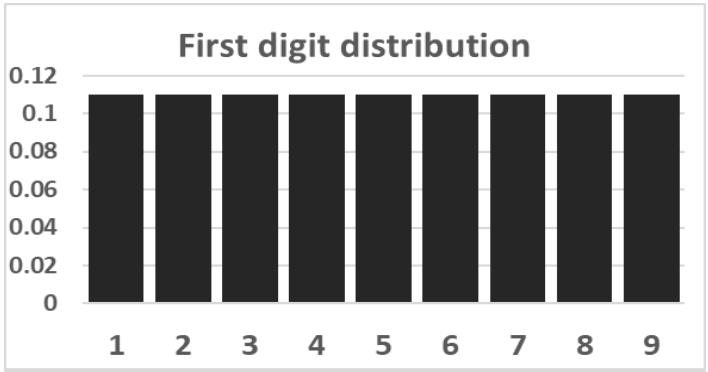
Term	Rank	Frequency	r x freq
the	1	5,134,790	5,134,790
of	2	3,102,474	6,204,948
in	3	2,607,875	7,823,625
a	4	2,492,328	9,969,312
is	5	2,181,502	10,907,510
and	6	1,962,326	11,773,956
was	7	1,159,088	8,113,616
to	8	1,088,396	8,707,168
by	9	766,656	6,899,904
an	10	566,970	5,669,700
it	11	557,492	6,132,412
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

Practical

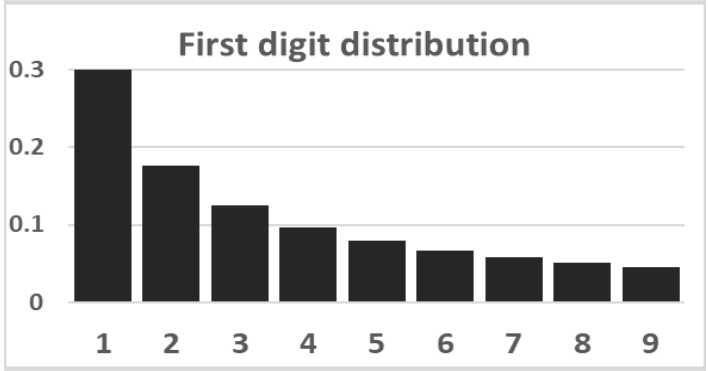
Collection	# words	File size
Bible	824,054	4.24 MB
Wiki abstracts	80,460,749	472 MB

Distribution of first digit in frequencies?

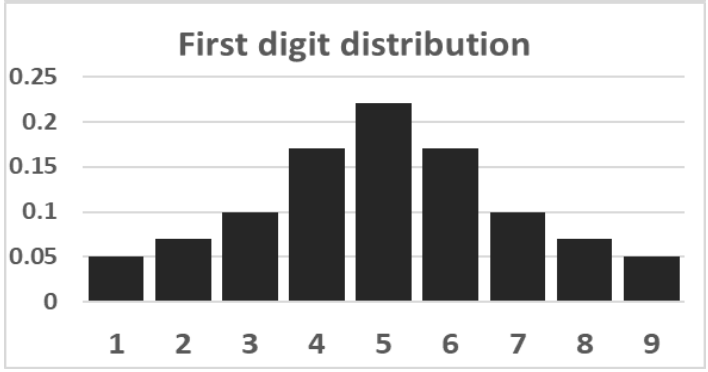
1) Uniform →



2) Exp decay →



3) Normal →



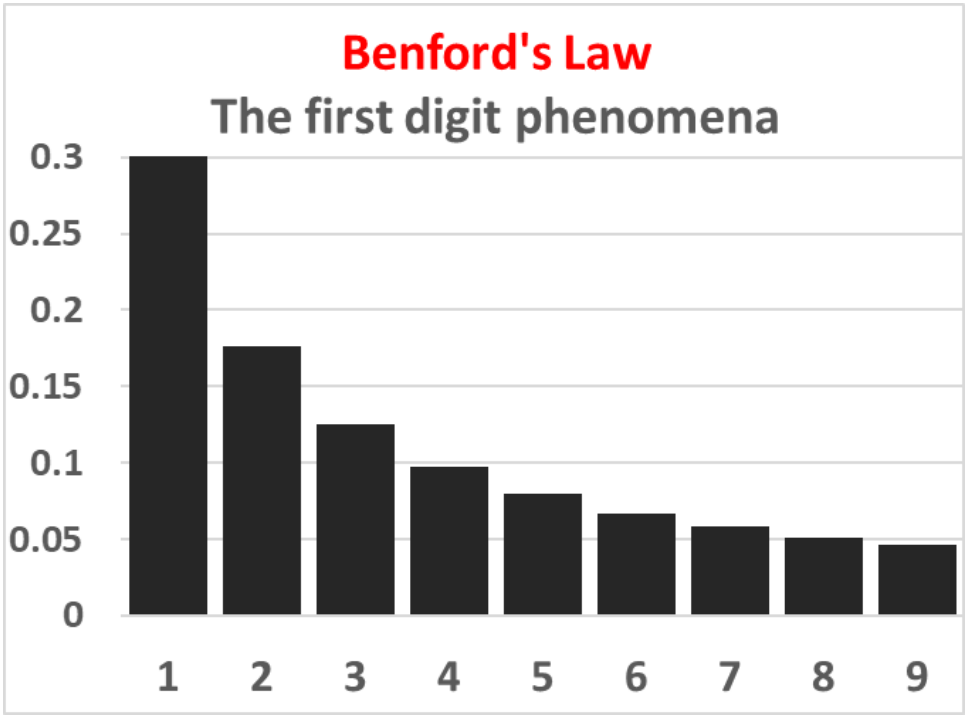
Term	Rank	Frequency
the	1	5 134,790
of	2	3 102,474
in	3	2 607,875
a	4	2 492,328
is	5	2 181,502
and	6	1 962,326
was	7	1 159,088
to	8	1 088,396
by	9	766,656
an	10	566,970
it	11	557,492
for	13	493,374
as	14	480,277
on	15	471,544
from	16	412,785

Benford's Law:

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

- Benford's law:

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



Practical

Heap's Law:

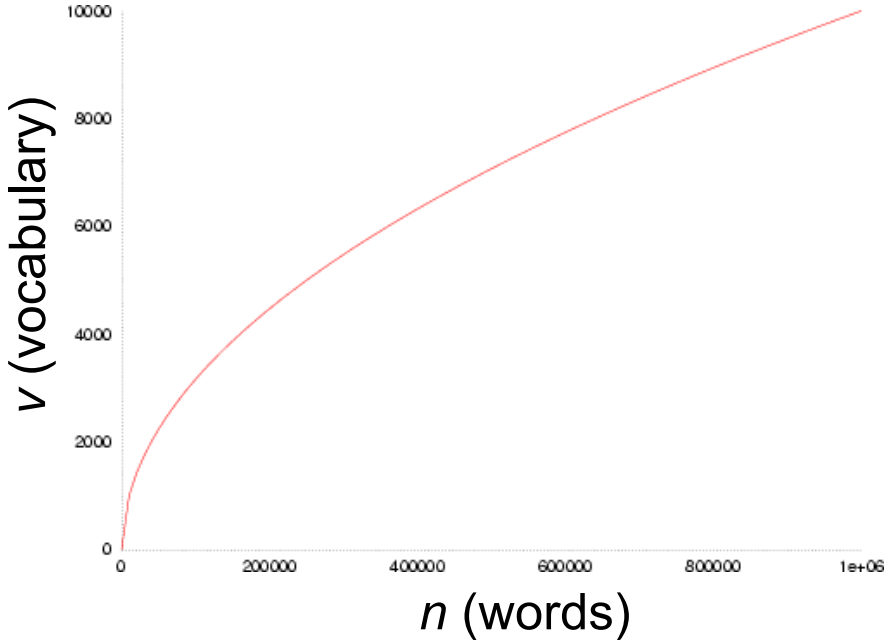
- While going through documents, the number of new terms noticed will reduce over time
- For a book/collection, while reading through, record:
  - n: number of words read
  - v: number of news words (unique words)

- Vocabulary growth:

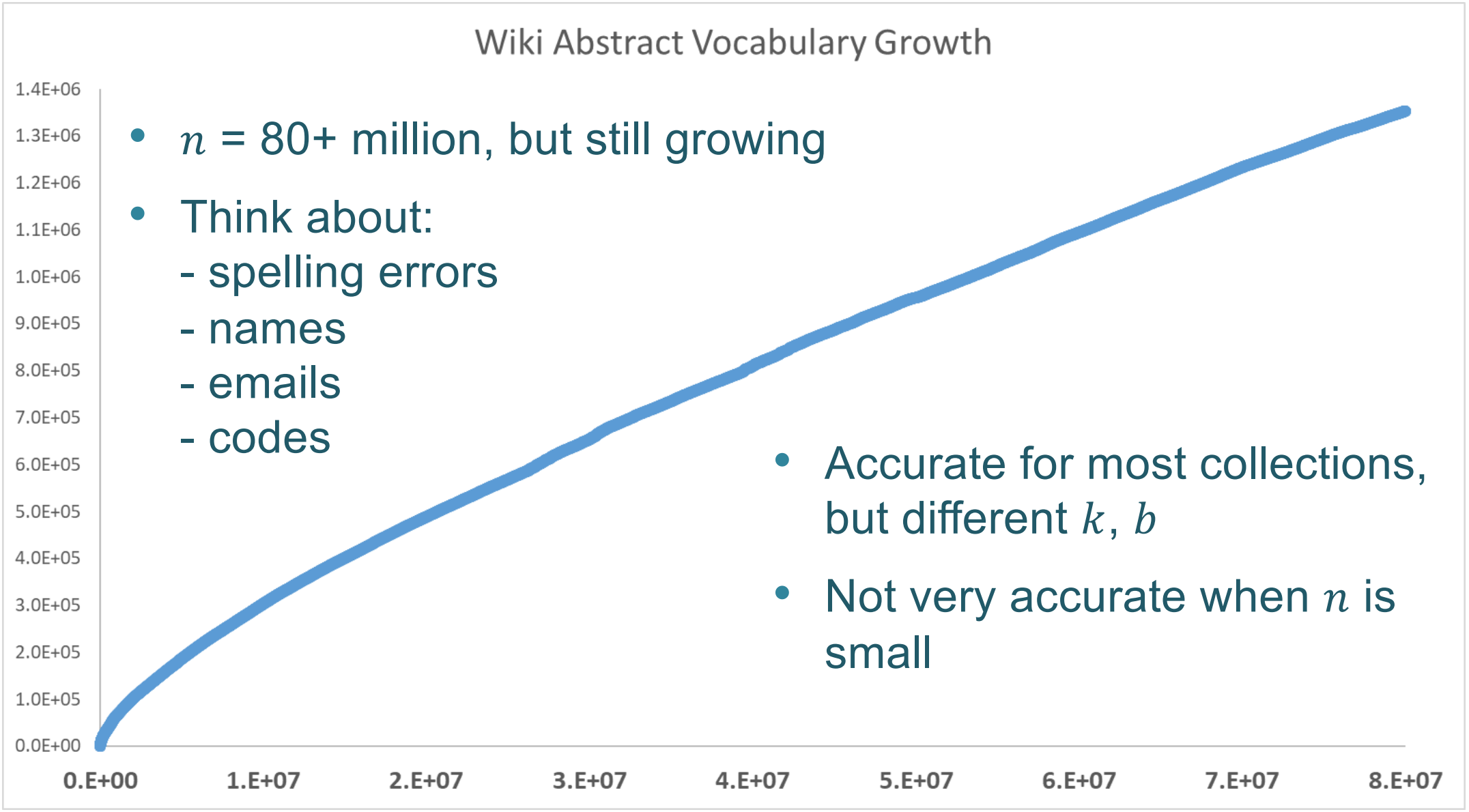
$$v(n) = k \times n^b$$

where, $b < 1$

typically, $0.4 < b < 0.7$



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

Practical

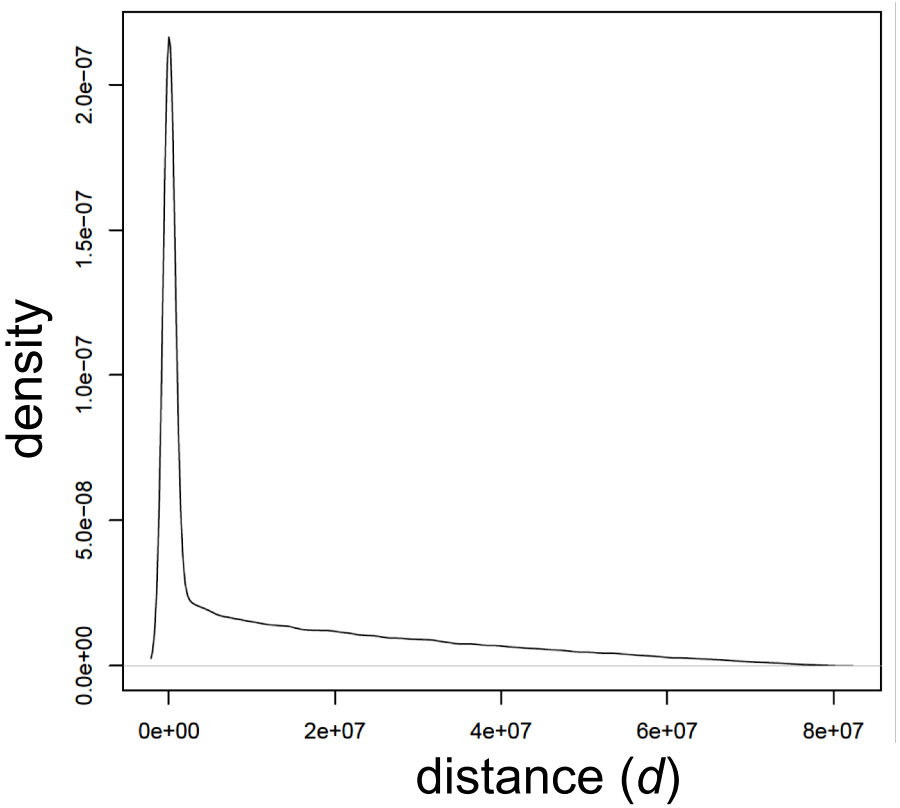
Clumping/Contagion in text

- From Zipf's law, we notice:
  - Most words do not appear that much!
  - Once you see 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

Summary

- Text follows well-known phenomena
- Text Laws:
  - Zipf
  - Heap
  - Benford
  - Contagion in text
- Try it on another language …

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
- Tools:
  - Unix commands for windows https://sourceforge.net/projects/unxutils

Next Lecture

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

- Reminder: 5-10 mins break after L1
  - Have a break, stretch, get food … etc.
  - Ask questions on chat
  - Questions on L1 are allowed before starting L2
  - Mind teaser math problem (for fun)