CS565: Intelligent Systems and Interfaces

Words

Semester: Jan – May 2019

Ashish Anand

Associate Professor, Dept of CSE

IIT Guwahati

Announcements

- Scribe for Next two lectures
  - Ayush Jaiswal, Saloni Rathi, Sayantan Basu: 22nd Lec
  - Vaibhav Pandey, Dhananjay, Susrita: 23rd Lec
- Extra Class on this Thursday, 24th Jan at 2 PM

Recap

- Essential Resources and basic pre-processing
  - Corpora
  - Word and Sentence Segmentation – focus on heuristics and issues associated with them

Objective

- Word
  - Basic statistics and inference: Zipf’s law
  - Collocation

Word

Basic Questions

- What is the length of the corpus?
- How many distinct words are used?
- What are the most common words?

Terminology

- Word Tokens: individual occurrences of words
- Word Types: distinct word tokens

Answering the basic questions and making some inference

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Table 1.1 Common words in *Tom Sawyer*.

- Corpus: Tom Sawyer by Mark Twain
- Basic Statistics:
  - 71,370 word tokens
  - 8,018 word types
- Observation: Domination of function words (determiner, prepositions etc.)
- Function words vs. Content Words
- Stop words: https://code.google.com/archive/p/stop-words/

Uneven distribution with long tail phenomena

Word Frequency	Frequency of Frequency
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11–50	540
51–100	99
> 100	102

Table 1.2 Frequency of frequencies of word types in *Tom Sawyer*.

- Some words are very common
  - Individual word type contributed 1% of all word tokens [12 such words]
- Vast majority of the words occurred very infrequently
  - Over 90% of the word types occur 10 times or less
- Many rare words
  - 12% of the text occurred 3 times or less

Empirical observation leading to Zipf's Law

Word	Freq. (f)	Rank (r)	$f \cdot r$	Word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10440	Could	2	4000	8000
two	104	100	10400	Applausive	1	8000	8000

Table 1.3 Empirical evaluation of Zipf's law on *Tom Sawyer*.

- Establish the relationship between frequency f of word type and its rank r based on frequency
  - $f \propto \frac{1}{r}$
- Good description of frequency distribution of words in natural languages
- Principle of Least Effort

Zipf's Law: Bad fit for low and high ranks

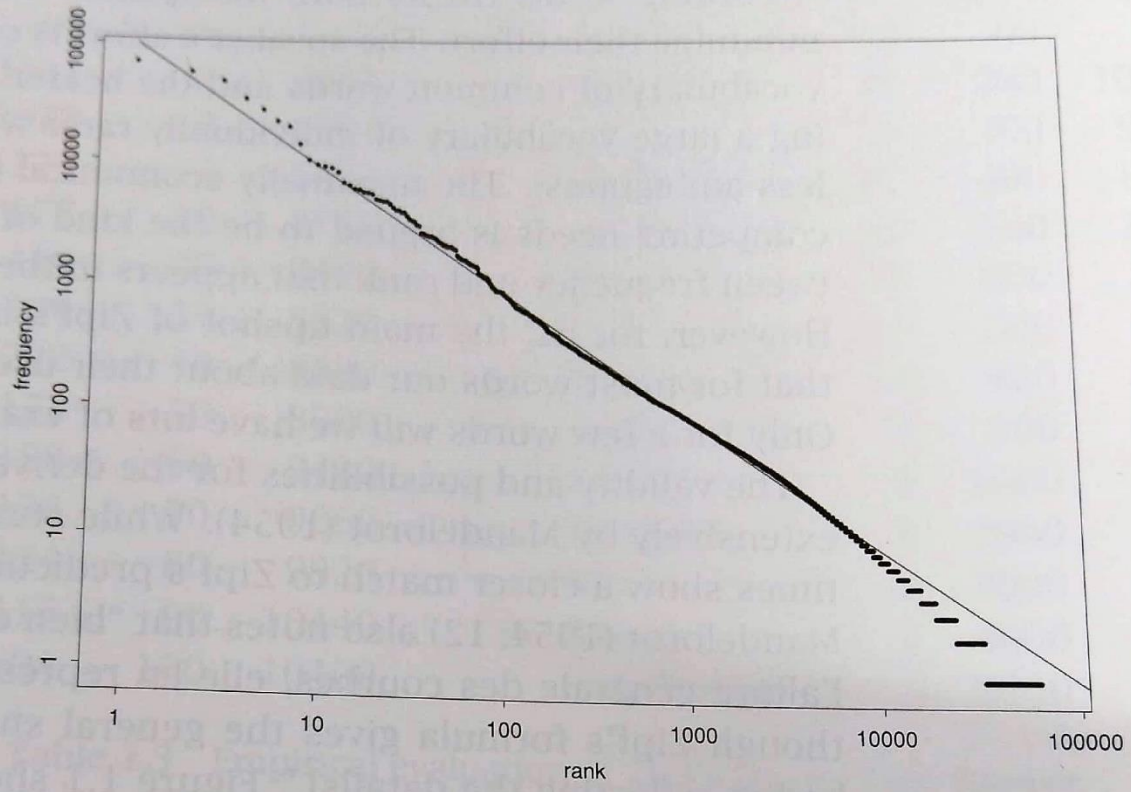


Figure 1.1 Zipf's law. The graph shows rank on the X-axis versus frequency on the Y-axis, using logarithmic scales. The points correspond to the ranks and frequencies of the words in one corpus (the Brown corpus). The line is the relationship between rank and frequency predicted by Zipf for $k = 100,000$, that is $f \times r = 100,000$.

Mandelbrot's Formula: More general relationship

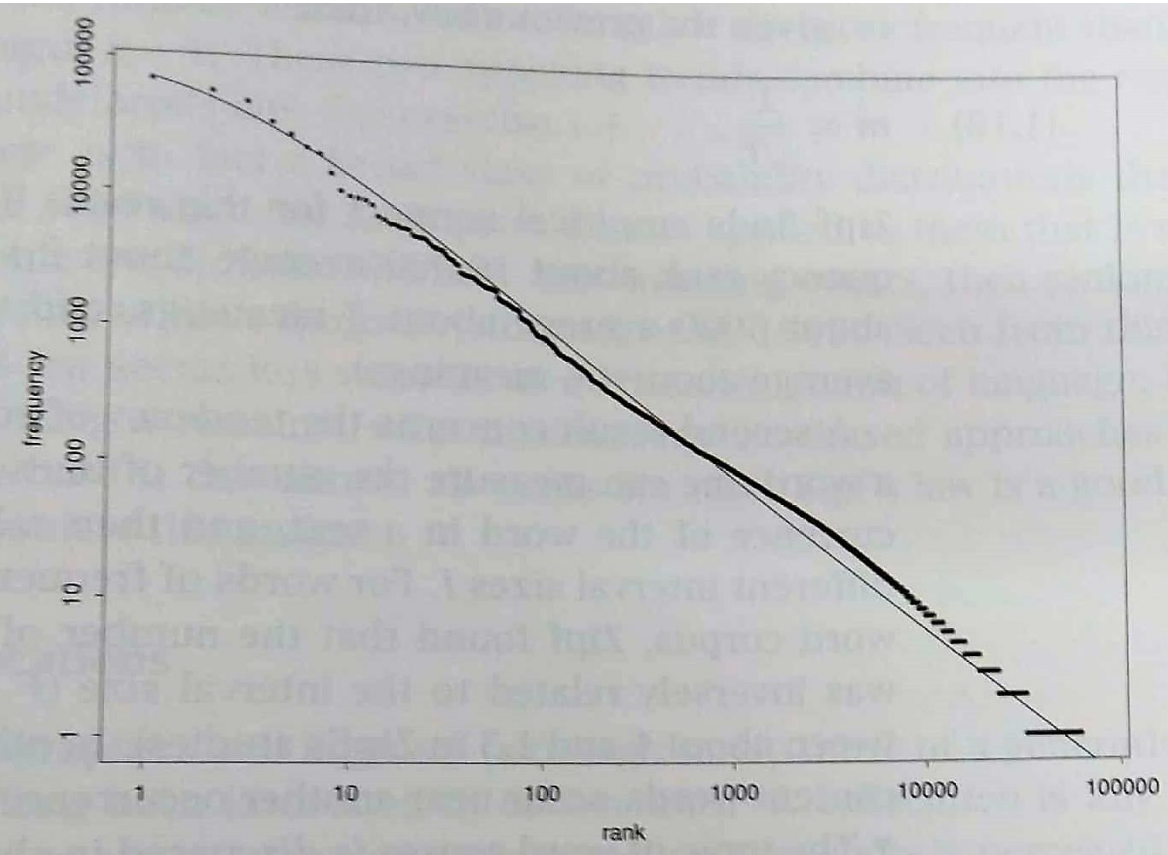


Figure 1.2 Mandelbrot's formula. The graph shows rank on the X-axis versus frequency on the Y-axis, using logarithmic scales. The points correspond to the ranks and frequencies of the words in one corpus (the Brown corpus). The line is the relationship between rank and frequency predicted by Mandelbrot's formula for $P = 10^{5.4}$, $B = 1.15$, $\rho = 100$.

$$f = P(r + \rho)^{-B}$$

$$\log f = \log P - B \log(r + \rho)$$

P, B, ρ are text parameters, collectively measure the richness of text's use of words.

Collocation: Whole is bigger than the sum of parts

Collocations: Examples

Strong Tea, Stiff breeze, Take a risk, Start up, New Delhi, Fly High

Vs

Last class, Next lecture, New companies

Collocations: Definition

- [Choueka, 1988]: "*A sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact, unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components*"

- Limitation
  - We may do away with the requirement of words being consecutive.

- Example
  - *Knocked* on the *door*
  - *Knocked* at the class-room *door*

Characteristics: subtle and not easily explainable

- *“Strong tea”* but not “*Powerful tea”*
- *“Stiff breeze”* but not *“Stiff wind”*
- *“White wine”* but not *“Yellow wine”*
- *“Broad daylight”* but not *“Bright daylight”*

Characteristics

- Limited compositionality
  - Example: Strong Tea
  - Example: White wine, white woman and white hair all refer to different colors and not exactly the white color.

- Non substitutability
  - Example: *yellow* cannot replace *white* in "*white wine*".

- Non-modifiability: can't be modified using additional lexical materials or through grammatical transformations.
  - Example: *people as poor as church mice; to get an ugly frog in one's throat.*

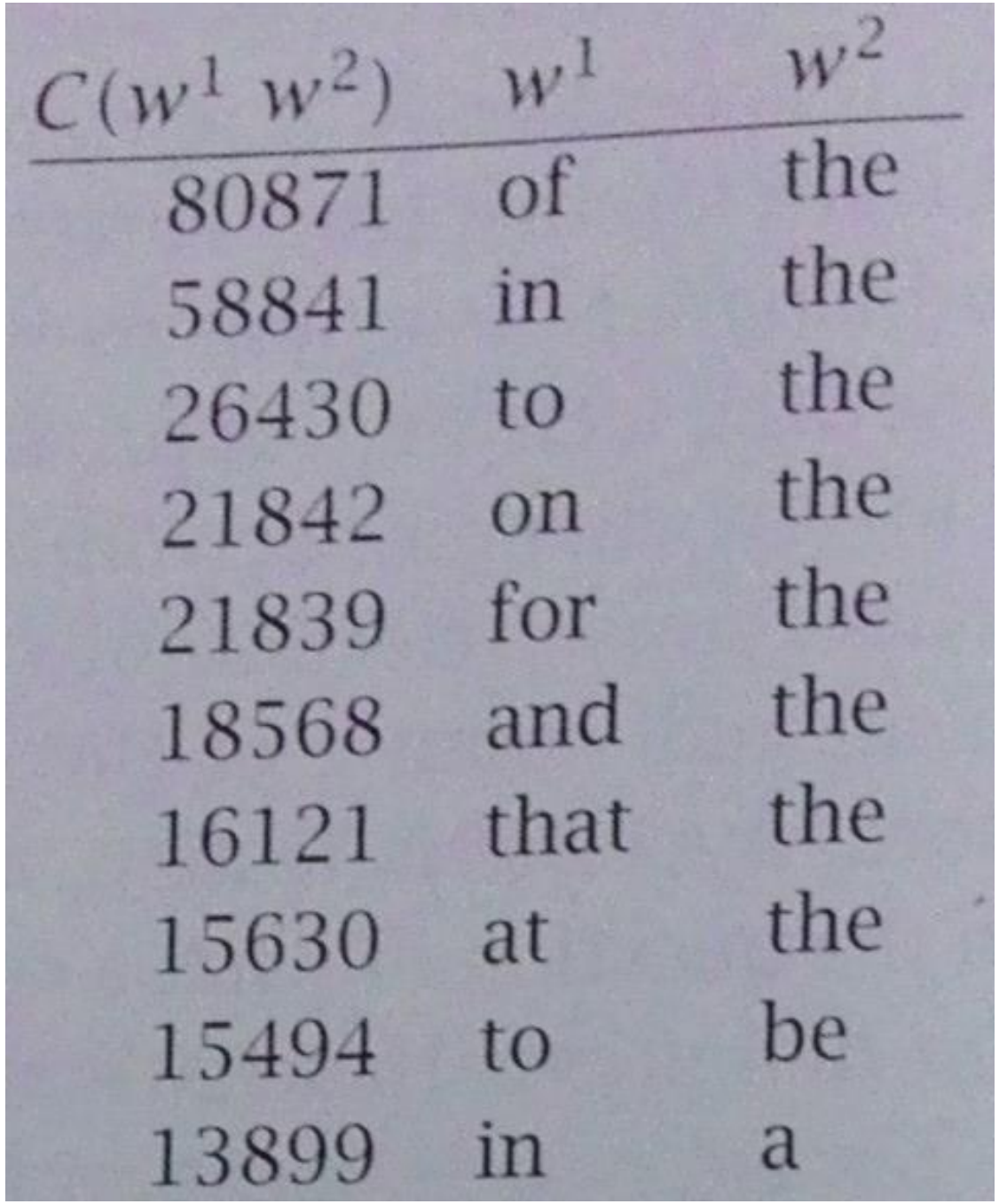
Why it is important?

- Computational lexicography
- Parsing
- Semantics
- Natural Language Generation
- Machine Translation
- Linguistic research

Finding Collocations

Frequency

- Assumption: More frequent occurrence of two words together may imply special function or property which can’t be simply explained



$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a

Frequency based methods for finding collocations

Source: Table 5.1[FSNLP: Page 154]

Corpus: New York Times newswire-Aug to Nov 1990.
Statistics: 115 MB text with roughly 14 million words

Adding linguistic knowledge to Frequency

Tag Pattern
A N
N N
A A N
A N N
N A N
N N N
N P N

1. Part of Speech (PoS) tag patterns for collocation filtering.
2. Patterns were proposed by *Justeson and Katz* (1995).
3. [*A*]djective; [*N*]oun; [*P*]reposition

Source: Table 5.2 [FSNLP: 154]

$C(w^1\ w^2)$	w^1	w^2	Tag Pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

Table 5.3 Finding Collocations: Justeson and Katz' part-of-speech filter.

Source: Table 5.3 [FSNLP: Page 155]

Pros and Cons of Frequency+PoS Filter

- Advantages
  - Simple method

- Disadvantages
  - Too much dependency on hand-designed filter
  - High frequency can be random without any specific meaning
  - Works well for fixed phrases but will not work for cases where variable number of words may exist between two words
  - Example
    - She *knocked* on his *door*
    - They *knocked* at the *door*
    - 100 women *knocked* on Donaldson's *door*
    - a man *knocked* on the metal front *door*

Sliding window could be savior

Sentence:

man knocked on the front door

Bigrams:

man knocked	*man on*	*man the*	*man front*	
	knocked on	*knocked the*	*knocked front*	*knocked door*
		on the	*on front*	*on door*
			the front	*the door*
				front door

Four word collocational window to capture bigrams at a distance

References

- Section 1.4 – 1.4.4 [FSNLP]
- Chapter 5 [FSNLP]