Content Analysis & Stemming

Yaşar Tonta

Hacettepe Üniversitesi

tonta@hacettepe.edu.tr

yunus.hacettepe.edu.tr/~tonta/

DOK324/BBY220 Bilgi Erişim İlkeleri

Note: Slides are taken from Prof. Ray Larson's web site (www.sims.berkeley.edu/~ray/

Content Analysis

- Automated Transformation of raw text into a form that represent some aspect(s) of its meaning
- Including, but not limited to:
  - Automated Thesaurus Generation
  - Phrase Detection
  - Categorization
  - Clustering
  - Summarization

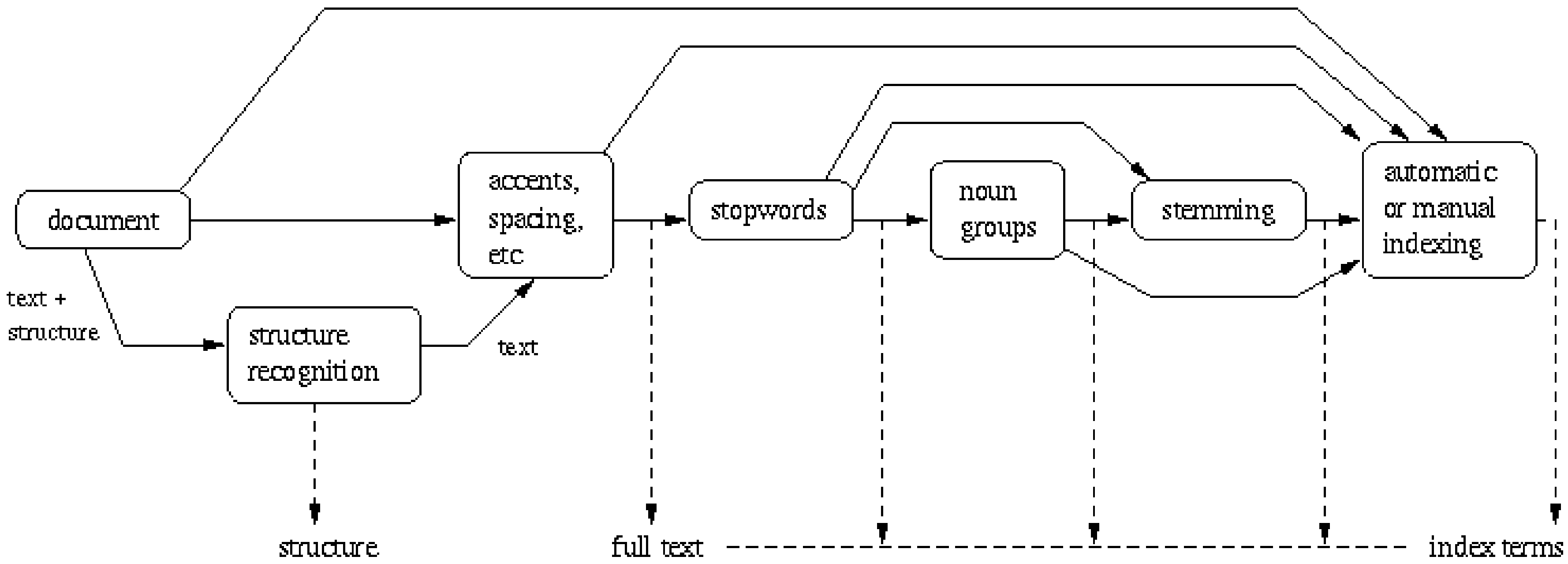
Techniques for Content Analysis

- Statistical
  - Single Document
  - Full Collection
- Linguistic
  - Syntactic
  - Semantic
  - Pragmatic
- Knowledge-Based (Artificial Intelligence)
- Hybrid (Combinations)

Text Processing

- Standard Steps:
  - Recognize document structure
    - titles, sections, paragraphs, etc.
  - Break into tokens
    - usually space and punctuation delineated
    - special issues with Asian languages
  - Stemming/morphological analysis
  - Store in inverted index (to be discussed later)

Document Processing Steps

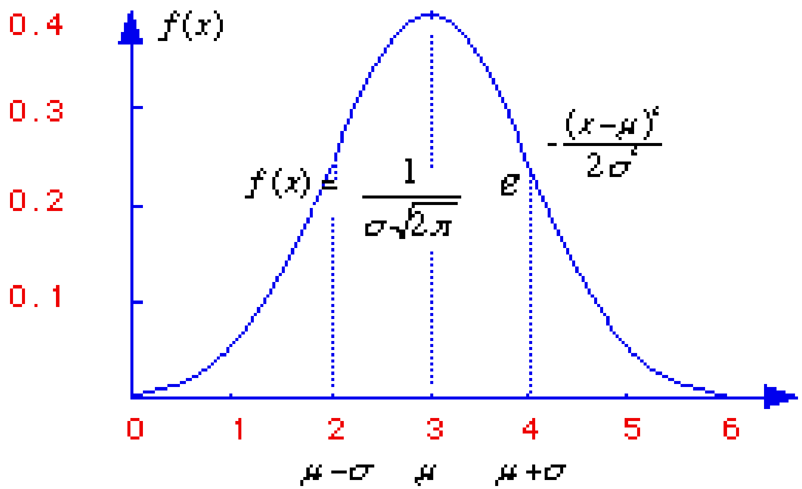


Stemming and

- Goal: “normalize” similar words
- Morphology (“form” of words)
  - Inflectional Morphology
    - E.g,. inflect verb endings and noun number
    - Never change grammatical class
      - *dog, dogs*
      - *tengo, tienes, tiene, tenemos, tienen*
  - Derivational Morphology
    - Derive one word from another,
    - Often change grammatical class
      - *build, building; health, healthy*

Statistical Properties of Text

- Token occurrences in text are not uniformly distributed
- They are also not normally distributed



- They do exhibit a Zipf distribution

Plotting Word Frequency by Rank

- Main idea: count
  - How many tokens occur 1 time
  - How many tokens occur 2 times
  - How many tokens occur 3 times …
- Now rank these according to how of they occur. This is called the rank.

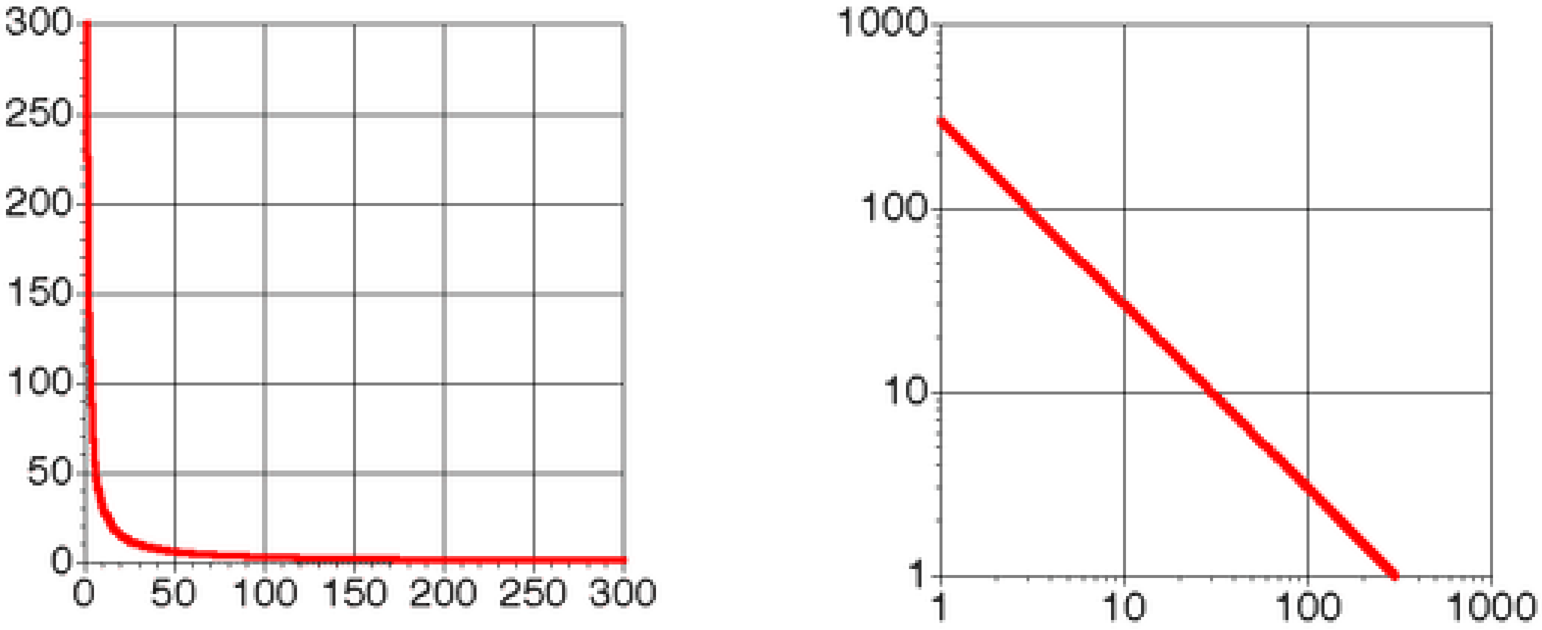
Plotting Word Frequency by Rank

- Say for a text with 100 tokens
- Count
  - How many tokens occur 1 time (50)
  - How many tokens occur 2 times (20) …
  - How many tokens occur 7 times (10) …
  - How many tokens occur 12 times (1)
  - How many tokens occur 14 times (1)
- So things that occur the most often share the highest rank (rank 1).
- Things that occur the fewest times have the lowest rank (rank n).

Observation: MANY phenomena can be characterized this way.

- Words in a text collection
- Library book checkout patterns
- Bradford's and Lotka's laws.
- Incoming Web Page Requests (Nielsen)
- Outgoing Web Page Requests (Cunha & Crovella)
- Document Size on Web (Cunha & Crovella)

Zipf Distribution (linear and log scale)



Zipf Distribution

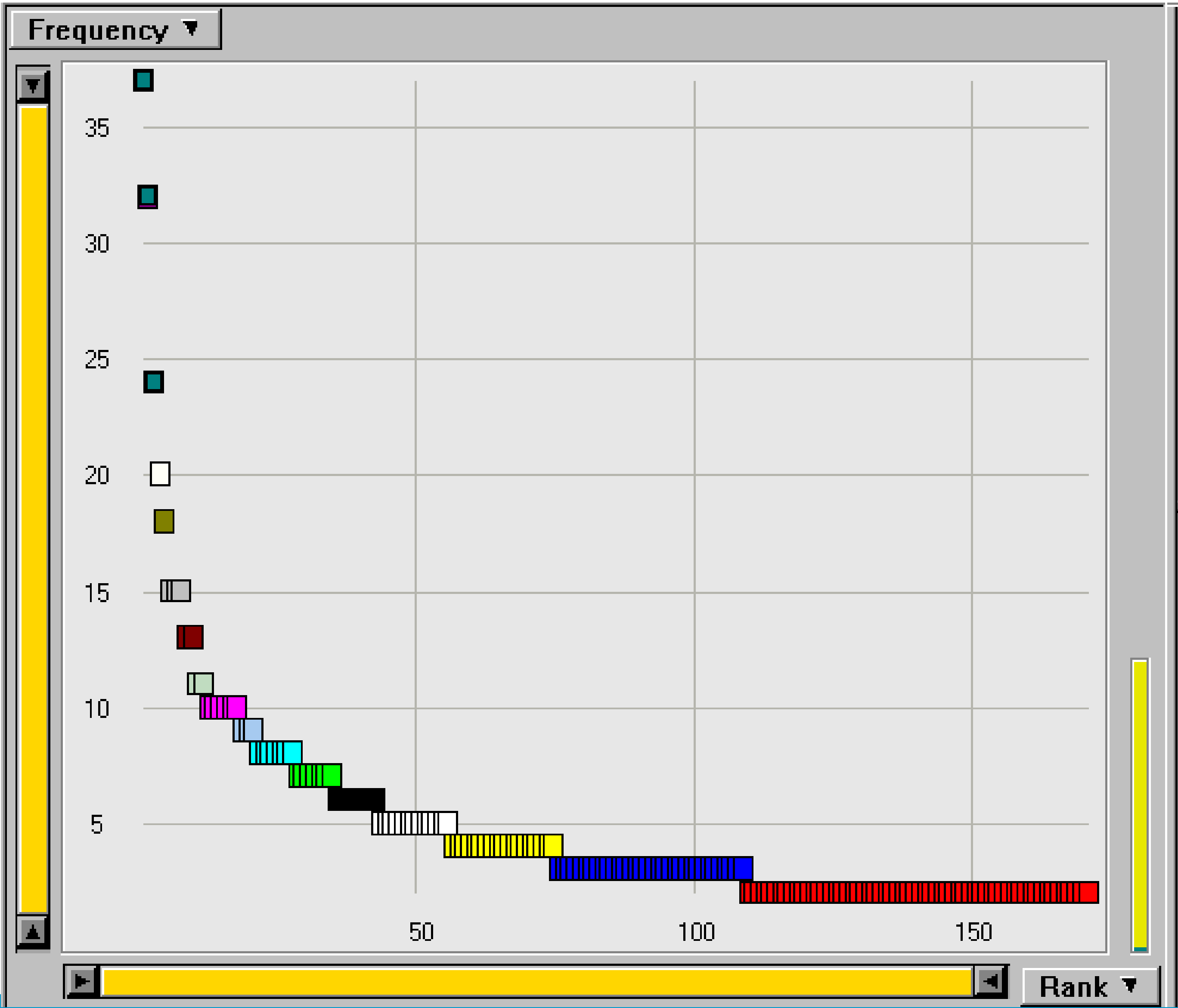
- The product of the frequency of words (f) and their rank (r) is approximately constant
  - Rank = order of words' frequency of occurrence

$$f = C * 1/r$$
$$C \cong N/10$$

- Another way to state this is with an approximately correct rule of thumb:
  - Say the most common term occurs C times
  - The second most common occurs C/2 times
  - The third most common occurs C/3 times
  - …

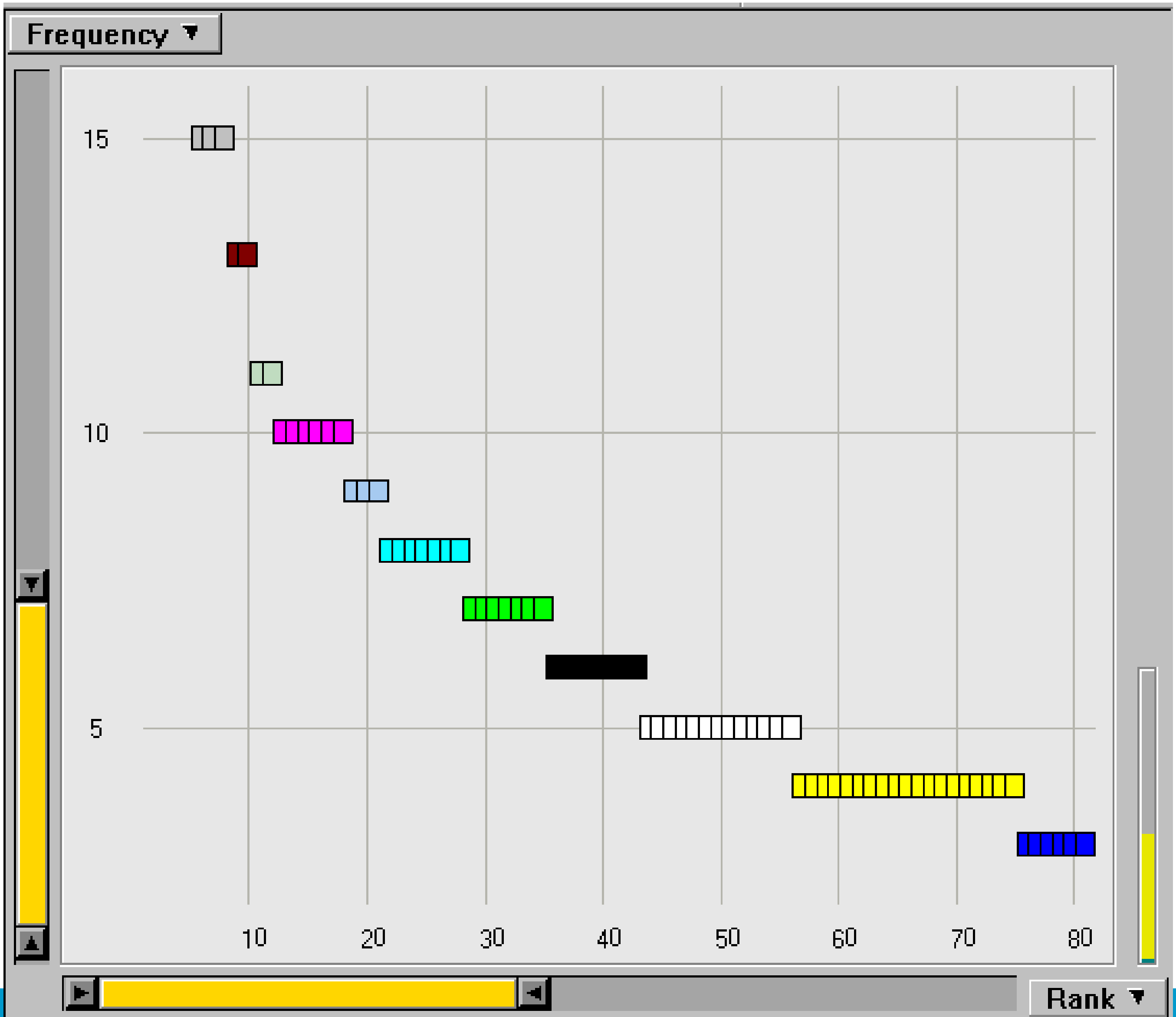
The Corresponding Zipf Curve

Rank	Freq	
1	37	system
2	32	knowledg
3	24	base
4	20	problem
5	18	abstract
6	15	model
7	15	languag
8	15	implem
9	13	reason
10	13	inform
11	11	expert
12	11	analysi
13	10	rule
14	10	program
15	10	oper
16	10	evalu
17	10	comput
18	10	case
19	9	gener
20	9	form



Zoom in on the Knee of the Curve

43	6	approach
44	5	work
45	5	variabl
46	5	theori
47	5	specif
48	5	softwar
49	5	requir
50	5	potenti
51	5	method
52	5	mean
53	5	inher
54	5	data
55	5	commit
56	5	applic
57	4	tool
58	4	technolog
59	4	techniqu



Zipf Distribution

- The Important Points:
  - a few elements occur *very frequently*
  - a medium number of elements have medium frequency
  - many elements occur *very infrequently*

Most and Least Frequent Terms

Rank	Freq	Term
1	37	system
2	32	knowledg
3	24	base
4	20	problem
5	18	abstract
6	15	model
7	15	languag
8	15	implem
9	13	reason
10	13	inform
11	11	expert
12	11	analysi
13	10	rule
14	10	program
15	10	oper
16	10	evalu
17	10	comput
18	10	case
19	9	gener
20	9	form

150	2	enhanc
151	2	energi
152	2	emphasi
153	2	detect
154	2	desir
155	2	date
156	2	critic
157	2	content
158	2	consider
159	2	concern
160	2	compon
161	2	compar
162	2	commerci
163	2	clause
164	2	aspect
165	2	area
166	2	aim
167	2	affect

A Standard Collection

Government documents, 157734 tokens, 32259 unique

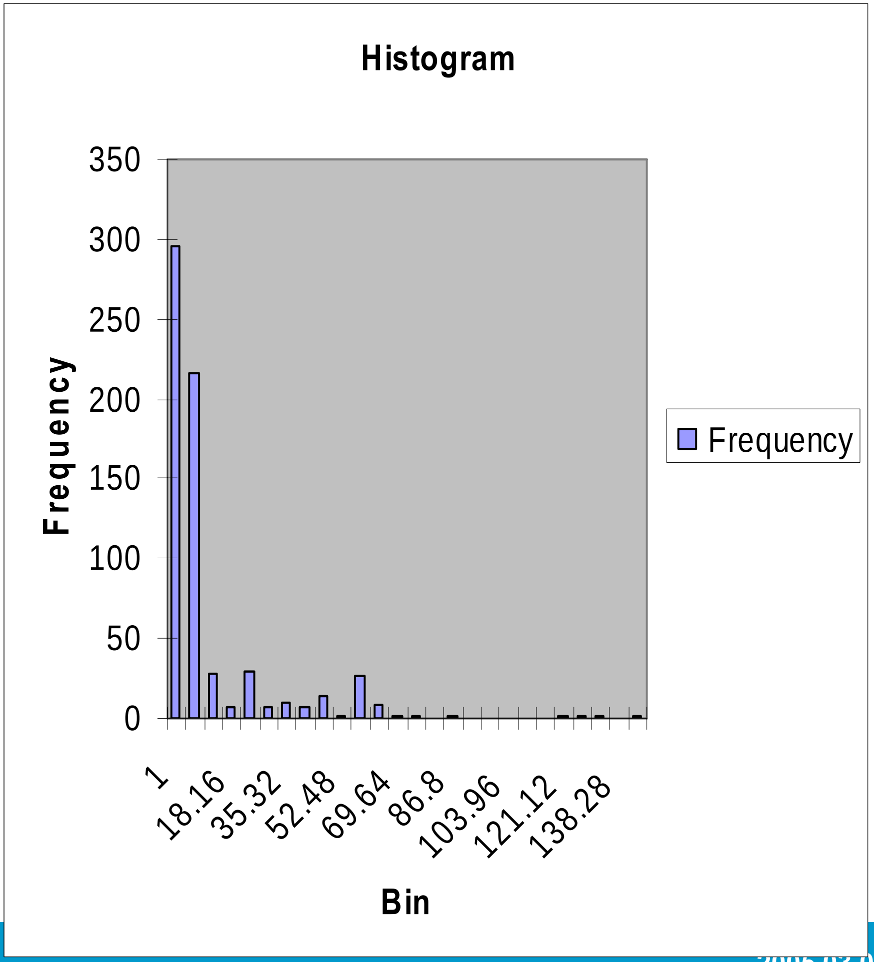
8164 the	969 on	1 ABC
4771 of	915 FT	1 ABFT
4005 to	883 Mr	1 ABOUT
2834 a	860 was	1 ACFT
2827 and	855 be	1 ACI
2802 in	849 Pounds	1 ACQUI
1592 The	798 TEXT	1 ACQUISITIONS
1370 for	798 PUB	1 ACSIS
1326 is	798 PROFILE	1 ADFT
1324 s	798 PAGE	1 ADVISERS
1194 that	798 HEADLINE	1 AE
973 by	798 DOCNO	

Housing Listing Frequency Data

Bin	*Frequency*
1	295
6.72	216
12.44	28
18.16	7
23.88	29
29.6	7
35.32	10
41.04	7
46.76	14
52.48	2
58.2	26
63.92	9
69.64	1
75.36	1
81.08	0
86.8	2
92.52	0
98.24	0
103.96	0
109.68	0
115.4	0
121.12	1
126.84	1
132.56	1
138.28	0
More	1

6208 tokens,

1318 unique (very small collection)



Very frequent word stems

WORD	FREQ
u	63245
ha	65470
california	67251
m	67903
1998	68662
system	69345
t	70014
about	70923
servic	71822
work	71958
home	72131
other	72726
research	74264
1997	75323
can	76762
next	77973
your	78489
all	79993
public	81427
us	82551
c	83250
www	87029
wa	92384
program	95260

not	100204
http	100696
d	101034
html	103698
student	104635
univers	105183
inform	106463
will	109700
new	115937
have	119428
page	128702
messag	141542
from	147440
you	162499
edu	167298
be	185162
publib	189334
librari	189347
i	190635
lib	223851
that	227311
s	234467
berkelei	245406
re	272123
web	280966
archiv	305834

Words that occur few times (Cha-Cha Web Index)

WORD	FREQ
agenda augu	1
an electronic	1
center janu	1
packard equi	1
system july	1
systems cs1	1
today mcb	1
workshops fi	1
workshops th	1
lollini	1
0+	1
0	1
00summary	1
35816	1
35823	1
01d	1
35830	1
35837	1
02-156-10	1
35844	1
35851	1
02aframst	1
311	1
313	1
03agenvchm	1
401	1
408	1

408	1
422	1
424	1
429	1
04agrcecon	1
04cklist	1
05-128-10	1
501	1
506	1
05amstud	1
06anhist	1
07-149	1
07-800-80	1
07anthro	1
08apst	1

Word Frequency vs. Resolving

The most frequent words are not the most descriptive.

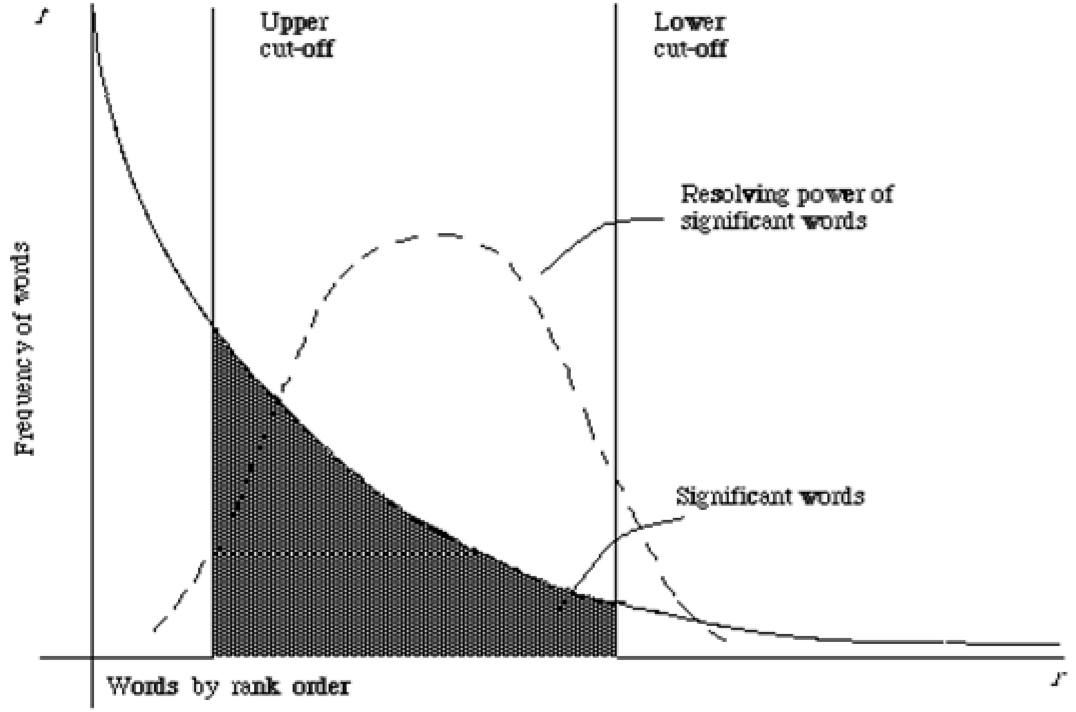


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adaped from Schultz[44] page 120)

Stemming and

- Goal: “normalize” similar words
- Morphology (“form” of words)
  - Inflectional Morphology
    - E.g,. inflect verb endings and noun number
    - Never change grammatical class
      - *dog, dogs*
      - *tengo, tienes, tiene, tenemos, tienen*
  - Derivational Morphology
    - Derive one word from another,
    - Often change grammatical class
      - *build, building; health, healthy*

Simple "S" stemming

- IF a word ends in "ies", but not "eies" or "aies"
  - THEN "ies" → "y"
- IF a word ends in "es", but not "aes", "ees", or "oes"
  - THEN "es"→ "e"
- IF a word ends in "s", but not "us" or "ss"
  - THEN "s" → NULL

Harman, JASIS 1991

Errors Generated by Porter Stemmer (Krovetz 93)

Too Aggressive	Too Timid
organization / organ	european / europe
policy / police	cylinder / cylindrical
execute / executive	create / creation
arm / army	search / searcher

Automated Methods

- Stemmers:
  - Very dumb rules work well (for English)
  - Porter Stemmer: Iteratively remove suffixes
  - Improvement: pass results through a lexicon
- Powerful multilingual tools exist for morphological analysis
  - PCKimmo, Xerox Lexical technology
  - Require a grammar and dictionary
  - Use “two-level” automata
  - Wordnet “morpher”

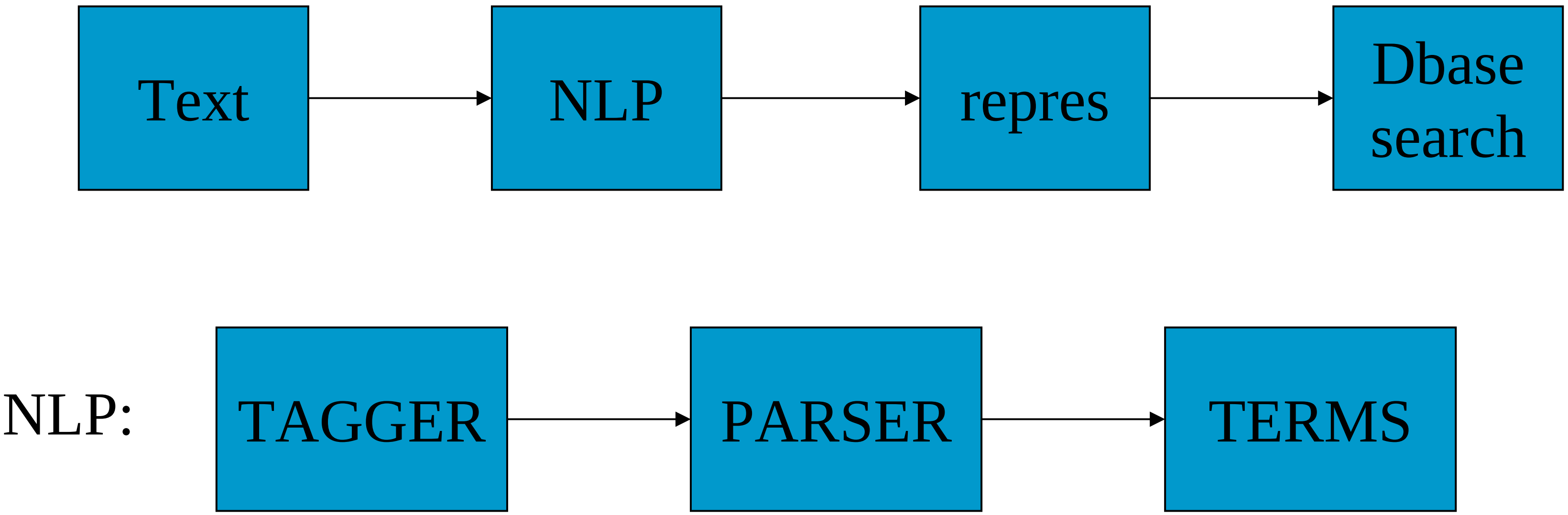
Wordnet

- Type “wn word” on irony.
- Large exception dictionary:
- Demo

aardwolves aardwolf
abaci abacus
abacuses abacus
abbacies abbacy
abhenries abhenry
abilities ability
abkhaz abkhaz
abnormalities abnormality
aboideaus aboideau
aboideaux aboideau
aboiteaus aboiteau
aboiteaux aboiteau
abos abo
abscissae abscissa
abscissas abscissa
absurdities absurdity
…

Using NLP

- Strzalkowski (in Reader)



Using NLP

INPUT SENTENCE

The former Soviet President has been a local hero ever since a Russian tank invaded Wisconsin.

TAGGED SENTENCE

The/dt former/jj Soviet/jj President/nn has/vbz been/vbn a/dt local/jj hero/nn ever/rb since/in a/dt Russian/jj tank/nn invaded/vbd Wisconsin/np ./per

Using NLP

TAGGED & STEMMED SENTENCE

the/dt former/jj soviet/jj president/nn have/vbz be/vbn a/dt local/jj hero/nn ever/rb since/in a/dt russian/jj tank/nn invade/vbd wisconsin/np ./per

Using NLP

PARSED SENTENCE

[assert

[[perf [have]][[verb[BE]]

[subject [np[n PRESIDENT][t_pos THE]

[adj[FORMER]][adj[SOVIET]]]]

[adv EVER]

[sub_ord[SINCE [[verb[INVADE]]

[subject [np [n TANK][t_pos A]

[adj [RUSSIAN]]]]

[object [np [name [WISCONSIN]]]]]]]]

Using NLP

EXTRACTED TERMS & WEIGHTS

President	2.623519	soviet	5.416102
President+soviet	11.556747	president+former	14.594883
Hero	7.896426	hero+local	14.314775
Invade	8.435012	tank	6.848128
Tank+invade	17.402237	tank+russian	16.030809
Russian	7.383342	wisconsin	7.785689

Other Considerations

- Church (SIGIR 1995) looked at correlations between forms of words in texts

	hostages	null
hostage	619(a)	479(b)
null	648(c)	78223(d)

Assumptions in IR

- Statistical independence of terms
- Dependence approximations

Statistical Independence

Two events x and y are statistically independent if the product of their probability of their happening individually equals their probability of happening together.

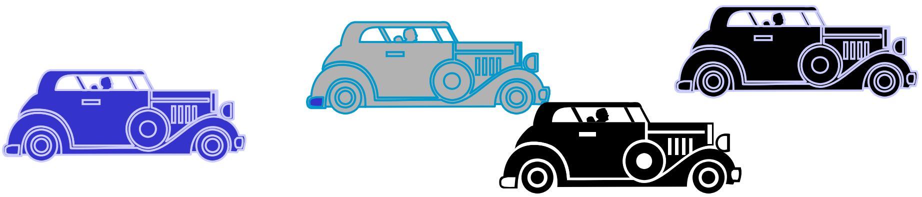
$$P(x)P(y) = P(x, y)$$

Statistical Independence and Dependence

- What are examples of things that are statistically independent?
- What are examples of things that are statistically dependent?

Statistical Independence vs. Statistical Dependence

- How likely is a red car to drive by given we've seen a black one?



- How likely is the word "ambulence" to appear, given that we've seen "car accident"?
- Color of cars driving by are independent (although more frequent colors are more likely)
- Words in text are not independent (although again more frequent words are more likely)

Lexical Associations

- Subjects write first word that comes to mind
  - doctor/nurse; black/white (Palermo & Jenkins 64)
- Text Corpora yield similar associations
- One measure: Mutual Information **(Church and Hanks 89)**

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x),P(y)}$$

- If word occurrences were independent, the numerator and denominator would be equal (if measured across a large collection)

Interesting Associations with “Doctor”
(AP Corpus, N=15 million, Church & Hanks 89)

I(x,y)	*f(x,y)*	*f(x)*	*x*	*f(y)*	*y*
11.3	12	111	Honorary	621	Doctor
11.3	8	1105	Doctors	44	Dentists
10.7	30	1105	Doctors	241	Nurses
9.4	8	1105	Doctors	154	Treating
9.0	6	275	Examined	621	Doctor
8.9	11	1105	Doctors	317	Treat
8.7	25	621	Doctor	1407	Bills

Un-Interesting Associations with

I(x,y)	*f(x,y)*	*f(x)*	*x*	*f(y)*	*y*
0.96	6	621	doctor	73785	with
0.95	41	284690	a	1105	doctors
0.93	12	84716	is	1105	doctors

These associations were likely to happen because the non-doctor words shown here are very common and therefore likely to co-occur with any noun.