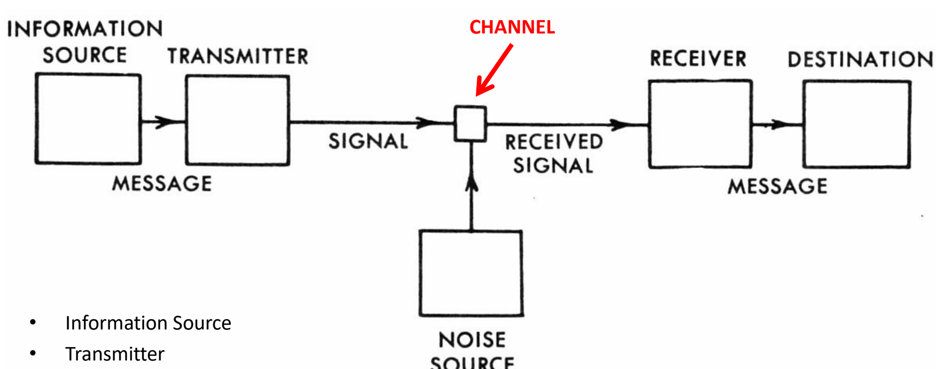
Introduction to Information Theory

Part 3

A General Communication System



- Information Source
- Transmitter
- Channel
- Receiver
- Destination

REVIEW…

Shannon's First Theorem

- By coding sequences of independent symbols (in S^n), it is possible to construct codes such that

$$\lim_{n \to \infty} \frac{L_n}{n} = \mathrm{H}$$

The price paid for such improvement is increased coding complexity (due to increased n) and increased delay in coding.

REVIEW…

Entropy & Coding: Central Ideas

- Use short codes for highly likely events. This shortens the average length of coded messages.
- Code several events at a time. Provides greater flexibility in code design.
- Shannon's Source Coding Theorem
- Algorithms
- Applications

REVIEW…

Source Coding

- **Efficient** codes
- **Singular** codes
- **Nonsingular** codes
- **Instantaneous** codes
- **Universal** Codes
  Codes that do not require knowledge of probability distribution but act in the limit as if they did have the knowledge.

REVIEW…

Huffman Codes

- Nonsingular
- Instantaneous
- Efficient
- Non-unique
- Powers of a source lead closer to H
- Requires knowledge of symbol probabilities

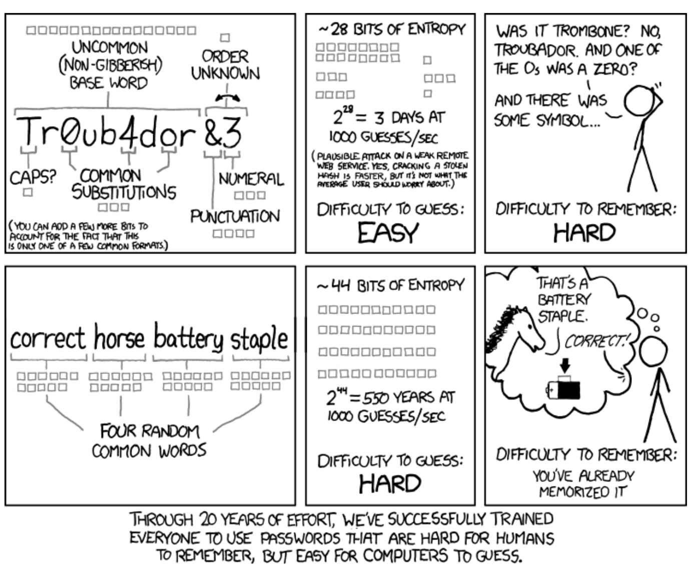
REVIEW…

Entropy & Coding: Central Ideas

- Use short codes for highly likely events. This shortens the average length of coded messages.
- Code several events at a time. Provides greater flexibility in code design.
- Shannon's Source Coding Theorem
- Algorithms: Huffman Encoding, …
- Applications: Compression…

REVIEW…

xkcd(#936): Password Strength



Some Observations

- Successive symbols from a source are not always independent. E.g. in english,

  h is more likely to occur following a *t*.

- This **intersymbol dependency** must be accounted for in an accurate measure of entropy.

Lossless Compression: English Text

- How much can we compress a given text?
- What is the accurate measure of entropy of English texts?
- Zeroth-Order entropy: S

$$H_0 \leq \log\left(\frac{1}{27}\right)$$
$$\leq 4.755\ bits/letter$$

- First-Order Entropy: S^2

$$H_1 = 3.3$$

- Second-Order Entropy: S^3

$$H_2 = 3.1$$

Zipf's Law, 1949

$$P_n \sim 1/n^a$$

P_n is the frequency of occurrence of the n^{th} ranked item and a is close to 1.

- The most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc.
- For example, in a corpus of over 1 million words:

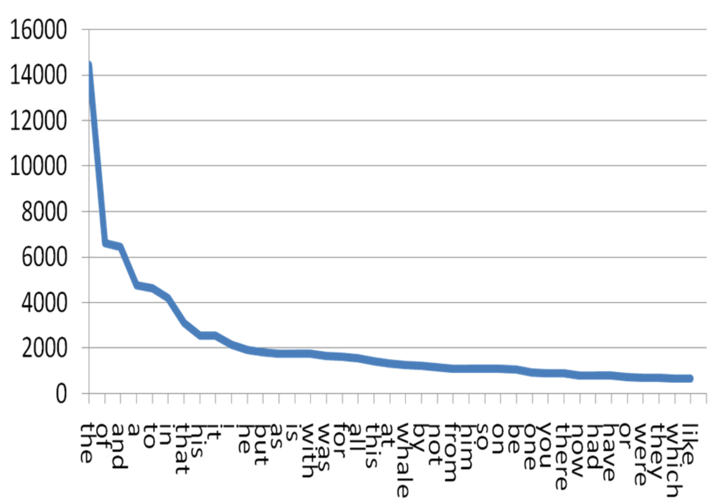
the	69,971	7%
of	36.411	3.5%
and	28,852	2.9%

- For English text:

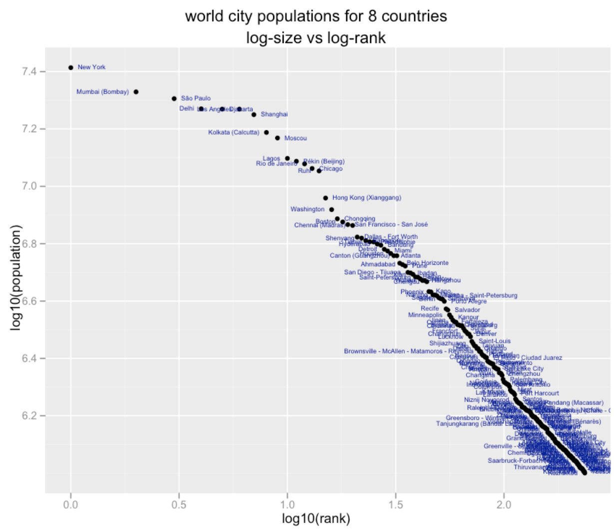
$$p_m = \frac{A}{m}$$

where m is the word's rank, p_m is the probability of the m^{th} rank word,
A is a constant that depends on the number of active words in the language.

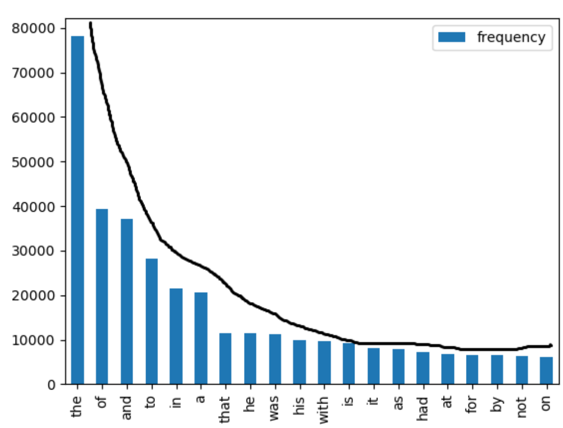
The "Long Tail" of Moby Dick



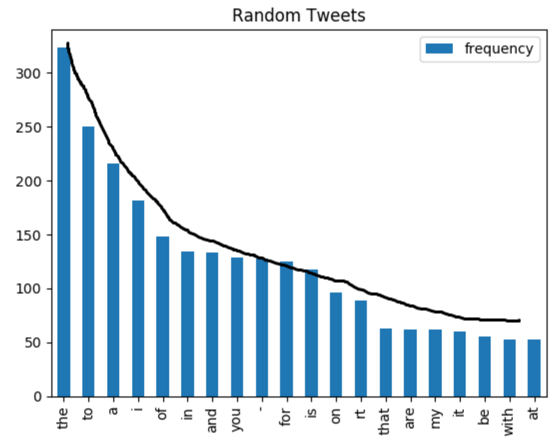
Zipf's Law and Populations



The Adventures of Sherlock Holmes – A. C. Doyle

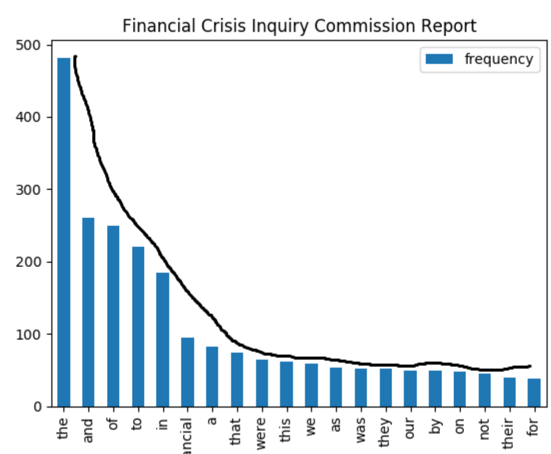


Random Tweets



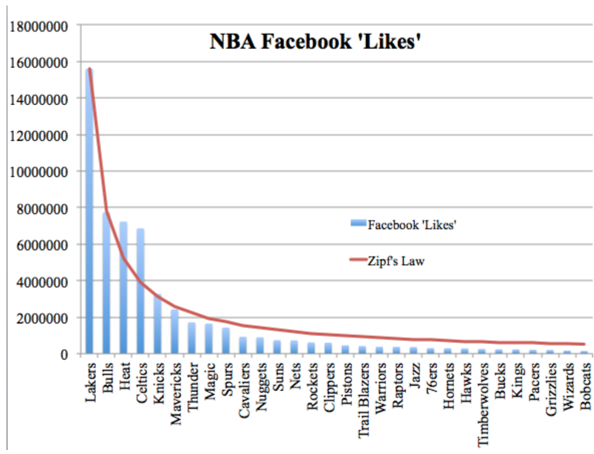
9/25/2019 15

Financial Crisis Enquiry Commission Report



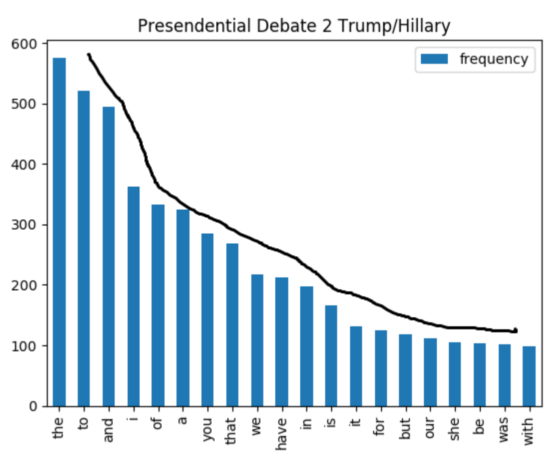
9/25/2019 16

NBA Facebook Likes



9/25/2019 17

2nd Presidential Debate between Clinton & Trump



9/25/2019 18

Estimating Entropy of English Text

- For English text, W with M words:

$$p_m = \frac{A}{m}$$

where m is the word's rank, p_m is the probability of the m^{th} rank word,
A is a constant that depends on the number of active words in the language.

- If $A = 0.1$, so that

$$\sum_{m=1}^{M} p_m = 1$$

We need, $M = 12{,}368$.

$$H_W = \sum_{m=1}^{m=12,368} \frac{.1}{m} \log\left(\frac{m}{.1}\right) = 9.716391\ bits/word$$

If $\overline{w}$=4.7 letters/word

$$H = \frac{9.716391}{4.7} = 2.067 \text{ bits}$$

Shannon Redundancy

$$R = 1 - \frac{H}{logM}$$

Where H is the per letter entropy, M is the size of the source alphabet. Thus redundancy of English is

$$1 - \frac{2.067}{\log 27} = 56.5\%$$

With an entropy of 1.5 we get 67% redundancy.

I.e. Huffman coding (even with an entropy of 3.3 or 3.1) will not get close to the theoretical limit.

Can we achieve compression rates close to 33%???

Lempel-Ziv Coding

- Sequences of text repeat patterns (words, phrases, etc)
- Construct a dictionary of common patterns
- Send references to patterns as triples (x, y, z)

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T		THIS-THESIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T		THIS-THESIS-IS-THE-THESIS.
0	0	H	H	T	HIS-THESIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T	T	HIS-THESIS-IS-THE-THESIS.
0	0	H	H	TH	IS-THESIS-IS-THE-THESIS.
0	0	I	I	THI	S-THESIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T	T	HIS-THESIS-IS-THE-THESIS.
0	0	H	H	TH	IS-THESIS-IS-THE-THESIS.
0	0	I	I	THI	S-THESIS-IS-THE-THESIS.
0	0	S	S	THIS	-THESIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T	T	HIS-THESIS-IS-THE-THESIS.
0	0	H	H	TH	IS-THESIS-IS-THE-THESIS.
0	0	I	I	THI	S-THESIS-IS-THE-THESIS.
0	0	S	S	THIS	-THESIS-IS-THE-THESIS.
0	0	-	-	THIS-	THESIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T	T	HIS-THESIS-IS-THE-THESIS.
0	0	H	H	TH	IS-THESIS-IS-THE-THESIS.
0	0	I	I	THI	S-THESIS-IS-THE-THESIS.
0	0	S	S	THIS	-THESIS-IS-THE-THESIS.
0	0	-	-	THIS-	THESIS-IS-THE-THESIS.
5	2	E	E	THIS-THE	SIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T		THIS-THESIS-IS-THE-THESIS.
0	0	H	H	T	HIS-THESIS-IS-THE-THESIS.
0	0	I	I	TH	IS-THESIS-IS-THE-THESIS.
0	0	S	S	THI	S-THESIS-IS-THE-THESIS.
0	0	-	-	THIS	-THESIS-IS-THE-THESIS.
5	2	E	THE	THIS-THE	SIS-IS-THE-THESIS.
5	1	I	SI	THIS-THESI	S-IS-THE-THESIS.

9/25/2019 29

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T		THIS-THESIS-IS-THE-THESIS.
0	0	H	H	T	HIS-THESIS-IS-THE-THESIS.
0	0	I	I	TH	IS-THESIS-IS-THE-THESIS.
0	0	S	S	THI	S-THESIS-IS-THE-THESIS.
0	0	-	-	THIS	-THESIS-IS-THE-THESIS.
5	2	E	THE	THIS-THE	SIS-IS-THE-THESIS.
5	1	I	SI	THIS-THESI	S-IS-THE-THESIS.
7	2	I	S-I	THIS-THESIS-I	S-THE-THESIS.

9/25/2019 30

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T	T	HIS-THESIS-IS-THE-THESIS.
0	0	H	H	TH	IS-THESIS-IS-THE-THESIS.
0	0	I	I	THI	S-THESIS-IS-THE-THESIS.
0	0	S	S	THIS	-THESIS-IS-THE-THESIS.
0	0	-	-	THIS-	THESIS-IS-THE-THESIS.
5	2	E	THE	THIS-THE	SIS-IS-THE-THESIS.
5	1	I	SI	THIS-THESI	S-IS-THE-THESIS.
7	2	I	S-I	THIS-THESIS-I	S-THE-THESIS.
10	5	-	S-THE-	THIS-THESIS-IS-THE-	THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T	T	HIS-THESIS-IS-THE-THESIS.
0	0	H	H	TH	IS-THESIS-IS-THE-THESIS.
0	0	I	I	THI	S-THESIS-IS-THE-THESIS.
0	0	S	S	THIS	-THESIS-IS-THE-THESIS.
0	0	-	-	THIS-	THESIS-IS-THE-THESIS.
5	2	E	THE	THIS-THE	SIS-IS-THE-THESIS.
5	1	I	SI	THIS-THESI	S-IS-THE-THESIS.
7	2	I	S-I	THIS-THESIS-I	S-THE-THESIS.
10	5	-	S-THE-	THIS-THESIS-IS-THE-	THESIS.
14	6	.	THESIS.	THIS-THESIS-IS-THE–THESIS.	

Lempel-Ziv Coding

- Sequences of text repeat patterns (words, phrases, etc)
- Construct a dictionary of common patterns
- Send references to patterns as triples (x, y, z)
  e.g. $(5, 3, F)$
  go back 5 received chars
  take the next 3 from there
  add F to the end
- Size of Search Buffer and Look-Ahead Buffer is finite.
- Used by ZIP, PKSip, Lharc, PNG, gzip, ARJ
- Extended to LZ78 (uses dictionary), LZW (+Terry Welch)
- Achieves optimal rate of transmission in the long run w/o using probability dist.

Decode

Message

0	0	I	
0	0	-	
0	0	M	
3	1	S	
1	1	-	
5	5	L	
5	3	Y	

Decode

Message

0	0	I	I
0	0	-	
0	0	M	
3	1	S	
1	1	-	
5	5	L	
5	3	Y	

Decode

Message

0	0	I	I
0	0	-	I -
0	0	M	
3	1	S	
1	1	-	
5	5	L	
5	3	Y	

Decode

Message

0	0	I	I
0	0	-	I -
0	0	M	I - M
3	1	S	
1	1	-	
5	5	L	
5	3	Y	

Decode

Message

0	0	I	I
0	0	-	I -
0	0	M	I - M
3	1	S	I - M I S
1	1	-	
5	5	L	
5	3	Y	

Decode

Message

0	0	I	I
0	0	-	I -
0	0	M	I - M
3	1	S	I - M I S
1	1	-	I - M I S S -
5	5	L	
5	3	Y	

9/25/2019 39

Decode

Message

0	0	I	I
0	0	-	I -
0	0	M	I - M
3	1	S	I - M I S
1	1	-	I - M I S S -
5	5	L	I - M I S S - M I S S - L
5	3	Y	

9/25/2019 40

Decode

Message

0	0	I	I
0	0	-	I -
0	0	M	I - M
3	1	S	I - M I S
1	1	-	I - M I S S -
5	5	L	I - M I S S - M I S S - L
5	3	Y	I - M I S S - M I S S - L I S S Y

References

- Eugene Chiu, Jocelyn Lin, Brok Mcferron, Noshirwan Petigara, Satwiksai Seshasai: *Mathematical Theory of Claude Shannon: A study of the style and context of his work up to the genesis of information theory. MIT* 6.933J / STS.420J The Structure of Engineering Revolutions
- Luciano Floridi, 2010: *Information: A Very Short Introduction*, Oxford University Press, 2011.
- Luciano Floridi, 2011: *The Philosophy of Information*, Oxford University Press, 2011.
- James Gleick, 2011: *The Information: A History, A Theory, A Flood*, Pantheon Books, 2011.
- Zhandong Liu , Santosh S Venkatesh and Carlo C Maley, 2008: *Sequence space coverage, entropy of genomes and the potential to detect non-human DNA in human samples*, *BMC Genomics* 2008, **9:**509
- David Luenberger, 2006: *Information Science*, Princeton University Press, 2006.
- David J.C. MacKay, 2003: *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.
- Claude Shannon & Warren Weaver, 1949: *The Mathematical Theory of Communication*, University of Illinois Press, 1949.
- W. N. Francis and H. Kucera: *Brown University Standard Corpus of Present-Day American English*, Brown University, 1967.
- Edward L. Glaeser: A Tale of Many Cities, New York Times, April 10, 2010. Available at: http://economix.blogs.nytimes.com/2010/04/20/a-tale-of-many-cities/
- Alan Rimm-Kaufman, The Long Tail of Search. Search Engine Land Website, September 18, 2007. Available at: http://searchengineland.com/the-long-tail-of-search-12198
- Deval Shah: Exploration of one of the most enigmatic mathematical law through lens of data science https://medium.com/@devalshah1619/a-mysterious-law-so-simple-and-yet-so-universal-aa9f1c8903d1