Edit distance and information theory units:

- The edit distance problem and algorithm (Monday)
- A quick tour of information theory (**today**)
- Lab: Autoregressive text generation (Friday)
- [Begin *n*-gram language models (next week Monday)]

Today:

- Information entropy
- The entropy of English
- The noisy channel model

Summary from Stone, *Information Theory*, pg 2.

> *In 1948, Claude Shannon published a paper called A Mathematical Theory of Communication. This paper heralded a transformation in our understanding of information. Before Shannon's paper, information had been viewed as a kind of poorly defined miasmic fluid. But after Shannon's paper, it became apparent that information is a well-defined and, above all, measurable quantity.*
>
> *[. . .]*
>
> *Shannon's theory of information provides a mathematical definition of information, and describes precisely how much information can be communicated between different elements of a system.*
>
> *[. . .]*
>
> *In this internet age, it is easy for us to appreciate the difference between information and data, and we have learned to treat the information as a useful 'signal' and the rest as distracting 'noise.'*

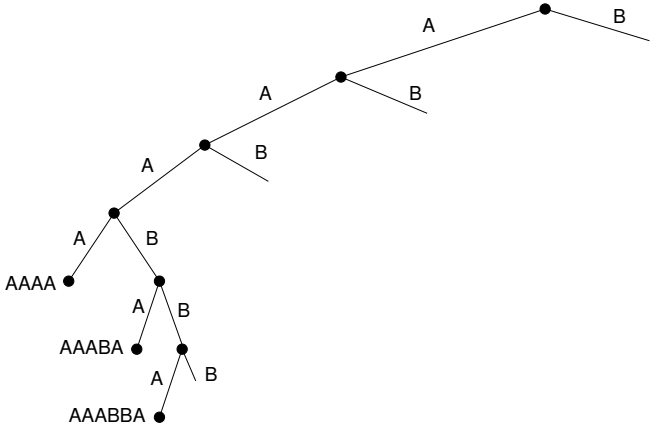
Example from Cover and Thomas, *Elements of Information Theory*, Pg 44.

> *The World Series is a seven-game series that terminates as soon as either team wins four games. Let X be the random variable that represents the outcome of a World Series between teams A and B; possible values of X [include] AAAA, BABABAB, BBBAAAA. Assuming that A and B are equally matched and that the games are independent, calculate $H(X)$.*

4-game series	$AAAA$ and $BBBB$	2
5-game series	${}^{B}A^{B}A^{B}A^{B}A \times 2$	8
6-game series		20

$$\left.\begin{array}{ll} B^{B}A^{B}A^{B}A^{B}A & 4 \\ AB^{B}A^{B}A^{B}A & 3 \\ AAB^{B}A^{B}A & 2 \\ AAABBA & 1 \end{array}\right\} 10\times 2$$

7-game series	number of 6-game series $\times 2$	40



$$2 \cdot \left(\tfrac{1}{2}\right)^4 \cdot 4 + 8 \cdot \left(\tfrac{1}{2}\right)^5 \cdot 5 + 20 \cdot \left(\tfrac{1}{2}\right)^6 \cdot 6 + 40 \cdot \left(\tfrac{1}{2}\right)^7 \cdot 7 = 5.8125$$

$$\tfrac{2}{2^4} + \tfrac{8}{2^5} + \tfrac{20}{2^6} + \tfrac{40}{2^7} = 1$$

The **entropy** of X is

$$H(X) = \sum_{i=1}^{m} p(x_i) \lg \frac{1}{p(x_i)}$$

where X is a **random variable**, (a function mapping the outcome of an experiment to a numerical value summarizing that outcome) and $x_1, x_2, \ldots x_m$ are the values that X can take on.

Note that

$$\lg \frac{1}{\left(\frac{1}{2}\right)^n} = \lg 2^n = n$$

The meaning of *entropy*

> *The word entropy had of course been used before Shannon. In 1864 Rudolf Clausius introduced the term... to represent a "transformation" that always accompanies a conversion between thermal and mechanical energy. ...*
>
> *[One of the authors] asked Shannon what he had thought about when he had finally confirmed his famous measure. Shannon replied: "My greatest concern was what to call it. I thought of calling it 'information,' but that word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy is, so in a debate you will always have the advantage.' "*
>
> *Tribus and McIrvine, "Energy and Information", Scientific American # 224, Sept 1971, pg 178–184*

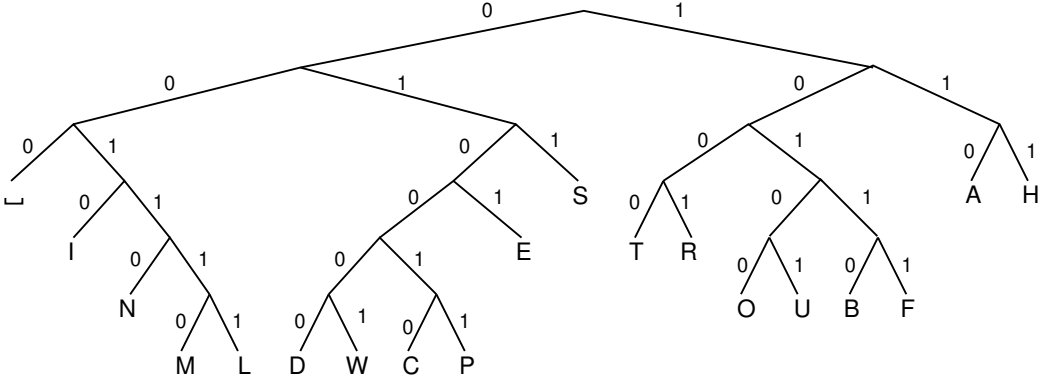
Entropy is the amount of choice involved on average in selecting an event.

Entropy is the average amount of surprise experienced by a person who knows the probability distribution.

Entropy is a measure of disorder or uncertainty.

Entropy is an upperbound on the average number of bits needed to communicate an outcome.

␣	000	E	0101	M	001110	S	011
A	110	F	10111	N	00110	T	1001
B	10110	H	111	O	10100	U	10101
C	010010	I	0010	P	010011	W	010000
D	010001	L	001111	R	1000		



Entish

'Hoom, hmm! Come now! Not so hasty! You call yourselves hobbits? But you should not go telling just anybody. You'll be letting out your own right names if you're not careful.'
'We aren't careful about that,' said Merry. 'As a matter of fact I'm a Brandybuck, Meriadoc Brandybuck, though most people call me just Merry.'
'And I'm a Took, Peregrin Took, but I'm generally called Pippin, or even Pip.'
'Hmm, but you are hasty folk, I see,' said Treebeard. 'I am honoured by your confidence; but you should not be too free all at once. There are Ents and Ents, you know; or there are Ents and things that look like Ents but ain't, as you might say. I'll call you Merry and Pippin if you please— nice names. For I am not going to tell you my name, not yet at any rate.' A queer half-knowing, half-humorous look came with a green flicker into his eyes. 'For one thing it would take a long while: my name is growing all the time, and I've lived a very long, long time; so my name is like a story. Real names tell you the story of the things they belong to in my language, in the Old Entish as you might say. It is a lovely language, but it takes a very long time to say anything in it, because we do not say anything in it, unless it is worth taking a long time to say, and to listen to. Tolkien, *TLotR III.4*


INSTRUCTIONS
1
PULL PIN. HOLD UNIT UPRIGHT. HALAR.
2
6ft
STAND BACK 6 FEET. AIM AT BASE OF FIRE. APUNTAR.
3
SQUEEZE LEVER & SWEEP SIDE TO SIDE. PRESIONAR Y APLICAR.
A TRASH • WOOD • PAPER
B LIQUIDS
C ELECTRICAL EQUIP.
BASURA • MADERA • PAPEL
LIQUIDOS
EQUIPO ELECTRICO
MULTIPURPOSE DRY CHEMICAL
AGENTE QUÍMICO SECO MULTIUSO
Kidde
Kidde Residential and Commercial
1394 South Third St.
Mebane, NC 27302-8711, U.S.A.
1-800-880-6788
www.kiddeco.com

Language technology	**Source**	**Channel**	**Observation**
Text decompression	Original text	Compressor	Compressed text
Sentiment analysis	Writer's sentiment	Writing process	Text whose sentiment is to be determined
Spelling correction	Correctly spelled word	Typing process	Possibly misspelled word
POS tagging	POS	Writing process	Word
Machine translation	Text in target language	Writing process	Text in original language

Coming up:

- Edit distance assignment (Mon, Sept 15)

- Reading: Stone (technical) or Doyle (fiction) (see Canvas) (Wed, Sept 10)
- Information theory quiz (Thurs, Sept 11)
- Huffman encoding assignment (Wednesday, Sept 17)

- Reading from J&M, Sections 3.(0-8) (Mon, Sept 15) (See Canvas for guidance)

Next time: Autoregressive text generation using character-based language models (lab)

Next week: Big unit on language models