Information and Entropy

STATISTICS

An introduction to Shannon information and entropy, with word- and token-level measurements from a Wikipedia article.

PUBLISHED
August 10, 2026

Series: A Statistical View of Data

4.4 min read

Information

There are many definitions of information - this post focuses on Shannon Information. One thing that might be somewhat unintuitive is that Shannon Information is not about meaning but rather about surprise.

Here's an example to make this more clear:

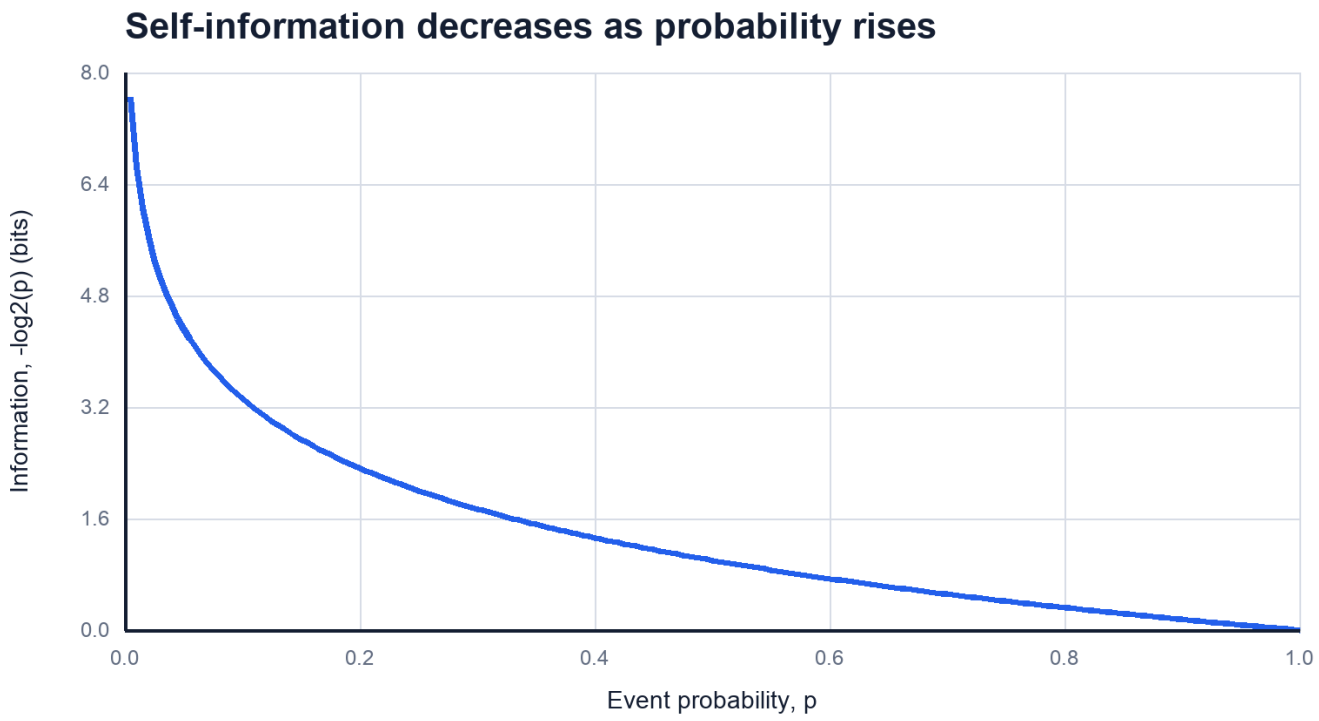
- Suppose I tell you that the sun rose this morning. Technically, you were exposed to information semantically, but it didn't surprise you much because you probably already knew this happened/would happen.
- On the other hand, suppose I told you that it snowed in SF. You'd probably be more surprised to hear that.

Claude Shannon's insight was that you can quantify this surprise using a term called information without caring about the semantics of what the actual event is. The unintuitive part is that information is not an intrinsic property of the event; it depends on the probability distribution we assign to it. To see what this means, consider I had the following message/event: $x = \text{It rained}$. I conveyed this same message to two people, one in Seattle and the other in Death Valley. The person in Seattle, would probably not be very surprised with this, so they got less information out of the message. Same semantics, different distributions, different information. In other words, the less likely an event is to occur the more information we gain (the more surprised we are) when we learn that it occured. We can measure this mathematically,

$$I = log(1/p)$$
$$I = -log(p)$$

where p is the is the probability an event occurs.

Now, you might be wondering why we use a log function here. One reason is that the negative log function is a decreasing function and captures the inverse relationship well. Also, when $p = 1$ the log equals 0 which is exactly what we want for information.



Self-information decreases as an event becomes more probable.

But another reason we use the log function here is to satisfy the multiplication of probabilities (we want information to be additive for independent events). Suppose we have two independent events x, y then $P(x \text{ and } y) = P(x).P(y)$. Now if, I told you two completely unrelated pieces of information - say for example "the weather is 23 degrees" and "my cat ate breakfast" - knowing one doesn't really interfere with the other so the information should add up. Mathematically,

$$\begin{aligned} I(x, y) &= -\log P(x, y) \\ &= -\log \Big(P(x)P(y) \Big) \\ &= -\log P(x) - \log P(y) \\ &= I(x) + I(y) \end{aligned}$$

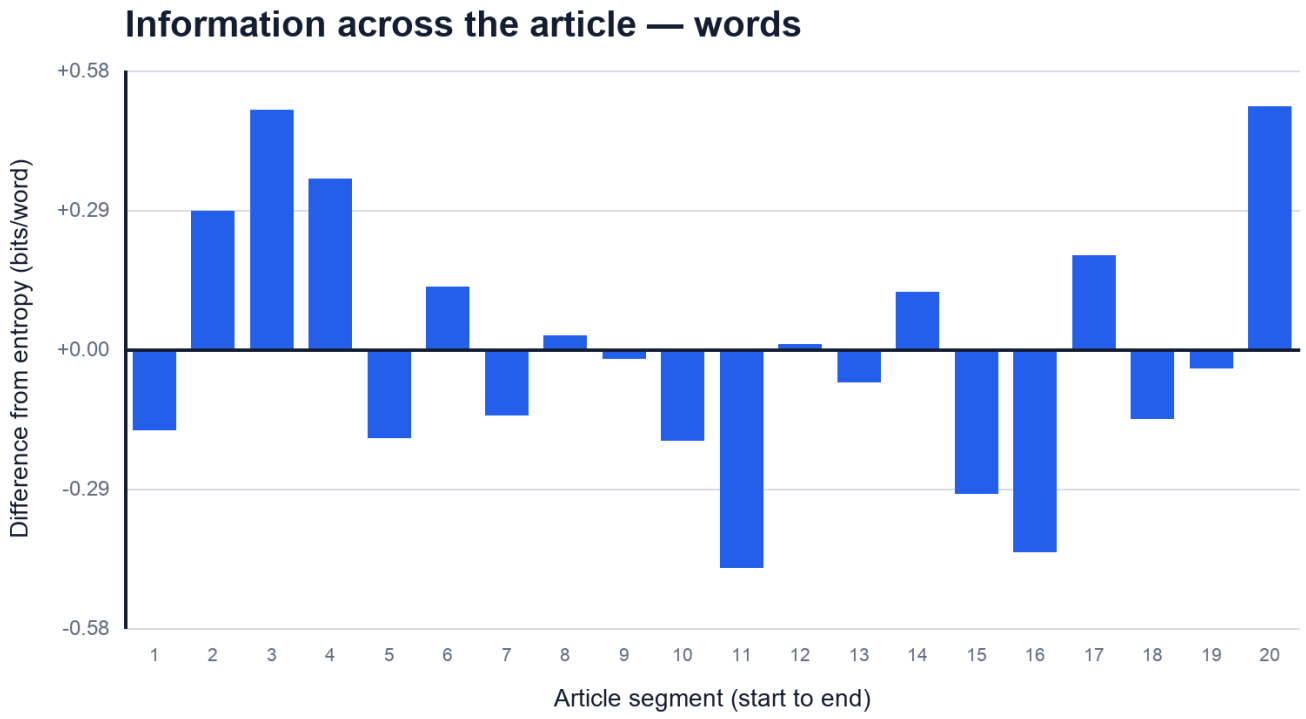
> ⓘ **Note on Log Base**
>
> You could choose any base for the logarithms but a common choice is 2, and the units of information are bits.

Calculating information of a Wikipedia Article

For the Black hole article, splitting on whitespace produces 8,582 words, of which 2,387 are unique. The sum of the observed words' self-information is 78,620.625 bits. Encoding the same article with the GPT-2 tokenizer produces 11,636 tokens, of which 2,447 are unique, for a total of 105,634.450 bits of self-information.

Information in terms of words

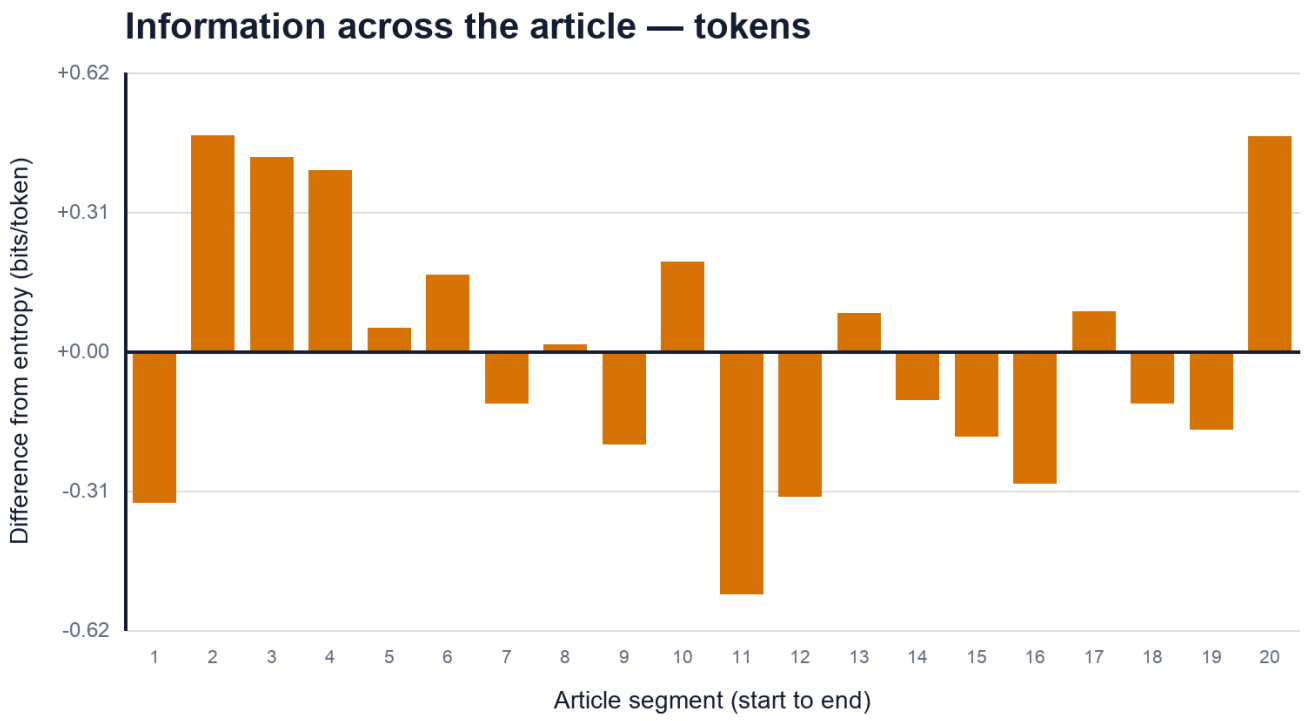
The article is divided into 20 equal sequential segments below. Each bar shows that segment's average word information minus the document-wide word entropy. Positive bars contain higher-information language than the article average; negative bars contain more repetitive or predictable language.



Average word information across sequential segments of the Black hole article.

Information in terms of tokens

The token-level view uses the same segmentation and baseline, but treats GPT-2 tokens rather than whitespace-delimited words as the possible outcomes.



Average GPT-2 token information across sequential segments of the Black hole article.

Entropy

Entropy is the expected (average) information contained in a series of mutually exclusive events.

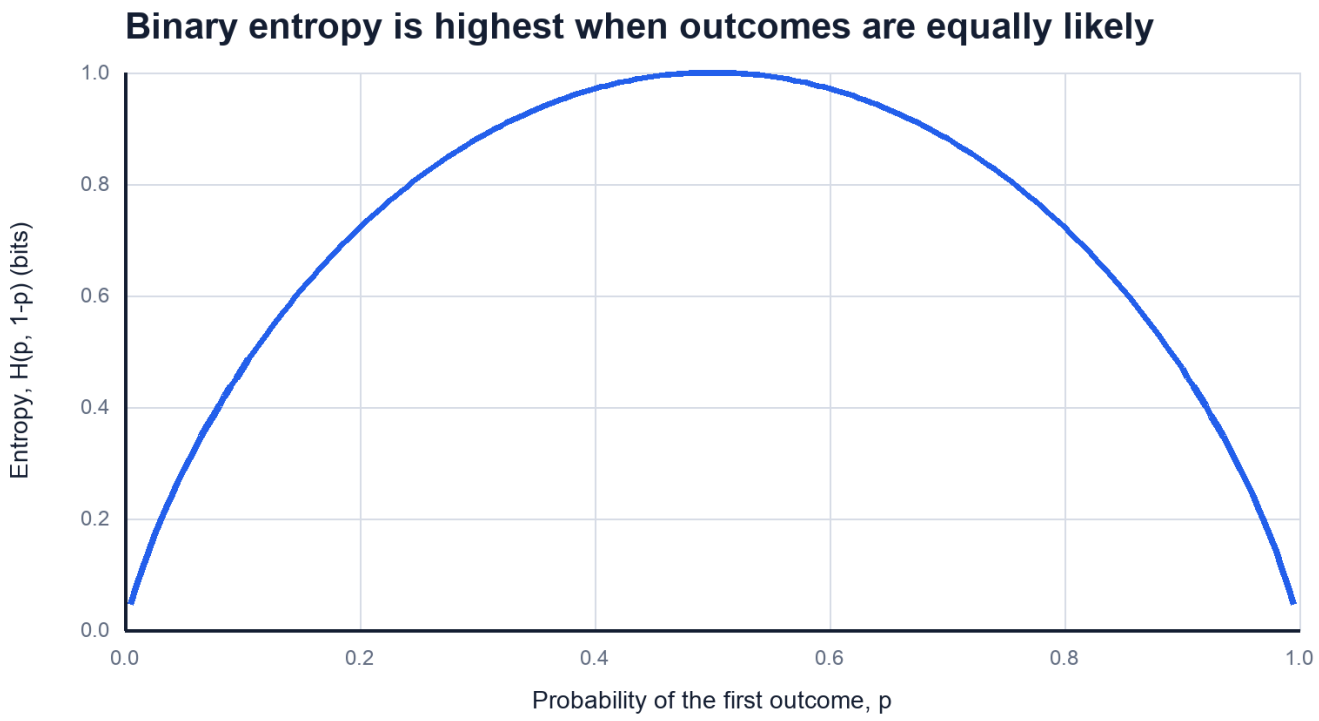
$$\begin{aligned} H(p_1, p_2, \dots, p_n) &= \Sigma_{i=1}^{n} p_i I(p_i) \\ &= -\Sigma_{i=1}^{n} p_i log(p_i) \end{aligned}$$

where p_i is the probability of an event.

If there are only two outcomes, then we get a special case:

$$H(p, 1-p) = -p\log(p) - (1-p)\log(1-p)$$

which has a nice graph that maximizes when $p = 0.5$



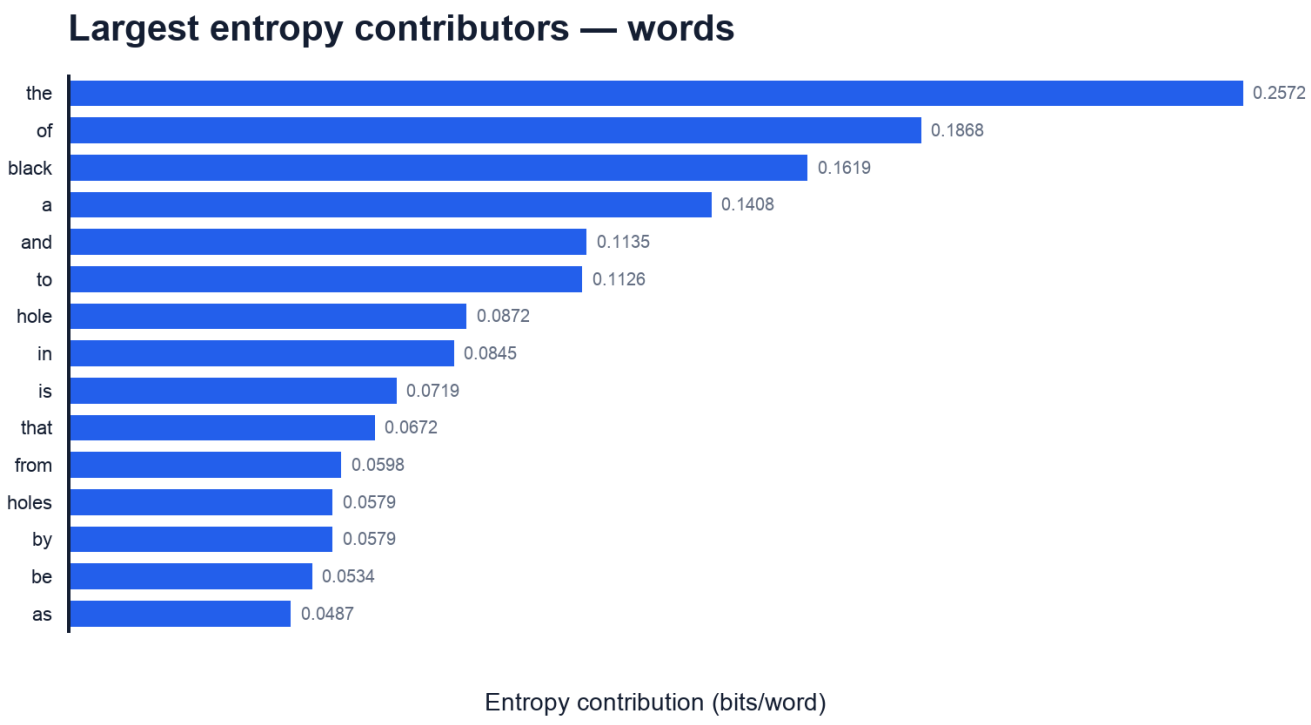
Binary entropy reaches its maximum of one bit at p = 0.5.

Note that these are defined only for discrete random variables. Continuous random variables have a separate type of entropy called differentiable entropy.

Calculating Information of a wikipedia article

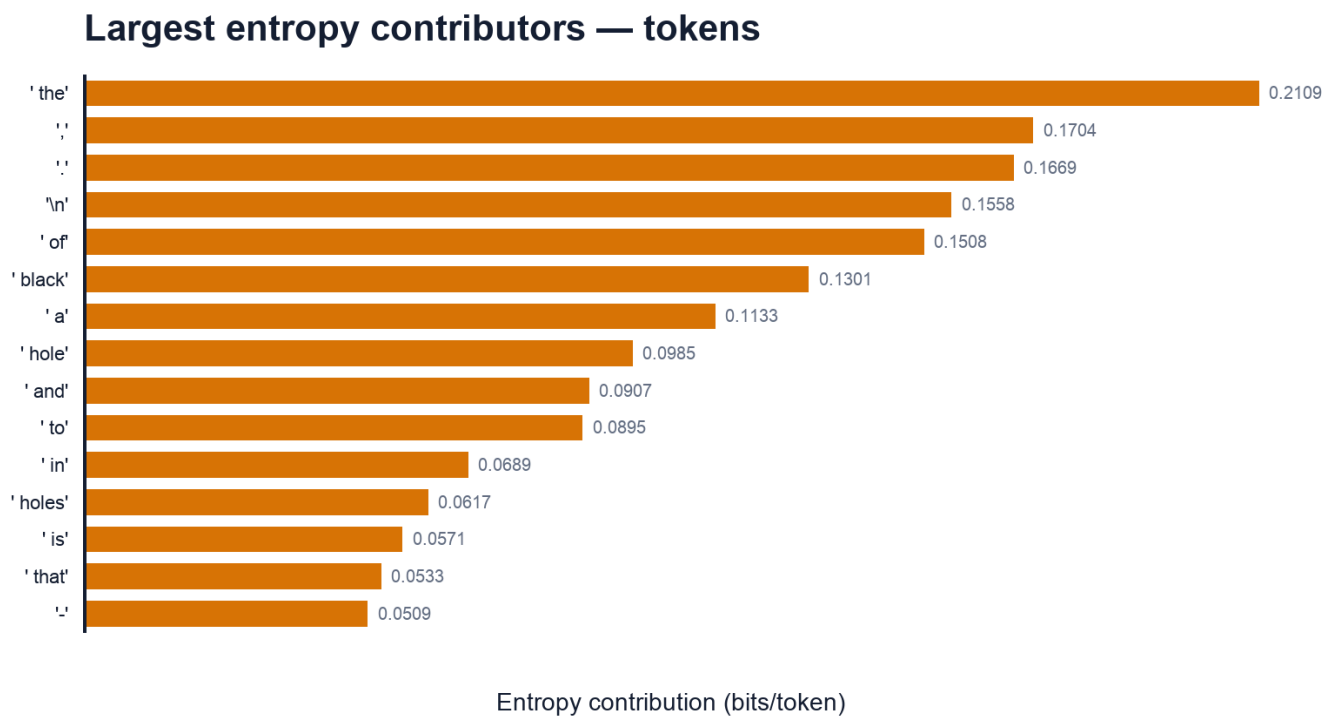
The empirical word distribution has an entropy of 9.161108 bits per word, while the empirical GPT-2 token distribution has an entropy of 9.078244 bits per token. The plots below rank outcomes by $p(x)[-\log_2 p(x)]$, their contribution to entropy, rather than by rarity alone.

Entropy in terms of words



The words making the largest contributions to the article's word-level entropy.

Entropy in terms of tokens



The GPT-2 tokens making the largest contributions to the article's token-level entropy.

Coming Up

1. Conditional Entropy
2. Differential Entropy (TBD)

Sources

1. MIT 6.02 Lecture Notes
2. Wikipedia article on Black holes

Appendix

1. Shannon information only sees distinctions that are encoded in the outcome space. If two experiments produce the same reported outcome and we discard which experiment produced it, they collapse to the same outcome. If they also have the same probability under the relevant distribution, they carry the same Shannon information.

For example, suppose we observe rain in two different cities but report only **“it is raining.”** If the city is omitted, both observations resolve to the same outcome. If knowing *which city* matters, then that provenance must be encoded explicitly, e.g. (Seattle,rain) versus (Death Valley,rain).