Are you sure you've won yet?

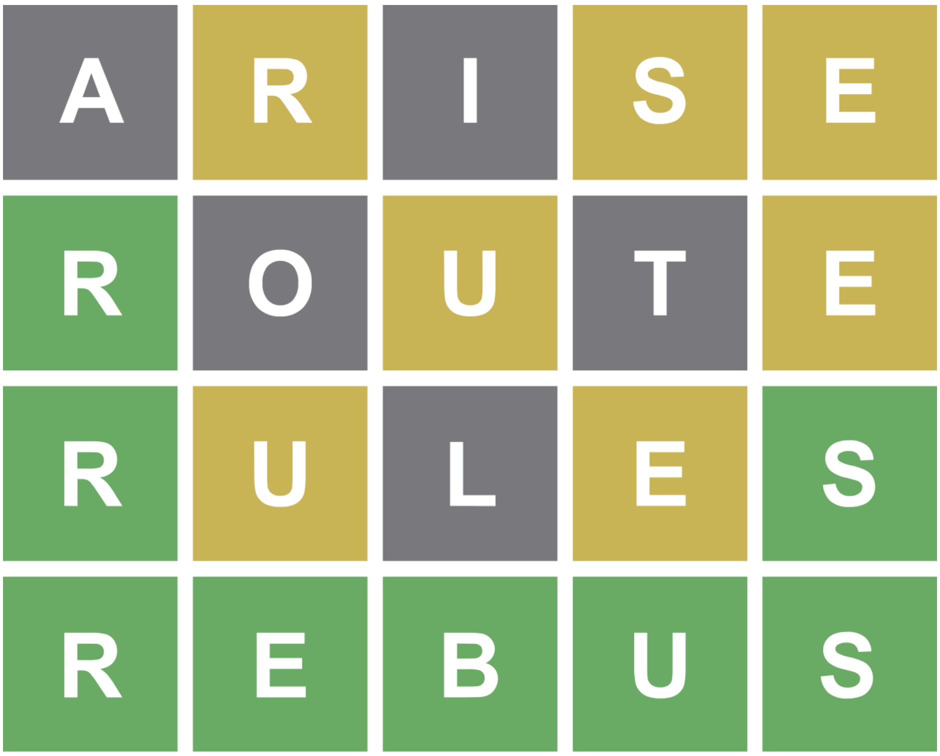
By: Ye Xu

You may be familiar with the game Wordle, you know, the game by the New York Times, the one where you have to guess the five-letter word of the day within six guesses.

WORDLE

A DAILY WORD GAME [1]

Ever since its release in 2022, Wordle has been one of my favourite games ever, even going so far as to post the daily Wordle onto my WhatsApp status for 60 consecutive days, despite having my friends constantly make jokes about me daily for using WhatsApp statuses instead of Instagram stories. And as a self-proclaimed and avid fan of Wordle, someone who has completed more than 600 Wordles and many other versions of this game, and hereby Wordle expert, I'm going to be exploring the mathematics behind Wordle and how to dominate!



[2]

[1] https://www.nytimes.com/games/wordle/index.html
[2] https://en.wikipedia.org/wiki/WordlGamementionede

The Rules of the Game

Other than the constraints mentioned in the introduction, each guess will give you information about the word of the day; once a valid word has been submitted, the square around individual letters will start turning into different colours, which are:

- ➔ A grey colour indicates that the letter within the word you have submitted as a guess is not within the word of the day.

[3]

- ➔ A yellow colour indicates that the letter is within the word of the day, but isn't in the correct place within the five slots

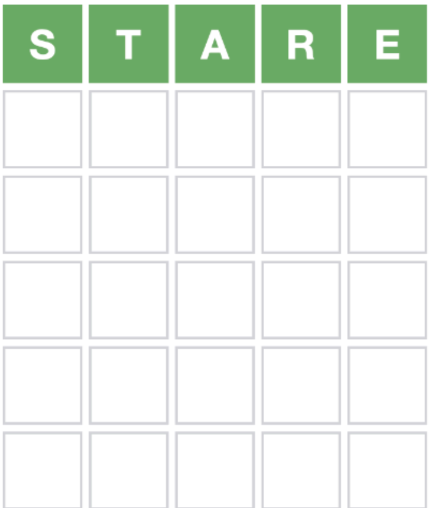
- ➔ A green colour indicates that the letter is within the word of the day and is also in the correct place within the five-letter slots

Here's a custom Wordle to see if you managed to get the gist of what I said: https://mywordle.strivemath.com/?word=iokkd

[3] https://www.nytimes.com/games/wordle/index.html, taken from the 'How To Play' section of the Wordle page.

Optimal word

As you're reading this, what do you think is, on average, the most optimal starting word? If you get it correctly, you win an imaginary egg McMuffin



[4]

In Wordle, there are around 2,300 words on the 'solution list', words that can be answers in Wordle, meaning that usually, the quality of your first guess is based on luck, or you can just get lucky on your following guesses. The only way to get down to 1 word in one guess would be if every pattern had exactly one pattern behind, meaning that for all 2,300 words, you'd need 2,300 distinct patterns, but we only have 243, because no matter how much you've narrowed it down, and no matter how sure, there can always be a nicher word that you didn't know of – hence the title.

[4] Yes these are mine ;)

By guessing any one of those 2,300 words, I'm certain that you will get one out of 243 fixed colour patterns. How am I so sure of this number?

Well, we mentioned that in Wordle there are only 3 colours - grey, yellow, and green. We can then say that in the first slot of the word, only one of 3 colours can go there, meaning that if we only have one square, we can only get a total of 3 possible patterns, either a grey square, a yellow square, or a green square.
Now, if we increase the number of squares to 2, we again have 3 possibilities for the first square, but also 3 different possibilities for the second square. This is because the squares are conditionally independent; the outcome of the first square doesn't impact the outcome of the second square, and vice versa. And if you were to find all the patterns of two squares by just counting, you would end up with 9 different patterns, which is just:

3×3 or 3^2

And to find it for all 5 squares, I just think of it as, 3 can go in the first square, 3 in the second, 3 in the third, 3 in the fourth, and 3 in the fifth, so just

$3 \times 3 \times 3 \times 3 \times 3$ or 3^5

which is equal to 243

Next, since every time you play, the word is different, the globally optimal word is going to be the average best starting word every time you play a new game, and to figure it out, you'd need a few axioms, laws or rules that we assume to be correct:

1. The rarest outcomes(or results) give us the most information, such as having 4 greens and a grey, as you'd be very close to guessing the word.
2. As a result, a certain outcome, such as all greens, gives us no information. Although this may not seem important from a player's point of view, as they would've won the game, so who cares, from a mathematical standpoint, it is important when trying to create a general formula.
3. As discussed before, since the colours of individual letters are independent, we will be able to add up these independent variables.

What I’m trying to do here is to create a general formula to measure how good of a starting word it is.

So to translate this into mathematics, let's say we have a function, a sort of ‘machine’, when you input something, you get one output,

And we’ll call this $J(p)$
Where:
p is our probability of something happening, in this case, it’ll be
$p = \frac{how\ \ many\ words\ will\ give\ pattern\ i[1\ out\ of\ 243\ possible\ ones]\ when\ guessing\ word\ W}{2300[the\ total\ number\ of\ solutions]}$
And
J being our ‘machine’

The nerve-wracking thing with p, however, is that it’s the probability that we will get pattern i by guessing one out of 2300 solutions, so we’d need to find the probability of getting every single one of those patterns by counting how many words give us that specific pattern.
So, for example, let’s say we used the word ‘MATHS’ as our starting guess(sorry!), and the pattern we got back was 1 green at the beginning, and 4 greys. And let's say the number of words that give us this exact pattern is 46, that would mean 2% of all words will give us pattern i because

$$\frac{46}{2300} = 0.02 = 2\%$$

Since all of these still fit under the description of independent variable, as they don't affect each other, we can add them all together, and what this helps us do is to find, on average, the best starting words, meaning that our equation now has to become

$$\sum_{i=1}^{243} J(p_i)$$

Now, according to the second axiom, we know that if an outcome is 100% certain, we will get nothing(0).

So

$$J(1) = 0$$

Then, according to axiom 3, since the colours of the letters are independent, we’ll be able to add the information about the letters up, in probability. When combining the probabilities of 2 or more things happening, we always have to multiply, and a function that does that very well is the log family. This is because you may be familiar with the general rule that when you multiply exponents with the same base, you add the powers, and since the log is like a ‘reverse exponent’ where instead of trying to find the answer of the base raised to a power you’re trying to find the power you need to raise a base to get an answer.

So

$$J(p_i) = \sum_{i=1}^{243} log_a p_i$$

Where a is an integer

And since a unitless probability will always be a number between 0 and 1, we have to multiply the equation by -1, because:

$$a^{-b} = \frac{1}{a^b}$$

So in reality

$$J(p_i) =- \sum_{i=1}^{243} log_a p_i$$

However, to find the average, we also have to multiply $log_a p_i$ by p_i. This is because $log_a p_i$ measures your 'surprise' element of when you get a pattern, or in other words information - let's say you start with a word that gives you 4 greens right off the bat, the number of words in the solution list that have that pattern will be extremely low, like 5 words maximum within the solution list, thus when putting this specific p_i into your $log_a p_i$ function, you'll get a very big number as long as a is an integer. This may skew your results, so to make up for it you have to multiply it by the probability of you getting 4 greens right off the bat, so that your favourite starting word doesn't always have to rely on you getting 4 greens right off the bat.

So

$$J(p_i) = \quad - \sum_{i=1}^{243} p_i \, log_a p_i$$

And this equation is equal to entropy, the average quantitative measure for uncertainty, and the common letter for entropy is H

$$H = - \sum_{i=1}^{243} p_i \, log_a p_i$$

Now, regarding your a, this doesn't really matter, as long $a > 1$, the way you're going to assess your words will be the same. And why not 1, you may ask, that's because 1 raised to any power is still one, meaning that no matter what you'll get as your p_i, it won't work, because $1^n = 1$ where n is any real number. Whereas a decimal may ruin your results because every p_i will be between 0 and 1, so you wouldn't want to take your chances with that.

I’d use 2 as the base, why? Because 2 digits can easily represent ‘yes’ and ‘no’, after all, that's how computers work, and if you’re still confused, I might clear it up later.

$$H = -\sum_{i=1}^{243} p_i \, log_2 p_i$$

You can’t find the most optimal starting word using this equation, but you can measure how much a word narrows down the solution on average.

If you were to use the equation for every single word in the solution list, the word with the highest score would be SALET, scoring 5.89 bits(the unit we’re using since we’re using 2 as our base for log).
The second-best word is CRANE, scoring 5.67 bits.

You could think of the value of entropy as an x amount of yes or no questions being answered at the same time with one guess, which is why using 2 as our base makes sense, and asking 5.89 yes or no questions in one single guess is more efficient than asking 5.67 questions, yes or no questions in one single guess.

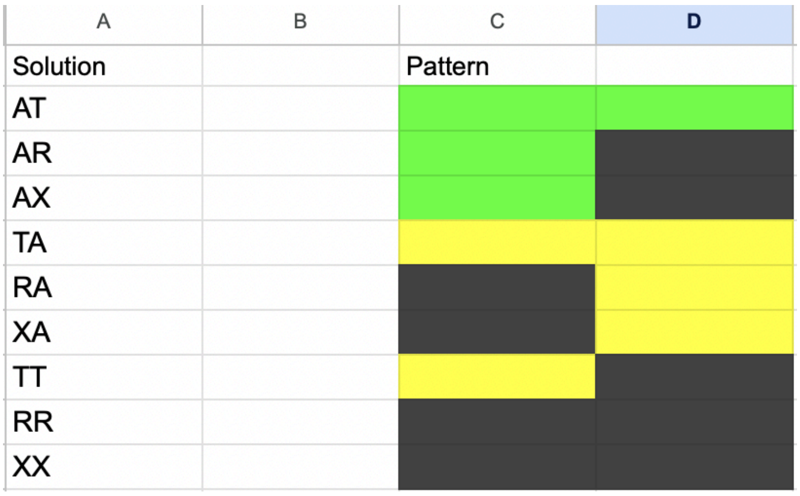
It is also worth mentioning that after the first guess, entropy is recalculated because of the new information we have gained, and the thing about SALET is that a letter such as A, or S are some of the most common letters among all words, even getting a grey A will drastically narrow down the number of words that can be the solution, then the entropy framework picks the best follow-up guess from the remaining

A Small Example

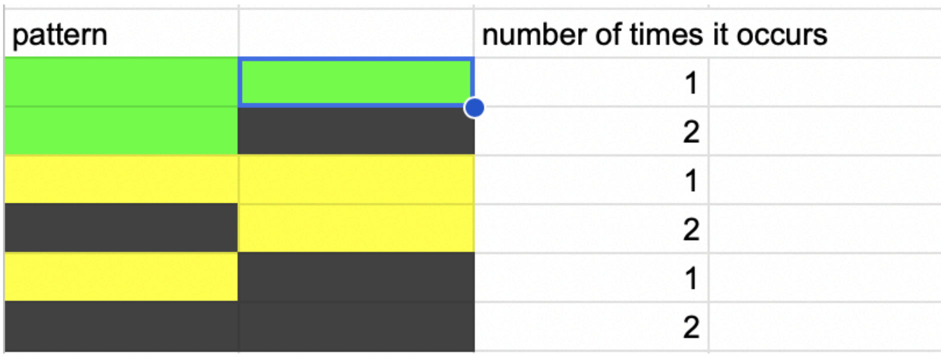
Let’s say we have a two-letter Wordle, with a solution list of 9 words, and these words are:

- AT
- AR
- AX
- TA
- RA
- XA
- TT
- RR
- XX

And let’s say you guess AT, the patterns for each solution would be:



And the number of times each pattern occurs out of 9 is:



So, using the equation $H = -\sum_{i=1}^{243} p_i \, log_2 p_i$

The calculation would be:

$$= -\left[\frac{1}{9}log_2\frac{1}{9} + \frac{2}{9}log_2\frac{2}{9} + \frac{1}{9}log_2\frac{1}{9} + \frac{2}{9}log_2\frac{2}{9} + \frac{1}{9}log_2\frac{1}{9} + \frac{2}{9}log_2\frac{2}{9}\right]$$

$$= -\left[3 \times \frac{1}{9}log_2\frac{1}{9} + 3 \times \frac{2}{9}log_2\frac{2}{9}\right]$$

$$\frac{1}{9}log_2\frac{1}{9} \approx -\ 0.352$$

$$\frac{2}{9}log_2\frac{2}{9} \approx -\ 0.482$$

$$H = -\ [3(-\ 0.352) + 3(-\ 0.482)]$$

$$= [-\ 1.056 + (-\ 1.446)]$$

$$= -\ [-\ 2.502]$$

$= 2.502$ bits

$$Max\, H = log_2 9 \approx 3.17$$

To calculate maximum entropy, you have to treat every pattern as equally, and by doing this, you have to spread the words easily across patterns. So by comparing the entropy of AT and the maximum entropy, AT is a pretty good starting word.

In comparison, the maximum entropy for the original Wordle is:

$$log_2 243 \approx 7.92$$

The reason why the difference between the entropy for SALET and the maximum entropy for the original Wordle is so big is due to the size of the solution list. The solution list being so big means that it leaves more space for uncertainty, meaning that having a universal optimal word is harder as words cut down a smaller proportion of words.

Conclusion

The best starting word for Wordle on average is SALET.

Personally, I still wouldn't start with SALET. Maybe once in a while I would, but usually what I do is start with a word on theme, for example if I'm feeling very flashy, or I want to start with a word that has flair, I would start with PZAZZ, one of the worst starting words since it has a score of nearly 0 bits and isn't even on the solutions list!

Anyway, thank you for liking my essay. I hope you've learned something, and I really, really hope you've enjoyed reading something, and all I'm asking you to do after reading this essay is to go win some games at Wordle.