Let’s talk hangman

Credit goes to Freakenomics episode 246 “How to Win Games and Beat People” and http://www.datagenetics.com/blog/april2012/ for providing the initial inspiration for this method, and my brother Daniel for suggesting improvements which made the program much more formidable.

First, for those of you that might not be familiar with this game, here are the rules.

Hangman is a two player game. One player, the hider, thinks of a secret word. The other player, the guesser tried to guess the hider’s word. At the beginning of the game, the guesser is given the length of the hider’s word, but nothing else.

Once the guesser shares the length of their word, then the guesser starts guessing! The guesser calls out a letter, and the hider reveals if their word contains the called out word or not. If the hider’s word does contain the called out letter, then they tell the guesser where in the word the called out letter is.

If the guesser calls out a letter which isn’t in the hider’s word, then the guesser gets one strike. 9 strikes and the guesser loses! If the guesser guesses all of the letters before 9 strikes, then they win.

For example, if the hider was thinking of the word “Happenstance”, and the guesser called out the letter ‘a’, then the hider would say “My word contains the letter ‘a’ in positions 2 and 9.”

Ok, so now that we all have an understanding of how the game works, the natural question is “What strategy should the guesser have in order to maximize their chances of guessing the word?” Let’s think of simple strategies, and try and improve on them step by step in order to get the best one that we can.

**Strategy 0**

The most basic strategy would be to call out random letters. If you get lucky and happen to guess the letters. After all, the hider could have guessed any word, how could you do better?

**Strategy 1**

Well, you could do a lot better. For example, if know that the hider’s word is 10 letters long, the first letter is unknown, and the last nine letters are “antaloupe”, then guessing the letter ‘q’ next is probably not a good idea.

Ok, so a better strategy would be to call out random letters, until you get enough letters that you have a pretty good idea of what the word is, and then call out letters in order to rule out possible words.

For example, if you found that the hider was thinking of a four letter word ending in “ing”, then they must\(^1\) be thinking of one of the following words: “ding, king, ping, ring, sing, ting, wing, zing.”

This is a reasonable strategy, and clearly an improvement over strategy 0, but it can still be made better.

---

\(^1\) According to my dictionary anyways, yours might have more!
Strategy 2

If you have spent any time actually using the English language, then you will know that some letters are much more common than others. If you take a quick trip to Wikipedia (https://en.wikipedia.org/wiki/Letter_frequency) then you can get a very nice table listing the letters in order from most common, to least common in English. The list is as follows:

“e t a o l n s h r d c u w f g y p b v k j q z”

So a better strategy would be to use strategy 1, but instead of calling out random letters, call letters according to the above list. That way, you’ll likely get more letters sooner.

Notice that the letter ‘h’ comes before ‘p’ in this list. That will come up later.

Strategy 3

Strategy 2 is a good one, but the list above can be improved in a number of ways. First, that list above only reflects the frequency of letters in English writing. For example, in English words like “the, then, a, is” are a lot more common than words like “juxtapose, rhythmic, doxy.” According to google’s ngram viewer (https://books.google.com/ngrams/graph?content=the%2C+juxtapose&year_start=1800&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t1%3B%2Cthe%3B%2Cc0%3B.t1%3B%2Cjuxtapose%3B%2Cc0), the word ‘the’ makes up anywhere from 4.5% to 6% of English words used in writing over the last 200 years, whereas the word juxtapose makes up at most 0.00003252% of words during 1998, it’s most popular year. This means, that the word ‘the’ is at least 100,000 times more common than the word ‘juxtapose’ in English.

This makes sense if you have read much English writing, but if you were playing hangman, do you think that a person would be 100,000 times more likely to choose the word ‘the’ over ‘juxtapose’? Probably not. In fact, if you were playing with someone who was trying to trick you, then they would probably be more likely to choose a word like “juxtapose” rather than “the.” This means the letters ‘h’ gets a huge boost in frequency as compared to a letter like ‘p’, not because it is used in more words, but because words like “the” and “then” are used so frequently.

A more useful list of letter frequencies wouldn’t weigh words that are used very commonly. It would instead weigh all words equally, and list letters in order of how common they are in those words.

That list would look like:

“e s l a r n t o i c d u p m g h b y f v k w z x q j”

Notice that in this list, ‘p’ is higher than ‘h,’ in this list.

Strategy 4

This new and improved list is certainly better, but it can actually be improved even more. Think about it. How many 2 letters words can you think of that contain the letter ‘g’? Probably not that many. My dictionary only has 3 (“mg”, “ga” and “go”) out of the 140 total 2 letter words, around 1.5%. Now, how many 4 letter words can you think of that contain the letter ‘g’? Probably a LOT more (e.g. “ring, sign,
king, hogs...”). In my dictionary, 323 out of 3130 of the four letter words contain a ‘g,’ around 10%. The frequency of letter usage changes a lot as the length of the word changes.

The point here, is that if you are guessing in hangman, then you know exactly how long the hider’s word is, and so you should really have a different list of frequencies for every word length. Here is what that frequency list would look like for words of length 2:

a o s m n e c h i d r t p l b f v w x y k g j q

And the list for words oh length 4:

e a s o i r l t n u d p m c h b g y k w f v j x z q

Certain letters (like ‘m’ and ‘h’) are much better guesses at length 2, as compared to length 4, whereas others (like ‘g’ and ‘r’) are better guesses at length 4.

That does all this say? Well, you should really have a bunch of different list of letter frequencies, one for each possible word length.

**Strategy 5**

At this point you have a pretty solid strategy, but still it can be improved. Strategy 3 relies on the knowledge that you can do much better than just calling out random letters in order to be as likely to hit a letter in the hider’s word as possible. Strategy 4 is an improvement on that idea, but it turns out that we can do even better.

In order to motivate this improvement it helps to look at an example. Let’s suppose that we had a 6 letter word, and so far we have the characters “_ i _ i n _”. Well, there are a lot of six letter words that end with the three letters “i n g,” and so even if among 6 letter ‘g’ isn’t very common, if our word end with “i n _” then ‘g’ suddenly becomes a very good guess.

To take the above example a little further, if we saw the above word, then we would also know that our word would not have a very high chance of having any other vowels. Out of the 65 words which could fit into “_ i _ i n _” only 8 of them have any other vowels.

Just for good measure, Here is the list of letter frequencies for all 6 letter words:

e s r a i o n l t d u c m p g h y b f k w v z x j q

And here is the list of all letter frequencies which could fit into “_ i _ i n _”:

i n g r v l t m s k p w b e f h a c j x o y z

Not only that, but if you knew that the word was “_ i _ i n _” and you knew that the word didn’t contain an ‘r’, then that would again change the probabilities completely. Every time that you guess a letter, weather you were right or wrong, you would have to change your frequency table.

What does all this say? Well, you either need to have a LOT of tables of letter frequencies, or you need to think about the problem in a different way.

Imagine that you had before you, you had a list of every English word. Once you figured out how long the hider’s word is, you cross out every single word on the list that isn’t of that length. Let’s pretend that
their word is of length 8. Then, you look through the words that are left for the most common letter. Let’s pretend that letter is ‘a.’ Than call out the letter ‘a!’

If you guess incorrectly, well then now you know that you can cross out every letter in your list which contains the letter ‘a.’

If you guessed correctly and the hider said that the word that they were thinking of had an ‘a’ in positions 1, 3 and 8. Then their word must look like:

“a _ a _ _ _ _ a”

And you can cross out every word on your list that doesn’t have an ‘a’ in positions 1, 3 and 8. That eliminates every single word except 4 (“academia,” “anaconda,” “analemma” and “anathema”).

In either case, every time you guess a letter you can make the list of possible words that you have to guess shorter and shorter, and every time you can just look at the most common letter in the letters that remain. That way, you don’t have to keep track of lists at all.

This also suggests that the best way to think about hangman isn’t as a game where you guess letter until you figure out their word, but as a game where you cross words off of a list, until there is only 1 possibility left.

This strategy of keeping a list of every possible English word would be very difficult for a human to keep track of in their head, but totally manageable for a computer.

**Strategy 6**

We are almost at the strategy used by the program. The last bit of insight used by the program, is that you don’t want to guess the most common letter. Let’s take the above example where we know that the word the hider is thinking of must be one of:

“academia, anaconda, analemma, anathema”

What is the next best letter to guess?

Well, the letters ‘e, m, n’ are all used in three out of the four words, and so they are good candidates. But, notice that if we guess ‘m,’ then we will know what word the hider is thinking of no matter what they answer. The same is not true if we guess e or n.

How can you see this quickly? Well, even though the words “academia, analemma, anathema” all have the letters ‘m’ and ‘a,’ they all contain ‘m’ in different spots. Remember, we not only know that their word contains an ‘a’ and an ‘m’, but also in what positions.

If you think about it, this means that even though those three words all contain the letters ‘m,’ they all have a very different relationship to the letter ‘m.’

Let’s pick a smaller example in order to make things easier.

Suppose that the hider’s word was 3 letters long. And we were considering calling out the letter ‘a.’ There are 8 possibilities.
• ‘a’ nowhere
• ‘a’ in position 1
• ‘a’ in position 2
• ‘a’ in position 3
• ‘a’ in positions 1 2
• ‘a’ in positions 1 3
• ‘a’ in positions 2 3
• ‘a’ in positions 1 2 3

Every single word must fall into one of those 8 cases. If you call out the letter ‘a,’ you figure out which case the hider’s word falls into.

Here’s the idea, before you decide what letter to call out, take every word on your list and figure out which case it belongs to. That is if it has no ‘a’s, an ‘a’ in position 1, etc. Then ask yourself the following question, “what case has the most words in it?” Let’s call that number $n_a$. $n_a$ is the maximum possible number of words that you could be left with after calling that letter.

If $n_a$ is small, then guessing ‘a’ is a great idea. No matter what happens, calling out ‘a’ will reduce the number of possible words left to check by a lot. If $n_a$ is large, then it’s probably not such a great idea. Even if you are correct, you don’t eliminate very many words off of your list.

First compute $n_a$, then $n_b$, then $n_c$, and on and on all the way out to $n_z$. Then, call out the letter which has the smallest $n$ value. For example, if $n_a = 1343$, and $n_e = 1555$, then it might be a good idea to call out ‘a’ even if ‘e’ is more common in your list of possible words.

Repeat this process until you only have one word left.