1.9 Hangman

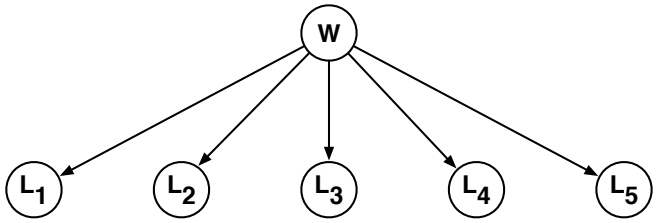
Consider the belief network shown below, where the random variable W stores a five-letter word and the random variable $L_i \in \{\texttt{A}, \texttt{B}, \ldots, \texttt{Z}\}$ reveals only the word's ith letter. Also, suppose that these five-letter words are chosen at random from a large corpus of text according to their frequency:

$$P(W = w) \;=\; \frac{\text{COUNT}(w)}{\sum_{w'} \text{COUNT}(w')},$$

where $\text{COUNT}(w)$ denotes the number of times that w appears in the corpus and where the denominator is a sum over all five-letter words. Note that in this model the conditional probability tables for the random variables L_i are particularly simple:

$$P(L_i = \ell | W = w) \;=\; \begin{cases} 1 & \text{if } \ell \text{ is the } i\text{th letter of } w, \\ 0 & \text{otherwise.} \end{cases}$$

Now imagine a game in which you are asked to guess the word w one letter at a time. The rules of this game are as follows: after each letter (A through Z) that you guess, you'll be told whether the letter appears in the word and also where it appears. Given the *evidence* that you have at any stage in this game, the critical question is what letter to guess next.



Let's work an example. Suppose that after three guesses—the letters D, I, M—you've learned that the letter I does *not* appear, and that the letters D and M appear as follows:

M ___ D ___ M

Now consider your next guess: call it ℓ. In this game the best guess is the letter ℓ that maximizes

$$P\Big(L_2 = \ell \textbf{ or } L_4 = \ell \;\Big|\; L_1 = \mathtt{M}, L_3 = \mathtt{D}, L_5 = \mathtt{M}, L_2 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}, L_4 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}\Big).$$

In other works, pick the letter ℓ that is most likely to appear in the blank (unguessed) spaces of the word. For any letter ℓ we can compute this probability as follows:

$$\begin{aligned}
&P\Big(L_2 = \ell \textbf{ or } L_4 = \ell \;\Big|\; L_1 = \mathtt{M}, L_3 = \mathtt{D}, L_5 = \mathtt{M}, L_2 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}, L_4 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}\Big) \\
&= \sum_w P\Big(W = w, L_2 = \ell \textbf{ or } L_4 = \ell \;\Big|\; L_1 = \mathtt{M}, L_3 = \mathtt{D}, L_5 = \mathtt{M}, L_2 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}, L_4 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}\Big), \quad \textbf{marginalization} \\
&= \sum_w P(W = w | L_1 = \mathtt{M}, L_3 = \mathtt{D}, L_5 = \mathtt{M}, L_2 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}, L_4 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}\Big)\, P(L_2 = \ell \textbf{ or } L_4 = \ell | W = w) \quad \textbf{product rule \& CI}
\end{aligned}$$

where in the third line we have exploited the conditional independence (**CI**) of the letters L_i given the word W. Inside this sum there are two terms, and they are both easy to compute. In particular, the second term is more or less trivial:

$$P(L_2 = \ell \textbf{ or } L_4 = \ell | W = w) \;=\; \begin{cases} 1 & \text{if } \ell \text{ is the second or fourth letter of } w \\ 0 & \text{otherwise.} \end{cases}$$

And the first term we obtain from Bayes rule:

$$\begin{aligned}
&P(W = w | L_1 = \mathtt{M}, L_3 = \mathtt{D}, L_5 = \mathtt{M}, L_2 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}, L_4 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}\Big) \\
&= \frac{P(L_1 = \mathtt{M}, L_3 = \mathtt{D}, L_5 = \mathtt{M}, L_2 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}, L_4 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\} | W = w\Big) P(W = w)}{P(L_1 = \mathtt{M}, L_3 = \mathtt{D}, L_5 = \mathtt{M}, L_2 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}, L_4 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\})} \quad \textbf{Bayes rule}
\end{aligned}$$

In the numerator of Bayes rule are two terms; the left term is equal to zero or one (depending on whether the evidence is compatible with the word w), and the right term is the prior probability $P(W = w)$, as determined by the empirical word frequencies. The denominator of Bayes rule is given by:

$$\begin{aligned}
&P(L_1 = \mathtt{M}, L_3 = \mathtt{D}, L_5 = \mathtt{M}, L_2 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}, L_4 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}) \\
&= \sum_w P(W = w, L_1 = \mathtt{M}, L_3 = \mathtt{D}, L_5 = \mathtt{M}, L_2 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}, L_4 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}), \quad \textbf{marginalization} \\
&= \sum_w P(W = w) P(L_1 = \mathtt{M}, L_3 = \mathtt{D}, L_5 = \mathtt{M}, L_2 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\}, L_4 \notin \{\mathtt{D}, \mathtt{I}, \mathtt{M}\} | W = w), \quad \textbf{product rule}
\end{aligned}$$

where again all the right terms inside the sum are equal to zero or one. Note that the denominator merely sums the empirical frequencies of words that are compatible with the observed evidence.

Now let's consider the general problem. Let E denote the evidence at some intermediate round of the game: in general, some letters will have been guessed correctly and their places revealed in the word, while other letters will have been guessed incorrectly and thus revealed to be absent. There are two essential computations. The first is the *posterior* probability, obtained from Bayes rule:

$$P(W=w|E) \;=\; \frac{P(E|W=w)\,P(W=w)}{\sum_{w'} P(E|W=w')\,P(W=w')}.$$

The second key computation is the *predictive* probability, based on the evidence, that the letter ℓ appears somewhere in the word:

$$P\Big(L_i=\ell \text{ for some } i\in\{1,2,3,4,5\}\Big|E\Big) \;=\; \sum_w P\Big(L_i=\ell \text{ for some } i\in\{1,2,3,4,5\}\Big|W=w\Big)P\Big(W=w\Big|E\Big).$$

Note in particular how the first computation feeds into the second. Your assignment in this problem is implement both of these calculations. **You may program in the language of your choice.**

(a) Download the file *hw1_word_counts_05.txt* that appears with the homework assignment. The file contains a list of 5-letter words (including names and proper nouns) and their counts from a large corpus of Wall Street Journal articles (roughly three million sentences). From the counts in this file compute the prior probability $P(w) = \text{COUNT}(w)/\sum_{w'}\text{COUNT}(w')$. **As a sanity check, print out the fifteen most frequent 5-letter words, as well as the fourteen least frequent 5-letter words. Do your results make sense?**

(b) Consider the following stages of the game. For each of the following, indicate the best next guess—namely, the letter ℓ that is most likely (probable) to be among the missing letters. Also report the probability $P(L_i=\ell \text{ for some } i\in\{1,2,3,4,5\}|E)$ for your guess ℓ. Your answers should fill in the last two columns of this table. (Some answers are shown so that you can check your work.)

correctly guessed	incorrectly guessed	best next guess ℓ	$P(L_i=\ell$ for some $i\in\{1,2,3,4,5\}\|E)$
_ _ _ _ _	{}		
_ _ _ _ _	{E, A}		
A _ _ _ S	{}		
A _ _ _ S	{I}		
_ _ O _ _	{A, E, M, N, T}		
_ _ _ _ _	{E, O}	I	0.6366
D _ _ I _	{}	A	0.8207
D _ _ I _	{A}	E	0.7521
_ U _ _ _	{A, E, I, O, S}	Y	0.6270

(c) Turn in a scanned **printout** of your source code. **Do not forget the source code**. It is worth many points on this assignment.

Just to be perfectly clear, you are **not** required in this problem to implement a user interface or any general functionality for the game of hangman. You will only be graded on your word lists in (a), the completed table for (b), and your source code in (c).