Sometimes a sequence of two words occurs more often in a prose text than one would expect on the basis of their separate frequencies as individual words. For example, if we considered the two words 'Siamese' and 'cat' separately, determined the relative frequency of each in a large corpus, and multiplied those relative frequencies together, as if the words were chosen independently and at random from all the tokens in the corpus, the product would probably be much smaller than the relative frequency of the two-word sequence ‘Siamese cat’ (the number of occurrences of that two-word sequence divided by the number of pairs of adjacent words in the corpus).

Specifically, the Google n-grams viewer reports that, in books published in 2004, the relative frequency of the word ‘Siamese’ was \(1.00 \cdot 10^{-6}\) (that is, on the average, ‘Siamese’ occurred once in every million words of text). In the same year, the relative frequency of ‘cat’ was \(2.18 \cdot 10^{-5}\). If the distribution of the two words was independent, the relative frequency of ‘Siamese cat’ would be roughly \(2.18 \cdot 10^{-11}\) — the product of the relative frequencies of the individual words. However, the actual relative frequency of ‘Siamese cat’ was \(4.9 \cdot 10^{-8}\) — more than two thousand times greater than if the words were independent.

When the ratio of the actual relative frequency of a two-word sequence is much greater than the product of the relative frequencies of the separate words, as it is in this case, it hints that the two words are likely to belong to the same constituent at a fairly low level. A smaller ratio suggests that the sequence may be crossing a constituent boundary, with the first word at the end of a constituent and the second word at the beginning of a different constituent. We might be able to use hints like these as aids to parsing or at least as confirming or disconfirming indicators of the correctness or incorrectness of a proposed parse tree.

With this in mind, let’s study the two-word sequences in Alcott’s *Eight Cousins*, determining their relative frequencies as well as the relative frequencies of the individual words, computing the ratio described in the previous paragraph for each two-word sequence, and sorting the two-word sequences according to this ratio. We’ll be particularly interested in the two-word sequences for which this ratio is highest and those for which it is lowest.

0. Correct, adapt, and improve the tokenizer presented in class and use your version to break up the text of *Eight Cousins* into individual words. Construct a list containing each of the word tokens from that text, in order.

1. From the word list, tally the number of occurrences of each word and compute the total number of word tokens. From the results, compute the relative frequency of each word.

2. From the same word list, tally the number of occurrences of each two-word sequence and compute the relative frequency of each two-word sequence.

3. For each two-word sequence, compute the ratio between its relative frequency and the product of the relative frequencies of the words it contains. Prepare a list of the two-word sequences sorted by the value of this ratio (from least to greatest).
4. Write the list to a file, displaying on each line one two-word sequence from *Eight Cousins* and the corresponding ratio. Inspect this file to determine the ten two-word sequences for which the ratio is least and the ten for which it is greatest.

This exercise will be due at the beginning of class on Wednesday, October 12.