Problem 11.2

A black text on a white background

Description automatically generated

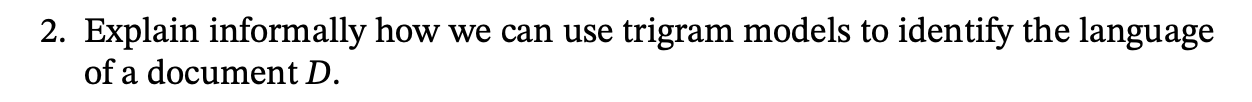
A model of the probability distribution of 3 letter sequences is thus called an tri**gram model**.

An trigram model is defined as a **Markov chain** of order 2. In a Markov chain the probability of character ci depends only on the immediately pre- ceding characters, not on any other characters. So in a trigram model, we have

P (ci | c1:i−1) = P (ci | ci−2:i−1) .

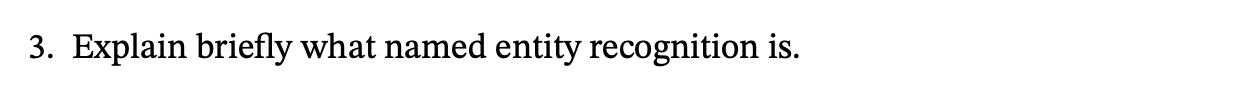
We can define the probability of a sequence of characters P (c1:N ) under the trigram model by first factoring with the chain rule and then using the Markov assumption:

P (c1:N ) = ∏i = 1to N P(ci |c1:i−1) = ∏i = 1to N P(ci |ci−2:i−1)



For a trigram character model in a language with 100 characters, **P**(Ci|Ci−2:i−1) has a million entries, and can be accurately estimated by counting character sequences in a body of text (corpus) of 10 million characters or more.

Language can be identified in a document D by first building a trigram character model of each candidate language and then count these trigrams in a corpus of that language. That gives us a P(text | language), to which further Bayes rule is applied followed by Markov assumption to find most probable language.



Named-entity recognition is the task of finding names of things in a document and deciding what class they belong to. For example, in the text “Mr. Sopersteen was prescribed aciphex,” we should recognize that “Mr. Sopersteen” is the name of a person and “aciphex” is the name of a drug.