Bigram language model

So that's what we get for n = 2:

$$p(\mathbf{w}) = \underset{p(w_1|start)}{\xcancel{p(w_1)}}p(w_2|w_1)\ldots p(w_k|w_{k-1})$$

It's normalized separately for each sequence length!

p(this) + p(that) = 1.0

p(this this) + p(this is) + … + p(built built) = 1.0

…