Estimation of N-Gram LMs

- Text Data: D
- Question: $p(w_m | w_{m-n+1},\ldots,w_{m-1})=?$

$P(X|Y)=p(X,Y)/p(Y)$

$$p(w_m | w_{m-n+1},\ldots,w_{m-1}) = \frac{p(w_{m-n+1},\ldots,w_{m-1},w_m)}{p(w_{m-n+1},\ldots,w_{m-1})}$$

- **Boils down to estimate $p(w_1,w_2,\ldots,w_m)$, ML estimate is:**

$$p(w_1,w_2,\ldots,w_m) = \frac{c(w_1 w_2 \ldots w_m, D)}{\sum_{u_i \in V} c(u_1 u_2 \ldots u_m, D)}$$

Count of word sequence "$w_1 w_2 \ldots w_m$"

Total counts of all word sequences of length m