Naive Bayes

Given sentence: "Wikipedia español el"

Predict its language: {EN, SP}

$P(Y \mid$ "Wikipedia", "español", "el"$)$

$P(EN,$ "Wikipedia", "español", "el"$)$

$= P(EN)\, P(\text{"Wikipedia"} \mid EN)\, P(\text{"español"} \mid EN)\, P(\text{"el"} \mid EN)$

$3/4 \quad 3/8 \quad \frac{3+1}{8+6} \quad \frac{0+1}{8+6} \quad 0/8 \quad \frac{0+1}{8+6} \quad 0/8$

$= 3/4 \times 4/14 \times 1/14 \times 1/14 = 0.00109$

$P(SP,$ "Wikipedia", "español", "el"$)$

$= P(SP)\, P(\text{"Wikipedia"} \mid SP)\, P(\text{"español"} \mid SP)\, P(\text{"el"} \mid SP)$

$1/4 \quad 1/3 \quad \frac{1+1}{3+6} \quad 1/3 \quad \frac{1+1}{3+6} \quad 0/3 \quad \frac{0+1}{3+6}$

$= 1/4 \times 2/9 \times 2/9 \times 1/9 = 0.00137$

$$\hat{P}(y) = \frac{\text{count}(y)}{\sum_{y \in Y} \text{count}(y)}$$

add-one (Laplace) smoothing

$$\hat{P}(w \mid y) = \frac{\text{count}(w, y)}{\sum_{w \in V} (w, y)}$$

$$\hat{P}(w \mid y) = \frac{\text{count}(w, y) + 1}{\sum_{w \in V} \text{count}(w, y) + |V|}$$

V: vocabulary

"English"
"Wikipedia"
"editor"
"free"
"español"
"de"

$|V| = 6$