Побудова моделі та обчислення велися за наступними формулами

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

where A and B are events and $P(B) \neq 0$.

- $P(A)$ and $P(B)$ are the probabilities of observing A and B without regard to each other.
- $P(A \mid B)$, a conditional probability, is the probability of observing event A given that B is true.
- $P(B \mid A)$ is the probability of observing event B given that A is true.

$$P(spam|w1 \cap w2 \cap \ldots \cap wn) = \frac{P(w1 \cap w2 \cap \ldots \cap wn|spam).\,P(spam)}{P(w1 \cap w2 \cap \ldots \cap wn)}$$

$$\frac{P(w1|spam).\,P(w2|spam)\ldots P(wn|spam).\,P(spam)}{P(w1).\,P(w2)..P(wn)}$$

$$P(spam|w1 \cap w2 \cap \ldots \cap wn)\ versus\ P(\sim spam|w1 \cap w2 \cap \ldots \cap wn)$$

Мішок слів

$$P(w) = \frac{Total\ number\ of\ occurrences\ of\ w\ in\ dataset}{Total\ number\ of\ words\ in\ dataset}$$

$$P(w|spam) = \frac{Total\ number\ of\ occurrences\ of\ w\ in\ spam\ messages}{Total\ number\ of\ words\ in\ spam\ messages}$$

TF-IDF

$$IDF(w) = \log \frac{Total\ number\ of\ messages}{Total\ number\ of\ messages\ containing\ w}$$

$$P(w) = \frac{TF(w) * IDF(w)}{\sum_{\forall\ words\ x \in train\ dataset} TF(x) * IDF(x)}$$

$$P(w|spam) = \frac{TF(w|spam) * IDF(w)}{\sum_{\forall\ words\ x \in train\ dataset} TF(x|spam) * IDF(x)}$$

$$P(w|spam) = \frac{TF(w|spam) \ + \ \alpha}{\sum_{\forall \ words \ x \in \ spam \ in \ train \ dataset} TF(x) \ + \ \alpha \sum_{\forall \ words \ x \in \ spam \ in \ train \ dataset} 1}$$

$$P(w|spam) = \frac{TF(w|spam) * IDF(w) \ + \ \alpha}{\sum_{\forall \ words \ x \in \ train \ dataset} TF(x) * IDF(x) \ + \ \alpha \sum_{\forall \ words \ x \in \ spam \ in \ train \ dataset} 1}$$

Більш детальну інформацію можна знайти за посиланнями

- https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering
- https://www.kdnuggets.com/2020/07/spam-filter-python-naive-bayes-scratch.html
- https://www.researchgate.net/publication/325270587_Ham_and_Spam_E-Mails_Classification_Using_Machine_Learning_Techniques
- https://www.kaggle.com/benvozza/spam-classification