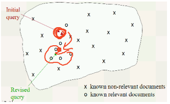
# Relevance feedback + Text classification (Chapter 9+13)

**Definition 1 (Rocchio relevance feedback)**
*Rocchio relevance feedback has the form*

$$q_m = \alpha q_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_r \in D_r} \vec{d}_r - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_{nr} \in D_{nr}} \vec{d}_{nr}$$

*where $q_0$ is the original query vector, $D_r$ is the set of relevant documents, $D_{nr}$ is the set of non-relevant documents and the values $\alpha$, $\beta$, $\gamma$ depend on the system setting.*




PSEUDO-RELEVANCE FEEDBACK

### Exercise 9/1

What is the main purpose of Rocchio relevance feedback?

### Exercise 9/2

A user's primary query is *cheap CDs cheap DVDs extremely cheap CDs*. The user has a look on two documents: doc1 a doc2, marking doc1 *CDs cheap software cheap CDs* as relevant and doc2 *cheap thrills DVDs* as non-relevant. Assume that we use a simple *tf* scheme without vector length normalization. What would be the restructured query vector after considering the Rocchio relevance feedback with values $\alpha = 1$, $\beta = 0.75$, and $\gamma = 0.25$?

---

We rewrite the exercise to the table for an easier processing.

| | relevant | non-relevant | |
|---|---|---|---|
| terms | doc1 | doc2 | query |
| CDs | 2 | 0 | 2 |
| cheap | 2 | 1 | 3 |
| software | 1 | 0 | 0 |
| thrills | 0 | 1 | 0 |
| DVDs | 0 | 1 | 1 |
| extremely | 0 | 0 | 1 |

Table 1:

$q = [2\ 3\ 0\ 0\ 1\ 1]$

$doc_1 = [2\ 2\ 1\ 0\ 0\ 0]$

$q_m = \alpha q_0 + \beta\, doc_1 - \gamma\, doc_2$

$= 1 \cdot q_0 + 0{,}75\, doc_1 - 0{,}25\, doc_2$

$= [3{,}5\ \ 4{,}25\ \ 0{,}75\ \ -0{,}25\ \ 0{,}75\ \ 1]$

$\cos sim(q_m, \ldots) \in [-1; 1]$

## Text classification and Naive Bayes (Chapter 13)

**Definition 2 (Naive Bayes Classifier)**
*Naive Bayes (NB) Classifier assumes that the effect of the value of a predictor $x$ on a given class $c$ is class conditional independent. Bayes theorem provides a way of calculating the posterior probability $P(c|x)$ from class prior probability $P(c)$, predictor prior probability $P(x)$ and probability of the predictor given the class $P(x|c)$*

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

*and for a vector of predictors $X = (x_1, \ldots, x_n)$*

$$P(c|X) = \frac{P(x_1|c)\ldots P(x_n|c)P(c)}{P(x_1)\ldots P(x_n)}.$$

*The class with the highest posterior probability is the outcome of prediction.*