TF ⇒ value over max in column

IDF → $\log_2$ (# docs / # docs word is in)

$$w_{iq} = \left(0.5 + \frac{0.5 f_{iq}}{\max(f_{iq})}\right) \times \log \frac{N}{df_i}$$

Dot product and cosine similarity is a similarity check

Query Calc = item TF IDF · $\left(\frac{(\#\text{ times word in query})}{(\#\text{ words in query})}\right)$

Dot product: Sim (item in doc, query) → sum((item · query))

Elucidean Distance: $\sqrt{\text{sum}((\text{doc item} - \text{query})^2}$

Cosine Angles: 0-90° map 0-1, 1 means same
a totally different

Cosine Similarity: $\frac{\text{Sum of (doc item} \times \text{query)}}{\sqrt{\text{sum of square items in doc}} \sqrt{\text{sum of square items in query}}}$

$$\text{Precision} = \frac{\#\text{ Relavant items retrieved}}{\#\text{ retrieved items}} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{\#\text{ relavant items retrived}}{\#\text{ relavant items}} = \frac{TP}{TP+FN}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

Confusion matrix

	The truth: Relevant	Nonrelevant	Total
Retrieved	True positives (TP)	False positives (FP)	TP + FP
Not Retrieved	False negatives (FN)	True negatives (TN)	FN + TN
Total	TP + FN	FP + TN	TP+FP+FN+TN

"hat the tem thinks

Total # of retrieved docs

Total # of relevant docs

Total # of nonrelevant docs

Total # of docs in collection

IDCG changes ordering

F measure (P): the weighted harmonic mean of precision and recall, trading off precision versus recall.

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}, \text{where } \beta^2 = \frac{1-\alpha}{\alpha}$$

where $\alpha \in [0, 1]$ and thus $\beta^2 \in [0, \infty]$.

The default balanced F measure equally weights precision and recall, which means making $\boldsymbol{\alpha = 1/2}$ **or** $\boldsymbol{\beta = 1}$. → F_1 measure $\boldsymbol{F_1} = \frac{2PR}{P+R}$

- $\beta < 1$: emphasize precision;
- $\beta > 1$: emphasize recall.

$$F_1@k = \frac{2\times(\text{Precision@}k)\times(\text{Recall@}k)}{(\text{Precision@}k)+(\text{Recall@}k)}$$

$$\text{Recall@}k = \frac{\#\text{ relevant items @}k}{\text{total \#relevant items}} = \frac{\text{TP@}k}{\text{TP + FN}}$$

All @k

Precision @k = # relavant items @k / # retrived items @k = TP / TP+FN

Mean reciprocal rank is defined as: $MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$

- $|Q|$: the total number of queries
- $rank_i$: the rank of the first relevant result retrieved in query i.

→ only 1 result

AP = sum (Precision @k when relavant) / total # relavant item)

Cumulative gain @ k is sum of relavancee until @k

$$MAP = \frac{1}{|Q|}\sum_{q=1}^{|Q|} AP(q)$$

Q: total # of queries

$AP(q)$: average precision for query q.

DCG @k is sum of (relavance score / $\log_2(k+1)$)

NDCG = DCG / IDCG