termID → documentID
(docID)

<u>reducing the dimension</u>
→ implies loss of info

<u>Compression</u>
→ loosy
→ loseless

esperanto

<u>Heaps' Law</u>

rate of growth

Jaccard Coefficient

A = This is a test

B = A test is conducted.

$J(A,B) = \frac{A \cap B}{A \cup B} = 3/5$

Term frequency
(Document freq.)

~~Log weight (frequenc~~

Log freq. weighting.

$$w_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d} & , \text{ if } tf_{t,d} > 0 \\ 0 & , \text{ otherwise} \end{cases}$$

log freq. wt. in document d

D = An apple a day keep you away from docto Apple is good for health. Nit one pr in Apple.

(apple)

$$= 1 + \log_{10} 3$$
$$= 1 + 0.4771$$
$$= \underline{\underline{1.4771}}$$

Log ~~term~~ term freq. weighting

$$\boxed{Score = \sum_{t \in (q \cap d)} (1 + \log_{10} tf_{t,d})}$$

IDF ⟶ Inverse doc. freq.

$$\boxed{idf_t = \log_{10}\left(\frac{N}{df_t}\right)}$$

tf TF-IDF

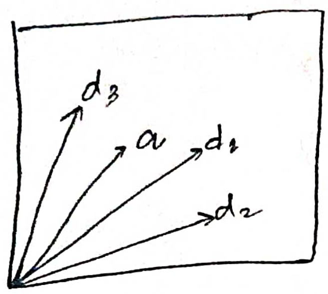
IDF weight of a term t in doc d is

$$W_{t,d} = \log(1 + tf_{t,d}) \times \log_{10}(N/df_t)$$

Score of a doc. for a given query –

$$\text{Score}(q,d) = \sum_{t \in q \cap d} tf.idf_{t,d}$$

Document as vector



Term as diff. axes

similarity.

1) distance

distance $\frac{1}{\alpha}$ similarity

less dist. more simila

Cosine Similarity.

cosine similarity betn 3 docs.

term	Sas	PaP	WH
affection	115	58	20
jealous	10	7	11
gossip	2	0	6
wuthering	0	0	38

PaP → pride & prejudice

Sas → sense & sensibility

WH → Wuthering height

awp

Log frequency weighting

term	Sas	PaP	WH
affection	3.06	2.76	2.3
jealous	2	1.84	2.04
gossip	1.3	0	1.77
wuthering	0	0	2.57

Cosine similarity for length normalized vector -

$$\cos(\vec{q},\vec{d}) = \vec{q}\cdot\vec{d}$$

$$= \sum_{i=1}^{|V|} q_i d_i$$

Length Normalization

$$\frac{3.06}{\sqrt{(3.06)^2+2^2+(1.30)^2}}$$

$$\frac{2.76}{\sqrt{(2.76)^2+(1.84)^2}}$$

term	SaS	PaP	WH
affection	0.79	0.03	0.524
jealous	0.515	0.55	0.465
gossip	0.335	0	0.405
wuthering	0	0	0.508

cos (sas, pap)

$\Rightarrow$ $0.79 \times 0.03 + 0.515 \times 0.55 + 0.+0$

$= 0.93$

cos (pap, WH) $= 0.03 \times 0.524 + 0.55 \times 0.465 + 0 + 0$

$= 0.6697$

cos (sas, WH) $= 0.79 \times 0.524 + 0.515 \times 0.465 + 0.335 \times 0.405 + 0$

$= 0.706$

item	D1	D2	D3
college	100	57	12
Election	50	30	70
farewell	10	20	00

Calculate cosine similarity of $\cos(D_1, D_2) = ?$

$\cos(D_1, D_3) = ?$

$\cos(D_2, D_3) = ?$

$\cos(D_1, D_2) = 100\times57 + 50\times30 + 10\times20$

$= 5700 + 1500 + 200$

$= 7400$

$\cos(D_1, D_3) = 100\times12 + 50\times70 + 10\times80$

$= 1200 + 3500 + 800$

$= 5500$

$\cos(D_2, D_3) = 57\times12 + 30\times70 + 20\times80$

$= 684 + 2100 + 1600$

$\Rightarrow \frac{2.07}{\sqrt{(2.07)^2 + (2.04)^2 + (2.9\ldots}}$

$\Rightarrow \frac{2.07}{10.54}$

$= \frac{2.04}{10.54}$

$= \frac{2.9}{10.54}$

$\Rightarrow \frac{3}{\sqrt{3^2+(2.69)^2+2^2}}$

$\Rightarrow \frac{3}{14}$

$= \frac{2.69}{14}$

$\Rightarrow \frac{2.69}{\sqrt{3^2+(2.69)^2+2^2}}$

$\Rightarrow \frac{2}{14}$

$\Rightarrow \frac{2}{\sqrt{3^2+(2.69)^2+2^2}}$

$\Rightarrow \frac{2.78}{\sqrt{(2.78)^2+(2.47)^2+(2.3)^2}}$

$\Rightarrow \frac{2.47}{}$

$\Rightarrow \frac{2.3}{}$

$= \frac{2.78}{14.17}$

log freq. weighting

↓

Length normalization

↓

~~normalized~~

Cosine

log freq. weight -

item	D_1	D_2	D_3
College	3	2.75	2.07
Election	2.69	2.47	2.04
farewell	2	2.30	2.90

Length normalization

item	D_1	D_2	D_3
College	0.2	0.19	0.11
Election	0.19	0.17	0.15
farewell	0.14	0.16	0.16

Cosine

$$\cos(D_1, D_2) = 0.2 \times 0.19 + 0.19 \times 0.17 + 0.14 \times 0.16$$
$$= 0.0927$$

$$\cos(D_1, D_3) = 0.2 \times 0.11 + 0.19 \times 0.15 + 0.14 \times 0.16$$
$$= 0.0729$$

$$\cos(D_2, D_3) = 0.19 \times 0.11 + 0.17 \times 0.15 + 0.16 \times 0.16$$
$$= 0.072$$

D_1 & D_2 has maximum relevance.

Probabilistic information Retrieval System (Model)

$P(R|D) \longrightarrow$ Relevant

This is a test doc.

doc

$P(NR|D) \longrightarrow$ non Relevant

→ examine the relevance of the document.

→ probabilistic Ranking principle.

→ decreasing probability of relevance.

→ highest probability at the top.

Probabilistic Ranking principle

N documents

probability of Relevance-

$P(R = X|D)$

$X \in \{0, 1\}$

$X = 1 \rightarrow$ Relevant

$X = 0 \rightarrow$ Non Relevant

$P(R=1|D)$

$P(R=0|D)$

The Score of Matching

$$\frac{P(R=1|D)}{P(R=0|D)}$$

$$\boxed{P(R|D) = \frac{P(D/R) \cdot P(R)}{P(D)}}$$

Okapi Model

→ Probabilistic IR Model

IDF → $\log_{10}\left(\frac{N}{df_t}\right)$

$$\boxed{\log_e\left(\frac{N}{df_t}\right) \cdot \frac{(K_1+1)\, tf_{t,d}}{K_1\left[(1-b) + b\left(L_d/L_{avg.}\right)\right] + tf_{t,d}}}$$

tf — term frequency

L_d → Length of the doc.

L_{avg} → avg. length of the doc.

K1 → Tuning parameter controlling the doc. term frequency scaling

b → Tuning para for doc. length

$$\sum_{i \in q} \log_e \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(K_1+1) f_i}{K + f_i}$$

$$\frac{(K_2+1)\, qf_i}{K_2 + qf_i}$$

r_i → relevance of i^{th} term

Q. Query: "President" & "Lincoln"

$R = r = 0$

$N = 5000,000$ doc

"president" word occurs in 40000 doc. $qf_i (i=1)$ = president occurs in Query = 1

"lincoln" word occurs in 300 doc $qf_i (i=2)$ = lincoln occurs in Query = 1

In a particular doc (D) that we are scoring "president" occurs 15 times ($f_1 = 15$) & "lincoln" occurs 25 times ($f_2 = 25$).

The doc length is 90% of the avg. length. i.e. $L_d / L_{avg} = 0.9$

$K_1 = 1.2 \quad b = 0.75 \quad \& \quad K_2 = 100$

$$K = K_1((1-b) + b(L_d / L_{avg}))$$

$$K = 1.11$$

$$\Rightarrow \log_e \frac{(0+0.5)/(0-0+0.5)}{(40000-0+0.5)/(5000000-40000-0+0+0.5)} \cdot \frac{(1.2+1)\times 15}{1.11+15} \cdot \frac{(100+1)\times 1}{(100+1)}$$

$$+ \log_e \frac{(0+0.5)/(0-0+0.5)}{(300-0+0.5)/(5000000-300-0+0+0.5)} \cdot \frac{(1.2+1)\times 25}{1.11+25} \cdot \frac{(100+1)\times 1}{100+1}$$

$\Rightarrow$ ~~log~~

$$\Rightarrow \log_e \frac{1}{0.008} \cdot \frac{2.2\times 15}{16.11} \times \frac{101}{101} + \log_e \frac{1}{0.00006} \times \frac{2.2\times 25}{26.11} \times \frac{101}{101}$$

$$\Rightarrow \log_e \frac{\cancel{1000}\,125}{\cancel{8}} \cdot 2.04 + \log_e \frac{100000}{6} \times 2.10$$

$$\Rightarrow \log_e 125 \times 2.04 + 9.72 \times 2.106$$

$$= 4.82 \times 2.04 + 20.47$$

$$= 9.84 + 20.47$$

$$= 30.300$$

Language Modelling

→ If query is more refine.
→ Simplest model is unigram model.
n-gram

5 words → In a document collection
→ (0.2, 0.1, 0.35, 0.25, 0.1)
"cat", "rain", "dog", "jump", "the"

Probability of occuring 2 words.

"cat rain" $= 0.2 \times 0.1 = 0.02$

"cat jump" $= 0.2 \times 0.25 = 0.5$

query = 'Apple' ⇒ q → $qf_1 = 1$

$N = 100$

Apple word occurs in 37 doc
In a particular document (D) that we want to score.
apple accurs 12 ~~apple~~ times.

The doc-length is 90% of the average length
$(L_d / L_{avg}) = 0.9$

$K_1 = 1.2$, $b = 0.75$

$K_2 = 100$

$K = K_1 \left((1-b) + b\,(L_d / L_{avg})\right)$

0.25 + 0.75 × 0.9

$= 1.11$

cal. score (q, D) using Okapi Method -

$$\log_e\left(\frac{N}{df_t}\right) \cdot \frac{(k_1+1)\, tf_{t,d}}{K_1\left[(1-b) + b\,(L_d / L_{avg})\right] + tf_{t,d}}$$

$$\Rightarrow \log_e\left(\frac{100}{37}\right) \cdot \frac{(1.2+1) \times 12}{1.11 + 12} \Rightarrow 0.99 \times \frac{2.2 \times 12}{13.11} \Rightarrow 0.99 \times 2.01 = \underline{1.9935}$$

$$\log_e \frac{(0+0.5)/(0-0+0.5)}{(37-0+0.5)/(100-37-0+0+0.5)} \times \frac{(1.2+1)12}{1.11+12} \times \frac{(100+1)1}{(100+1)}$$

$$\Rightarrow \log_e \frac{63.5}{37.5} \times \frac{26.4}{13.11} \times 1$$

$$\log_e(1.69)$$

$$= 0.52 \times 2.01$$

$$\Rightarrow \underline{1.05}$$

Performance measure

	Relevant	N.R
Retrieved	TP	FP
Not Retrieved	FN	TN

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

P ⟶ fraction of retrieved document that are actually relevant.

$$\boxed{P = \frac{TP}{TP+FP}}$$

R ⟶ fraction of relevant document

	Rev.	NRev
Ret	5	3
NR	2	5

$$Acc = \frac{10}{15}$$

$$= 2/3 = 66.67\%$$

P = # Rev. item / # Ret. $= \frac{5}{8} = 0.625$

R = # Rev item retrieved / Rev $= \frac{5}{7} = 0.71$

q - query
relevant doc = 20

relevant retrieved =

15 doc N. Rel
↓
5 (N.R.)

Harmonic mean of P & R is known as
$f_1 \rightarrow$ measure

$$F1 = 2 \frac{P * R}{P + R}$$