23-08-24

TF-IDF

#score of a doc for a given query.

$$score(q, d) = \sum_{t \in q \cap D} tf \cdot idf_t$$

where
tf = term frequency
idf = inverse doc. freq.

Document as vector

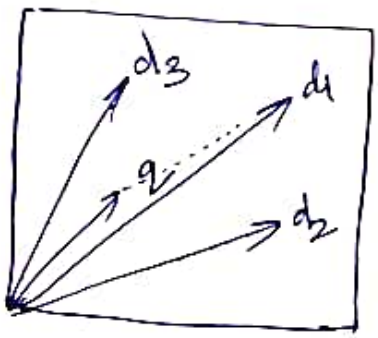
terms – axis/axes

doc – points in space

query – vectors in space

similarity – Rank the doc. based on the proximity.

mag + direction



distance $\propto \frac{1}{\text{similarity}}$

28-08-24

ECLUDIAN DISTANCE

Distance b/w (x_1, y_1) & (x_2, y_2)

$$= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- Drawback : distance will not be the only parameter to find out the similarity.
∵ large distance can also have similarity.

Cosine Similarity b/w 3 docs

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

Abbr

SaS : Sense & Sensibility } jane austen
PaP : Pride & Prejudice }
WH : Wuthering Height - Emily Bronte

28.08.24

$w = 1 + \log_{10} tf$

i) Log. freq weighting.

term	SaS	PaP	WH
affection	3.06	2.76	2.8
jealous	2	1.84	1.81
gossip	1.3	0	1.17
wuthering	0	0	2.57

ii) cosine similarity. of length normalized vector

$$\cos(\vec{q}, \vec{d}) = \vec{q} \cdot \vec{d} = \sum_{i=1}^{|V|} q_i d_i$$

where,
|V| = no. of terms

|V| = 4. (in above example).

ii) To normalize. $\frac{\text{wt}}{\sqrt{\sum w^2}}$

eg: $\frac{3.06}{\sqrt{3.06^2 + 2^2 + 1.3^2}} = \frac{3.06}{\sqrt{15.0656}} = \frac{3.06}{3.88}$

$\sqrt{19.1894} = 4.38$

$\sqrt{11.0052} = 3.32$

23-05-24

ii) length-normalized vector

term	SaS (Wi/3.87)	PaP (Wi/3.32)	WH (Wi/4.38)
Affection	0.789	0.83	0.525 = 0.53
jealous	0.517 = 0.52	0.55	0.466 = 0.47
gossip	0.3136 = 0.34.	0	0.404 = 0.40
wuthering	0	0	0.59

$cos(SaS, PaP) = 0.6557 + 0.286 + 0 + 0$

$= 0.9417$

$= \underline{0.94.}$

can assume written by same author. ∵ similarity is highest.

$cos(SaS, WH) = 0.4187 + 0.2444 + 0.136 + 0$

$= 0.7991$

$= \underline{0.80}$

$cos(PaP, WH) = 0.4399 + 0.2585 + 0 + 0$

$= 0.6984$

$= \underline{0.70}$

26-08-24.

Q. Calc the cosine similarity of D_1, D_2, D_3 for the given term freq

term	D_1	D_2	D_3
College	100	57	12
Farewell	10	30	70
Election	50	20	80.

i) log freq weighing. $w = 1 + \log_{10} tf$

term	D_1	D_2	D_3
College	3	2.76	2.07
Farewell	2	2.48	2.85
Election	2.69	2.30	2.90
$\sqrt{\Sigma w^2}$	4.49	4.37	4.56.

ii) Length-normalized vector. $w_i / \sqrt{\Sigma w^2}$

term	D_1	D_2	D_3
College	0.67	0.63	0.45
Farewell	0.45	0.57	0.625
Election	0.59	0.53	0.64

26-08-24

$\cos(D_1, D_2)$

$= 0.4221 + 0.2565 + 0.3127$

$= 0.9913$

$= 0.99.$

$\cos(D_2, D_3)$

$= 0.2835 + 0.35625 + 0.3392$

$= 0.97895$

$= \cancel{0.97} = 0.98.$

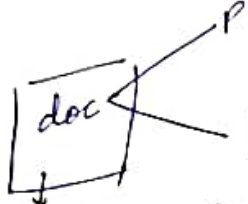
$\cos(D_1, D_3)$

$= 0.3015 + 0.28125 + 0.3776.$

$= 0.96.$

26-08-24

PROBABILISTIC RETRIEVAL MODEL



P(R|D) → Prob that a given doc is relevant

P(NR|D) → Prob that a given doc is not relevant

i) A doc is either relevant / non relevant

ii) Relevance of a doc does not convey any info about other doc being relevant.

Probabilistic Ranking principle (PRP)

- Ranking documents based on the ~~decreasing~~ decreasing probability of relevance to a query for the available data.

- Probability of relevance of a doc D for a given query. Q is

$P(R_Q = x \mid D)$ $\quad x \in \{0, 1\}$ $\quad 1 \rightarrow$ relevant, $0 \rightarrow$ Not relevant

$\frac{P(R_Q = 1 \mid D)}{P(R_Q = 0 \mid D)} \Rightarrow$ Matching score of query & document.

$P(D|Q)$ → Prob retrieval of relevant doc

$= \frac{P(R_Q=1)\, P(D \mid R_Q=1)}{P(R_Q=0)\, P(D \mid R_Q=0)}$

28-08-24

OKapi Retrieval Model

It is a sophisticated ranking func that builds on tradition TF-IDR method but emphasis the importance of less common word. It rank doc based on their relevant score, which is determined by foll^{n} eq^{n}:

$$\sum_{i \in q} \log_e \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(K_1+1) f_i}{K + f_i} \cdot \frac{(K_2+1) qf_i}{K_2 + qf_i}$$

where,

r_i - no. of relevant doc containing item i

n_i - no. of doc containing item i

N - total no. of doc in collection

R - no. of relevant doc for this query

f_i - freq of item i in the doc.

qf_i - freq of item i in the query

K_1, K_2, K - parametric parameter value set empirically.

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

L_d ↓ length of doc; L_{avg} ↓ avg. length of doc.

28-08-24

8) Query = "president" & "lincoln" $k_1 = 1.2$, $k_2 = 100$, $b =$

$N = 500\,000$ doc.

"president" occurs in 40,000 doc ($n_1 = 40\,000$)

"lincoln" occurs in 300 doc ($n_2 = 300$)

For a particular doc (D) that we are scoring "president" occurs 15 times ($f_1 = 15$) & "lincoln" occurs 5 times ($f_2 = 5$).

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

$K = k_1((1-b) + b(L_d / L_{avg}))$.

$= 1.11$

b=0 ; no normalization
b=1 ; normalized.

$R_1 = R_2 = R = 0$ (Assumig)

$n_1 = 40\,000$

$n_2 = 300$

$N = 500\,000$

$k_1 = 1.2$

$k_2 = 100$

$b = 0.75$

$K = 1.11$

$f_1 = 15$

$f_2 = 25$

$qf_1 = 1$

$qf_2 = 1$

score

$$= \log_e \frac{(r_1+0.5)/(R-r_1+0.5)}{(n_1-r_1+0.5)/(N-n_1-R+r_1+0.5)} \frac{(K_1+1)f_1}{K+f_1} \frac{(K_2+1)qf_1}{K_2+qf_1}$$

$$+ \log_e \frac{(r_2+0.5)/(R-r_2+0.5)}{(n_2-r_2+0.5)/(N-n_2-R-r_2+0.5)} \left(\frac{(K_1+1)f_2}{K+f_2}\right) \frac{K_2+1)qf_2}{K_2+qf_2}$$

$$= \log_e \frac{0.5/0.5}{40000.5/460000.5} \left(\frac{2.2 \times 15}{1.11+15}\right) \left(\frac{101}{101}\right)$$

$$+ \log_e \frac{0.5/0.5}{300.5/4999700.5} \left(\frac{2.24 \times 25}{1.11+25}\right) \left(\frac{101}{101}\right).$$

$$= \left(\frac{2.44}{1.1076} \times 2.05\right) + \left(\frac{7.42}{3.3333} \times 2.11\right)$$

$$= 2.178 + 6.7942 = 5.002 + 15.66$$

$$= 8.9672 = 20.662$$

$$= 8.97 = 20.66$$

30-08-24

8. Consider

query = "Information"

N = 100

"Information" word occurs in 37 doc

The doc length is 90% of avg length (Ld/Lavg) = 0.9

$k_1 = 1.2$, $b = 0.75$, $k_2 = 100$, $k_6 = 1.11$

In a particular doc, D that we are scoring we want to score, "Inf..." occurs 12 times = $f_1 = 12$.

Calc score of doc using Okapi method.

$$\text{score} = \log_e \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_1 - r_i + 0.5)/(N - n_1 - R + r_i + 0.5)} \times \frac{(k_1+1)f_1}{K + f_1} \times \frac{(k_2+1)}{k_2}$$

$$= \log_e \frac{63.5}{37.5} \times \frac{2.2 \times 12}{13.11} \times \frac{101}{101}$$

$$= 0.5266 \times 2.013$$

$$= 1.06.$$

30-08-24

Performance Measure

IR system

TP ⟶ doc retrieved is relevant
(true +ve)

TN ⟶ doc not retrieved & it is not relevant
(true -ve)

FP ⟶ doc retrieved & not relevant as relevant
(false +ve)

FN ⟶ Not retrieved & relevant (relevant but not retrieved)
false -ve)

	Relevant	Not Rel
Retrieved	TP	FP
Not Ret	FN	TN

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

Precision = fraction of retrieved doc that are relevant

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

30.08.24

* Recall = fraction of relevant doc that is retrieved.

$$R = \frac{TP}{TP+FN}$$

* f1 measure

↓

Harmonic mean of P & R

$$f1 = 2\left(\frac{P \times R}{P+R}\right)$$

8.

	Rev	N/Rev
Ret	7	2
N/Ret	5	6.

$$P = \frac{7}{7+2} = \frac{7}{9} = 0.777 = 0.78$$

$$R = \frac{7}{7+5} = \frac{7}{12} = 0.5833 = 0.583$$

$$Acc = \frac{7+6}{7+5+2+6} = \frac{13}{20} = \frac{6.5}{10} = 0.65$$

$$f1 = 2\left(\frac{7 \times 7}{9 \times 12}\right) \quad 2\left(\frac{0.454}{1.363}\right) = 2 \times 0.333 = 0.666$$