COMP90049 Knowledge Technologies

2019_S1_T06

Boolean / Ranked Query

Boolean Query

Ranked Query

TF-IDF

TF – term frequency

IDF – inversed document frequency

TF-IDF

TF – term frequency

More weight is given to documents where the query terms appear many times.

IDF – inversed document frequency

Less weight is given to terms that appear in many documents.

TF-IDF

TF – term frequency

More weight is given to documents where the query terms appear many times.

IDF – inversed document frequency

Less weight is given to terms that appear in many documents.

Less weight is given to documents that have many terms.

TF-IDF Model

$$w_{d,t} = \begin{cases} 1 + log_2 f_{d,t} & if\ f_{d,t} > 0 \\ 0 & otherwise \end{cases}$$

$$w_{q,t} = \begin{cases} log\left(1 + \frac{N}{f_t}\right) & if\ f_{q,t} > 0 \\ 0 & otherwise \end{cases}$$

$w_{d,t}$	Weight of a term *t* in document *d*
$f_{d,t}$	Frequency of term *t* in document *d*
$w_{q,t}$	Weight of a term *t* in query
N	Number of documents
f_t	Number of documents containing term *t*
$f_{q,t}$	Frequency of term *t* in query (0 or 1)

TF-IDF Model

$$w_{d,t} = \begin{cases} 1 + log_2 f_{d,t} & if\ f_{d,t} > 0 \\ 0 & otherwise \end{cases}$$

Term Frequency

Doc ID	apple	ibm	lemon	sun
Doc 1	4	0	0	1
Doc 2	5	0	5	0
Doc 3	2	5	0	0
Doc 4	1	2	1	7
Doc 5	1	1	3	0

TF-IDF Model

$$w_{d,t} = \begin{cases} 1 + log_2 f_{d,t} & if\ f_{d,t} > 0 \\ 0 & otherwise \end{cases}$$

Term Frequency -> Term Weight

Doc ID	apple	ibm	lemon	sun
Doc 1	$1 + log_2 4 = 3$	0	0	$1 + log_2 1$=1
Doc 2	5	0	5	0
Doc 3	2	5	0	0
Doc 4	1	2	1	7
Doc 5	1	1	3	0

TF-IDF Model

$$w_{d,t} = \begin{cases} 1 + log_2 f_{d,t} & if\ f_{d,t} > 0 \\ 0 & otherwise \end{cases}$$

Term Frequency -> Term Weight

Doc ID	apple	ibm	lemon	sun
Doc 1	$1 + log_2 4 = 3$	0	0	$1 + log_2 1 = 1$
Doc 2	$1 + log_2 5 \approx 3.32$	0	$1 + log_2 5 \approx 3.32$	0
Doc 3	$1 + log_2 2 = 2$	$1 + log_2 5 \approx 3.32$	0	0
Doc 4	$1 + log_2 1 = 1$	$1 + log_2 2 = 2$	$1 + log_2 1 = 1$	$1 + log_2 7 \approx 3.81$
Doc 5	$1 + log_2 1 = 1$	$1 + log_2 1 = 1$	$1 + log_2 3 \approx 2.58$	0

TF-IDF Model

$$w_{d,t} = \begin{cases} 1 + log_2 f_{d,t} & if\ f_{d,t} > 0 \\ 0 & otherwise \end{cases}$$

Term Weight

Doc ID	apple	ibm	lemon	sun
Doc 1	3	0	0	1
Doc 2	3.32	0	3.32	0
Doc 3	2	3.32	0	0
Doc 4	1	2	1	3.81
Doc 5	1	1	2.58	0

TF-IDF Model

Term Weight

Doc ID	apple	ibm	lemon	sun
Doc 1	3	0	0	1
Doc 2	3.32	0	3.32	0
Doc 3	2	3.32	0	0
Doc 4	1	2	1	3.81
Doc 5	1	1	2.58	0

As Vector

Doc 1 : < 3 , 0 , 0 , 1 >

Doc 2 : < 3.32 , 0 , 3.32 , 0 >

Doc 3 : < 2 , 3.32 , 0 , 0 >

Doc 4 : < 1 , 2 , 1 , 3.81 >

Doc 5 : < 1 , 1 , 2.58 , 0 >

TF-IDF Model

$$w_{q,t} = \begin{cases} log\left(1 + \frac{N}{f_t}\right) & if\ f_{q,t} > 0 \\ 0 & otherwise \end{cases}$$

Term Weight (Query)

$$w_{apple,q} = log_2\left(1 + \frac{5}{5}\right) = 1$$

$$w_{ibm,q} = 0$$

$$w_{lemon,q} = log_2\left(1 + \frac{5}{3}\right) \approx 1.42$$

$$w_{sun,q} = 0$$

Doc 1 : < 3 , 0 , 0 , 1 >

Doc 2 : < 3.32 , 0 , 3.32 , 0 >

Doc 3 : < 2 , 3.32 , 0 , 0 >

Doc 4 : < 1 , 2 , 1 , 3.81 >

Doc 5 : < 1 , 1 , 2.58 , 0 >

TF-IDF Model

$$w_{q,t} = \begin{cases} log\left(1+\frac{N}{f_t}\right) & if\ f_{q,t} > 0 \\ 0 & otherwise \end{cases}$$

Term Weight (Query)

$$w_{apple,q} = log_2\left(1+\frac{5}{5}\right) = 1$$

$$w_{ibm,q} = 0$$

$$w_{lemon,q} = log_2\left(1+\frac{5}{3}\right) \approx 1.42$$

$$w_{sun,q} = 0$$

Doc 1 : < 3 , 0 , 0 , 1 >

Doc 2 : < 3.32 , 0 , 3.32 , 0 >

Doc 3 : < 2 , 3.32 , 0 , 0 >

Doc 4 : < 1 , 2 , 1 , 3.81 >

Doc 5 : < 1 , 1 , 2.58 , 0 >

Query: < 1 , 0 , 1.42 , 0 >

TF-IDF Model

Doc 1 : < 3 , 0 , 0 , 1 >

Doc 2 : < 3.32 , 0 , 3.32 , 0 >

Doc 3 : < 2 , 3.32 , 0 , 0 >

Doc 4 : < 1 , 2 , 1 , 3.81 >

Doc 5 : < 1 , 1 , 2.58 , 0 >

Query: < 1 , 0 , 1.42 , 0 >

Cosine Similarity : $cos(Doc, q) = \frac{Doc\cdot q}{|Doc|\cdot|q|}$

TF-IDF Model

Doc 1 : < 3 , 0 , 0 , 1 >

Doc 2 : < 3.32 , 0 , 3.32 , 0 >

Doc 3 : < 2 , 3.32 , 0 , 0 >

Doc 4 : < 1 , 2 , 1 , 3.81 >

Doc 5 : < 1 , 1 , 2.58 , 0 >

Query: < 1 , 0 , 1.42 , 0 >

$cos(Doc1, q) =$ 0.55

$$cos(Doc1, q) = \frac{Doc1 \cdot q}{|Doc1| \cdot |q|} = \frac{3 \times 1 + 0 \times 0 + 0 \times 1.42 + 1 \times 0}{\sqrt{3^2 + 0^2 + 0^2 + 1^2} \times \sqrt{1^2 + 0^2 + 1.42^2 + 0^2}} \approx 0.55$$

TF-IDF Model

Doc 1 : < 3 , 0 , 0 , 1 > $cos(Doc1, q) =$ 0.55

Doc 2 : < 3.32 , 0 , 3.32 , 0 > $cos(Doc2, q) =$ 0.99

Doc 3 : < 2 , 3.32 , 0 , 0 > $cos(Doc3, q) =$ 0.30

Doc 4 : < 1 , 2 , 1 , 3.81 > $cos(Doc4, q) =$ 0.31

Doc 5 : < 1 , 1 , 2.58 , 0 > $cos(Doc5, q) =$ 0.91

Query: < 1 , 0 , 1.42 , 0 >

TF-IDF Model

Doc 1 : < 3 , 0 , 0 , 1 > $cos(Doc1, q) = 0.55$

Doc 2 : < 3.32 , 0 , 3.32 , 0 > $cos(Doc2, q) = 0.99$

Doc 3 : < 2 , 3.32 , 0 , 0 > $cos(Doc3, q) = 0.30$

Doc 4 : < 1 , 2 , 1 , 3.81 > $cos(Doc4, q) = 0.31$

Doc 5 : < 1 , 1 , 2.58 , 0 > $cos(Doc5, q) = 0.91$

Query: < 1 , 0 , 1.42 , 0 >

Document ranking: Doc 2 > Doc 5 > Doc 1 > Doc 4 > Doc 3

P@k

Precision at top k results

$$P@k = \frac{TP(in\ top\ k)}{k}$$

Document ranking: Doc 2 > Doc 5 > Doc 1 > Doc 4 > Doc 3

Recall

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$