Term-Frequency (tf) weighting

Raw Term-Frequency (tf)

- Count of occurance of a term in a document

Raw Term-Frequency (tf)

- Count of occurance of a term in a document

$$tf_{t,d}$$

Raw Term-Frequency (tf)

- Count of occurance of a term in a document

$tf_{t,d}$ = Number of times, term t occurs in document d

Raw Term-Frequency (tf)

- Count of occurance of a term in a document

Document:
Big-Data Analytics is the process of examining large volume of data

Analytics Vidhya

Raw Term-Frequency (tf)

- Count of occurance of a term in a document

Document:
Big-Data Analytics is the process of examining large volume of data

Term	Frequency
Analytics	1
Big-Data	1
data	1
examining	1
is	1
large	1
of	2
process	1
the	1
volume	1

Challenges: Raw Term-Frequency (tf)

- Relevance of a document increases with the term frequency

Challenges: Raw Term-Frequency (tf)

- Relevance of a document increases with the term frequency
    - But this relevance is not proportional to the term frequency

Log-normalized Term-Frequency

$W_{t,d}$

Log-normalized Term-Frequency

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

Log-normalized Term-Frequency

$$W_{t,d} = \begin{cases} 1 + \log_{10}(tf_{t,d}) & , \text{if } tf_{t,d} > 0 \\ 0 & , \text{Otherwise} \end{cases}$$

Log-normalized Term-Frequency

$$W_{t,d} = \begin{cases} 1 + \log_{10}(tf_{t,d}) & \text{, if } tf_{t,d} > 0 \\ 0 & \text{, Otherwise} \end{cases}$$

$tf_{t,d}$	$W_{t,d}$
0	0

Log-normalized Term-Frequency

$$W_{t,d} = \begin{cases} 1 + \log_{10}(tf_{t,d}) & \text{, if } tf_{t,d} > 0 \\ 0 & \text{, Otherwise} \end{cases}$$

$tf_{t,d}$	$W_{t,d}$
0	0
1	1

Log-normalized Term-Frequency

$$W_{t,d} = \begin{cases} 1 + \log_{10}(tf_{t,d}) & , \text{if } tf_{t,d} > 0 \\ 0 & , \text{Otherwise} \end{cases}$$

$tf_{t,d}$	$W_{t,d}$
0	0
1	1
5	1.7

Log-normalized Term-Frequency

$$W_{t,d} = \begin{cases} 1 + \log_{10}(tf_{t,d}) & \text{, if } tf_{t,d} > 0 \\ 0 & \text{, Otherwise} \end{cases}$$

$tf_{t,d}$	$W_{t,d}$
0	0
1	1
5	1.7
10	2
100	3
1000	4

Term Frequency for Ranked Retrieval

Query (q): Analytics book

Term Frequency for Ranked Retrieval

Query (q): Analytics book

Doc 1 (d1): This book is on Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term Frequency for Ranked Retrieval

Query (q): Analytics book

Doc 1 (d1): This book is on Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term	Doc1	Doc2
Analytics	1	1
Big-Data	0	1
book	1	0
data	0	1
examining	0	1
is	1	1
large	0	1
of	0	2
on	1	0
process	0	1
the	0	1
this	1	0
volume	0	1

Term Frequency for Ranked Retrieval

Query (q): Analytics book

Doc 1 (d1): This book is on Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term	Doc1	Doc2
Analytics	1	1
Big-Data	0	1
book	1	0
data	0	1
examining	0	1
is	1	1
large	0	1
of	0	1.3
on	1	0
process	0	1
the	0	1
this	1	0
volume	0	1

Term Frequency for Ranked Retrieval

Query (q): Analytics book

Doc 1 (d1): This book is on Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term	Doc1	Doc2
Analytics	1	1
Big-Data	0	1
book	1	0
data	0	1
examining	0	1
is	1	1
large	0	1
of	0	1.3
on	1	0
process	0	1
the	0	1
this	1	0
volume	0	1

Score (d1) = Sum over terms in both query and doc 1

Term Frequency for Ranked Retrieval

Query (q): Analytics book

Doc 1 (d1): This book is on Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term	Doc1	Doc2
Analytics	1	1
Big-Data	0	1
book	1	0
data	0	1
examining	0	1
is	1	1
large	0	1
of	0	1.3
on	1	0
process	0	1
the	0	1
this	1	0
volume	0	1

Score (d1) = 1 + 1 = 2

Term Frequency for Ranked Retrieval

Query (q): Analytics book

Doc 1 (d1): This book is on Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term	Doc1	Doc2
Analytics	1	1
Big-Data	0	1
book	1	0
data	0	1
examining	0	1
is	1	1
large	0	1
of	0	1.3
on	1	0
process	0	1
the	0	1
this	1	0
volume	0	1

Score (d1) = 2

Score (d2) = Sum over terms in both query and doc 2

Analytics Vidhya

Term Frequency for Ranked Retrieval

Query (q): Analytics book

Doc 1 (d1): This book is on Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term	Doc1	Doc2
Analytics	1	1
Big-Data	0	1
book	1	0
data	0	1
examining	0	1
is	1	1
large	0	1
of	0	1.3
on	1	0
process	0	1
the	0	1
this	1	0
volume	0	1

Score (d1) = 2

Score (d2) = 1

Analytics Vidhya

Term Frequency for Ranked Retrieval

Query (q): Analytics book

Doc 1 (d1): This book is on Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term	Doc1	Doc2
Analytics	1	1
Big-Data	0	1
book	1	0
data	0	1
examining	0	1
is	1	1
large	0	1
of	0	1.3
on	1	0
process	0	1
the	0	1
this	1	0
volume	0	1

Score (d1) = 2

Score (d2) = 1

Challenge: Term Frequency

- Rare terms are more informative than frequent terms (like stop words: the, is, of)

Challenge: Term Frequency

Query (q): book of Analytics

Doc 1 (d1): This book is on Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term	Doc1	Doc2
Analytics	1	1
Big-Data	0	1
book	1	0
data	0	1
examining	0	1
is	1	1
large	0	1
of	0	1.3
on	1	0
process	0	1
the	0	1
this	1	0
volume	0	1

Challenge: Term Frequency

Query (q): book of Analytics

Doc 1 (d1): This book is on Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term	Doc1	Doc2
Analytics	1	1
Big-Data	0	1
book	1	0
data	0	1
examining	0	1
is	1	1
large	0	1
of	0	1.3
on	1	0
process	0	1
the	0	1
this	1	0
volume	0	1

Score (d1) = 1 + 1 = 2

Analytics Vidhya

Challenge: Term Frequency

Query (q): book of Analytics

Doc 1 (d1): This book is on Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term	Doc1	Doc2
Analytics	1	1
Big-Data	0	1
book	1	0
data	0	1
examining	0	1
is	1	1
large	0	1
of	0	1.3
on	1	0
process	0	1
the	0	1
this	1	0
volume	0	1

Score (d1) = 2

Score (d2) = 1 + 1.3 = 2.3

Analytics Vidhya

Challenge: Term Frequency

Query (q): book of Analytics

Doc 1 (d1): This book is on Analytics

Doc 2 (d2): Big-Data Analytics is the process of examining large volume of data

Term	Doc1	Doc2
Analytics	1	1
Big-Data	0	1
book	1	0
data	0	1
examining	0	1
is	1	1
large	0	1
of	0	1.3
on	1	0
process	0	1
the	0	1
this	1	0
volume	0	1

Score (d1) = 2

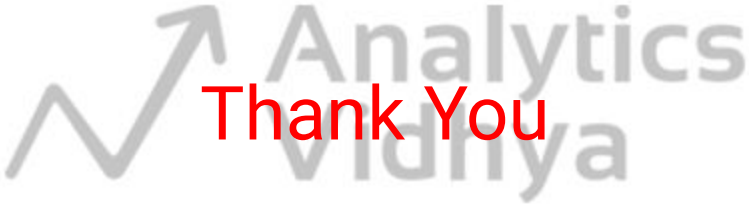
Score (d2) = 2.3

Analytics Vidhya

Challenge: Term Frequency

- Rare terms are more informative than frequent terms (like stop words: the, is, of)

- More weight to rare terms than frequent terms



Thank You

Analytics
Vidhya