Compute TFIDF scores and build inverted index

TF_t ➔ Frequency of *t* in a document

$IDF_t = \log (N/DF_t)$

TF-IDF for the word 'hadoop' in Doc1:

$TF_{how} = 1$

$N = 5$

$TF_{how} = 1$

$IDF_{how} = log(5/1) = 0.699$

$TF\text{-}IDF_{how} = TF_{how} \times IDF_{how} = 0.699$

Doc1	hits	TF	IDF	TF-IDF
hadoop	1	1	0.699	0.699
is	3	1	0.222	0.222
taking	1	1	0.699	0.699
the	3	1	0.222	0.222
big	2	1	0.398	0.398
data	2	1	0.398	0.398
world	2	1	0.398	0.398
by	1	1	0.699	0.699
storm	2	1	0.398	0.398

Doc2	hits	TF	IDF	TF-IDF
there	1	1	0.699	0.699
is	3	1	0.222	0.222
a	1	1	0.699	0.699
big	2	1	0.398	0.398
storm	2	1	0.398	0.398
coming	1	1	0.699	0.699
this	2	1	0.398	0.398
weekend	2	1	0.398	0.398

Doc3	hits	TF	IDF	TF-IDF
data	2	1	0.398	0.398
is	3	1	0.222	0.222
the	3	1	0.222	0.222
new	1	1	0.699	0.699
oil	1	1	0.699	0.699

Doc4	hits	TF	IDF	TF-IDF
how	1	1	0.699	0.699
does	1	2	0.699	1.398
the	3	3	0.222	0.666
weather	1	4	0.699	2.796
look	1	5	0.699	3.495
like	1	6	0.699	4.194
this	2	7	0.398	2.786
weekend	2	8	0.398	3.184

Doc5	hits	TF	IDF	TF-IDF
hello	1	1	0.699	0.699
world	2	2	0.398	0.796

There are variations to computing TFIDF scores. You can use any variant

https://en.wikipedia.org/wiki/Tf%E2%80%93idf

Retrieve the matching documents based on TFIDF scores

Query (Q): the big data

$$\text{Score(Query, Doc)} = \frac{|q \cap Doc|}{|q|} \sum_{q \in Q} TFIDF(q, Doc)$$

Score(Query, Doc1) = (0.398 + 0.398 + 0.222) * 3/3

Score(Query, Doc3) = (0.398 + 0.222) * 2/3

Score(Query, Doc2) = (0.398) * 1/3

Score(Query, Doc4) = (0.222) * 1/3

Doc1	hits	TF	IDF	TF-IDF
hadoop	1	1	0.699	0.699
is	3	1	0.222	0.222
taking	1	1	0.699	0.699
the	3	1	0.222	0.222
big	2	1	0.398	0.398
data	2	1	0.398	0.398
world	2	1	0.398	0.398
by	1	1	0.699	0.699
storm	2	1	0.398	0.398

Doc3	hits	TF	IDF	TF-IDF
data	2	1	0.398	0.398
is	3	1	0.222	0.222
the	3	1	0.222	0.222
new	1	1	0.699	0.699
oil	1	1	0.699	0.699

Doc2	hits	TF	IDF	TF-IDF
there	1	1	0.699	0.699
is	3	1	0.222	0.222
a	1	1	0.699	0.699
big	2	1	0.398	0.398
storm	2	1	0.398	0.398
coming	1	1	0.699	0.699
this	2	1	0.398	0.398
weekend	2	1	0.398	0.398

Doc4	hits	TF	IDF	TF-IDF
how	1	1	0.699	0.699
does	1	2	0.699	1.398
the	3	3	0.222	0.666
weather	1	4	0.699	2.796
look	1	5	0.699	3.495
like	1	6	0.699	4.194
this	2	7	0.398	2.786
weekend	2	8	0.398	3.184

Doc5	hits	TF	IDF	TF-IDF
hello	1	1	0.699	0.699
world	2	2	0.398	0.796

There are variations to combining the score.