Feedback — Information Retrieval

Help

You submitted this quiz on **Sun 29 Apr 2012 1:23 AM PDT**. You got a score of **4.00** out of **4.00**.

Question 1

Given the two documents:

- ***We all live in that yellow submarine.***
- ***The yellow mustard in that submarine sandwich was not yellow!***

What is the Jaccard similarity between them (ignoring punctuation)?

Your Answer		Score	Explanation
○ $\frac{4}{17}$			
◉ $\frac{1}{3}$	✔	1.00	Correct. Exactly 4 distinct words (***in***, ***that***, ***yellow***, and ***submarine***) occur in both documents, and there are 12 distinct word types, so we get the Jaccard coefficient is $4/12 = 1/3$.
○ $\frac{9}{17}$			
○ $\frac{8}{17}$			
Total		1.00 / 1.00	

Question Explanation

The Jaccard coefficient between two documents is given by $\frac{|D_1 \cap D_1|}{|D_1 \cup D_2|}$, where D_1 is the set of words occurring in document 1 and D_2 is the set of words occurring in document 2 - i.e. it is the number of distinct word types occurring in both documents divided by the number of distinct word types occurring in either of the documents.

Question 2

In a set of $806,791$ documents, we get the following data on a few terms and a few documents:

term	document frequency	Doc 1	Doc 2	Doc 3
car	$18,165$	27	4	24
auto	6723	3	33	0
insurance	$19,241$	0	39	29
best	$25,235$	14	0	17

What is the tf-idf value for the term ***insurance*** in Document 2?

(Please answer to 3 decimal places - e.g. $5.18254 \rightarrow 5.183$ or $1.128333 \rightarrow 1.128$).

Note: Be sure to use $\log_{10}$ (log base 10) in your calculations.

You entered:

4.204

Your Answer		Score	Explanation
4.204	✔	1.00	Correct. The TF-IDF is given by $(1+\log_{10} 39) \times \log_{10}(806,791/19,241)$ $\approx 2.591 \times 1.623$
Total		1.00 / 1.00	

Question Explanation

TF-IDF for a term t in a document d is given by
$w_{t,d} = (1+\log_{10} \mathrm{tf}_{t,d}) \times \log_{10}(N/\mathrm{df}_t)$

Question 3

Given the following term frequencies (counts) for a few words in a collection of 4 documents,

term	Dawn	Beatrice	She	Regeneration
happiness	37	30	0	3
surprise	40	10	6	0
family	31	0	12	17
adventure	0	5	13	0

What is the cosine similarity between *Beatrice* and *She*? Use tf-idf weighting and assume that these are the only documents and words in the collection.

Your Answer		Score	Explanation
0.49			
0.57			
0.81	✔	1.00	Correct!
0.72			
Total		1.00 / 1.00	

Question 4

What is the average precision for the following sequence of retrieved documents, where **R** denotes a relevant document and **N** denotes an irrelevant document?

R N R R N N N R R N N R

Your Answer		Score	Explanation
0.497			
0.580			
0.500			
0.662	✔	1.00	Correct!
Total		1.00 / 1.00	

Question Explanation

Recall that average precision is calculated by scanning through the list of retrieved documents (in the order that they are reported) and computing the precision at every point that we retrieve a relevant document. We then get the Average Precision by averaging over those precisions. Also recall that precision in this context is the number of relevant documents retrieved so far divided by the number of documents retrieved so far.