2.4 RAT

1. Given the two documents:

   q = *To be or not to be*.

   d = *To think and therefore to be*.

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

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	6,723	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?

   Recall: $w_{t,d} = (1 + \log_{10}\mathrm{tf}_{t,d}) \times \log_{10}(N/\mathrm{df}_t)$

   1. 39/19241
   2. (1+log10(39))*log10(806791/19241.)
   3. (1+log10(806791/19241.))*log10(39)
   4. (1+log10(19241))*log10(806791/39.)

3. What is the cosine similarity between the query and document? Use tf-idf weighting (lnc.ltc variation, see table below):

Term Frequencies & Document Frequencies

Term	Query						Document				Prod
	tf-raw	tf-wt	df	idf	wt	n'lize	tf-raw	tf-wt	wt	n'lize	
happiness	0	0.00	3,000	0.12	0.00	0.000	0	0.00	0.00	0.000	0.000
surprise	0	0.00	3,000	0.12	0.00	0.000	6	1.78	1.78	0.514	0.000
family	1	1.00	3,000	0.12	0.12	0.383	12	2.08	2.08	0.601	0.231
adventure	1	1.00	2,000	0.30	0.30	0.924	13	2.11	2.11	0.611	0.565

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 N R N

Term frequency		Document frequency		Normalization	
n (natural)	$tf_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(tf_{t,d})$	t (idf)	$\log \frac{N}{df_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5 + \frac{0.5 \times tf_{t,d}}{\max_t(tf_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-df_t}{df_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases} 1 & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha < 1$
L (log ave)	$\frac{1+\log(tf_{t,d})}{1+\log(ave_{t \in d}(tf_{t,d}))}$				