1. [25 points]
   (a) (10 points)No standard solution for results listing

   (5 points)For evaluation, we can use precision and recall. Students may not be able to use specific term "precision" and "recall", but they may have a general and intuitive idea about that. As long as students can provide solutions related to the concept of precision and recall, we can give them full mark.

   (5 points)Google is the best since the first few results are exactly the information I want to find. However, both Bing and Yahoo don't return the page we expect. No Standard example.

   (b) (5 points)Sometimes, user may not be able to type the queries clear and clean, thus most of the time search engine need to preprocess the queries first. Open Questions.

2. [15 points]
   (a) (1.5 points*6)max(tf) = 5 for every term

   $w_{1,D}$ = 2/5 * $\log_2$ (100/10) = 1.33
   $w_{2,D}$ = 0
   $w_{3,D}$ = 0
   $w_{4,D}$ = 5/5* $\log_2$ (100/20) = 2.32
   $w_{5,D}$ = 2/5 * $\log_2$ (100/1) = 2.66

   D = ⟨ 1.33, 0, 0, 2.32, 2.66 ⟩

   [I use the notation $w_{1,D}$ to represent 'the weight of term 1 in document D', which is consistent with the notation used in the lecture notes.]

   (b) (2 points)Inner product: Sim (Q, D) = 1.33 + 2.66 = 3.99

   (2 points)Cosine: Sim (Q, D) = 3.99 / [$(1.33^2 + 2.32^2 + 2.66^2)^{0.5}$ * $(4)^{0.5}$] = 3.99/7.54 = 0.53

   (2 points)Most contributed term: T5, or T1&T5.

3. [40 points]

   (a) (0.5 points*21)

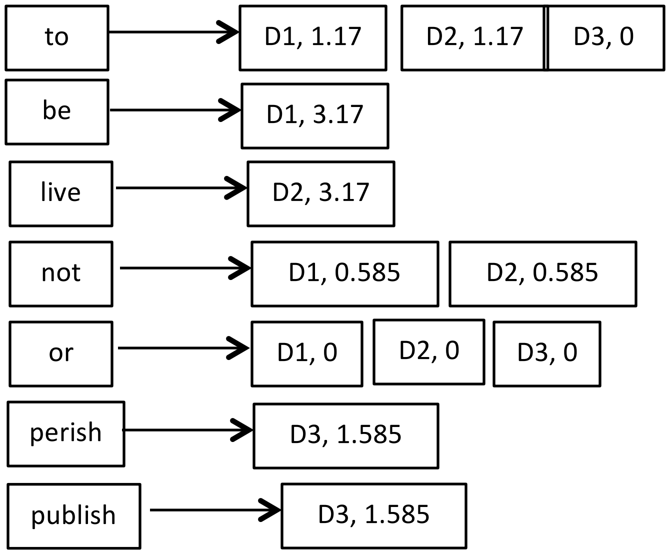
	be	live	not	or	perish	publish	to
D1	2*$\log_2$(3)	0	1*$\log_2$ (3/2)	1*$\log_2$ (3/3)	0	0	2*$\log_2$ (3/2)
D2	0	2*$\log_2$ (3)	1*$\log_2$ (3/2)	1*$\log_2$ (3/3)	0	0	2*$\log_2$ (3/2)
D3	0	0	0	1*$\log_2$ (3/3)	1*$\log_2$(3/1)	1*$\log_2$(3/1)	0

D1 = <3.17, 0, 0.585, **0**, 0, 0, **1.17**>

D2 = <0, **3.17**, 0.585, **0**, 0, 0, **1.17**>

D3 = <0, 0, 0, **0**, 1.585, 1.585, 0>

(b) (1 points*7)Design an inverted list for the seven terms above, in the form of:



(c) (12.5 points)The pseudo code:

```
For each term $Q_i$ in Q
Look up $Q_i$ from inverted index
If not found: continue
•    If found: retrieve postings list for $Q_i$
        For each document $D_j$ on the postings list
                Compute partial score between $D_j$ and $Q_i$
                score ($D_j$, Q) += partial score ( $D_j$ , $Q_i$ )
        end {for}
end {for}
Perform normalization if needed
```

(d) D1 = <1.057,1.057,0.39,0, 0.528, 0.528, 0.78>

Jaccard coefficient=2.503/(3.55+4-2.503)=0.4959

4. [20 points]

tf: (1.5points*8)

	apple	juice
D1	10	0
D2	4	3

$w_{x,j}$

D1	1	0

D2	1	0.75

Extended Boolean
sim(q_{and}, d_j) = $1 - \{ [(1-x)^2 + (1-y)^2]/2 \}^{0.5}$
sim(q_{and}, D1) = 1 – (((1-1)**2 + 1)/2)**0.5 = 0.293(2 points)
sim(q_{and}, D2) = 1 – (((1-1)**2 + (1-0.75))/2)**0.5 = 0.823(2 points)

(2 points)Vector space model with inner product and tf as weights:
sim(q, D1) = 10
sim(q, D2) = 7

The extended Boolean model ranks D2 higher than D1, whereas the vector space model with inner product and tf weights ranks D1 higher than D2 and D2 does not talk about apple juice.