Paper Code	Examiner	Department	Office
INT309	Fangyu Wu	Department of Intelligence Science	SD555

1. What is the difference between a term and a token? (L1. Introduction)

> **Solution:**
>
> Given a character sequence and a defined document unit, chopping it up into pieces, called tokens. However terms, rather than being exactly the tokens that appear in the document, they are usually derived from them by various normalization processes.

2. What are the two key statistics to describe the effectiveness of an information retrieval system? Also give the definitions of these two statistics. (L3. Vocabulary)

> **Solution:**
>
> Precision ratio : Fraction of retrieved docs that are relevant to the users information need.
> Recall ratio: Fraction of relevant docs in collection that are retrieved.

3. Explain the effect of skip pointers. What are the implications of short skip spans and long skip spans? (L2. Query Processing)

Solution: To skip postings that will not figure in the search results.

More skips - shorter skip spans - more likely to skip - But lots of comparisons to skip pointers;
Fewer skips - few pointer comparison, - but then long skip spans - few successful skips.

4. In a Boolean retrieval system, how does stemming affect the precision and recall? (L3. Vocabulary)

Solution: 1)Stemming can increase the retrieved set without increasing the number of relevant documents. So precision may increase/remain the same/decrease.
2)Stemming can only increase the retrieved set, which means increased or unchanged recall.

5. Consider these documents:
   Doc 1 breakthrough drug for schizophrenia
   Doc 2 new schizophrenia drug
   Doc 3 new hopes for schizophrenia patients

   Draw the inverted index representation for this collection. (L2. Query Processing)

Solution:

Inverted Index:

[2] breakthrough → 1

[2] drug → 1 → 2

[2] for → 1 → 3

[2]hope → 3

[2] new →2 → 3

[2]patient → 3

[2]schizophrenia → 1→ 2 → 3

6. What is stemming? Will stemming increase the size of a vocabulary? (L3. Vocabulary)

Solution:

[3]Stemming usually chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.

[3]Stemming decreases the size of the vocabulary.

7. Explain the effect of the size of the document unit. What is the implication of a too large or too small document unit? (L3. Vocabulary)

Solution:

It becomes clear that there is a precision/recall tradeoff here. If the units get too small, we are likely to miss important passages because terms were distributed over several mini-documents;While if units are too large we tend to get spurious matches and the relevant information is hard for the user to find.

8. How could we answer phrase queries like **"school closed"**? (L3. Vocabulary)

Solution: Biword indexes: dictionary blow up, infeasible for more than biwords; Positional indexes: expands postings storage substantially.

9. Compute the edit distance between **form** and **from**. Write down the 4×4 array of distances between all prefixes as computed by the algorithm showing in Fig Q9. (L4. Index Construction)

```
EDITDISTANCE(s1, s2)
 1  int m[i, j] = 0
 2  for i ← 1 to |s1|
 3  do m[i, 0] = i
 4  for j ← 1 to |s2|
 5  do m[0, j] = j
 6  for i ← 1 to |s1|
 7  do for j ← 1 to |s2|
 8     do m[i, j] = min{m[i − 1, j − 1] + if (s1[i] = s2[j]) then 0 else 1fi,
 9                      m[i − 1, j] + 1,
10                      m[i, j − 1] + 1}
11  return m[|s1|, |s2|]
```

Fig Q9

Solution:

	f	o	r	m
f	0	1	2	3
r	1	1	1	2
o	2	1	2	2
m	3	2	2	2

End of paper