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

1. What is the difference between a term and a token? (Lecture 3)

   A term is a normalized token.
   Tokens are defined as the smallest unit possible that can be processed by a NLP retrieval system, whereas terms are specific words that go through normalization steps including stemming and lemmatization from tokens.

2. What are the two key statistics to describe the effectiveness of an information retrieval system? Also give the definitions of these two statistics. (Lecture 3)

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False negatives}}$$

3. Explain the effect of skip pointers. What are the implications of short skip spans and long skip spans? (Lecture 2)

   Skip pointers are the tools in the posting list retrieval structures that are used to simplify the comparison steps. They jump into a bigger index for the simplicity of comparison. Shorter skip spans means more comparisons needed, whereas longer skip spans may lead to fewer successful skips.

(Lecture 3)

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

   Stemming may reduce the precision, since it may increase the amount of false positives. On the other hand, it usually increase the recall since TP+FN (the amount of all the relevant docs) are fixed, whereas the amount of TP either increases or stays the same.

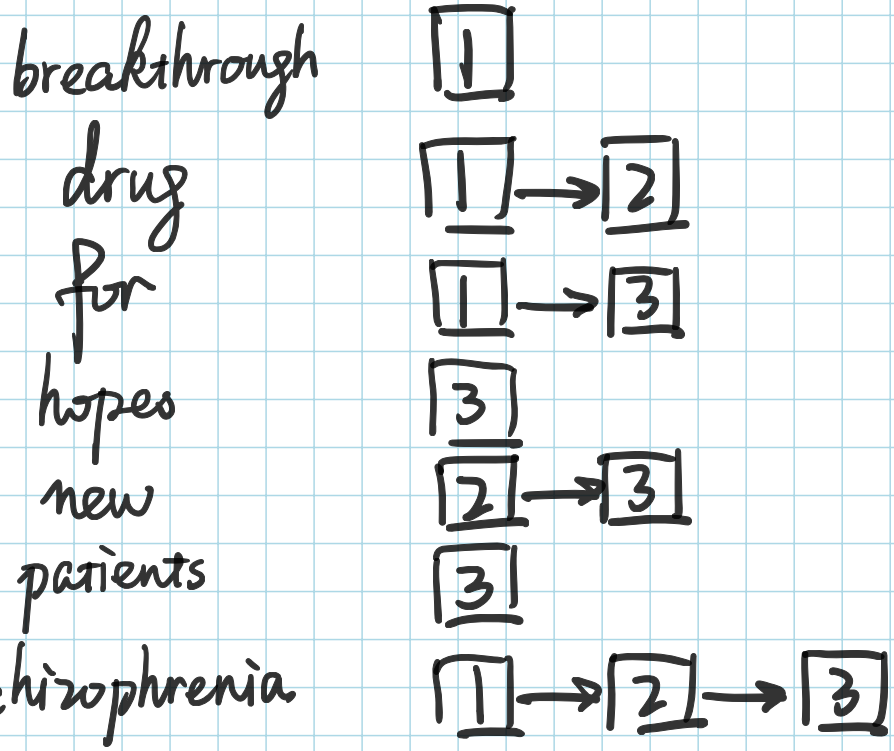
5. Consider these documents:
   Doc 1 breakthrough drug for schizophrenia
   Doc 2 new schizophrenia drug
   Doc 3 new hopes for schizophrenia patients
   Draw the inverted index representation for this collection. (Lecture 1)

5.

Term-incidence matrix:

term	doc1	doc2	doc3
breakthrough	1	0	0
drug	1	1	0
for	1	0	1
hopes	0	0	1
new	0	1	1
patients	0	0	1
schizophrenia	1	1	1

Inverted index representation



6. What is stemming? Will stemming increase the size of a vocabulary? (Lecture 3)

Stemming is the process that reduces a variant form into its base form. It doesn't increase the size of a vocabulary. It's always the case that multiple words share the same base form. (am, is, are -> be, canto, cantas, canta, cantamos, cantáis, cantan -> cantar...)

7. Explain the effect of the size of the document unit. What is the implication of a too large or too small document unit? (Lecture 3)

The size of the document unit affects the information contained and the amount of the units. If a document is too small, it becomes easy to miss the important messages, whereas if it's too large it makes it hard for the users to search for information.

8. How could we answer phrase queries like **"school closed"**? (Lecture 3)

2 approaches:
1. We can use biword indices, tokenizing every two consecutive words as a single term and perform query based on that.
2. Or we can also use positional indices to mark the position of appearance for each word, and look for the positions where the phrase terms occur in the documents.

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.

(Lecture 4)

```
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

9.

		f	r	o	m
	0	1	2	3	4
f	1	0	1	2	3
o	2	1	1	1	2
r	3	2	1	2	2
m	4	3	2	2	2