## Week 2 Quiz



 $1_{\circ}$  Let  $w_1$ ,  $w_2$ , and  $w_3$  represent three words in the dictionary of an inverted index. Suppose we have the following document frequency distribution:

Word	Document Frequency
$\mathbf{W}_1$	1000
$W_2$	100
$W_3$	10

Assume that each posting entry of document ID and term frequency takes exactly the same disk space. Which word, if removed from the inverted index, will save the **most** disk space?

	$W_1$
	$W_2$
	$W_3$
$\bigcirc$	We cannot tell from the given information.

1 point

2.

the query $Q = w_1 w_2$ then the <b>minimum</b> possible number of accumulators needed to score all the matching documents is:
1100
1000
<u> </u>
100
1 point
3. The gamma code for the term frequency of a certain document is <b>1110010</b> . What is the term frequency of the document?
<u> </u>
<u> </u>
9
<ul><li>10</li></ul>
1 point
4. When using an inverted index for scoring documents for queries, a shorter query always uses fewer score accumulators than a longer query.
True
False

Assume we have the same scenario as in Question 1. If we enter

1 point	
5. What is the before inde:	advantage of tokenization (normalize and stemming) x?
✓ Red	uces the number of terms (size of vocabulary)
	roves performance by mapping words with similar anings into the same indexing term
Extr	acts words as lexical units from strings of text
1 point	
•	an inverted index alone do for fast search?
Sea	rch document contains "A" and "B"
Sean	rch document contains "A" or "B"
Retr	ieve documents that are relevant to the query
1 point  7. If Zipf's law slower?  Fast Slove	

1 point

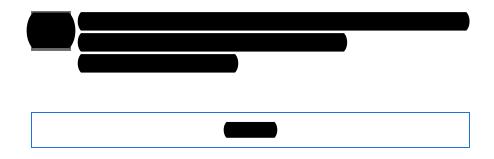
In BM25, the TF after transformation has upper bound
○ k
<ul><li>k +1</li></ul>
<u> </u>
1 point
9. Which of the following are weighing heuristics for the vector space model?
✓ TF weighting and transformation
✓ Document length normalization
✓ IDF weighting
1 point 10.
point
10。 Which of the following integer compression has equal-length
10. Which of the following integer compression has equal-length coding?
10. Which of the following integer compression has equal-length coding?  Binary
point $10_{\circ}$ Which of the following integer compression has equal-length coding? $\qquad \qquad \qquad$

11.

Consider the following retrieval formula:

$$score(Q,D) = \sum_{w \in Q,D} \frac{\log(c(w,D)+1)}{1 + \frac{avdl}{dl}} \log \frac{df(w)}{N+1}$$

Where c(w, D) is the count of word w in document D,			
dl is the document length,			
avdl is the average document length of the collection,			
N is the total number of documents in the collection,			
and df (w) is the number of documents containing word w.			
In view of TF, IDF weighting, and document length normalization, which part is missing or does not work appropriately?			
○ TF			
IDF			
O Document length normalization			
1 point			
12. Suppose we compute the term vector for a baseball sports news article in a collection of general news articles using <b>TF-IDF weighting</b> . Which of the following words do you expect to have the highest weight in this case?			
<ul><li>baseball</li></ul>			
computer			
( ) the			



r p