## CSCI 5417-IR Fall 2009 Quiz 1

CSCI	3417-1K Faii 2009 Quiz 1
Name	·
	honor, as a University of Colorado at Boulder student, I have neither given nor ed unauthorized assistance on this work.
	(5 Points) What is your favorite movie? (5 Points) True/False: Dividing inverted indexes into separate dictionary and postings files is done in order to speed up the process of creating an index.
3.	(5 Points) In what way should postings be ordered in order to facilitate the processing of boolean queries? Why?
4.	<ul> <li>(10 Points) The formula on the attached page describes Lucene's default document scoring mechanism. Think about this formula in the context of the basic vector space cosine measure and answer the following questions.</li> <li>a) Give a motivation as to why idf term is being squared.</li> </ul>
	<b>b)</b> What's the likely role of the <i>lengthNorm</i> factor in this equation?
	by what is the fixery fore of the tenguirtoim factor in this equation:

5.	(10 points) Describe two ways to speed up the cosine algorithm on the attached page. For each of your suggested speedups, characterize the impact that your method has on the output of the algorithm.

**6. (5 points)** What aspect of ad hoc retrieval system performance is *relevance feedback* primarily intended to address?

7.

## Lucene scoring equation

```
\sum_{t \in q} (tf(t \in d) \times idf(t)^2 \times boost(t.field \in d) \times lengthNorm(t.field \in d))
\times coord(q, d) \times queryNorm(q)
```

## Cosine algorithm

```
COSINESCORE (q)

1 float Scores [N] = 0

2 Initialize Length [N]

3 for each query term t

4 do calculate w<sub>t,q</sub> and fetch postings list for t

5 for each pair (d, tf<sub>t,d</sub>) in postings list

6 do Scores [d] += wf<sub>t,d</sub> × w<sub>t,q</sub>

7 Read the array Length [d]

8 for each d

9 do Scores [d] = Scores [d] / Length [d]

10 return Top K components of Scores []
```