# CSE 6240 - Spring 2015
# Web Search & Text Mining

Homework 1

01/12/2015

**Due: 01/18/2015 23:59**

1. Positional Indexing (25pt)

Shown below is a portion of a positional index in the format: term: doc1: ⟨position1, position2, . . . ⟩; doc2: ⟨position1, position2, . . . ⟩; etc.

angels: 2: ⟨36,174,252,651⟩; 4: ⟨12,22,102,432⟩; 7: ⟨17⟩;
fools: 2: ⟨1,17,74,222⟩; 4: ⟨8,78,108,458⟩; 7: ⟨3,13,23,193⟩;
fear: 2: ⟨87,704,722,901⟩; 4: ⟨13,43,113,433⟩; 7: ⟨18,328,528⟩;
in: 2: ⟨3,37,76,444,851⟩; 4: ⟨10,20,110,470,500⟩; 7: ⟨5,15,25,195⟩;
rush: 2: ⟨2,66,194,321,702⟩; 4: ⟨9,69,149,429,569⟩; 7: ⟨4,14,404⟩;
to: 2: ⟨47,86,234,999⟩; 4: ⟨14,24,774,944⟩; 7: ⟨199,319,599,709⟩;
tread: 2: ⟨57,94,333⟩; 4: ⟨15,35,155⟩; 7: ⟨20,320⟩;
where: 2: ⟨67,124,393,1001⟩; 4: ⟨11,41,101,421,431⟩; 7: ⟨16,36,736⟩;

Which document(s) if any meet each of the following queries, where each expression within quotes is a phrase query?

a. “fools rush in”
b. “where angels fear”
c. “fools rush in” AND “angels fear to tread”

2. TF-IDF (25pt)

Consider idf as the most commonly used version: $idf_t = \log\left(\frac{N}{df_t}\right)$.

a. Why is the idf of a term always finite?
b. What is the idf of a term that occurs in every document? Compare this with the use of the stop word lists.
c. Consider the table of tf, df for 3 documents denoted Doc1, Doc2, Doc3 in Doc3 in Figure 1 and Figure 2. Compute the tf-idf weights for the four terms, where the total number of documents N = 806,791.
d. Can the tf-idf weight of a term in a document exceed 1?

| term | Doc1 | Doc2 | Doc3 |
|---|---|---|---|
| car | 27 | 4 | 24 |
| moto | 3 | 33 | 0 |
| insurance | 0 | 33 | 29 |
| rent | 14 | 0 | 17 |

Figure 1. Table of tf values for Question 2.

| term | df |
|---|---|
| car | 18,165 |
| moto | 6,723 |
| insurance | 19,241 |
| rent | 25,235 |

Figure 2. Table of df values for Question 2.

3. Evaluation - Example (25pt)
There are two indexing systems, S1 and S2. Both of them make two queries Q1 and Q2 on a document set, and return top 5 results as is shown below.

| system, query | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| S1, Q1 | d3 | d5 | d8 | d10 | d11 |
| S1, Q2 | d1 | d2 | d7 | d11 | d13 |
| S2, Q1 | d6 | d7 | d2 | d9 | d8 |
| S2, Q2 | d1 | d2 | d4 | d11 | d14 |

Figure 3. Table of query results for Question 3.

Suppose the relevant documents of Q1 is {d3, d6, d7, d8}, and Q2 {d1, d4, d11}. Calculate precision, recall, F Measure (harmonic mean), Average Precision and Mean Average Precision of S1 and S2 on Q1 and Q2, respectively (no need to consider interpolation). Which values consider ranks? Which don't?

4. Evaluation - General (25pt)
Answer the following questions:
a. The balanced F measure (a.k.a. F1) is defined as the harmonic mean of precision and recall. What is the advantage of using the harmonic mean rather than "averaging" (using the arithmetic mean)?
b. What are the possible values for interpolated precision at a recall level of 0?
c. Must there always be a break-even point between precision and recall? Either show there must be or give a counter-example.