# Web Information Management: Homework #1

Due on April 14, 2015

*Professor Fang TTh 12:10*

**Rick Sullivan**

# Problem 1

Assume that a search engine returns a ranked list of 10 total documents for a given query. According to the ground truth labeling, there are 7 relevant documents for this query, and that the relevant documents in the ranked list are in the 1st, 3rd, 5th, 8th, and 10th positions in the ranked results.

1. Calculate Precision, Recall, F-measure, nDCG, at the 10 retrived documents.
2. Calculate the interpolated precision value for each of the following standard recall levels:

$$\{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$$

   for this individual query.
3. Calculate the average precision.

**Part 1**
Precision is determined by

$$\frac{\#\ of\ returned\ relevant\ documents}{total\ \#\ of\ returned\ documents}.$$

We have 5 relevant documents returned, and 10 total returned documents, so $P = 0.5 = 50\%$.

Recall is determined by

$$\frac{\#\ of\ returned\ relevant\ documents}{total\ \#\ of\ relevant\ documents}.$$

We have 5 relevant documents returned, and 7 total relevant documents, so $R = 5/7 \approx 0.714 = 71.4\%$.

F-measure is the harmonic mean of recall and precision,

$$\begin{aligned} F &= \frac{2PR}{P+R} \\ &= \frac{2(0.5)(0.714)}{0.5+0.714} \\ &\approx 0.5881 \end{aligned}$$

nDCG

# Problem 2

The *San Jose Mercury News* repository from 2000 to 2005 (i.e., 5 years) contains about 400 million word tokens, with the vocabulary size about 1 million. What would be a good estimation of the vocabulary size one would get in indexing the *San Jose Mercury News* repository from 2000 to 2010 (i.e., 10 years)?

**Solution**