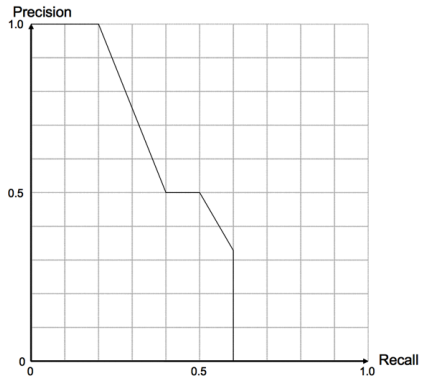
Sample Questions for Midterm Exam

1. Are the following statements true or false? Explain why if your answer is false.
   (1) Consider the vector space model and TF-IDF as its weighting scheme. The inverse of Euclidian distance is appropriate for the measure of the similarity between a query and a document.
   (2) Enlarging skip intervals (or skip spans) can always reduce the number of postings to be accessed if our goal is to locate a specific posting in a posting list with skip pointers.
   (3) Stemming always improves retrieval performance.
   (4) According to Zipf's law, the frequency of occurrence of the fifth most common term is 0.2.

2. Suppose we have a query with a total of 10 relevant documents in the whole collection. Given the query, an IR system returns 20 documents in the order of ranking and produces the interpolated recall-precision curve.



   (1) What is the precision after the system has retrieved 3 relevant documents? Show the calculation.
   (2) Suppose the following is the ranking list of the retrieved 20 documents. The number stands for its ranking. Mark a relevant document with a '+' in the corresponding box. Leave irrelevant documents unmarked.

1	2	3	4	5	6	7	8	9	10
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

   (3) Calculate MAP (Mean Average Precision) for the system. Show the calculation.
   (4) F-measure is defined as the weighted harmonic mean of precision and recall; the Dice coefficient of two sets X and Y is defined as

$$\text{Dice}(X,Y) = \frac{2|X \cap Y|}{|X|+|Y|} \quad \text{Dice}(X,Y) = \frac{2|X \cap Y|}{|X|+|Y|},$$

   where | A | is the size of set A. Please prove that the F1-measure is equal to the Dice coefficient of the retrieved and relevant document sets.

3. Language modeling (LM) has been widely used in IR.
   (a) Please describe what "smoothing" is and explain why LM often needs smoothing in IR.

   Consider the following four documents.

   Document 1 (d_1): search search search search where the engine is
   Document 2 (d_2): Google engine like search search search
   Document 3 (d_3): search engine
   Document 4 (d_4): Google Yahoo search engine

In the following, q is a query, d is a document, w_i is a word, and C is the corpus $\{d_1, d_2, d_3, d_4\}$.

(b) Consider the uni-gram query likelihood LM without smoothing:.

$$p(q \mid d) = \prod_{w_i \in q} p(w_i \mid d)$$

Suppose query q is *search engine*. Calculate $p(\textit{search engine} \mid d_1)$ and $p(\textit{search engine} \mid d_4)$, respectively. Which document (d_1 or d_4) is more relevant? Show your calculations.

(c) Consider the following uni-gram query likelihood LM with the corpus smoothing:

$$p(q \mid d) = \prod_{w_i \in q} \left[\lambda\, p(w_i \mid d) + (1 - \lambda)\, p(w_i \mid C) \right]$$

where λ is a weight varied from 0 to 1. Given that q is *search engine* and λ is set to be 0.5, calculate $p(\textit{search engine} \mid d_1)$ and $p(\textit{search engine} \mid d_4)$, respectively. Which document (d_1 or d_4) is more relevant? Show your calculations.

(d) Compare the results obtained from (b) and (c). How does the collection frequency of a common word in the smoothing version affect the probabilities?