36-350: Data Mining

Homework 1
Date: August 25, 2003 **Due: start of class September 3, 2003**

---

1. What is the bag-of-words representation of the sentence "to be or not to be"?

2. Suppose we search for the above sentence via the keyword "be". What is the bag-of-words representation for this query, and what is the Euclidean distance from the sentence?

3. Describe how weighting words by inverse-document-frequency (IDF) should help when making a Web query for "The Principles of Data Mining."

4. Describe a simple Web page search that could not be carried out effectively using a bag-of-words representation.

5. (a) What is the Euclidean distance between each of the vectors $(1, 0, 0)$, $(1, 4, 5)$, and $(10, 0, 0)$?
   (b) Divide each vector by its sum. Roughly, how do the distances change?
   (c) Divide each vector by its Euclidean length. Roughly, how do the distances change?

6. In this problem, you will interpret the results of the previous question.
   (a) Suppose we're using the bag-of-words representation for similarity searching with a Euclidean metric. Describe how the previous question illustrates a potential problem if we do not normalize for document length.
   (b) In a conventional database, one cannot search by similarity, but only for the set of documents containing particular keywords. Describe how the previous question illustrates a potential problem with this type of search.

7. In the computer lab, you will have computed three different distance matrices for a collection of documents. Suppose you performed a search for documents similar to `politics3`, using each of the three distance metrics in turn. For each metric, what would be the first document returned? For each metric, is this a good or bad result?