FRS 117: Google and Ye Shall Find???
Fall 2007
Assignment 2
due Friday, October 5 at 5PM

Blog entry: none this week.

Technical Exercises:

Problem 1: In class we discussed the technique of building a table of documents by words. A numeric score for each word-document pair would be entered in the table to represent the significance of the word in the document. We discussed scoring a word based on the frequency with which the word appears in the document[1]. Below is a fragment of such a table for a collection of documents. Scores are between 0.0 and 1.0; a larger score indicates a more significant word.

Part a. Suppose to score a document with respect to a query, the simple strategy of adding up the scores for each query word in the document is used. For the given table and the query *apple ant*, what are the scores and rankings of the four documents shown?

Part b. Does the ranking you found in **Part a** agree with your intuition of what the ranking of documents should be based on the entries in the table? Explain your intuition. If the ranking does not agree, can you suggest a better scoring strategy?

Score of **word** → in **document** ↓	**actor**	**apple**	**ant**	. . .
DOC1	0.04	0.2	0.2	
DOC2	0.4	0.004	0.4	
DOC3	0.4	0.18	0.2	
DOC4	0.0	0.28	0.08	
. . .				

---

1. In class we saw one historically important formula (reproduced below). This exercise does ***not*** depend on that formula:

score of a word in a document (doc.) based on word frequency (freq.) =

$$\frac{\text{(freq. of the word in the doc)}}{\text{(max freq. of any word in the doc)}} \times \log\left(\frac{\text{(\# of docs in the collection)}}{\text{(\# of docs containing the word in the collection)}}\right)$$

Problem 2:
In class we also discussed using an *inverted index* organized by *words* to record information about each word in each document of a collection. An inverted index uses a nested outline structure; each entry is labeled with a word and has the following structure:

Word
- ID of 1st document containing word
  - Position of 1st occurrence in this document:
    - information about occurrence: (e.g. italicized? In title? …)
  - Position of 2nd occurrence in this document:
    - information about occurrence: (e.g. italicized? In title? …)
  - …
- ID of 2nd document containing word
  - Position of 1st occurrence in this document:
    - information about occurrence: (e.g. italicized? In title? …)
  - Position of 2nd occurrence in this document:
    - information about occurrence: (e.g. italicized? In title? …)
  - …

Now consider the use of words in *anchor text.* *Anchor text* is a group of words that are used to label a link to another document, which we call the target document. For example the FRS117 Schedule and Assignments page has anchor text "Assignment 2 page" that points to this assignment. The Schedule and Assignments page should appear as one of the documents in the inverted index entry for each of "assignment", "2" and "page", but the inverted index entry for each of these words should also record this assignment page as a target page of anchor text containing the word.

Explain how to include target documents in the inverted index entries for words contained in anchor text pointing to the target documents. Remember that a word can be in the anchor text of many links, and a document can have a huge number of links into it. Give details, and show how this page would appear in the index entry for "assignment".

Problem 3: The position of words has many uses in calculating the score of a document with respect to a query. We discussed some in class: appearing early in a document is an indication of importance; the close proximity of two query words in a document can indicate higher relevance of the document for the query; phrases can be recognized by word position. Give another way in which word position can be used to help determine relevance of a document with respect to a query. It is fine to suggest a way that uses position in conjunction with another property of word occurrences. Explain your idea in enough detail that someone could apply it to a sample of documents and queries.

Problem 4:
In class, we saw examples of graphs from An Atlas of Cyberspaces - Topology Maps and "Map: Welcome to the Blogosphere" on the DISCOVER Magazine site. The concept of a graph as a set of nodes connected by edges is simple, but the richness of types of graphs and properties they can possess makes it an important concept in mathematics and computer science.

You have surely encountered many other graphs. Two examples that are probably familiar are family trees and airline route maps. These two types of graphs have very different properties. Family trees are directed (parent to child) and have a sense of levels (generations). They tend to fan out uniformly from generation to generation. Airline route maps usually have a small portion of the nodes that have many edges connected to them - the airline hubs.

Describe two examples of graphs not mentioned in class or above. These examples should represent different kinds of information. Try to find two examples that different from each other in their properties. Either draw a small example yourself or find an image on the Web and give the location of that image. What properties do you see in each type of graph?

We will discuss some graph properties in class Wednesday, October 3, but you need not wait until then to do this problem; it is fine to work with properties that are intuitive from the drawings of the graphs.