Problem Set 2

Assigned: Feb. 22
Due: Mar. 7

Problem 1

Suppose there is a collection of 4 documents with the following word occurrences:

	Doc1	Doc2	Doc3	Doc4
walrus	10	0	0	10
carpenter	8	0	40	0
bread	4	24	0	20
butter	1	16	0	0

Compute the scores and the rankings for the four documents for the queries "walrus", "walrus carpenter" and "walrus bread butter" under the following vector model:

Use the TF/IDF model described in MRS section 6.4.1: For each term t and document d:

- Let $f(t,d)$ = the number of occurrences of t in d.
- Let $o(t)$ = the number of documents in the collection containing term t.
- Let c = the total number of documents in the collection (in this case, 4.)

Define the function $w(t,d)$ as follows

$$w(t,d) = \begin{cases} 1 + \log_2(f(t,d)) & \text{if } f(t,d) > 0 \\ 0 & \text{if } f(t,d) = 0 \end{cases}$$

Let $i(t) = 1 + \log_2(c/o(t))$.

Finally define the coordinates of the document vector $\vec{d}$ as $\vec{d_t} = w(t,d) \cdot i(t)$.

For example, for d=Doc1, t="carpenter", we have $f(t,d) = 8$, $o(t) = 2$, $c = 4$, so $w(t,d) = 4$, $i(t) = 1 + \log_2(4/2) = 2$ and $\vec{d_t} = 4 \cdot 2 = 8$.

Then the score is computed as the similarity measure

$$Sim(\vec{d}, \vec{q}) = \frac{\vec{d} \bullet \vec{q}}{|\vec{d}| \cdot |\vec{q}|}$$

where $\vec{d}$ is the document vector and $\vec{q}$ is the query vector. For every word w in the query, $\vec{q}_w = 1$.

The length of a vector is the Euclidean length: $|\vec{v}| = \sqrt{\sum_{i=1}^{k} \vec{v}_i^2}$

Problem 2

A. The same document vector model can be used as a measure of the "similarity" of two documents. Compute the similarities between document Doc1 and the other three documents.

B. A similar model can be used to compute the similarity of words. Define a vector space in which each document is one dimension of the space. For each word w define a word vector $\vec{w}$ where $\vec{w}_d = f(w, d)$. Compute the similarity of the word "bread" to the other words. Again, the similarity of two vectors $\vec{u}$ and $\vec{v}$ is defined as $Sim(\vec{u}, \vec{v}) = \vec{u} \bullet \vec{v}/(|\vec{u} \cdot |\vec{v}|)$.

Problem 3

The following properties of ranking algorithms might be considered desirable. For each property determine whether it holds for the ranking algorithm in problem 1. If it holds, give an explanation (proof) why it holds; if it does not, construct a toy example where it fails to hold.

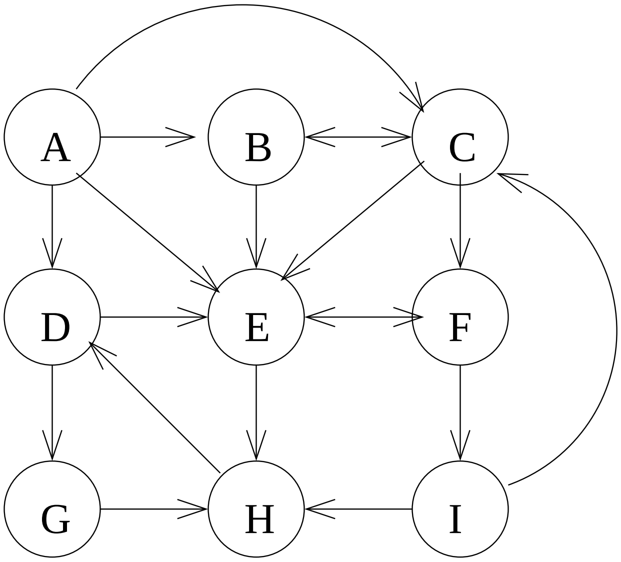
Property A: (Invariance under irrelevant words). If there are two documents d and e in the collection, and for every term t in query q, $f(t, d) = f(t, e)$, then $Sim(d, q) = Sim(e, q)$.

Property B: (Invariance under scaling.) If there are two documents d and e in the collection and there is a constant p such that, for every term t, $f(t, d) = p \cdot f(t, e)$, then $Sim(d, q) = Sim(e, q)$. For instance, if d consists of k exact copies of e, appended one after another, then this condition would hold with $p = k$.

Property C: (Order invariance under collection). Suppose that there are two documents d and e and two collections b and c, both of which contain both d and e. Suppose the rankings of the documents for query q are computed using collection b, and d is ranked higher than e. Then if you recompute the rankings using collection c, d will still be ranked higher than e.

Problem 4

Consider the graph of documents shown below.



A. Write the system of linear equations for PageRank. Use the parameter value $e = 0.3$ used in the notes (that is, in the random walk model there is probability 0.3 of jumping to a random new page and probability 0.7 of following a random outlink).

B. Compute the PageRank by solving the system in (A), and order the pages by PageRank. You may use any software you want to solve the equations.

Problem 5

Recompute the PageRank order for the graph in problem 4 with $e = 0.99$ and with $e = 0.01$.