CS838-1 Advanced NLP Homework 3

Due 3/13/2007 in class

Instructor: Jerry Zhu, jerryzhu@cs.wisc.edu

Type your answers and hand in a printed version to the instructor in class on the due date. The homework is worth 50% if it is no later than 48 hours (you may email me a pdf file), and worth nothing after that. I will not accept homeworks to the TA or in the physical mailbox.

Note: This homework involves computing the leading eigenvector of a transition matrix. Although you can use the power method and program it in any language, we recommend using a scientific computing language like Matlab or GNU Scientific Library.

1 Link Analysis

Social network is another kind of graph where link analysis can be useful. For this question the datasets can be downloaded from `http://www.cs.wisc.edu/~cs838-1/dataset/imdb_top500_comedy_actors/`. Consider the top 500 most productive movie comedians in IMDB (whose names can be found in `names.txt`, if you are curious). We can build a *co-star* graph, where each node is a comedian. An edge exists between the i-th and j-th comedians, if they co-starred in the same movies. The edges are *weighted*, with the weight being the number of movies they co-starred in.

The co-star graph is in `costar.txt`, where each line represents an edge. The format is (i, j, count). For example, the line (6,1,1) means Christopher Walken (the 6th line in `names.txt`) and Anthony Anderson co-starred in a total of one movie. The file is symmetric: both (6,1,1) and (1,6,1) appear in the file.

Now consider a "random reporter", who interviews movie stars similar to a random Web surfer: if the reporter is interviewing comedian i today, he will decide whom to interview tomorrow with the following rules:

1. The reporter flips a coin with head probability α. If the coin comes up head, he picks a comedian uniformly at random (teleporting).

2. Otherwise, he picks a comedian j who has co-starred with i, with probability proportional to the number of movies they co-starred in:

$$p(i \to j) = \frac{n_{ij}}{\sum_{k=1}^{500} n_{ik}} \tag{1}$$

Let P be the 500×500 transition matrix, with $P_{ji} = p(i \to j)$ (note the order of subscript).

Question 1 [10]. Write down all *non-zero* transition probability entries P_{ji} for i="Jackie_Chan", $j = 1, \ldots, 500$, $P_{ji} > 0$. You do not have to translate indices into names. Note this does not involve teleporting yet. Do these probability entries sum to one? Why?

Question 2 [5]. Similarly, write down all *non-zero* transition probability entries P_{ji} for j="Jackie_Chan", $i = 1, \ldots, 500$, $P_{ji} > 0$. Do these probability entries sum to one?

Question 3 [5]. Let $r^{(t)}$ be the probability vector that the reporter is interviewing each comedian on day t. Write down the iterative formula for $r^{(t+1)}$. Note this involves teleporting.

Question 4 [10]. Let M be any transition matrix, i.e. the entries are non-negative, and each column sums to one. Let r be a probability vector, i.e. the entries are non-negative and sum to one. Prove that $r' = Mr$ is a probability vector too.

Question 5 [20]. Let $\alpha = 0.1$. Compute the stationary distribution r with respect to the matrix in the iterative formula (call it M) in Question 3. Briefly describe how you compute it with your program. Write down the top 5 comedians with the largest stationary probability.

Question 6 [5]. What is the eigenvalue λ corresponding to the stationary distribution r?

Question 7 [5]. Verify that r is indeed an eigenvector (or fairly close to one) by computing $\max_{i=1}^{500} |\lambda r_i - (Mr)_i|$. What do you get?

2 Information Retrieval

Consider the following document collection, represented as a document-word count matrix:

	w_1	w_2	w_3	w_4
d_1	4	2	2	
d_2	1	1		
d_3	3		1	
d_4	6			2

Question 8 [10]. Compute the *tf.idf* representation of each document (use log base 2).

Question 9 [10]. Compute the cosine similarity of each document to the query "w_1 w_2 w_3".

Question 10 [20]. Let $cs(p, q)$ be the cosine similarity between vectors p, q. Let $d(p, q)$ be the Euclidean distance between vectors p, q. For any p and q that are normalized to length 1 (i.e., $p^\top p = 1$, $q^\top q = 1$), find the relation between $cs(p, q)$ and $d^2(p, q)$.