COMP 6000I Search Engines and Applications
Fall 2019 Homework 2

Due: Nov 10, 2019 11:59pm **Read LSI of stanford**

Submit your answers in a zip file via Canvas.

1. **[50]** The question refers to the paper "PageRank: Standing on the Shoulders of Giants" available from the course homepage. In the PageRank formula, all links are assumed to have weights equal to 1, so they are not explicit indicated in the webgraph. In the following questions, link weights can mean different things. Short (a few sentences) and precise answers are expected.

   (a) [20] In Bibliometric, (i) how is an entry in the link matrix H defined in Pinski and Narin's method and what does it mean in plain words? (ii) Write down the link matrix H for the citation graph in Fig. 3, the impact factor formulas (aka PageRank formulas), and compute the impact factors for the four journals in the table below, initializing the initial impact factors of the four journals to ¼. Note that there is no damping factor in the example and there no need to normalize the scores in each iteration.

Iteration	0	1	2
IF(A)	1/4	impactFactor(B)*weight_B_A / #citations_A 0.25*5 /8 = 0.15625	0.6136*5 /8 = 0.3835
IF(B)	1/4	(impactFactor(A)*weight_A_B + impactFactor(B)*weight_B_B + impactFactor(C)*weight_C_B + impactFactor(D)*weight_D_B) / #citations_B + (0.25*8 +0.25*1 + 0.25*8 + 0.25*10)/11= 0.6136	(11.25*8 +2.25*1 + 1.25*8 + 1.25*10)/11=0.3585
IF(C)	1/4	impactFactor(D)*weight_D_C / #citations_C 0.25*5 /8 = 0.15625	0.083*5 /8 =0.051875
IF(D)	1/4	impactFactor(B)*weight_B_D / #citations_D 0.25*5 /15 = 0.083	0.6136*5 /15 = 0.2045

   (b) [20] In sociometry, (i) how is an entry in the link matrix W defined in the Katz model and what does it mean in plain words? (ii) Katz defines the link matrix $W = \sum_{k=1}^{\infty}(aL)^k$. Assuming $a = 0.2$ and $k = 2$, write down the link matrix W for the example graph in Fig. 4.

k only 2

(c) [10] In econometrics, the link matrix representing the consumption/production relationship between the industries. Is the link matrix more similar to that of Pinski and Narin's model or Katz's model? Give a brief explanation.

2. **[50]** You are given the following document-term matrix A. The entries are the weights of the terms in the documents. No normalization on the weights are needed.

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11
d1	1	0	1	0	1	1	1	1	1	0	0
d2	1	1	0	1	0	0	1	1	0	2	1
d3	1	1	0	0	0	1	1	1	1	0	1

*** Answer this question using R. Submit your R code in a separate file, together with a pdf file containing the outputs for the parts of the question. **(Update: Although R is easier, you can use Python to get the same result.)**

(a) [10] Decompose A using Singular Value Decomposition, i.e., $A = USV^T$

(b) [10] Given the query $q = [0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 1]$, compute the inner product scores of $d1$, $d2$ and $d3$ using the decomposed matrices. Verify if the result is the same as performing the dot product of q and A.

(c) [10] Apply Rank-2 approximation to the decomposed matrices, i.e., give U_2, S_2 and V_2, V_2^T, and A_2 (The subscript "2" means Rank-2 approximation)

(d) [10] Obtain the document vectors and query vector $q = [0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 1]$ in the reduced 2-dimensional space, and plot them in 2-D coordinates.

(e) [10] Compute inner product scores of d1, d2 and d3 to q after Rank-2 approximation.