HW6, CS769

Nathanael Fillmore

Due April 11, 2008

1 Page Rank

Question 1. After one step, the walker will be at page 2, 11, 21, or 31, each with probability 1/4.

Question 2. After one step, the walker will be at page 2, 11, 21, or 31, each with probability $0.2275 = 9/10 * 1/4 + 1/10 * 1/40$, or at any other page, each with probability $0.0025 = 1/10 * 1/40$.

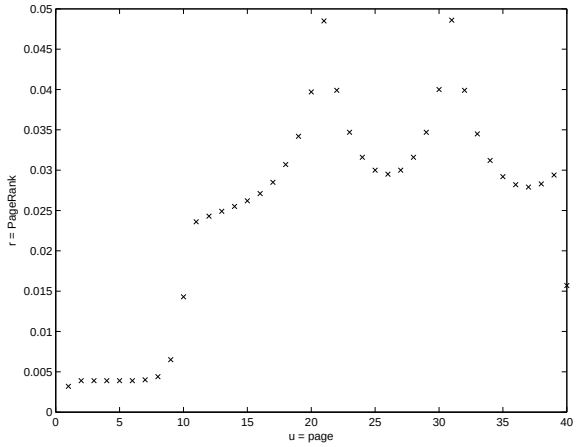
Question 3. The PageRank r for each page u is:

u	r	u	r	u	r	u	r
1	0.0032	11	0.0236	21	0.0485	31	0.0486
2	0.0039	12	0.0243	22	0.0399	32	0.0399
3	0.0039	13	0.0249	23	0.0347	33	0.0345
4	0.0039	14	0.0255	24	0.0316	34	0.0312
5	0.0039	15	0.0262	25	0.0300	35	0.0292
6	0.0039	16	0.0271	26	0.0295	36	0.0282
7	0.0040	17	0.0285	27	0.0300	37	0.0279
8	0.0044	18	0.0307	28	0.0316	38	0.0283
9	0.0065	19	0.0342	29	0.0347	39	0.0294
10	0.0143	20	0.0397	30	0.0400	40	0.0157

This was computed as follows:

```
% P was defined earlier
b = ones(40,1)/40 % uniform distribution
alpha = 0.9
M = alpha*P + (1-alpha)*b*ones(40,1)'
[V,D] = eig(M)
r = V(:,1)
r = r/sum(r)
```

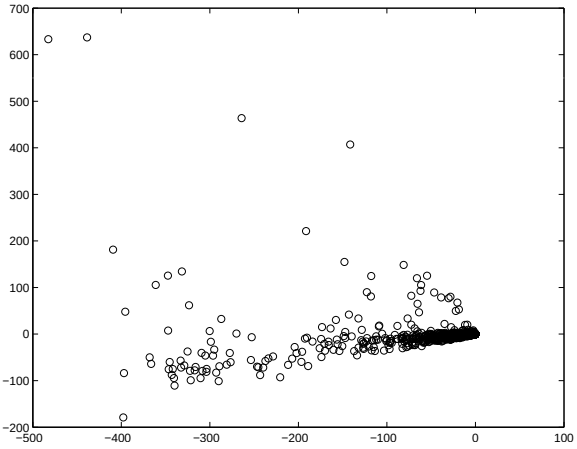
Question 4.



Pages 21 and 31 have the same PageRank (almost). This is because they have the same number of pages linking to them, and the linking pages are also ranked similarly. Page 11 is ranked lower. Why? Pages 21 and 31 are linked to by 1-10 as well as 20, 22 and 30, 32 respectively; page 11 is linked to by 1-10 as well as 12 (since 10 already links to it). Thus in effect page 11 has one less link than pages 21 and 31 do. Additionally, page 11 is near the 1-10 range, which has more outgoing links than the other pages; hence the incoming neighbor links 11 does have are worth less than those of 21 and 31.

2 Latent Semantic Indexing

Question 6.



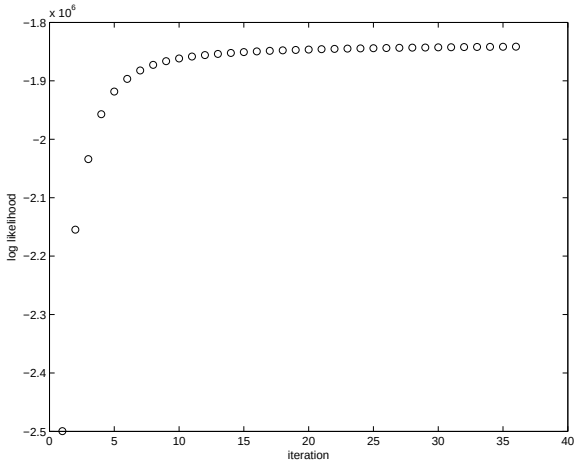
This was computed as follows:

```
load cs.dat;
X = spconvert(cs);
X = X';
[U,S,V] = svds(X);
U_hat = U(:,1:2);
S_hat = S(1:2,1:2);
```

```
V_hat = V(:,1:2);
X_hat = S_hat*V_hat';
plot(X_hat(1,:), X_hat(2,:), 'o')
```

3 Latent Dirichlet Allocation

Question 7.



Question 8.

topic	words
000	# java pdf 100 jasmin 4 programming ## cs 97
001	madison wi research sciences wisconsin 608 department dayton science 1210
002	class player frames will week use no package if all
003	was pictures jerel n#t one some back me but last
004	picture camera canon last first next atlantic#city previous am 30
005	was me #a href# #font font first tenaya n#t ###
006	madison wisconsin parallel links j# usa m# computing performance information
007	br #a href# http #td td #tr tr var class#
008	june rest 2006 29#30 cloud created bbgallerycloud n parallel sorting
009	learning xml machine other which one also reinforcement neural its
010	systems # database will project management material course minibase class
011	class t f # if points will array which 2
012	# 5 series 0 assignment 2 time analysis 4 models
013	science department madison research wisconsin wisconsin#madison sciences 2007 statistics wi
014	program profiling path warts new paths qpt research cache systems
015	statistics stat 2006 fall statistical 608 madison spring department 301
016	31 new image function if press e keypress keypresslistener next0
017	# image 2 c## will library your system standard
018	# cs cs#wisc#edu email algorithms os#9 theory list group
019	0 learning 2007 style# semi#supervised machine uw zhu cs width#
020	week project what 2007 # top clustering report information which
021	0 0#00 phase 0#0000 number rate first process proxy cache
022	memory code not x chapter time program p class int
023	#a li href# br http p href pdf postscript #
024	was david but me madison wisconsin web parter about if
025	discussion stat statistics model analysis y# 2 spring fall #
026	was me about n#t do he so but what ###
027	research systems learning doan arpaci#dusseau information a# c# anhai pdf
028	discussion # pdf homework 2 office 3 4 9 exam
029	p h5 #h5 devise user li puzzle program will our
030	wisconsin memory research architecture parallel more systems mark david hill
031	var false null # images parentindexpageurl firstindexpageurl lastindexpageurl u not
032	# 2007 b paper # # t# research abstract program
033	but do so was n#t not me if ### did
034	# td #td 2 all 3 photo e # center
035	function name value days savecookie readcookie album web next index
036	here ### homepage me web if click your links am
037	courier new #font your font program which function face# file
038	picture last next previous first #06 hawaii animation class simulation
039	cs document#write # me introduction dot ece about classes math
040	he theory his will complexity program science some cai may
041	cache web proxy sharing query protocol traces server client summary
042	è â width# #p p center align# 0 íå#td
043	condor linux project software computing code our if system your
044	analysis 2005 newton statistics 00 hours a# statistical methods 2006
045	server go last results updated configuration client back mn about
046	### java www#cs#wisc#edu##koconnor 2000 here some was 2 naming out
047	systems networks advanced prof# research fall network project system pdf
048	not do n#t was your but me so one all
049	http mangasarian 2005 mining # research l# pdf olvi wisconsin

Some of the topics seem sensible, for example topic 001 has to do with the mailing address of the CS department. Topic 007 has is all from inside html tags. Topic 004 combines photography with navigation, presumably because of photo galleries. Some of the topics are strange, for example topic 000 finds a connection between "java", "pdf", "100", and "##"—I guess from course webpages?

Question 9.

user	best topic-weight pairs
jerryzhu	(17, 905.6575810603), (19, 877.37349511399998)
shavlik	(9, 1853.0155380872)
miron	(7, 84.293074253699999), (43, 67.275698118700006), (1, 33.477841889099999)
sohi	(30, 153.01553808720001)
pb	(47, 1477.0950339067001), (28, 297.93604226769997)