CS246: Mining Massive Data Sets

Assignment number: 3

Fill in and include this cover sheet with each of your assignments. It is an honor code violation to write down the wrong time. Assignments and code are due at 5:00 PM on Scoryst and SNAP respectively. Failure to include the coversheet with you assignment will be penalized by 2 points. Each student will have a total of *two* free late periods. *One late period expires at the start of each class.* (Assignments are due on Thursdays, which means the first late period expires on the following Tuesday at 5:00 PM.) Once these late periods are exhausted, any assignments turned in late will be penalized 50% per late period. However, no assignment will be accepted more than one late period after its due date. (If an assignment is due to Thursday then we will not accept it after the following Thursday.)

Your name: Priyank Mathur
Email: priyankm@stanford.edu **SUNet ID:** priyankm

Collaborators: Emrah Budur, Shundan Xiao, Dibyajyoti Ghosh, Arkajyoti Misra

I acknowledge and accept the Honor Code.

(Signed) Priyank Mathur

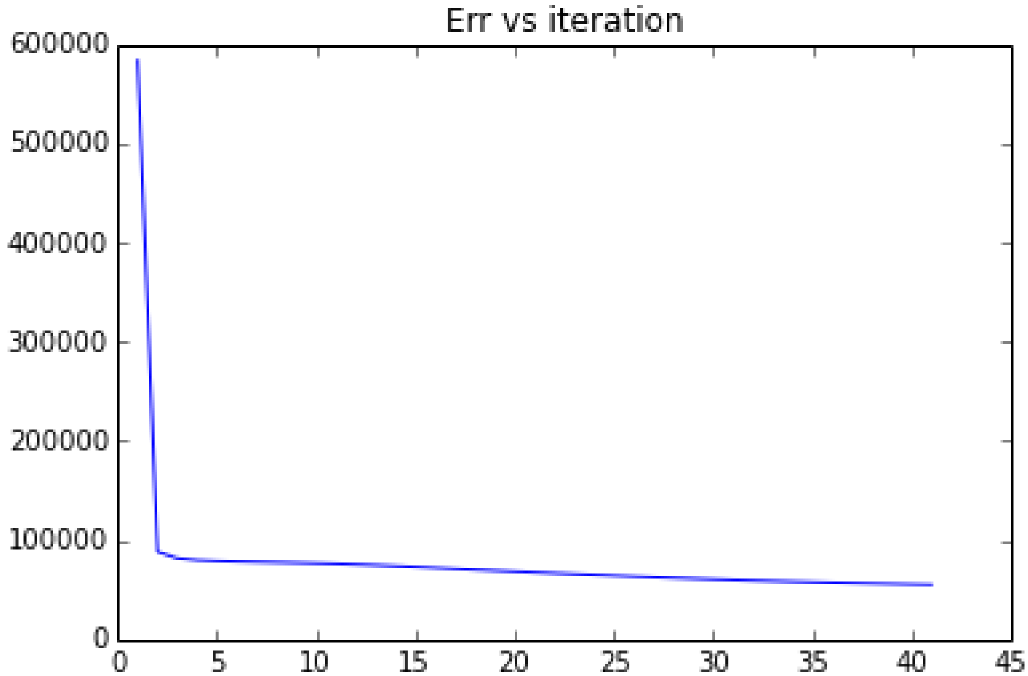
Answer to Question 1.a

$$\epsilon_{iu} = 2 \times (R_{iu} - q_i.p_u^T)$$
$$q_i = q_i + \eta(\epsilon_{iu} p_u - \lambda q_i)$$
$$p_u = p_u + \eta(\epsilon_{iu} q_i - \lambda p_u)$$

Note that the factor 2 obtained during differentiation w.r.t p_u and q_i has been consumed within the learning rate η.

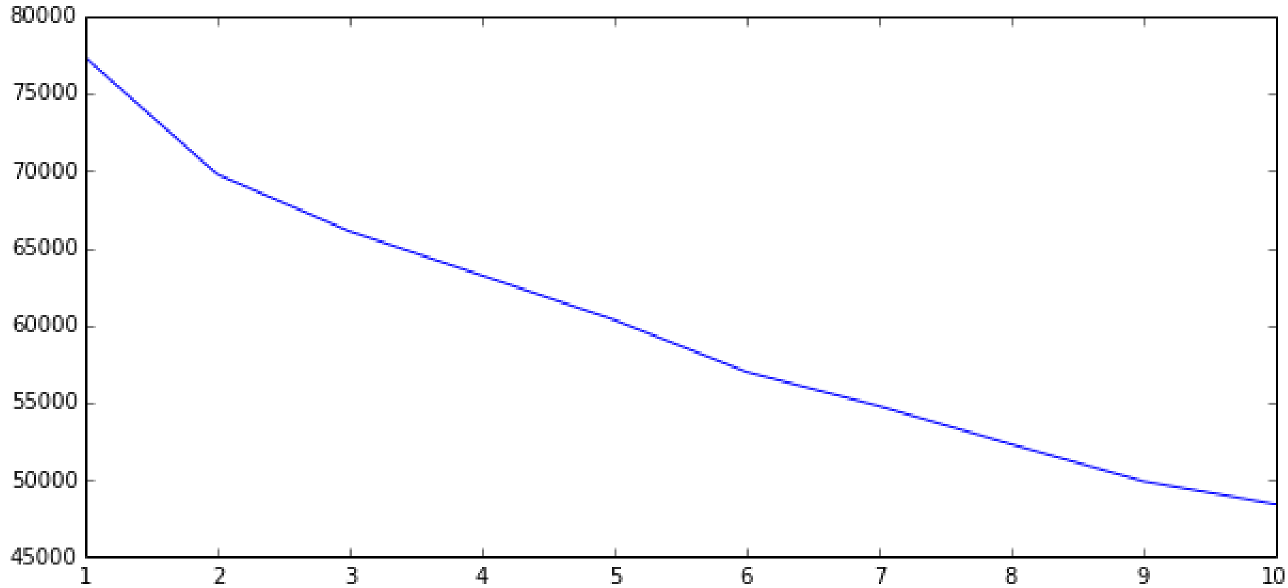
Answer to Question 1.b

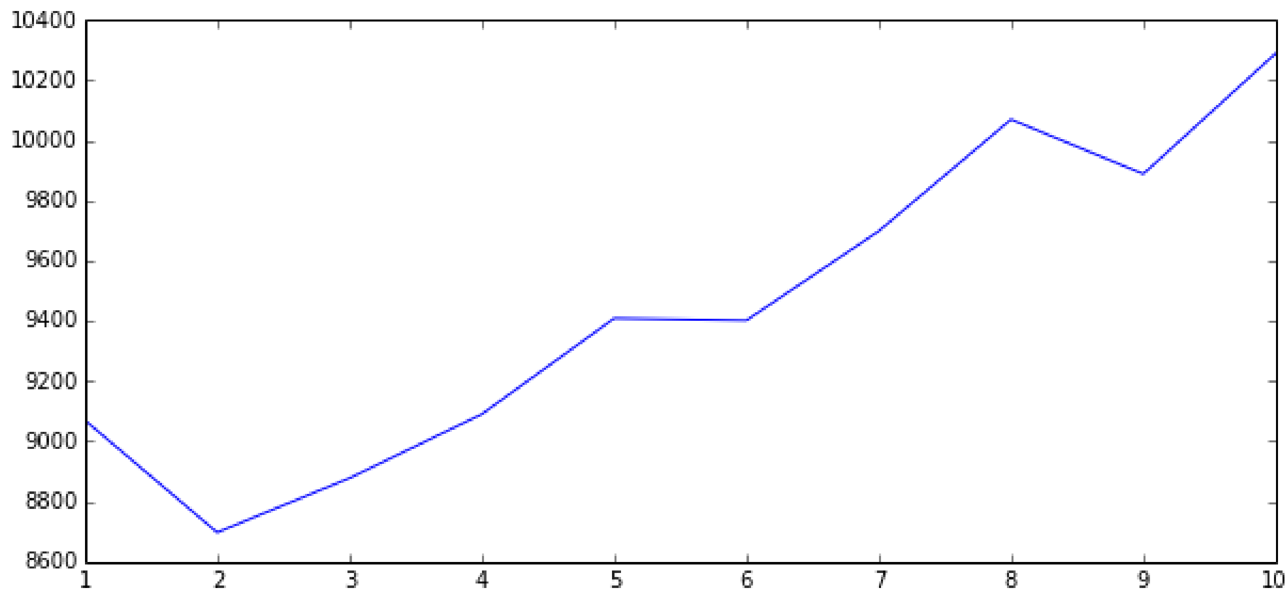
$\eta = 0.005$



Answer to Question 1.c

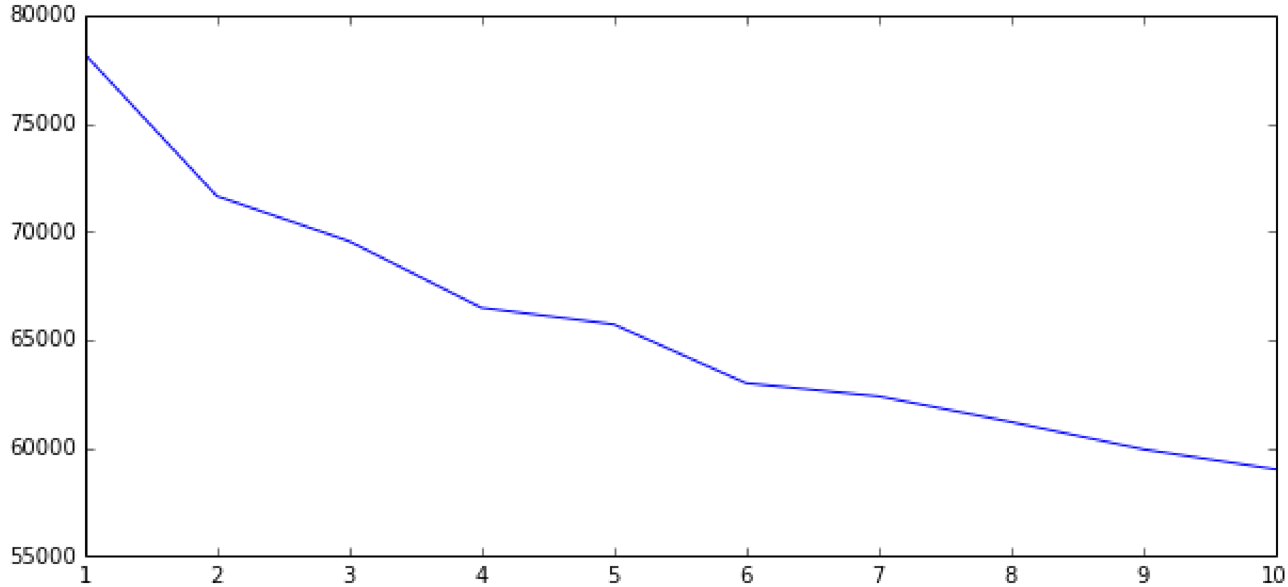
$Error_{tr}$ vs. k for $\lambda = 0$

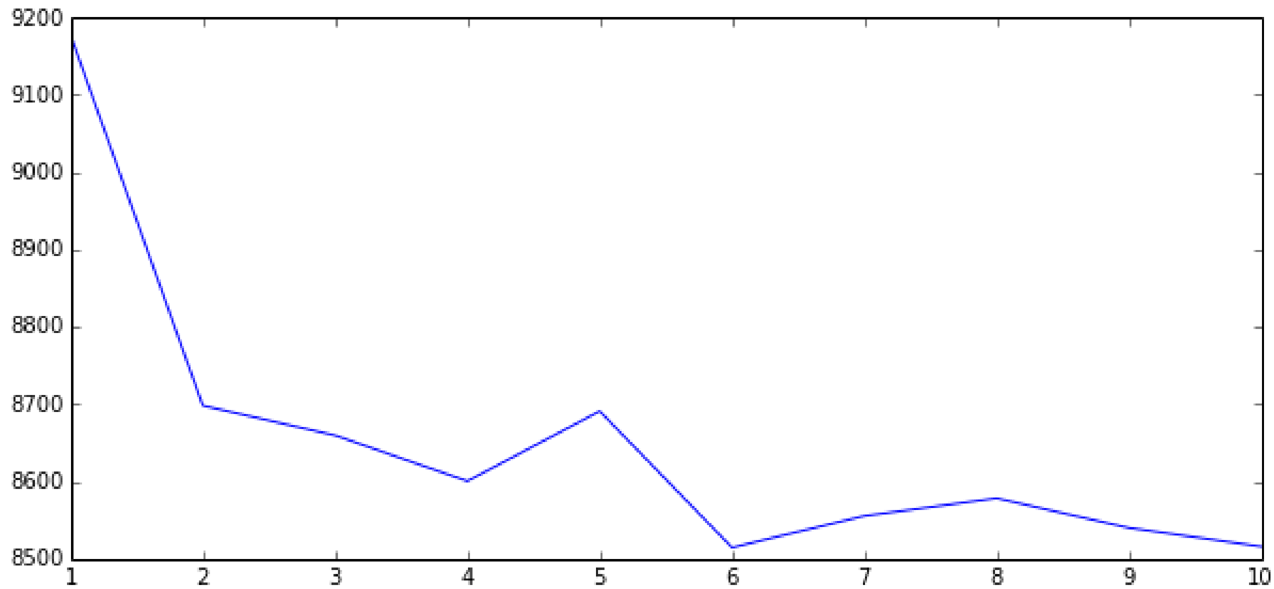




$Error_{te}$ vs. k for $\lambda = 0$

$Error_{tr}$ vs. k for $\lambda = 0.2$





$Error_{te}$ vs. k for $\lambda = 0.2$

Based on these graphs, we find the following to be true -

- B: Regularization decreases the test error for $k \geq 5$
- D: Regularization increases the training error for all (or almost all) k
- H: Regularization decreases overfitting

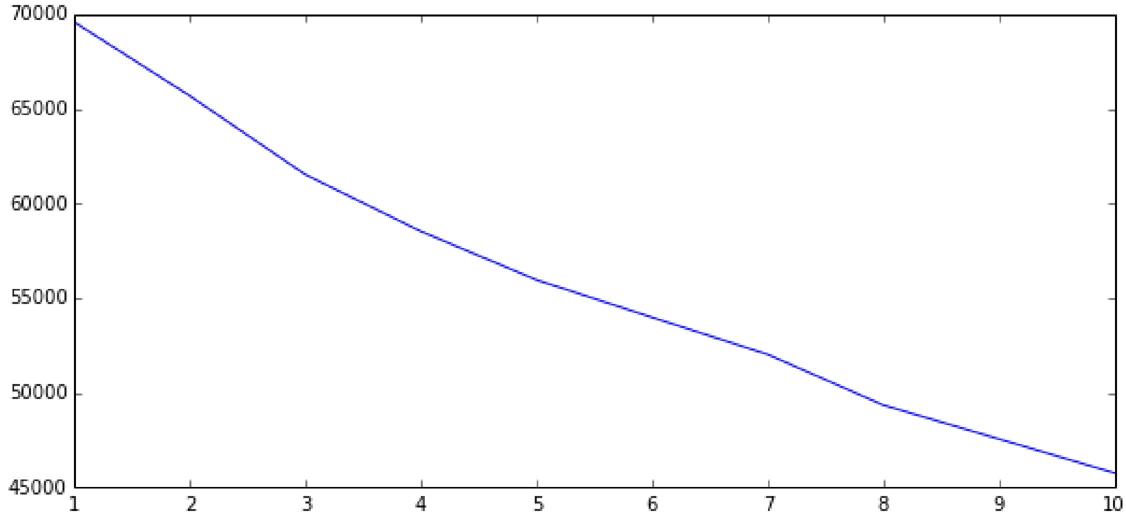
Answer to Question 1.d

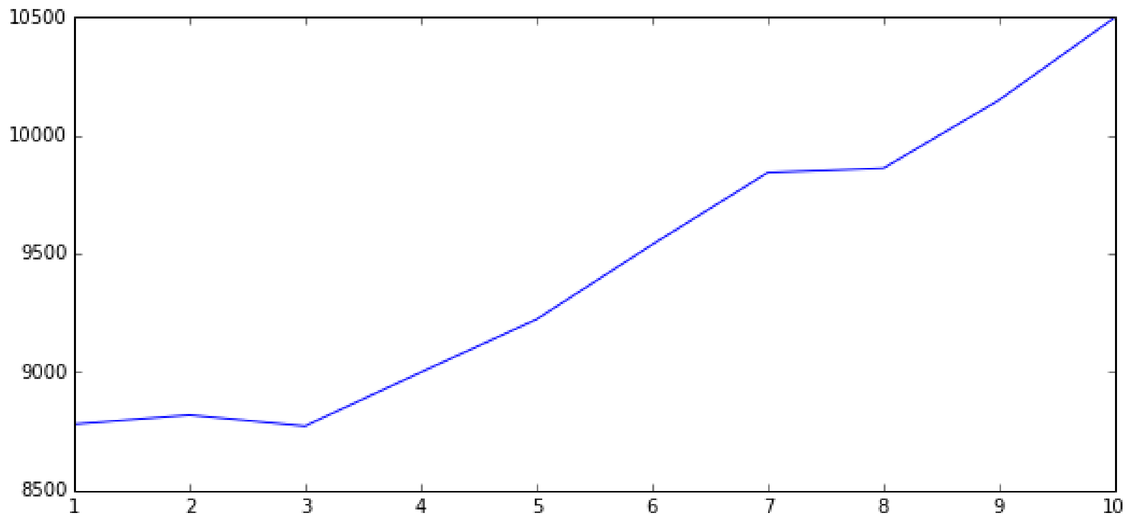
$$\epsilon_{iu} = 2 \times (R_{iu} - (\mu + b_u + b_i + q_i.p_u^T))$$
$$q_i = q_i + \eta_{LF}(\epsilon_{iu} p_u - \lambda q_i)$$
$$p_u = p_u + \eta_{LF}(\epsilon_{iu} q_i - \lambda p_u)$$

$$b_{i_i} = b_{i_i} + \eta_{b_i}(\epsilon_{iu} - \lambda b_{i_i})$$
$$b_{u_u} = b_{u_u} + \eta_{b_u}(\epsilon_{iu} - \lambda b_{u_u})$$

$$\eta_{LF} = 0.005$$
$$\eta_{b_i} = 0.01$$
$$\eta_{b_u} = 0.01$$

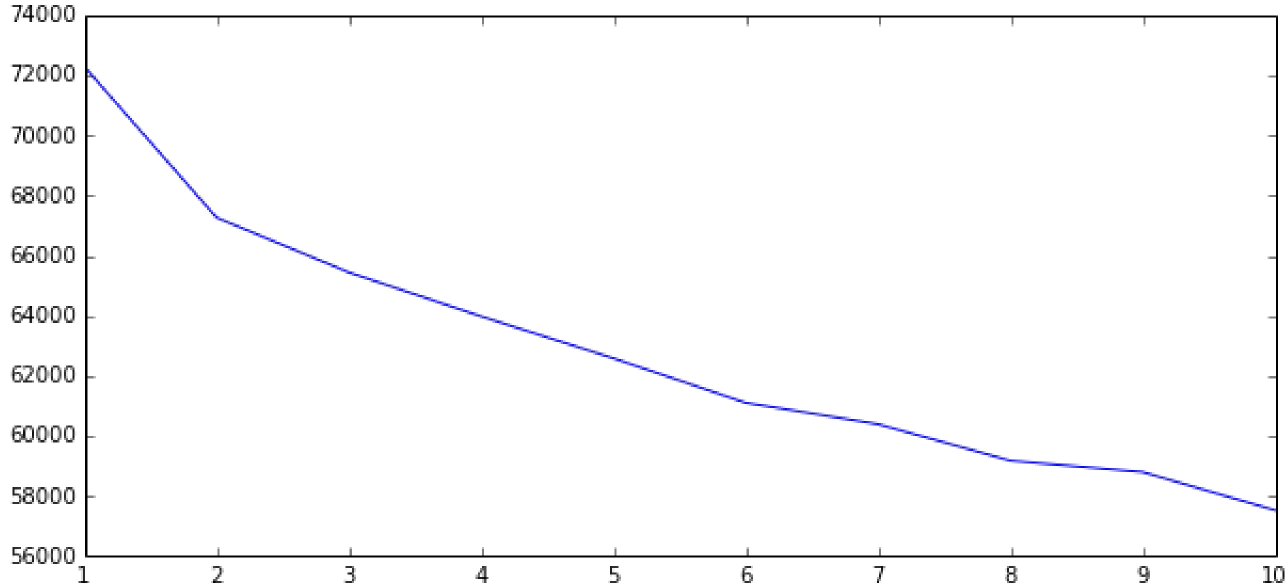
$Error_{tr}$ vs. k for $\lambda = 0$

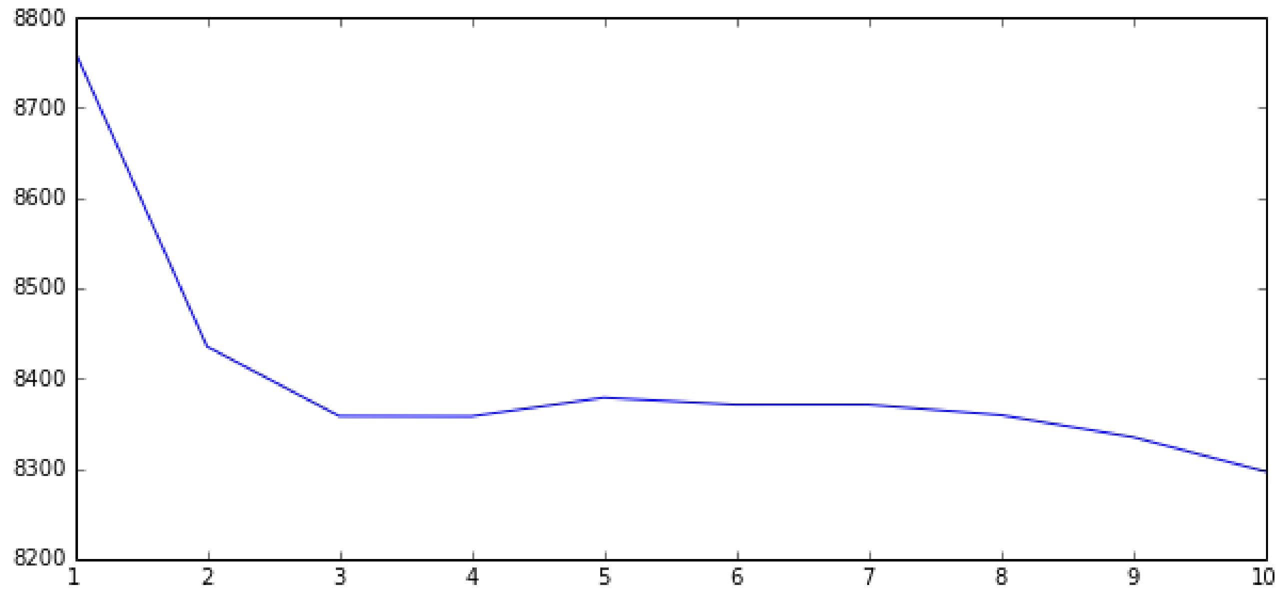




$Error_{te}$ vs. k for $\lambda = 0$

$Error_{tr}$ vs. k for $\lambda = 0.2$





$Error_{te}$ vs. k for $\lambda = 0.2$

Based on these graphs, we find the following to be true -

- B: Regularization decreases the test error for $k \geq 5$
- D: Regularization increases the training error for all (or almost all) k
- H: Regularization decreases overfitting

Answer to Question 2a

To prove -

$$||r - r^k||_1 \leq 2\beta^k$$

$$r - r^k = \frac{1-\beta}{n}\mathbf{1} + \beta M r - \frac{1-\beta}{n}\mathbf{1} - \beta M r^{k-1} = \beta M r - \beta M r^{k-1} = \beta(M\mathbf{k})$$

where $k = (r - r^{k-1})$

$||r - r^k||_1 = \beta||M\mathbf{k}||_1$

Expanding rhp, we get -
$\beta||M\mathbf{k}||_1$
$= \beta(M_{11}K_1 + M_{12}K_2 + \cdots + M_{1N}K_N +$
$M_{21}K_1 + M_{22}K_2 + \cdots + M_{2N}K_N||_1$
$\vdots$
$M_{N1}K_1 + M_{N2}K_2 + \cdots + M_{NN}K_N)$

$= \beta(K_1(M_{11} + M_{21} + \cdots + M_{N1})$
$+ K_2(M_{11} + M_{21} + \cdots + M_{N1})$
$\vdots$
$+ K_N(M_1 + M_{21} + \cdots + M_{N1}))$

$= \beta||K||_1 = \beta||r - r^{k-1}||$

Using the approach multiple times, we get -
$||r-r^k||_1 = \beta||r-r^{k-1}||_1 = \beta^2||r-r^{k-2}||_1 = \beta^3||r-r^{k-3}||_1 \cdots = \beta^k||r-r^0||_1$

Since r is a probability distribution, we know that -

$$||r - r^0||_1 \leq 2$$

Hence,

$$||r - r^k||_1 \leq 2\beta^k$$

Answer to Question 2b

From part a we know that the upper limit of L_1 error is given by

$$||r - r^k||_1 \leq 2\beta^k$$

To limit this to δ, we have $2\beta^k \leq \delta$.

Taking log on both sides - $\log 2 + k \log \beta \leq \log \delta$

$\implies k \leq \dfrac{\log \delta - \log 2}{\log \beta}$

$\implies k \leq \dfrac{\log \frac{\delta}{2}}{\log \beta} \implies k \leq \dfrac{\log \frac{2}{\delta}}{\log \frac{1}{\beta}}$

where k is the number of iterations. Thus we need to run at least $\dfrac{\log \frac{2}{\delta}}{\log \frac{1}{\beta}}$ times to ensure error $\leq \delta$.

In each iteration, we calculate the page rank of a node i as

$$r_i = \sum_{j \in N(i)} \frac{r_j}{deg_j} + C$$

where C is the constant amount of work needed to add the pagerank due to teleports. Therefore in each iteration, we cover all the m edges of the graph and do a constant amount of work for each edge.

Therefore,

$$runtime(iteration) = O(m)$$

Hence, total cost is

$$runtime = O\left(\frac{m \times \frac{2}{\delta}}{\log \frac{1}{\beta}}\right) = O\left(\frac{m}{\log \frac{1}{\beta}}\right)$$

Since δ is constant.

Answer to Question 2c

Let's consider that each iteration (each random walk) is a series of success/fail events. We define fail event if we transition to another node and success event when the random walk ends.

$$P(fail) = \beta$$
$$P(success) = 1 - \beta$$

This is a geometric series, where the first occurrence of success requires k independent events. Thus, the probability that we get success at trial k is

$$P(X = k) = \beta^{k-1}(1 - \beta)$$

The expected value of this geometric series is[1]

$$E(X) = \frac{1}{P(success)} = \frac{1}{1 - \beta}$$

Therefore, in each random walk, we perform approximately $\frac{1}{1-\beta}$ steps. In the total of nR iterations, we perform $\frac{nR}{1-\beta}$ steps.

Since each step takes unit time,

$$runtime = O\left(\frac{nR}{1 - \beta}\right)$$

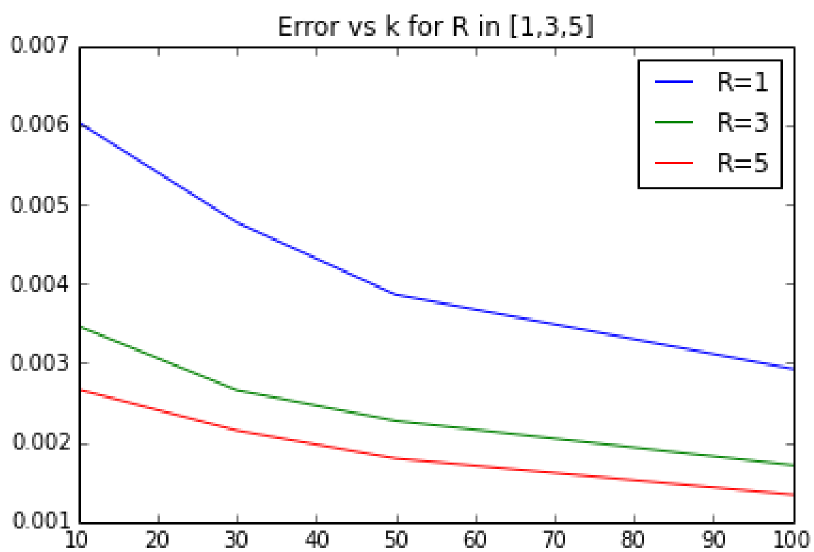
[1] http://en.wikipedia.org/wiki/Geometric_distribution

Answer to Question 2d

Runtime for 40 iterations of power iteration - 791 μs.

Avg runtime for Monte Carlo with R = 1 - 7.83 ms.
Avg runtime for Monte Carlo with R = 3 - 20.6 ms.
Avg runtime for Monte Carlo with R = 5 - 35.57 ms.

R	K	Error
1	10	0.0060365444268
1	30	0.0047717255253
1	50	0.0038620385493
1	100	0.0029277461094
3	10	0.0034659740412
3	30	0.0026577309646
3	50	0.0022720065398
3	100	0.0017132042790
5	10	0.0026678865383
5	30	0.0021502352960
5	50	0.0017972828358
5	100	0.0013425298758



Answer to Question 3.a

$$s_A(cameras, phones) = \frac{C1}{3 \times 2}[s_B(nokia.com, nokia.com) + s_B(nokia.com, apple.com) + s_B(kodak.com, nokia.com) + s_B(kodak.com, apple.com) + s_B(cannon.com, nokia.com) + s_B(cannon.com, apple.com)]$$

$$s_A(cameras, phones) = \frac{0.8}{6}[1 + 0 + 0 + 0 + 0 + 0] = 0.13333$$

Iteration 1 -

$s_A('cameras', 'cameras') : 1,$
$\mathbf{s_A('cameras', 'phones') : 0.13333333333333333},$
$s_A('printers', 'printers') : 1,$
$s_A('phones', 'phones') : 1,$
$\mathbf{s_A('cameras', 'printers') : 0.0},$
$s_A('phones', 'printers') : 0.0$

$s_B('cannon.com', 'nokia.com') : 0.4,$
$s_B('hp.com', 'nokia.com') : 0.0,$
$s_B('kodak.com', 'kodak.com') : 1,$
$s_B('apple.com', 'kodak.com') : 0.0,$
$s_B('cannon.com', 'kodak.com') : 0.8,$
$s_B('cannon.com', 'cannon.com') : 1,$
$s_B('hp.com', 'kodak.com') : 0.0,$
$s_B('apple.com', 'hp.com') : 0.0,$
$s_B('kodak.com', 'nokia.com') : 0.4,$
$s_B('apple.com', 'apple.com') : 1,$
$s_B('apple.com', 'nokia.com') : 0.4,$
$s_B('nokia.com', 'nokia.com') : 1,$
$s_B('cannon.com', 'hp.com') : 0.0,$
$s_B('hp.com', 'hp.com') : 1,$
$s_B('apple.com', 'cannon.com') : 0.0$

Iteration 2 -

$s_A('cameras', 'cameras') : 1,$
$\mathbf{s_A('cameras', 'phones') : 0.2933333333333333},$
$s_A('printers', 'printers') : 1,$
$s_A('phones', 'phones') : 1,$
$\mathbf{s_A('cameras', 'printers') : 0.0},$
$s_A('phones', 'printers') : 0.0$

$s_B('cannon.com', 'nokia.com') : 0.45333333333333337,$
$s_B('hp.com', 'nokia.com') : 0.0,$
$s_B('kodak.com', 'kodak.com') : 1,$
$s_B('apple.com', 'kodak.com') : 0.10666666666666667,$
$s_B('cannon.com', 'kodak.com') : 0.8,$
$s_B('cannon.com', 'cannon.com') : 1,$
$s_B('hp.com', 'kodak.com') : 0.0,$
$s_B('apple.com', 'hp.com') : 0.0,$

$s_B('kodak.com','nokia.com') : 0.45333333333333337,$
$s_B('apple.com','apple.com') : 1,$
$s_B('apple.com','nokia.com') : 0.45333333333333337,$
$s_B('nokia.com','nokia.com') : 1,$
$s_B('cannon.com','hp.com') : 0.0,$
$s_B('hp.com','hp.com') : 1,$
$s_B('apple.com','cannon.com') : 0.10666666666666667$

Iteration 3 -

$s_A('cameras','cameras') : 1,$

$\mathbf{s_A('cameras','phones') : 0.34311111111111114},$

$s_A('printers','printers') : 1,$

$s_A('phones','phones') : 1,$

$\mathbf{s_A('cameras','printers') : 0.0},$

$s_A('phones','printers') : 0.0$

$s_B('cannon.com','nokia.com') : 0.5173333333333333,$

$s_B('hp.com','nokia.com') : 0.0,$

$s_B('kodak.com','kodak.com') : 1,$

$s_B('apple.com','kodak.com') : 0.23466666666666663,$

$s_B('cannon.com','kodak.com') : 0.8,$

$s_B('cannon.com','cannon.com') : 1,$

$s_B('hp.com','kodak.com') : 0.0,$

$s_B('apple.com','hp.com') : 0.0,$

$s_B('kodak.com','nokia.com') : 0.5173333333333333,$

$s_B('apple.com','apple.com') : 1,$

$s_B('apple.com','nokia.com') : 0.5173333333333333,$

$s_B('nokia.com','nokia.com') : 1,$

$s_B('cannon.com','hp.com') : 0.0,$

$s_B('hp.com','hp.com') : 1,$

$s_B('apple.com','cannon.com') : 0.23466666666666663$

Final result after 3 iterations -
$\mathbf{s_A('cameras','phones') : 0.34311111111111114},$
$\mathbf{s_A('cameras','printers') : 0.0}$

Answer to Question 3.b

$$s_A(X,Y) = \frac{C1}{\sum_{i=1}^{|O(X)|}\sum_{j=1}^{|O(Y)|} W_{X,O_i(X)} \cdot W_{Y,O_j(Y)}} \times \sum_{i=1}^{|O(X)|}\sum_{j=1}^{|O(Y)|} W_{X,O_i(X)} \cdot W_{Y,O_j(Y)} \cdot s_B(O_i(X), O_j(Y)) \quad (3)$$

$$s_B(x,y) = \frac{C2}{\sum_{i=1}^{|I(x)|}\sum_{j=1}^{|I(y)|} \cdot W_{I_i(x),x} W_{I_j(y),y}} \times \sum_{i=1}^{|I(x)|}\sum_{j=1}^{|I(y)|} W_{I_i(x),x} \cdot W_{I_j(y),y} \cdot s_A(I_i(x), I_j(y)) \quad (4)$$

Answer to Question 3.c

Let the graph $K_{2,1}$ have nodes $A1, A2$ in A and $B1$ in B. Similarly, let the graph $K_{2,2}$ have nodes $A1, A2$ in A and $B1, B2$ in B.

Therefore, for iteration 1 -
$s_A(A1, A2)_{K_{2,1}} = \dfrac{C1}{1 \times 1} \times s_B(B1, B1) = 0.8$
$s_A(A1, A2)_{K_{2,2}} = \dfrac{C1}{2 \times 2} \times [s_B(B1, B1) + s_B(B1, B2) + s_B(B2, B1) + s_B\ (B2, B2)] = \dfrac{0.8}{2 \times 2} \times 2 = 0.4$

Results -
Iteration 1

$K_{2,1}$
$\mathbf{s_A('A1','A2') : 0.8}$,
$s_A('A1','A1') : 1$,
$s_A('A2','A2') : 1$
$s_B('B1','B1') : 1$

$K_{2,2}$
$\mathbf{s_A('A1','A2') : 0.4}$,
$s_A('A1','A1') : 1$,
$s_A('A2','A2') : 1$
$\mathbf{s_B('B1','B2') : 0.4}$,
$s_B('B1','B1') : 1$,
$s_B('B2','B2') : 1$

Iteration 2

$K_{2,1}$
$\mathbf{s_A('A1','A2') : 0.8}$,
$s_A('A1','A1') : 1$,
$s_A('A2','A2') : 1$
$s_B('B1','B1') : 1$

$K_{2,2}$
$\mathbf{s_A('A1','A2') : 0.5599999999999999}$,
$s_A('A1','A1') : 1$,
$s_A('A2','A2') : 1$
$\mathbf{s_B('B1','B2') : 0.5599999999999999}$,
$s_B('B1','B1') : 1$,
$s_B('B2','B2') : 1$

Iteration 3

$K_{2,1}$
$\mathbf{s_A('A1','A2') : 0.8},$
$s_A('A1','A1') : 1,$
$s_A('A2','A2') : 1$
$s_B('B1','B1') : 1$

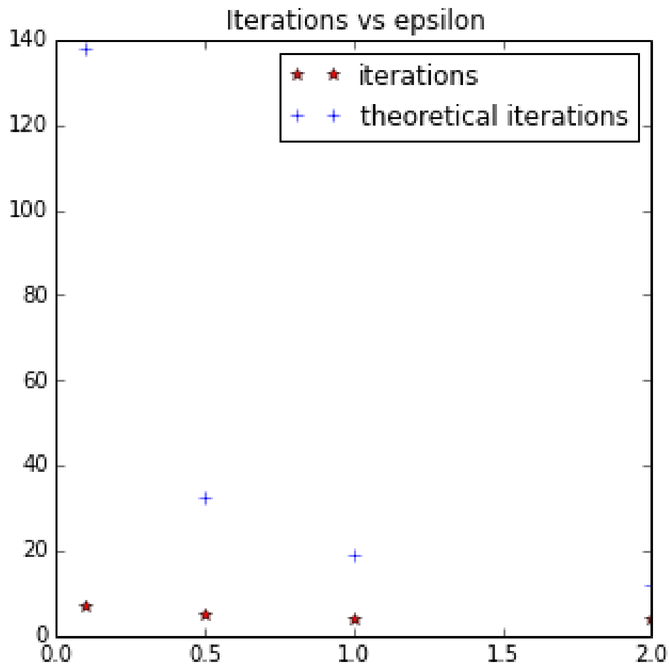
$K_{2,2}$
$\mathbf{s_A('A1','A2') : 0.6240000000000001},$
$s_A('A1','A1') : 1,$
$s_A('A2','A2') : 1$
$\mathbf{s_B('B1','B2') : 0.6240000000000001},$
$s_B('B1','B1') : 1,$
$s_B('B2','B2') : 1$

Answer to Question 4a

Answer to Question 4b

Answer to Question 4c

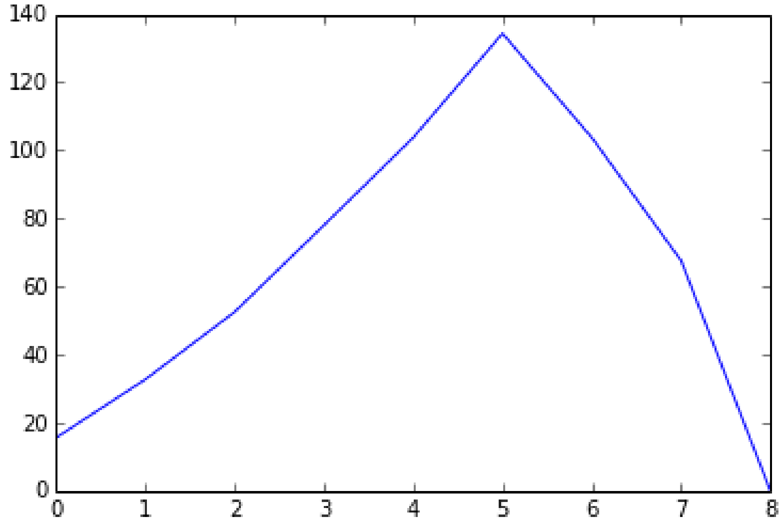
1



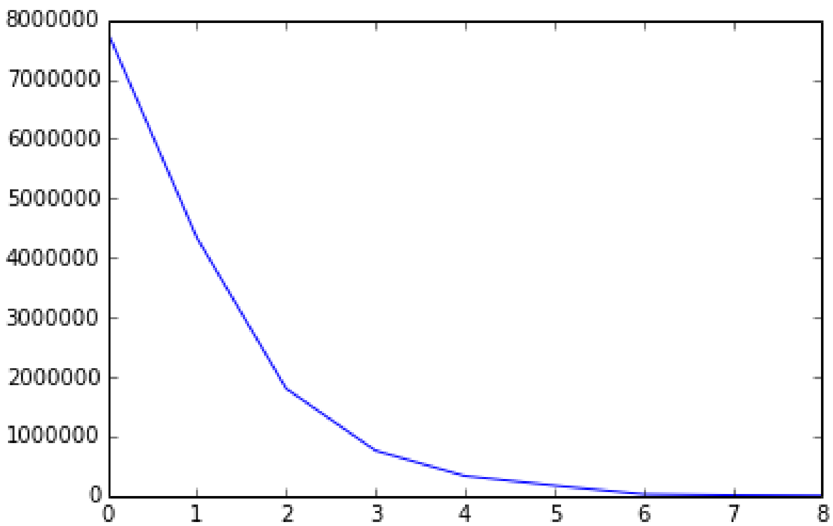
ϵ	Iterations	Theoretical iterations
0.1	7	137.67
0.5	5	32.36
1	4	18.93
2	3	11.94

2

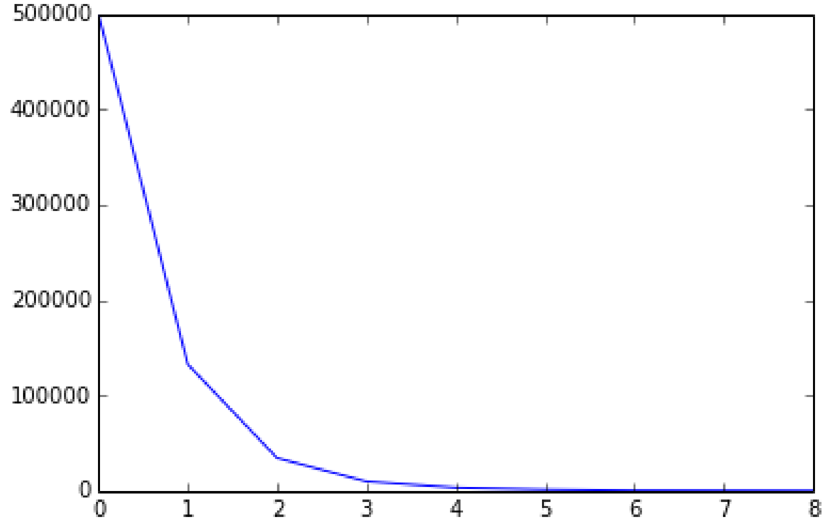
Density ($\rho(S_i)$) vs iteration



$|E(S_i)|$ vs iteration

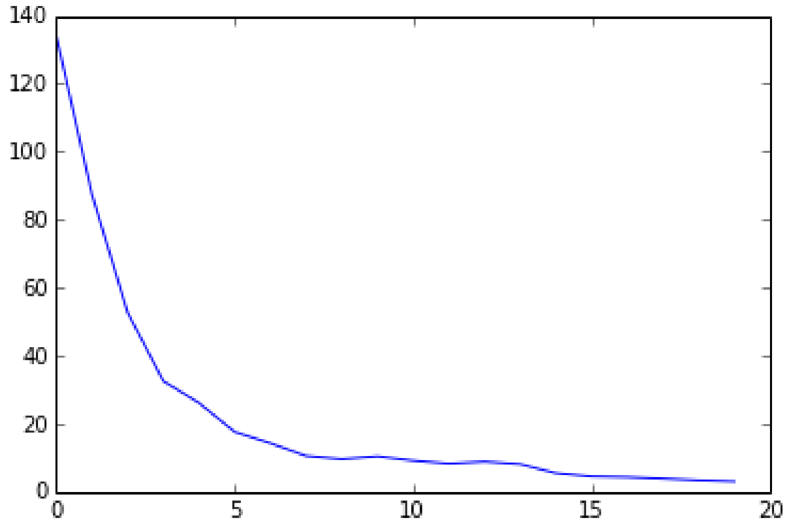


$|S_i|$ vs iteration

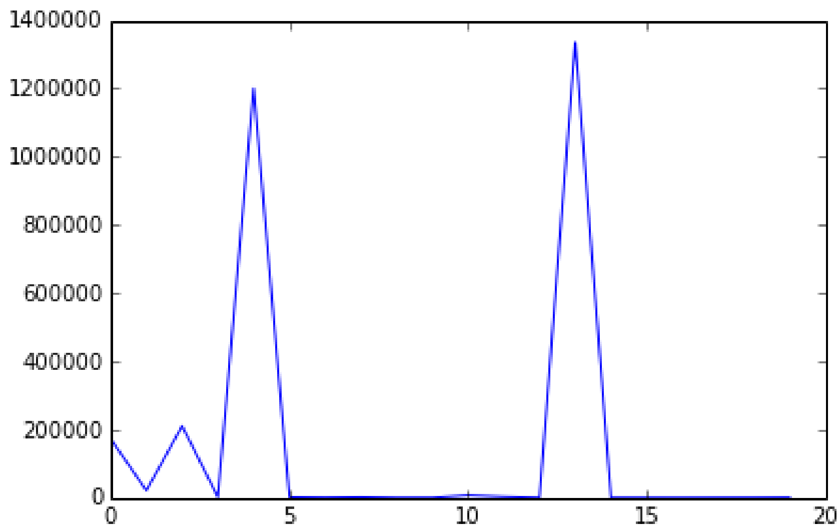


3

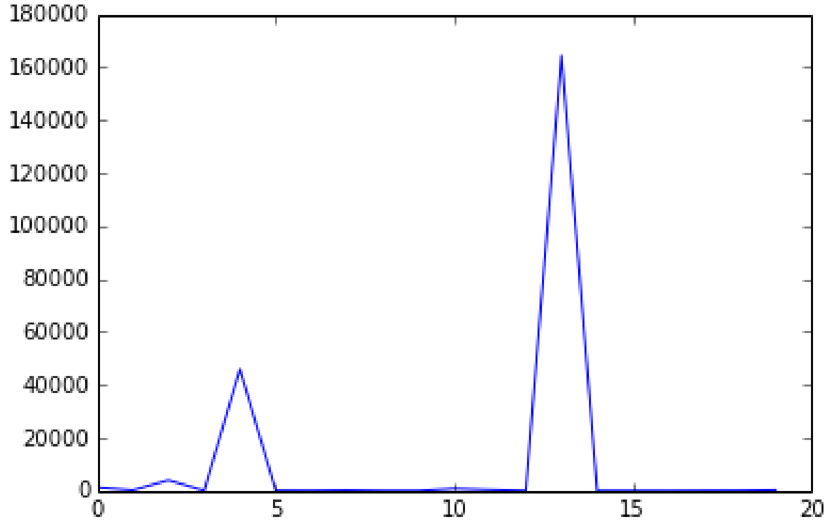
Density ($\rho(\bar{S}_j)$) vs iteration



$|E(\bar{S}_j)|$ vs iteration



$|\bar{S}_j|$ vs iteration



Code for Q1

Code for Q2

Code for Q4