Cover Sheet

Assignment Submission Fill in and include this cover sheet with each of your assignments. Assignments are due at 11:59pm. All students (SCPD and non-SCPD) must submit their homeworks via GradeScope (`http://www.gradescope.com`). Students can typeset or scan their homeworks. Make sure that you answer each question on a separate page. Students also need to upload their code at `http://snap.stanford.edu/submit`. Put all the code for a single question into a single file and upload it. Please do not put any code in your GradeScope submissions.

Late Day Policy Each student will have a total of *two* free late periods. *One late period expires at the start of each class.* (Homeworks are usually due on Thursdays, which means the first late periods expires on the following Tuesday.) 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.

Honor Code We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down their solutions independently i.e., each student must understand the solution well enough in order to reconstruct it by him/herself. Students should clearly mention the names of all the other students who were part of their discussion group. Using code or solutions obtained from the web (github/google/previous year solutions etc.) is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code very seriously and expect students to do the same.

Your name: MANOJ RAYI

Email: manoj1390@gmail.com **SUID:** manojr

Discussion Group: -

I acknowledge and accept the Honor Code.

CS246: Mining Massive Datasets
Homework 3

Answer to Question 1(a)

Derivative of error function

$$\varepsilon_{iu} = 2(R_{iu} - q_i * p_u^T)$$

Ignoring constant

$$\varepsilon_{iu} = R_{iu} - q_i * p_u^T$$

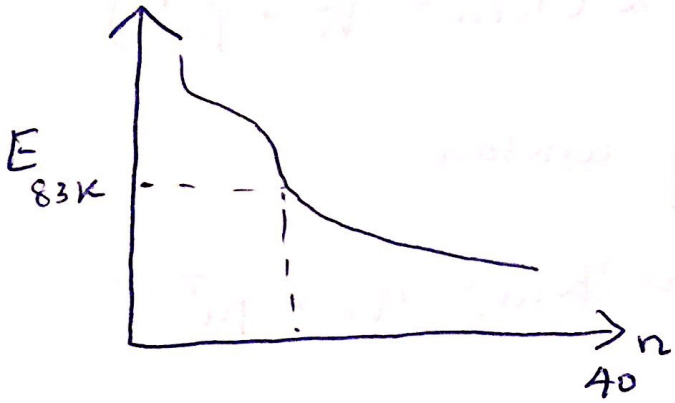
$$q_i \leftarrow q_i + \eta * (\varepsilon_{iu} * p_u - \lambda * q_i)$$

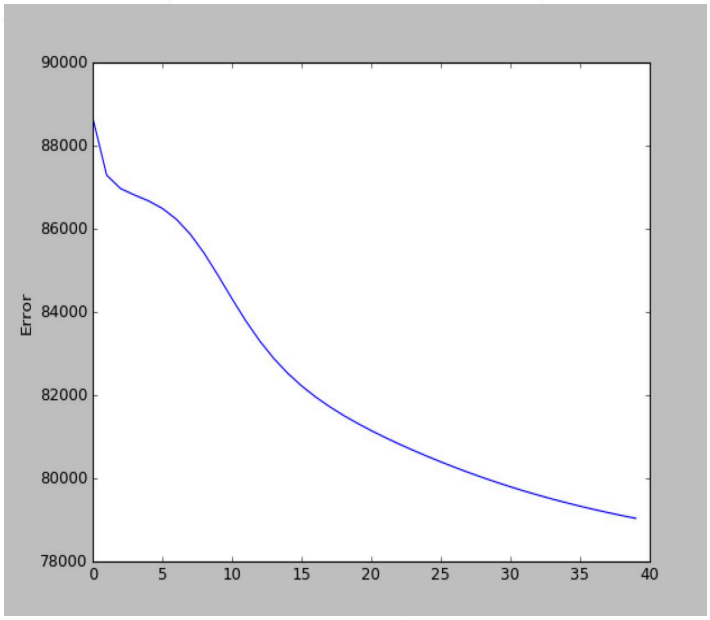
$$p_u \leftarrow p_u + \eta * (\varepsilon_{iu} * q_i - \lambda * p_u)$$

Answer to Question 1(b)

An η value of 0.025 reduces the error below 83000 within 40 iterations

(Graph attached)





Answer to Question 2(a)

(i) top 5 nodes with highest PR scores

27 , 1 , 14 , 40 , 53

0.02226 , 0.0223, 0.02474, 0.02505, 0.02608

(ii) bottom 5 node ids with lowest PR scores

85 , 59 , 81 , 23 , 37

0.003099 , 0.0032866, 0.00333, 0.003375, 0.0034469

Answer to Question 2(b)

(i) top 5 with highest hubbiness score

58 , 11 , 22 , 39 , 59

0.9574262, 0.957428, 0.97411071, 0.9810799, 1.0

(ii) bottom 5 with lowest hubbiness score

9 , 35 , 15 , 95, 53

0.209368 , 0.21233 , 0.221067, 0.22976 , 0.23548

(iii) node ids with highest authority score

1 53 27 40 66

0.8215458 0.8951798 0.9567024 0.9825375 1.

(iv) node ids with lowest authority score

54 33 24 67 50

0.04859 0.055604 0.065669 0.0676 0.06971

Answer to Question 3(a)

Ex: $M = \begin{bmatrix} 0 & 1/2 \\ 1 & 1/2 \end{bmatrix}$ $r^T = [1 \quad 1]$ } Sum = 2

$r_{new_1} = M r$

$= \begin{bmatrix} 0 & 1/2 \\ 1 & 1/2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

① ②

$= \begin{bmatrix} 1/2 \\ 1 + 1/2 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}$ } Sum=2

$r_{new_2} = M r_{new_1}$

$= \begin{bmatrix} 0 & 1/2 \\ 1 & 1/2 \end{bmatrix} \begin{bmatrix} 1/2 \\ 3/2 \end{bmatrix}$

$= \begin{bmatrix} 3/4 \\ 0.5 + 3/4 \end{bmatrix}$ } Sum=2

Proof :- For base case:

If M is the column stochastic,

1 is largest eigen value & other eigen values are smaller than 1

For eigen value 1 there exists a unique eigen vector with sum of entries equal to 1.

Hence by induction

$w(r_{new_n}) = \cdots\cdots w(r_{new_2}) = w(r_{new_1}) = w(r)$

Hence $w(r') = w(r)$

In another perspective, we know $r' = Mr$

M is column stochastic; column sum is 1

It is basically a transformation matrix for r. Which means it is just going to redistribute the weights of r in the new r'. Hence $w(r)$ is always constant despite ith iteration.

Answer to Question 3(b)

Page rank is the limit distribution of a stochastic process where states are Web pages. We need the convergence to occur for $w(r') = w(r)$. When β is too close to 1 numeric instability occurs.

$r' = \beta M r + (1-\beta)/n$ is irreducible and aperiodic for every $\beta \in [0,1)$

PR with high damping factor will result in total PR growing higher, the max limit is inbound PR $\leq \frac{\beta}{(1-\beta)}$

From previous example, if $\beta = 0.8$

$$r_{new_1} = 0.8 \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} + \frac{0.2}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 10/3 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5 \\ 0.43 \end{bmatrix}$$

All the pages of the web must be accounted for $w(r') = w(r)$

This is true unless the PR is scaled to 0 //

When we donot normalise the sum of pageRof pages will still sum to N and converse while $\beta \in (0.6, 1)$

Until M is column Stochastic and the distribution is stationary the sum of page rank will remain constant

w

Answer to Question 3(c)

without dead ends $r_i' = \beta \sum_{j=1}^{n} M_{ij} r_j + (1-\beta)/n$

Ex: $M \begin{bmatrix} 0 & 1/2 & 0 \\ 1 & 1/2 & 0 \end{bmatrix}$

with dead ends, $r_i' = \beta \sum_{j=1}^{n} M_{ij} r_j + \left(\sum_{j \in \text{live}} (1-\beta) r_j + \sum_{j \in \text{dead}} r_j \right) * \frac{1}{n} [1\ 1\ \ldots\ 1]^T$

$r_\xi' = \beta M r + \left[\left(\sum_{j \in \text{live}} (1-\beta) r_j + \sum_{j \in \text{dead}} r_i \right) * \frac{1}{n} [1\ 1\ \ldots\ 1]^T \right]$

Since β is $\in [0,1)$, M is colum stochastic

the converse of sum of PRs should holds are held when $w(r') = w(r)$

Also we could use the induction from 3(a) to prove this again.

Answer to Question 4(a)

Per graph theory,

$$\rho(S) = \frac{1}{2|S|} \sum_{v \in S} deg_S(v)$$

$$\Rightarrow 2|E(S)| = \sum_{v \in S} deg_S(v)$$

$$\geq \sum_{v \in S \setminus A(S)} deg_S(v)$$

$$\geq (|S| - |A(S)|) * 2(1+\epsilon)e(S)$$

$$\geq (|S| - |A(S)|) * 2(1+\epsilon)\frac{|E(S)|}{|S|}$$

$$\Rightarrow |E(S)|\,|S| \geq (|S| - |A(S)|)(1+\epsilon)$$

$$(1+\epsilon) \geq \epsilon |S| \quad \Rightarrow |A(S)| \geq \frac{\epsilon}{1+\epsilon} |S|$$

Suppose we have n nodes after each of the iterations, after ith iteration we will have atmost $n - \frac{\epsilon}{1+\epsilon} n$

Hence, the factor decrease is atleast $(1+\epsilon)$

$= \frac{n}{1+\epsilon}$ nodes

Thus algorithm terminates in atmost $\log_{1+\epsilon}(n)$ iterations

Answer to Question 4(b)

(i) for $v \in S^*$ let's assume $\deg_{S^*}(v) < \rho^*(G) = \rho(S^*)$

$$\Rightarrow \rho(S^*/v) = \frac{|E[S^*]| - \deg_{S^*}(v)}{|S^*| - 1} = \rho(S^*) + \frac{\rho(S^*) - \deg_{S^*}(v)}{|S^*| - 1}$$

$$> \rho(S^*)$$

This is in contradiction to the assumption that S^* is the densest subgraph

$\therefore$ by contradiction for any $v \in S^*$ $\boxed{\deg_{S^*}(v) \geq \rho^*(G)}$

(ii) $\because v \in A(S)$, $\deg_{S^*}(v) \not\leq 2(1+\epsilon)\rho(S)$

Since it is the first iteration such that $S^* \cap A(S) \neq \phi$

we can say $S^* \subseteq S$ and $\deg_S(v) \geq \deg_{S^*}(v)$

$\hookrightarrow$ Therefore $\rho^*(G) \leq \deg_{S^*}(v) \leq \deg_S(v) \leq 2(1+\epsilon)\rho(S)$

(iii) According to algo, we start with $S = V$ since $S^* \subseteq V$ and $2(1+\epsilon)\rho(S) \geq \rho^*(G)$ & (ii)

and continue removing vertices until $S = \phi$

We can see $\tilde{S}$ is a set which $\uparrow$ (maxes) $\rho(S)$ so,

$$\rho(S) \geq \frac{1}{2(1+\epsilon)} \rho^*(G)$$