CSE 547: Machine Learning for Big Data
Homework 3

Academic Integrity We take https://www.cs.washington.edu/academics/misconductacademic integrity extremely seriously. We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down the solutions and the code independently. In addition, each student should write down the set of people whom they interacted with.

Discussion Group (People with whom you discussed ideas used in your answers):
Wendan (Emily) Yan

On-line or hardcopy documents used as part of your answers:

https://www.math.ucdavis.edu/ strohmer/courses/180BigData/180lecture$_s pectral.pdf$
https://math.stackexchange.com/questions/3266709/prove-that-the-smallest-eigenvalue-of-a-symmetric-matrix-a-is-equal-to-the-min
https://people.eecs.berkeley.edu/ malik/papers/SM-ncut.pdf

I acknowledge and accept the Academic Integrity clause.

*(Dongyang Wang)*_____

Answer to Question 1(a)

Since there are no dead ends and by defnition, $w(r') = \sum_{i=1}^{n} r'_i = \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} r_j = \sum_{j=1}^{n} r_j \sum_{i=1}^{n} M_{ij} = \sum_{j=1}^{n} r_j * 1 = w(r)$

Answer to Question 1(b)

$w(r') = \sum_{i=1}^{n} r'_i$
$= \sum_{i=1}^{n} (\beta \sum_{j=1}^{n} M_{ij} r_j) + n * \frac{1-\beta}{n}$
$= \beta \sum_{j=1}^{n} r_j \sum_{i=1}^{n} M_{ij} + 1 - \beta$
$= \beta \sum_{j=1}^{n} r_j + 1 - \beta$
$= \beta w(r) + 1 - \beta$
Therefore, if $\beta = 1$ then the condition is satisfied.

Answer to Question 1(c)

$r' = \sum_{j \notin D}(\beta M_{ij} r_j + \frac{1-\beta}{n} r_j) + \sum_{j \in D} \frac{r_j}{n}$
Per the previous question,
$w(r') = \sum_{i=1}^{n}(\sum_{j \notin D}(\beta M_{ij} r_j + \frac{1-\beta}{n} r_j) + \sum_{j \in D} \frac{r_j}{n})$
$= \beta \sum_{j \notin D} \sum_{i=1}^{n} M_{ij} r_j + (1-\beta) \sum_{j \notin D} r_j + \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in D} r_j$
$= \beta \sum_{j \notin D} r_j + (1-\beta) \sum_{j \notin D} r_j + \sum_{j \in D} r_j$
$= \sum_{j=1}^{n} r_j$
$= w(r)$
$= 1$

Answer to Question 2(a)

PageRank:
The top 5 nodes are [263, 537, 965, 243, 285]
The bottom 5 nodes are [558, 93, 62, 424, 408]

Answer to Question 2(b)

The top 5 hubbiness score nodes are [840, 155, 234, 389, 472]
The bottom 5 hubbiness score nodes are [23, 835, 141, 539, 889]
The top 5 authority score nodes are [893, 16, 799, 146, 473]
The bottom 5 authority score nodes are [19, 135, 462, 24, 910]

Answer to Question 3(a)

For the first part,
By hint, $x_S^T L x_S = \sum_{\{i,j\}\in E}(x_i - x_j)^2 = 2\sum_{\{i,j\}\in E} A_{ij}(x_i - x_j)^2 = 2\sum_{i\in S, j\in \bar{S}} A_{ij}(\sqrt{\frac{vol(\bar{S})}{vol(S)}} - \sqrt{\frac{vol(S)}{vol(\bar{S})}})^2 = 2\sum_{i\in S, j\in \bar{S}} A_{ij}(\frac{vol(\bar{S})}{vol(S)} + \frac{vol(S)}{vol(\bar{S})} + 2)$ and by some algebra,
$x_S^T L x_S = 2 * |V| * cut(S) * \frac{1}{vol(S)vol(\bar{S})} = c * NCUT(S)$ where $c = 2 * |V|$

For the second part,
$x_S^T De = e * \sum_{i\in V} d_i x_s = e * (\sum_{i\in S} x_S^i d_i + \sum_{i\in \bar{S}} x_S^i d_i) = e * (\sqrt{\frac{vol(\bar{S})}{vol(S)}} * \sum_{i\in S} d_i - \sqrt{\frac{vol(S)}{vol(\bar{S})}} * \sum_{i\in \bar{S}} d_i) = e * (\sqrt{vol(S)vol(\bar{S})} - \sqrt{vol(S)vol(\bar{S})}) = 0$

For the third part,
$x_S^T D x_S = \sum_{i\in V} d_i x_s^2 = 2 * (\frac{vol(\bar{S})}{vol(S)} * \sum_{i\in S} d_i + \frac{vol(S)}{vol(\bar{S})} * \sum_{i\in \bar{S}} d_i) = 2 * (vol(\bar{S}) + vol(S)) = 2m$

Answer to Question 3(b)

Since we know that L is a symmetric matrix, then its smallest eigenvalue can be found by $min(u^T Lu) = \sum_{\{i,j\}}(u_i - u_j)^2$. Since the item is always non-negative, when $u_i = u_j$, we have the eigenvalue equal to 0. From results from Q3(a), we know that $x_S^T De = 0$ and so the eigenvector is e.

The minimization problem can be rewritten by solving the generalized eigenvalue system, as $(D-A)x = \lambda Dx$. With the constraints, we know that we are to solve $D^{-1/2}(D-A)D^{-1/2}z = \lambda z$. Since L is PSD, we also have Ł $= D^{-1/2}(D-A)D^{-1/2}$ to be PSD. Therefore, the smallest eigenvalue of Ł is a transformation of the smallest eigenvalue of L, which will be $D^{1/2}e$ given that $z = D^{-1/2}x$ and the smallest eigenvector in previous section is e.

We know all eigenvectors of are perpendicular to each other. Also, since $z^T z = 2m$, this is a constant term we can simply disregard. Then the problem becomes minimizing z^TŁz. Remembering the Rayleigh quotient, the quotient is minimized by the next smallest eigenvector and its minimum value is the corresponding eigenvalue. Therefore, the second smallest eigenvector is the solution, i.e., $z^* = v$ and by construction $x^* = D^{-1/2)}v$.

Answer to Question 3(c)

The number 1 user belongs to group 1
The number 2 user belongs to group 2
The number 3 user belongs to group 1
The number 4 user belongs to group 1
The number 5 user belongs to group 2

The number 1 feature is energy with value 327.78210857555644
The number 2 feature is danceability with value 37.6519559853
The number 3 feature is pitches with value 17.06088734106173

The clusters might represent two distinct extremes of music: one that represents really loud and music for dancing such as Tiktok music, but on the other extreme it would be quieter music for relaxation, etc.

The p-value for the feature energy is 0.0
The p-value for the feature danceability is 0.0001407643075458805
The p-value for the feature pitches is 0.08262747436647741

Pitches is not significant at the 0.05 level, but energy and danceability are. It's possible that when the music can be vastly different in whether they are energetic, but not so much when difference is small as in pitches.