Assignment 1

CS 512: Data Mining Principles (Spring 2021)
Instructor: Hanghang Tong

Release date: Feb. 1st, 2021
Due date: Mar. 5th, 2021

- This assignment will cover the content from slides #1 (Introduction), #2 (Frequent Pattern Mining), and #3 (Classification).
- Feel free to talk to other members in the class when doing the homework. We care more about whether you learn how to solve the problem entirely on your own than you demonstrate that you solved it. You should, however, write down your solution yourself. Please try to keep the solution brief and clear.
- Please use Piazza first if you have questions about the homework. Also feel free to send us e-mails and come to office hours.
- The homework is due at 11:59 PM on the due date. We will be using Compass (`http://compass2g.illinois.edu`) for collecting assignments. **Please do not hand in a scan of your handwritten solution, only the typed solution (e.g., Microsoft Word, Latex, etc) will be graded**. The datasets for HW1 is in **HW1_dataset.zip** on Compass. Contact the TAs if you are having technical difficulties in submitting the assignment. We do **NOT** accept late homework!
- The solution report should be submitted as a **single** pdf file using the name convention `yourNetid_HW1.pdf`. If you use additional source code (Python/Java/C++ are all acceptable) for solving problems, you are required to submit them and use the file names to identify the corresponding questions. For instance, '`yourNetid_HW1_problem1.py`' refers to the python source code for Problem 1 for HW 1. Compress all the files (pdf and source code files) into one zip file. Submit the compressed file ONLY.
- For each question, you will NOT get full credits if you only give out a final result. Necessary calculation steps are required. If the result is not an integer, round your result to 2 decimal places.

Problem 1. Short Answers. (8 points)

(a) (2 points) List two differences between Frequent Pattern Growth algorithm and Apriori algorithm (e.g., pattern generation, candidate generation, processing time).

Solution: (1) FP growth generates pattern by constructing a FP tree; Apriori generates pattern by pairing the items into singletons, pairs and triplets. (2) FP-growth does not have candidate generation process. Apriori uses candidate generation. (3) FP-growth is faster as compared to Apriori. The runtime of FP-growth increases linearly with increase in number of itemsets. The runtime of Apriori increases exponentially with increase in number of itemsets.

(b) (2 points) The Apriori algorithm make use of prior knowledge of subset support properties. Prove (1) all nonempty subsets of a frequent itemset also must be frequent; and (2) the support of any nonempty subset s' of itemset s must be at least as large as the support of s.

Solution: (1) let U be a frequent itemset and $support(U) = k$, this means in a given database, the number of transactions that contains all the items in U is k. If we remove any items V from those transactions, then the occurrences of items in $U \setminus V$ do not change in "these" transactions. Therefore the subset is also frequent. (2) for equality, it is obvious that any subsets of frequent itemset V is also frequent because the number of occurrences of the elements in subset is at least as V, for $>$, we can have an example, (a,b,c), (a,b), (a,b,c,d), let $min\_sup = 2$, compare (a,b,c) and (a,b).

(c) (2 points) Regarding measures of interestingness, what problem might *Lift* and χ^2 have? To address this problem, null-invariant measures are used, explain why Jaccard function is null-invariant? (hint: Venn diagram)

Solution: null transactions will impact the measure of interestingness;

using a Venn diagram, we can observe that Jaccard will not be affected by the background (i.e., null transactions $\neg A, \neg B$)

(d) (2 points) Overall, in sequential pattern mining, PrefixSpan has better performance compared to GSP and SPADE. Under what circumstance, PrefixSpan will also have poor performance?

Solution: When the number of frequent subsequences is very large because the major cost of PrefixSpan is constructing projection database.

Problem 2. Apriori vs Frequent Pattern Growth. (15 points)

TID	Items
1	$\{A, B, C, H, P\}$
2	$\{A, C, M, B, L\}$
3	$\{B, F, O, L\}$
4	$\{A, C, D, F, G, M, P\}$
5	$\{A, C, F, P, M\}$
6	$\{C, F, H, D\}$

Table 1: Transaction database.

A transactional database is given in Table 1.

(a) (6 points) Let $min\_sup = 3$ and apply Apriori algorithm on the dataset. Present the intermediate results of the first **three** scans and the final derived frequent itemsets.

Solution: (1) 1-itemset, A: 4, B: 3, C: 5, F: 4, M: 3, P: 3; (2) 2-itemset, AC: 4, AM: 3, AP: 3, CF: 3, CM: 3, CP: 3; (3) 3-itemset, ACM: 3, ACP: 3.

(b) (6 points) Let $min\_sup = 3$ and sort the frequent items first in frequency descending and then alphabetical order. Construct the Frequent Pattern-tree from the dataset in the same format as in the lecture.

Solution:

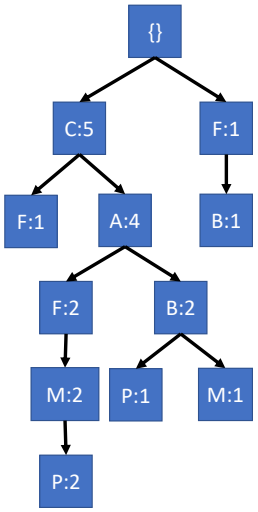

Figure 1: Frequent Pattern Tree for Problem 2(b).

(c) (3 points) List the top-3 association rules in terms of confidence from 2-itemsets.

Solution: A->C: 1, M->C: 1, P->C: 1, P->A: 1

Problem 3. Frequent Pattern Mining Implementation (12 points)
Implement the core algorithm of Apriori (direct calling Apriori from external packages is NOT allowed) on the grocery market transaction data (i.e., **groceries.csv**), and finish the following tasks.

(a) (2 points) Report data statistics, for example, #product types, average #products per transaction, 5 most popular products, etc.

Solution: the number of product types: 169; the average number of items per transaction: 4.41; top popular items: whole milk, other vegetables, rolls/buns, soda, yogurt, bottled water.

(b) (5 points) Manually define the threshold settings, e.g., $min\_support, min\_confidence$, record the settings and list the top-10 corresponding frequent itemsets if existed. Choose any pair of items (e.g., "yogurt" and "coffee"), construct the contingency table and evaluate the interestingness using *Lift*. Report the table and evaluation results.

Solution: The answers vary based on the settings and the code will be checked.

(c) (5 points) Report the running time of Apriori w.r.t. different data sizes (e.g., let the number of transactions be $1,000$, $2,000$, etc.). Compare the efficiency performance of Apriori and FP-Growth (You can directly call third-party libraries to run FP-Growth).

Solution: Line figures or tables of running time are preferred and the results vary based on the machine configuration.

Problem 4. Sequential and Graph Pattern Mining (10 points)

(a) A sequential database is presented in Table 2.

(1) (3 points) Given a set of prefixes, $\langle ab\rangle, \langle (ab)\rangle, \langle de\rangle$, $\langle abc\rangle, \langle acb\rangle, \langle acba\rangle$, list the corresponding projection(s) according to the definition of prefix and projection;
Solution: $\langle ab\rangle : (\_cf)(ac)b, (ce), (cd)f(abd)$;
$\langle (ab)\rangle : c(bc)ab(ce), (\_c)(ae), (\_d)df(ab), (\_c)eab(cd)f(abd)$;
$\langle de\rangle : (af)acb(acf)acd, (abc)(ae)$;
$\langle abc\rangle : (\_e), (\_d)f(abd)$;
$\langle acb\rangle : (acf)acd, (\_d)df(ab)$;
$\langle acba\rangle : (\_cf)acd$.

(2) (2 points) List the sequential pattern(s) where $support = 3$ and length ≥ 3;
Solution: Only 10 are needed in your answers. '(ab)b', '(ab)e', '(bc)ab', '(bc)b', '(bc)c', '(bc)e', 'a(ab)', 'a(bc)', 'a(bc)a', 'aaba', 'aabc', 'aabd', 'aabf', 'aabfa', 'aaca', 'aacb', 'aacd', 'aacf', 'aacfa', 'aad', 'aaf', 'aafa', 'abd', 'abf', 'abfa', 'acab', 'acba', 'acc', 'acd', 'acf', 'acfa', 'ad(ab)', 'ada', 'adb', 'afb', 'bab', 'bac', 'bca', 'bcb', 'bfa', 'caa', 'cac', 'cad', 'cbb', 'cbc', 'cbd', 'cbf', 'cbfa', 'cca', 'ccb', 'cfa', 'd(ab)', 'dfa', 'dfb', 'e(ab)', 'eab', 'eaba', 'eabd', 'eabf', 'eabfa', 'eac', 'eaca', 'eacb', 'eacd', 'eacf', 'eacfa', 'ead', 'eaf', 'eafa', 'ebd', 'ebf', 'ebfa', 'ecb', 'ecd', 'ecf', 'ecfa', 'efa', 'efb', 'fab', 'fcb'

SID	Sequence
1	$\langle a(bcf)(ac)b\rangle$
2	$\langle (ab)c(bc)ab(ce)\rangle$
3	$\langle de(af)acb(acf)acd\rangle$
4	$\langle (ad)de(abc)(ae)\rangle$
5	$\langle (aef)ac(abd)df(ab)\rangle$
6	$\langle (abc)eab(cd)f(abd)\rangle$

Table 2: Sequence Database.

(b) Given the input graphs (i.e., first row) and patterns (i.e., second row) in Figure 2.

(1) (3 points) Compute the support (in fraction) of the subgraph patterns and list the corresponding input graphs in which the patterns appear.
Solution: P1: 0.8, G2, G3, G4, G5 ;
P2: 0.6, G1, G2, G5;
P3: 0.6, G2, G3, G5

(2) (2 points) Let $min\_sup = 3$, show how Apriori-based method works to find all frequent graph patterns using either edge growing or vertex growing (present intermediate results/patterns at each iteration).

Solution:

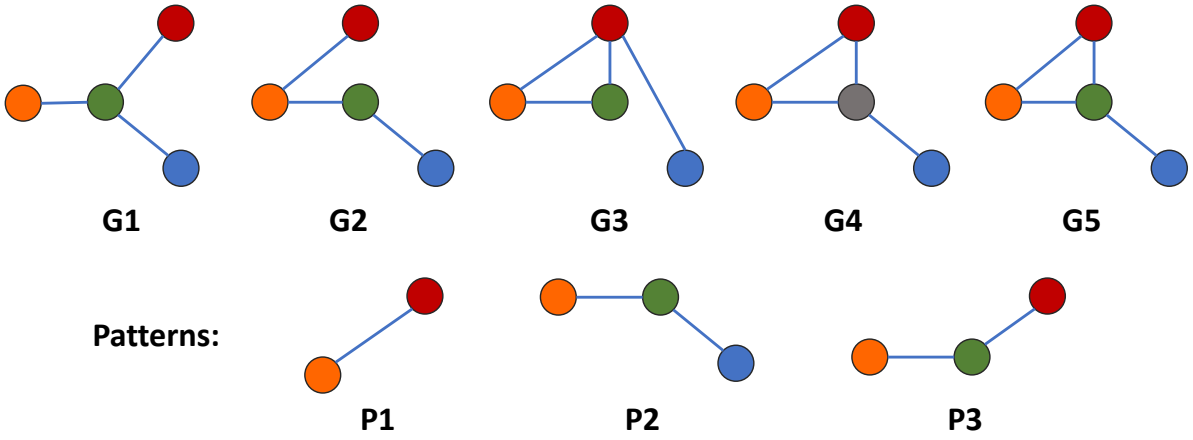

Figure 2: Input graphs and subgraph patterns.

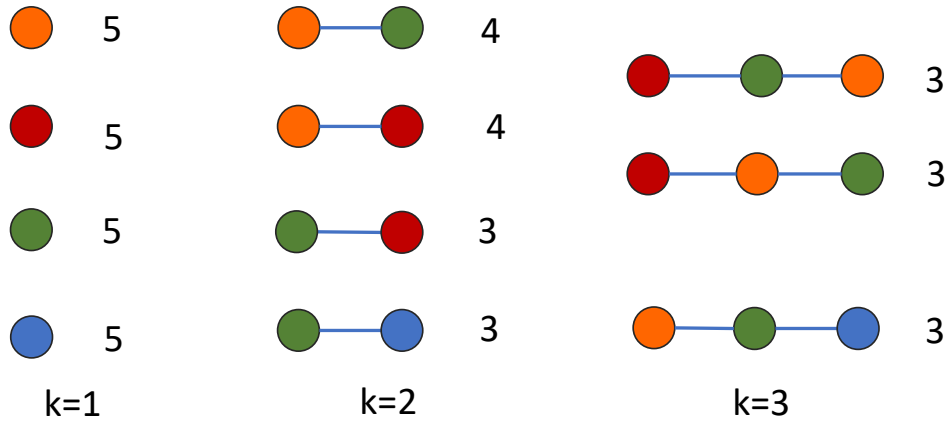

Figure 3: Solution for problem 4-(b).

Problem 5. SVM, RBF Kernel (15 points)
Given n training examples $(\mathbf{x}_i, y_i)(i = 1, 2, \ldots, n)$ where $\mathbf{x}_i$ is the feature vector of the i-th training example and y_i is its label, we train an support vector machine (SVM) with Radial Basis Function (RBF) kernel on the training data. Note that the RBF kernel is defined as $K_{RBF}(\mathbf{x}, \mathbf{y}) = \exp\left(-\gamma||\mathbf{x} - \mathbf{y}||_2^2\right)$.

(a) (5 points) Let $\mathbf{G}$ be the $n \times n$ gram matrix of RBF kernel, i.e., $\mathbf{G}[i, j] = K_{RBF}(\mathbf{x}_i, \mathbf{x}_j)$. Prove that all the eigenvalues of $\mathbf{G}$ are non-negative.

Solution: 'All the eigenvalues of $\mathbf{G}$ are non-negative' is equivalent to '$\mathbf{G}$ is positive semi-definite', i.e., for any vector $\mathbf{a}$, $\mathbf{a}^T\mathbf{G}\mathbf{a} \geq 0$. To prove that, we first prove that the gram matrix of linear kernel is positive semi-definite.

We have $K_{\texttt{linear}}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$. Hence the gram matrix of linear kernel can be written as:

$$\mathbf{G}_{\texttt{linear}} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \cdots \\ \mathbf{x}_n \end{bmatrix} \begin{bmatrix} \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \end{bmatrix} = \mathbf{X}^T\mathbf{X},$$

and

$$\mathbf{a}^T\mathbf{G}_{\texttt{linear}}\mathbf{a} = \mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a} \geq 0.$$

Hence, $\mathbf{G}_{\texttt{linear}}$ is a positive semi-definite matrix. Then, for the kernel function $K_{\texttt{exp}}(\mathbf{x}_i, \mathbf{x}_j) = \exp(\mathbf{x}_i^T\mathbf{x}_j)$ and its corresponding gram matrix $\mathbf{G}_{\texttt{exp}}$, based on the Taylor expansion, we have:

$$\mathbf{G}_{\texttt{exp}}[i, j] = \sum_{k=1}^{\infty} \frac{1}{k!}(\mathbf{G}_{\texttt{linear}}[i, j])^k. \tag{1}$$

Based on the properties of positive semi-definite matrix that: (1) the sum of two positive semi-definite matrices is a positive semi-definite matrix, and (2) the Hadamard product of two positive semi-definite matrices is a positive semi-definite matrix, we know that $\mathbf{G}_{\texttt{exp}}$ is a positive semi-definite matrix.

Then for the RBF kernel $K_{\texttt{RBF}}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma||\mathbf{x}_i - \mathbf{x}_j||_2^2\right)$ and its corresponding gram matrix $\mathbf{G}_{\texttt{RBF}}$, we have:

$$K_{\texttt{RBF}}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\mathbf{x}_i^T\mathbf{x}_i)\exp(2\gamma\mathbf{x}_i^T\mathbf{x}_j)\exp(-\gamma\mathbf{x}_j^T\mathbf{x}_j),$$

and

$$\mathbf{a}^T\mathbf{G}_{\texttt{RBF}}\mathbf{a} = \mathbf{b}^T\mathbf{G}_{\texttt{new}}\mathbf{b}$$

where $\mathbf{b}[i] = \mathbf{a}[i]\exp(-\gamma\mathbf{x}_i^T\mathbf{x}_i)$ and $\mathbf{G}_{\texttt{new}}[i, j] = \exp(2\gamma\mathbf{x}_i^T\mathbf{x}_j)$. Based on Eq. (1) it is clear that $\mathbf{G}_{\texttt{new}}$ is positive semi-definite. Hence, $\mathbf{b}^T\mathbf{G}_{\texttt{new}}\mathbf{b} \geq 0$ and $\mathbf{a}^T\mathbf{G}_{\texttt{RBF}}\mathbf{a} \geq 0$ and all the eigenvalues of $\mathbf{G}_{\texttt{RBF}}$ are non-negative

(b) (5 points) Prove that RBF kernel is the sum of infinite number of polynomial kernels.

Solution: By definition, we have

$$\begin{aligned} K_{\texttt{RBF}}(\mathbf{x}, \mathbf{y}) &= \exp(-\gamma||\mathbf{x} - \mathbf{y}||_2^2) \\ &= \exp(-\gamma(||\mathbf{x}||_2^2 + ||\mathbf{y}||_2^2 - 2\langle \mathbf{x}, \mathbf{y}\rangle)) \\ &= \exp(-\gamma||\mathbf{x}||_2^2 - \gamma||\mathbf{y}||_2^2) \exp(2\gamma\langle \mathbf{x}, \mathbf{y}\rangle) \end{aligned}$$

Let $C = \exp(-\gamma||\mathbf{x}||_2^2 - \gamma||\mathbf{y}||_2^2)$, we have $K_{RBF}(\mathbf{x}, \mathbf{y}) = Ce^{2\gamma\langle \mathbf{x}, \mathbf{y}\rangle}$. Then we have Taylor expansion on $e^{2\gamma\langle \mathbf{x}, \mathbf{y}\rangle}$ with

$$K_{\texttt{RBF}}(\mathbf{x}, \mathbf{y}) = Ce^{2\gamma\langle \mathbf{x}, \mathbf{y}\rangle} = C\sum_{i=0}^{\infty} \frac{(\langle \mathbf{x}, \mathbf{y}\rangle)^i}{i!} = C\sum_{i=0}^{\infty} \frac{K_{\texttt{poly(i)}}(\mathbf{x}, \mathbf{y})}{i!}$$

where $K_{\texttt{poly(i)}}$ is the polynomial kernel with degree i.

(c) (3 points) Suppose the distribution of training examples is shown in Figure 4 where '+' denotes positive example and '-' denotes negative example. If we set γ large enough (say 1000 or larger), what could possibly be the decision boundary of the SVM after training? Please draw it on Figure 4.

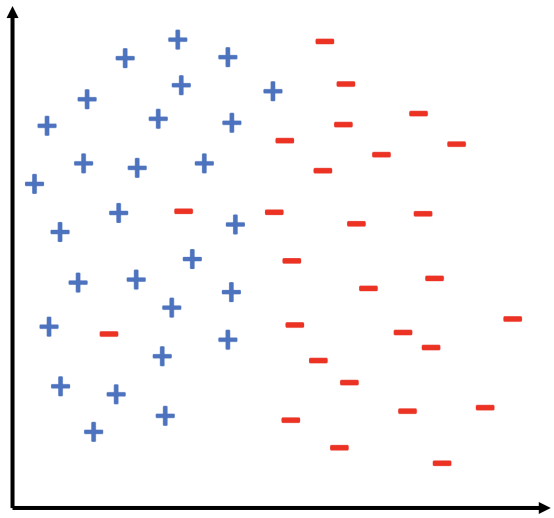

Figure 4: Distribution of Training Examples.

Solution: As Figure 5. It would lead into a perfect boundary on the training set.

(d) (2 points) if we set γ to be infinitely large, what could possibly happen when training this SVM?

Solution: It would lead to overfit on the training set.

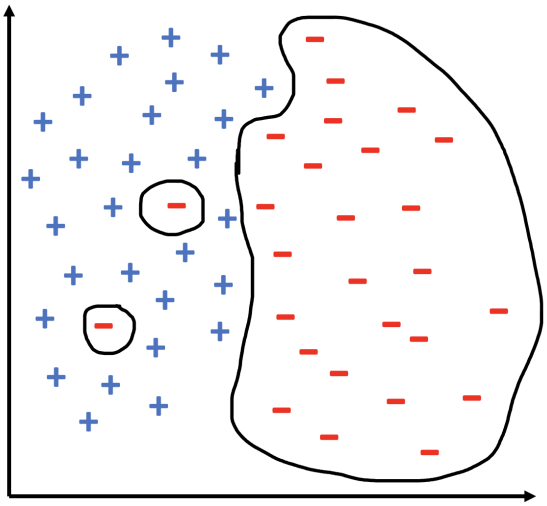

Figure 5: Solution of Problem 5-3.

Problem 6. Feature Selection (15 points)
Given the **risk auditing** dataset where every row refers to a data sample, the last column denotes the labels of data samples, and the remaining columns serve as the input features, try to:

(a) (5 points) implement an **SVM** classifier using the **train.csv**, **validation.csv**, and **test.csv**. Report the classification results over the test set. (3rd-party packages are **allowed**)

(b) (10 points) improve the performance of SVM classifier by implementing **Fisher Score** feature selection method. (3rd-party packages are **NOT allowed**)

Remarks that you should strictly follow the conventional machine learning paradigm to train your model on the training set, select the best hyper-parameter on the validation set and report the corresponding results on the test set. The source code you write should be submitted with the .pdf report in the compressed .zip file.

Solution: The Fisher score is computed as follows:

$$s = \frac{\sum_{j=1}^{c} n_j(\mu_j - \mu)^2}{\sum_{j=1}^{c} n_j \sigma_j^2}$$

where c is the number of classes, n_j is the number of samples of the j-th class, μ_j is the mean of feature value of samples from the j-th class, σ_j^2 is the variance of feature value of samples of the j-th class, μ is the mean of feature value of all the samples. Compute Fisher score of every feature and select the features with the k highest feature values (decide k based on the performance of model on validation set).

Problem 7. Self-Training (15 points)

Follow the improvement made from **Problem 6** via feature selection, try to further improve the classification results by self-training using the unlabelled data **unlabelled.csv**. Remarks that all the columns in unlabelled.csv should serve as input features. You should strictly follow the conventional machine learning paradigm to train your model on the training set, select the best hyper-parameter on the validation set and report the corresponding results on the test set. The source code you write should be submitted with the .pdf report in the compressed .zip file. [**HINT:** change the output of SVM into probability output and define the 'confidence' towards every unlabelled sample] (for the implementation of SVM, 3rd-party packages are **allowed** but for the implementation of self-training framework, 3rd-party packages are **NOT allowed**)

Solution: Follow the pseudo code from slides 'Classification' page 44. The confidence of every unlabelled sample can be defined based on 'the highest prediction probability', 'the highest prediction probability minus the second highest probability', 'entropy', and many more. The total number of augmented samples should be defined based on the performance on validation set.

Problem 8. Random Walk with Restart (10 points)

(a) (3 points) From its formula, try to explain why random walk with restart (RWR) can capture multiple weighted relationships between nodes?

Solution: The formula of RWR is as follows:

$$\mathbf{r} = c\tilde{\mathbf{A}}\mathbf{r} + (1-c)\mathbf{e}.$$

We obtain its closed-form solution as follows:

$$\mathbf{r} = (\mathbf{I} - c\tilde{\mathbf{A}})^{-1}(1-c)\mathbf{e}$$

where $(1-c)\mathbf{e}$ denotes the preference of the user, and $(\mathbf{I} - c\tilde{\mathbf{A}})^{-1}$ can be expanded as $\sum_n (c\tilde{\mathbf{A}})^n$. The $\tilde{\mathbf{A}}^n$ can present n-hop relationships. For example, $\tilde{\mathbf{A}}$ indicates 1-hop relationship (e.g., direct friendship), and $\tilde{\mathbf{A}}^2$ indicates 2-hop relationship (e.g. friend's friend). c^n is the weight on the n-hop relationship which means the importance of a path decreases with the increase of path length (i.e., number of hops). Hence, RWR score can capture multiple weighted relationships between nodes.

(b) (7 points) Implement random walk with restart (RWR) on the **email-EU-core** unweighted undirected graph. The given file is an edge list and since it is an unweighted undirected graph, when see 'i j' in the edge list, set both $\mathbf{A}[i,j]$ and $\mathbf{A}[j,i]$ as 1. Set preference vector $\mathbf{e}$ as a uniform vector to obtain global ranking and set the damping factor c as 0.9. Report the indices of top-10 nodes and the source code you write should be submitted with the .pdf report in the compressed .zip file. [**HINT:** normalize adjacency matrix before RWR: $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is an diagonal matrix s.t. $\mathbf{D}[i,i] = \sum_j \mathbf{A}[i,j]$] (3rd-party packages are **NOT allowed**)

Solution: Update $\mathbf{r}$ based on $\mathbf{r} \leftarrow c\tilde{\mathbf{A}}\mathbf{r} + (1-c)\mathbf{e}$ till convergence and report the indices of nodes with the 10 highest scores.