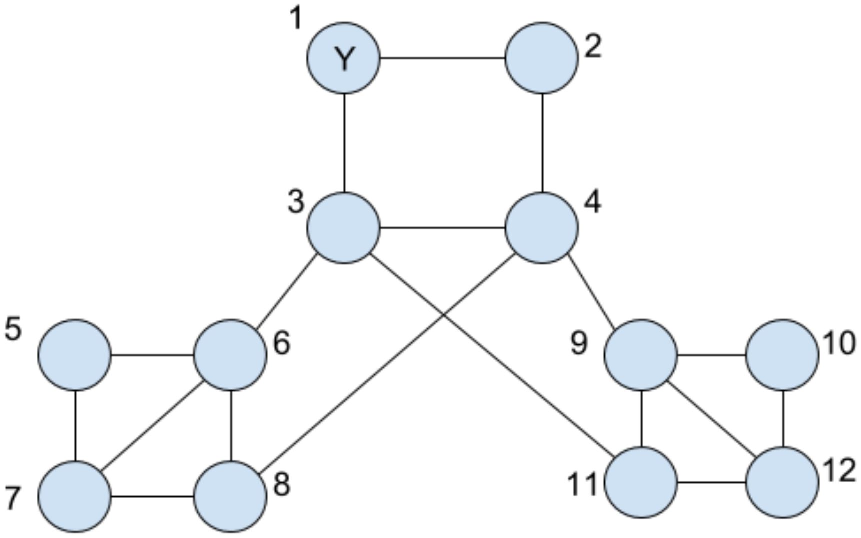
Big Data Analysis (Practice Final Exam)

Question 1

The parliament has organized a voting scheme for a new bill this summer. You are a strategic advisor in charge of vote forecasting and voter acquisition tactics. You have the following social graph of voters, which is undirected.



The undecided voters will go through a 3-day decision period where they choose a candidate based on the majority of their friends. The decision period works as follows:

1. The graphs are initialized with every voter's initial state as the above figure. (yes (Y), no (N), or undecided)
2. In each day, every undecided voter decides on a vote 'yes' or 'no'. Voters are processed in an increasing order of node ID. For every undecided voter, if the majority of their friends (>=50%) vote 'yes', they now vote 'yes'. Otherwise, they vote 'no'.
3. When processing the updates, use the values from the current day. For example, when update the votes for node 2, you should use the updated votes for nodes 1 and 4 from the current day.
4. There are 3 days of the process described above.
5. On the 4th day, the votes are counted.

a) Perform iterations of the voting process. How many votes each option has?

b) You have a public relation idea to increase the 'yes' voters by organizing a very classy $1000 per plate dinner event. Assume everyone that comes to your dinner is instantly persuaded to vote 'yes' regardless of his/her previous decision. This event will happen before the decision period.

Choose a minimum number of voters to invite for dinners such that all the voters in the graph vote 'yes'. Justify your strategy and compute the voting result.

c) You have another idea to increase the 'yes' voters by spending $1000 to make any two voters in the network become friends.

Choose a minimum number of connections you want to create such that all the voters in the graph vote 'yes'. Justify your strategy and compute the voting result.

Question 2

Schema reuse is a new trend in creating database schemas by allowing users to copy and adapt existing ones. The motivation behind schema reuse is the slight differences between schemas in the same domain; thus making schema design more efficient. Reusing existing schemas supports reducing not only the effort of creating a new schema but also the heterogeneity between schemas.

Finding related schemas is one of the core problems of schema reuse. You work as a data engineer at Oracle. Oracle has a large repository of schemas. Each database schema has a set of attributes. Some attributes are common among schemas, while others are not. Your task is to support database designers to create new schemas via the schema reuse paradigm. For instance, when a database designer wants to create a new schema, he wants to query the schema repository for references:

- He can start with a few attributes and query the schema repository for hints to finish his design.
- Alternatively, he can complete a schema and query the schema repository to check his design.

Example: we have a repository of schemas:

- *S1: {a1, a3, a7}*
- *S2: {a1, a4, a8}*
- *S3: {a2, a6, a9}*
- *S4: {a1, a5, a10}*

Given a query Q = {a1, a2}, we should rank these schemas as S3 > S1=S2=S4. S3 has the highest rank since attribute a1 occurs frequently in many schemas and thus has less discriminatory power (i.e. the more schemas contain an attribute, less information it provides).

Design an algorithm to find related schemas ranked by their similarity to the query.

a) How do you model the problem (input, output, etc.)? Justify your model.

b) What steps should be involved? Provide a quantitative measure for each step if needed. Justify the design choice for each step.

c) Apply your approach to the above example and calculate the quantitative results.

Question 3

As shown in the table below, the dataset contains the ratings from 4 users to 4 movies. The ratings range from 1 to 5 stars.

	Alice	Bob	Carol	Dave
Love at last	?	5	1	4
Romance forever	2	5	3	?
Nonstop car chases	4	?	5	4
Swords vs. karate	3	5	4	5

- ❖ Use cosine similarity to compute the missing rating in this table using user-based collaborative filtering (CF). Provide the detailed calculation.
- ❖ Similarly, computing the missing rating using item-based CF. Provide the detailed calculation.

The table below is the ground truth for the given data:

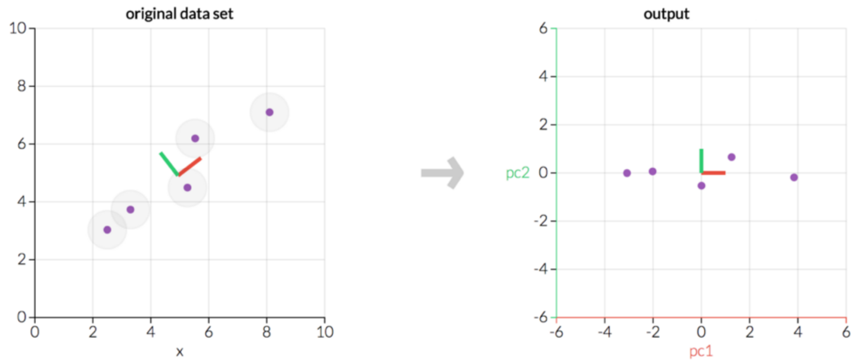
	Alice	Bob	Carol	Dave
Love at last	4	5	1	4
Romance forever	2	5	3	3
Nonstop car chases	4	4	5	4
Swords vs. karate	3	5	4	5

- ❖ Compute the predictive accuracy of the above recommendation.

Question 4

a) What are the differences between feature selection and feature reduction?

b) Given the following image, which technique is used? Justify your choice.

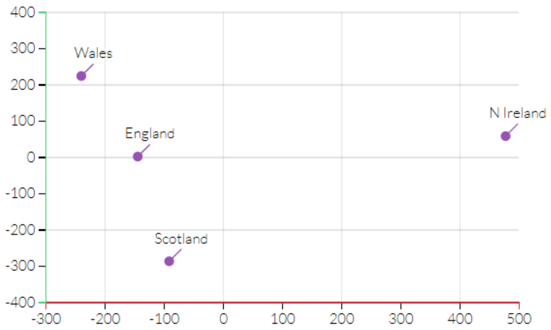


Question 5

The following table presents the average consumption of 17 food types in grams per person per week in the UK.

	England	Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175
Alcoholic drinks	375	135	458	475

After viewing the outlier presented in the below image, two members in the data analysis team argue about the technique used to get such a result. The first member, Tom, claims that feature selection is applied to get the optimal subset. Whereas the second member, Adam, thinks that the technique behind it must be feature reduction.



a. Who is correct? Explain why.

b. Compared with the table's data, do you think that the graph makes sense? Explain why.

Question 6

A company has collected quarterly sales for the past two years which are shown in the following table. The company wants to forecast the next year's seasonal sales.

Index	Time	Sales ($)	Index	Time	Sales ($)
1	Spring 2017	4836	5	Spring 2018	5412
2	Summer 2017	5890	6	Summer 2018	6138
3	Fall 2017	6510	7	Fall 2018	6666
4	Winter 2017	7564	8	Winter 2018	8184

a) Let A_1 and A_2 be the actual total sales (i.e., the sum of all four seasonal sales) in 2017 and 2018, respectively. Assume current time is t, the n-moving average (MV) technique makes forecast for the time $t+1$ by taking the average of previous n actual values where $n \leq t$. (The formula can be written as $F_{t+1} = \frac{\sum_{i=t-n+1}^{t} A_i}{n}$ where A_i is the i-th actual data). Predict the total sales for 2019 using MV with 2 actual values A_1 and A_2.

b) In general, we would expect the total sales gets increased in both 2018 and 2019 if the economy situation has been keeping going well since 2017. Use this and your answer to sub-question **a)** to explain the limitation of moving average (MV) method in forecasting.

c) Calculate the average seasonal sales for both 2017 and 2018.

d) Here are the steps of forecasting with seasonality. Please follow the steps to fill out the blank cells in the following form.

1. Calculate the average seasonal sales for each year;

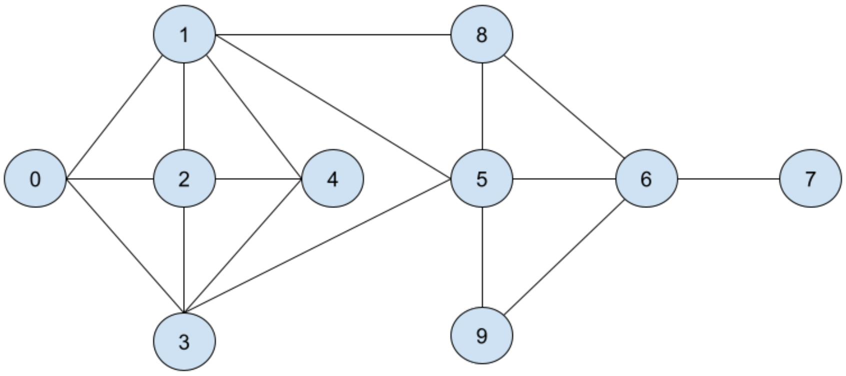
2. Calculate each seasonal index (by dividing the actual seasonal sales by the average seasonal sales);
3. Compute the average indexes;
4. Predict the average seasonal sales for the next year (i.e., 2019);
5. Multiple next year's average seasonal sales by each <u>average</u> seasonal index.

(**Note**, you have already done the step 1) by answering sub-question c) and please put those values into the table. Also, step 4 is already done for you (the red value). You can copy the form to your answer sheet then fill blank cells).

Quarter	2017	Seasonal Index	2018	Seasonal Index	Average Index	2019
Spring	4836		5412			
Summer	5890		6138			
Fall	6510		6666			
Winter	7564		8184			
Average						6900

Question 7

Given the following network:



a) Compute the adjacency matrix A of this network, where each cell A_{ij} is the number of edges from node i to node j.

b) Given an iterative algorithm that computes the eigenvector centrality for every node v_i at k-th iteration as follows:

$$C_k(v_i) = \frac{1}{10}\sum_{j=0}^{9} A_{ij} \times C_{k-1}(v_i)$$

The algorithm converges when the centrality values do not change significantly after two consecutive iterations.

At iteration 0, Alice initialises the nodes with the same centrality values:

$$v_0 = v_1 = v_2 = v_3 = v_4 = v_5 = v_6 = v_7 = v_8 = v_9 = 0.5$$

and Bob initialises the nodes with the following values:

$$v_0 = 3, v_1 = 5, v_2 = 4, v_3 = 4, v_4 = 3, v_5 = 5, v_6 = 4, v_7 = 1, v_8 = 3, v_9 = 2$$

Do they achieve the same ranking of nodes in terms of centrality values after the algorithm converges?

c) Between Alice and Bob, who reaches the final ranking faster (less iterations)? Explain your answer. (Note: we count the number of iterations excluding initialization and including the final iteration that reaches convergence)

d) Charles has another initialization strategy as follows.

$$v_0 = 3, v_1 = 1, v_2 = 2, v_3 = 2, v_4 = 3, v_5 = 1, v_6 = 2, v_7 = 5, v_8 = 3, v_9 = 4$$

Among Alice, Bob, and Charles, whose strategy has the least number of iterations? Whose strategy has the most number of iterations? Explain your answer.

Question 8

Alice wants to build a retrieval system for a movie database. Each movie is associated with a set of tags.

Consider the following movie collection C which has 4 movies: $C = \{M_1, M_2, M_3, M_4\}$.

Movies	Tags
M_1	Adventure, Action
M_2	Action, Comedy
M_3	Comedy, Adventure
M_4	Action, Fantasy, Adventure

Let Dict be a dictionary which consists of 5 tags: Dict = {t_1 = Adventure, t_2 = Action, t_3 = Comedy, t_4 = Fantasy, t_5 = Documentary}.

a) Denote by $tf(t, M)$ the tag frequency (TF) of the tag t in the movie M. Please fill out the blank cells in the following table, i.e., give the values $tf(t_i, M_j)$ for $i = 1,3,5$ and

$1 \leq j \leq 4$. The value $tf(t_i, M_j)$ should be put into the cell specified by t_i and M_j. (**Note,** you can copy the form to your answer sheet and the fill it out.)

	M_1	M_2	M_3	M_4
t_1				
t_3				
t_5				

b) Recall the inverse document frequency (IDF) is defined as $idf(t, C) = \ln \frac{|C|}{|C_t|}$. Here $|C|$ denotes the number of movies in the collection C, $|C_t|$ denotes the number of movies from C that contains tag t. Please compute $\text{idf}(t_1, C)$ and $\text{idf}(t_3, C)$.

c) Tag frequency – inverse document frequency (TF-IDF) takes both tag frequency (TF) and inverse document frequency (IDF) into consideration. For the movie collection C, the tag t's TF-IDF value on movie M_i is defined as $tfidf(t, M_i, C) = tf(t, M_i) \times idf(t, C)$, i.e., the product of t's TF value on M_i and t's IDF value. Please compute $tfidf(t_1, M_3, C)$ and $tfidf(t_3, M_3, C)$.

d) Using TF-IDF, a movie can be represented by a 5-dimension vector of TF-IDF values of 5 tags in the dictionary. Compute the vector for each movie.

e) Similarly, any given query can be represented by a 5-dimension vector. Compute the vector for the query “Action”.

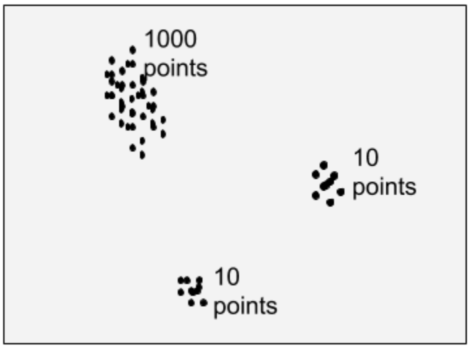
f) Using document vectors, we can compute the relevance score for each document to a given query using cosine similarity between the document vector and the query vector. A retrieval result of a query is the ranking of the movies in the decreasing order of relevance scores. What is the result of the query “Action”?

g) If each movie has an additional tag “Hollywood”, what is the result of the query “Action”? What is the result of the query “Hollywood”? And what is the result of the query “Hollywood Action”?

h) If the appearance order of tags in each movie matters, how do you propose to solve the retrieval problem? What is the result of the query “Action” now?

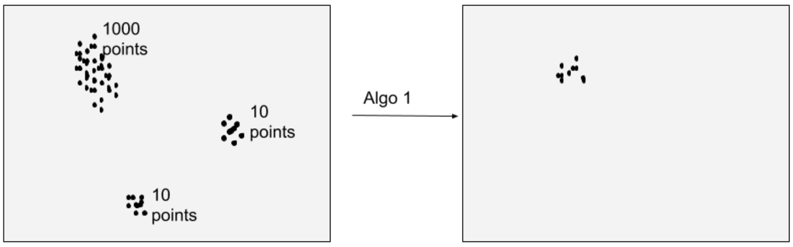
Question 9 (just for practice, data sampling is not a part of the final exam)

Corporation A wants to investigate the performance of its server swarm. Each server has 2 attributes: number of errors and number of users. The servers can be depicted in 2D as follows.



Corporation A hires Allen to do a sample testing on these servers. Their budget allows him to select only 10 servers for testing.

Write an algorithm in pseudo-code or Python to select the samples as the following image.


1000 points
10 points
10 points
Algo 1