COMP5434 – Big Data Computing

Assignment 1

Due Date: 11:59 PM, June 10th, 2025. (No Extension. Late policy applies.)

Submission: Please submit a PDF file to Blackboard.

Total: 3 Questions

You can use as many pages as you want.

Notes:

(1) Please include your name and student ID in your submission.

(2) Please submit your answer PDF file in advance, to avoid network congestion. Network, device, or related problems cannot be acceptable reasons for late submission.

(3) Please use **electronic editors** such as Word to draft your answers instead of taking photos or screenshots of your hand-written answers.

(4) Do NOT include question text in your submissions.

(5) Please provide **calculation details** in your answers, i.e., the complete logical reasoning path about how you get the final results from the given information. Answers with only final results are not acceptable.

(6) Do NOT write programs and post codes in your answers.

(7) If you get a result with more than 3 decimal places during your calculation, round values in **3 decimal places**. (e.g., 0.0423 -> 0.042, 1.3458 -> 1.346, -1.5389 -> -1.539)

Question 1: Basic Concepts of Big Data. (35/100)

Task 1 (6 Marks): What are the basic ideas of *Scale-up* method and *Scale-out* method respectively?

Task 2 (4 Marks): Which two kinds of information does Spatial Temporal data contain?

Task 3 (8 Marks): Given the following numbers, re-scale each number by using **"scaling to a range"** method.

$$[2, 4, 3, 10, 8, 5, 6, 7]$$

Task 4 (11 Marks): Please draw a graph whose matrix is as follows. Numbers outside the matrix are the IDs of nodes. (The graph can be drawn by using online tools, such as https://graphonline.top/en/)

	1	2	3	4	5
1	0	0	5	0	2
2	0	0	0	6	3
3	5	0	0	2	1
4	0	6	2	0	0
5	2	3	1	0	0

Task 5 (6 Marks): What is the main difference between Descriptive Analytics and Predictive Analytics?

Question 2: Text Data Processing. (35/100)

Here are processed bags of words in 5 documents.

$D_1 = \{apple, orange, fruit, fruit, pear\}$

$D_2 = \{apple, fruit\}$

$D_3 = \{orange, banana\}$

$D_4 = \{orange, apple\}$

$D_5 = \{fruit, banana, banana\}$

The vocabulary for this task is $V = \{apple, fruit, banana, orange, pear\}$

Task 1 (25 Marks): Generate TF-IDF vectors for these five documents. The order of components should be **the same as the order of words in the vocabulary**, i.e., the first component should be the weight of *apple*, the second component should be the weight of *fruit*, and so on.

In this task, TF is defined as $\frac{f_{t,d}}{\Sigma_{(t' \in d)} f_{t',d}}$, where $f_{t,d}$ is the raw occurrence of term t in the document d. IDF is defined as $\log_{10}\left(\frac{N}{1+n_t}\right) + 1$, where n_t is the number of documents that contain the term t and N is the total number of documents in this task.

Task 2 (10 Marks): Which document is the most similar document to D_5? Please use Cosine Similarity to find it.

Question 3: Association Rules. (30/100)

We have already got 6 transactions from a shop. Find all association rules by using *Apriori* algorithm. The minimum support is set as 0.4 and the minimum confidence is set as 0.8. Show your steps to find the rules.

Transaction ID	Items
1	A, B, D, E
2	A, D, E
3	A, B
4	C, D, E
5	A, C, D, E
6	C