1. (20 points) Working with strings/texts.

(a) Given two strings s1 = ATCGTACGTGTA, and s2 = TCGTACGTGTAA, what's their Hamming distance? (3 pts)

Answer -> s1 = ATC GTA CGT GTA
s2 = TCG TAC GTG TAA

We can see that 11 characters are mismatched in the two strings so the hamming distance between the two strings would be 11.

(b) Now represent s1 and s2 as vectors of 2-shingles (shingles of 2 letters), and compute their Jaccard similarity. (7 pts)

Answer -> s1 = { [A,T], [T,C], [C,G], [G,T], [T,A], [A,C], [T,G]}
s2 = { [T,C], [C,G], [G,T], [T,A], [A,C], [T,G], [A,A]}

Jaccard Similarity = | s1 ∩ s2 | / | s1 ∪ s2 |
= 6 / 8 = 0.75

(c) Based on your calculations above, which of the two metrics (Hamming distance or Jaccard similarity) do you think better captures the similarity/dissimilarity between these two strings? (3 pts)

Answer -> Since the strings are of the same length Hamming distance better captures the dissimilarity. But if the strings were of different lengths Jaccard similarity would have been optimal.

(d) Given two strings of average lengths of n letters represented as vectors of 2-shingles, what's the time complexity of computing the Jaccard similarity between the vectors? Use big O notation. You don't need to provide formal proof, but brief explanations are required. (7 pts)

Answer -> To represent the strings as vectors of 2-shingles we will have to iterate over the entire length of the string which takes O(n) time.

To find the intersection between the vectors in the worst case scenario all shingles could be similar so time complexity would be O(n).

So, Total time taken = O(n) + O(n) + c(constant time for addition and subtraction to calculate union)

So the time complexity to calculate Jaccard Similarity would be O(n).

2. (20 points total) Given a shingle(word)-document matrix as below,

Shingle_ID	**d1**	**d2**	**d3**	**d4**	**d5**	**d6**
0	**0**	**0**	**0**	**1**	**0**	**0**
1	**0**	**0**	**1**	**0**	**0**	**0**
2	**1**	**1**	**1**	**0**	**0**	**0**
3	**1**	**1**	**0**	**1**	**1**	**1**
4	**1**	**0**	**0**	**1**	**1**	**1**
5	**0**	**1**	**0**	**0**	**1**	**1**

(a) Compute the Jaccard similarity for all pairs of the six documents (5 points).
Answer -> J=(f11)/(f01 + f10 + f11)
J(d1,d2) = 2 / (1 + 1 + 2) = 2 / 4 = 0.5
J(d1,d3) = 1 / (1 + 2 + 1) = 1 / 4 = 0.25
J(d1,d4) = 2 / (1 + 1 + 2) = 2 / 4 = 0.5
J(d1,d5) = 2 / (1 + 1 + 2) = 2 / 4 = 0.5
J(d1,d6) = 2 / (1 + 1 + 2) = 2 / 4 = 0.5
J(d2,d3) = 1 / (1 + 2 + 1) = 1 / 4 = 0.25
J(d2,d4) = 1 / (2 + 2 + 1) = 1 / 5 = 0.2
J(d2,d5) = 2 / (1 + 1 + 2) = 2 / 4 = 0.5
J(d2,d6) = 2 / (1 + 1 + 2) = 2 / 4 = 0.5
J(d3,d4) = 0
J(d3,d5) = 0
J(d3,d6) = 0
J(d4,d5) = 2 / (1 + 1 + 2) = 2 / 4 = 0.5
J(d4,d6) = 2 / (1 + 1 + 2) = 2 / 4 = 0.5
J(d5,d6) = 3 / (0 + 0 + 3) = 3 / 3 = 1

b) Compute the minhash signatures for each column (document) if the following hash functions are used: h1(x) = (2x + 1)%6; h2(x) = (3x + 2)%6; h3(x) = (5x + 2)%6; and h4(x) = (7x + 3)%6 (8 points).
Answer ->

Shingle_ID	d1	d2	d3	d4	d5	d6	(2x + 1)%6	(3x + 2)%6	(5x + 2)%6	(7x + 3)%6
0	0	0	0	1	0	0	1	2	2	3
1	0	0	1	0	0	0	3	5	1	4
2	1	1	1	0	0	0	5	2	0	5
3	1	1	0	1	1	1	1	5	5	0
4	1	0	0	1	1	1	3	2	4	1
5	0	1	0	0	1	1	5	5	3	2

	d1	d2	d3	d4	d5	d6
h1	1	1	3	1	1	1
h2	2	2	2	2	2	2
h3	0	0	0	2	3	3
h4	0	0	4	0	0	0

(c) Compute the similarities for all pairs of the six documents using the signatures (5 points).

Answer ->
Sim(d1,d2) = 4/4 = 1
Sim(d1,d3) = 2/4 = 0.5
Sim(d1,d4) = 3/4 = 0.75
Sim(d1,d5) = 3/4 = 0.75
Sim(d1,d6) = 3/4 = 0.75
Sim(d2,d3) = 2/4 = 0.5
Sim(d2,d4) = 3/4 = 0.75
Sim(d2,d5) = 3/4 = 0.75
Sim(d2,d6) = 3/4 = 0.75
Sim(d3,d4) = 1/4 = 0.25
Sim(d3,d5) = 1/4 = 0.25
Sim(d3,d6) = 1/4 = 0.25

Sim(d4,d5) = 3/4 = 0.75
Sim(d4,d6) = 3/4 = 0.75
Sim(d5,d6) = 4/4 = 1

(d) Does minhash provide good signatures for computing the document similarity for this example (2 points)?
Answer ->
No, if we consider the difference of 0.25 is not good then the min hash signatures only match the true similarity once in case of d5,d6 and better hash functions should be used.
But if the difference of 0.25 is acceptable then its doing a good job for calculating the similarity with a small margin of error.

3. (10 points) Credit card fraud detection using KNN. Please use this code as the start code.
• Before calling KNN for classification, were there data processing steps applied in the start code? What distance metric was used in KNN in the start code? (5 points)
Answer -> Yes some data processing step were applied like ->

- Check for missing values
- Changing Class columns datatype to bool
- A new column was added called as AmountNormalized which contains transformed and scaled data of the column Amount

4. (25 points total) Link analysis
a) Given the above toy web (with 6 web pages, A, B, ···, F), derive its transition probability matrix (5 points).
Answer ->

	A	B	C	D	E	F
A	0	0.5	0	0	0.5	0
B	0.25	0	0.25	0.25	0	0.25
C	0	0	0	1	0	0
D	0	0	0	0	0.5	0.5
E	0	0	0.5	0	0	0.5

F	0.5	0	0.5	0	0	0

(b) Assume a surfer is on web page A, what’s the probability that the person will be next visiting B and then D (5 points)?

Answer ->

P(A->B) = 0.5

P(B->D) = 0.25

P(A->B->D) = P(A->B)*P(B->D) = 0.5 * 0.25 = 0.125

5. Write a summary for this review article about Precision Nutrition (PN): Precision nutrition: Maintaining scientific integrity while realizing market potential. [25 pts]

Answer - >

Precision Nutrition(PN) can be defined as an approach that uses individual data to predict how a person will respond to specific foods or dietary patterns and tailors dietary recommendations to their individual needs. One size fits all has been an older approach. Over the years of development ,promising research results support the predictive potential of assessments of the gut microbiome and metabolome—among other factors of our biology. PN is aiming to introduce a more scientific, predictive and personalized approach to understand how diet impacts health.

Key attributes of the PN approach are listed are -

- Personalized
- Reliable
- Evidence-based
- Complex
- Integrative
- Systematic
- Evolving

PN is personalized as it is user data dependent, it is based on scientific evidence and robust methodology, and it evolves based on the changes in the data.

Older methods of assessments like food questionnaires, diet records have inaccuracies and are not convenient. With active developments in sensor and wearable technologies along with user friendly apps, more and more data with high accuracy and frequency is being generated which can be utilized for accurate implementation of PN.Data mining techniques are used to discover and validate new bioactive ingredients, integration of dietary and health data, development of predictive models and recommendations to optimize health outcomes. Meaningful and hidden statistical associations can be found between specific factors and health-related outcomes.

Benefits of PN are personalized diet recommendations, accurate measurement of ones diet for a healthy lifestyle and user-friendly ways to collect important information about a person. But as good as it sounds there are risks involved with PN as well. Trust and privacy of data is a huge risk when it comes to applications which invasively monitor important health related measures. People need to invest extra time in order to update their day to day eating habits and activities. Limitation in predictive algorithms.

Digital twins approach can be used as future improvements to simulate individual dietary effects and generating recommendations for health optimization. This is an approach which tries to create a digital equivalent representation of an individuals health statistics which can undergo infinite combinations of nutritional combinations which can be used to provide very personalized recommendations.

n-of-1 PN is another such improving method which is a data-driven approach of assessing health that tailors dietary recommendations to individual needs. Aggregated coordinated n-of-1 studies cluster individuals with similar characteristics or responses, and can help present excellent opportunities to advance this field.