2. Answer the following:

a) What are the main differences between projects for software development and Data mining?

Sol^n

- Data scientists write code as a mean to an end, whereas software developers write code to build things.
- Data science is an analytical activity, whereas software development in ~~data~~ has much more in common with traditional engineering.
- Data scientists tackle problems such as identifying fraudulent transactions, predicting ~~c~~hurn etc. Software developers can take the data scientists model and turn them into fully functioning systems with production-quality code.
- Data mining deals with processing of data, ML model, evaluation of models etc. Software development deals with computer programming, web design, DevOps etc.
- Data science methodology is Extract, Transform, Load (ETL). Software development methodology is SDLC (software development life cycle).
- Data mining required skillsets are statistics, data visualization, ML, analytical thinking. Software development required skills are coding using various languages, object oriented design, problem solving.

b) Compare relative strengths of statistical outlier detection and distance based outlier detection.

Sol. Statistical outlier detection : It assumes that the objects in a dataset are generated by a stochastic process (a generative model). The effectiveness of statistical methods highly depends on whether the assumptions made for the statistical model hold true for the given data.

It identifies outliers with respect to the model using a discordancy test.

There are two basic types of procedures for detecting outliers:

- Blocked procedures: In this case, either the entire suspect objects are treated as outliers or all of them are accepted as consistent.
- Consecutive (sequential) procedures: The idea is that the object that is least "likely" to be an outlier is tested first. If it is found to be an outlier, then all of the more extreme values are also considered outliers; otherwise, the next most extreme object is tested and so on.

Distance-based outlier detection: An object o is an outlier if its neighborhood does not have enough other points. An object o is an outlier if most (taking π as fraction threshold) of the objects in D are far away o, i.e., not in the r-neighborhood of o.

An object o is a $DB(r, \pi)$ outlier if

$$\frac{\|\{o' \mid dist(o, o') \leq r\}\|}{\|D\|} \leq \pi$$

c) You are evaluating similarity of two organisms in terms of the no. of genes they share. Each organism is represented as a bit vector indicating presence (bit = 1) of a particular age. The bit vectors of two species s1, s2 are given by the bit vectors (0101010001) and (0100011000) respectively. Compute Hamming distance (no of different bits) and Jaccard similarity coefficient. Which one seems better?

Sol^n.

s1 :	0	1	0	1	0	1	0	0	0	1
s2 :	0	1	0	0	0	1	1	0	0	0

Hamming distance = $1+1+1 = 3$

$$\text{Jaccard similarity coefficient} = \frac{f_{11}}{f_{10}+f_{01}+f_{11}}$$

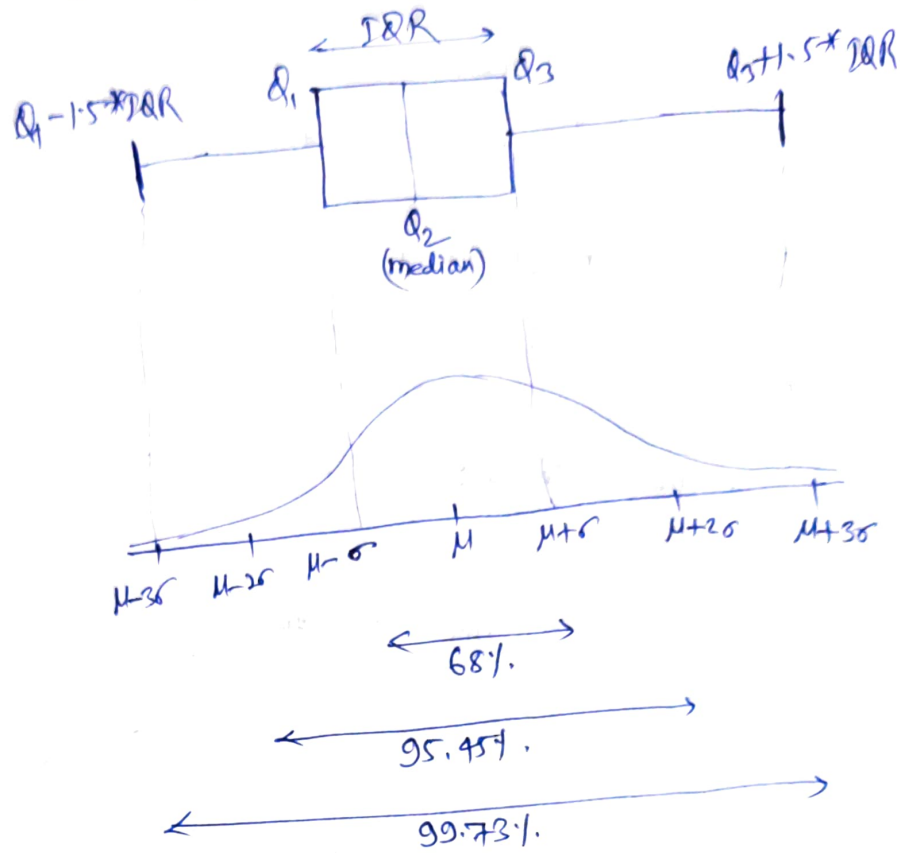
$$= \frac{2}{2+1+2} = \frac{2}{5} = 0.4$$

$$\text{Jaccard distance} = 1 - 0.4 = 0.6$$

Jaccard similarity coefficient seems better as it's normalized between 0 & 1 and easy to interpret.

4) A data scientist has one-dimensional data and she is checking if there are outliers in the data. Compare pros and cons of using statistical approach of ±3 sigma dispersion and boxplot approach with 1.5 * IQR.

Soln:



- Standard deviation is a little bit complex calculation than IQR but a more powerful measurement to analyse the dataset. So, measuring SD statistical parameter helps us to more easily summarize the data.
- Box plot is visual representation of statistical measures. It provides five points summary (Min, Max, Median, 25th percentile and 75th percentile).
- SD approach is not good for skewed data whereas Box plot doesnot get affected by skewed data.

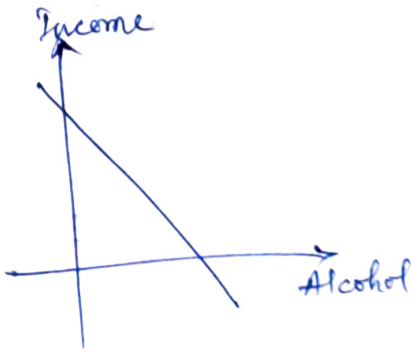
3. The data concerning alcohol consumption in each province of Canada, and income per household for each province are given below. From the given data find out the impact of income on alcohol consumption using linear regression method.

Province	Income	Alcohol consumption
Newfoundland	26.8	8.7
Prince Edward Island	27.1	8.4
Nova Scotia	29.5	8.8
New Brunswick	28.4	7.6
Quebec	30.8	8.9
Ontario	36.4	10.0
Manitoba	30.4	9.7
Saskatchewan	29.8	8.9
Alberta	35.1	11.1
British columbia	32.5	10.9

Sol^n

y = income, x = alcohol

$y = mx + c$ → linear regression



$$m = \frac{\sum x_i y_i - \left[\sum x_i \sum y_i / N\right]}{\sum x_i^2 - (\sum x_i)^2 / N}$$

$$c = \frac{\sum y_i - m \sum x_i}{N}$$

Income(y)	alcohol(x)	x^2	xy
26.8	8.7	75.69	233.16
27.1	8.4	70.56	227.64
29.5	8.8	77.44	259.6
28.4	7.6	57.76	215.84
30.8	8.9	79.21	274.12
36.4	10.0	100	364
30.4	9.7	94.09	294.88
29.8	8.9	79.21	265.22
35.1	11.1	123.21	289.61
32.5	10.9	118.81	354.25
$\Sigma y = 306.8$	$\Sigma x = 93$	$\Sigma x^2 = 875.98$	$\Sigma xy = 2778.32$

$$m = \frac{\sum x_i y_i - \left[\sum x_i \sum y_i / N\right]}{\sum x_i^2 - (\sum x)^2 / N}$$

$$= \frac{2778.32 - \frac{93 \times 306.8}{10}}{875.98 - (93)^2/10} = -\frac{74.92}{11.08} = -6.76$$

$$c = \frac{\sum y_i - m\sum x_i}{N} = \frac{306.8 - (-6.76) \times 93}{10} = -32.19$$

$\therefore y = mx + c \quad = -6.76x + (-32.19)$

1. A vegetable vendor sells three items: potatoes, onions, and tomatoes. Frequent pattern mining was done on her daily sales transaction database. Minimum support is 3. It is found that all frequent itemsets are also closed frequent itemsets. The transaction database has 3 transactions containing all 3 items (there are other transactions with fewer items). Describe the transaction database. Justify your answer.

Sol^n Given,

Transaction ID	List of items
1	P, O, T
2	P, O, T
3	P, O, T
4	P, T
5	P, O
6	O, T
⋮	⋮

min support = 3

P → Potato

O → Onion

T → Tomato

Also, all frequent itemsets given are closed frequent itemsets

1. Itemset

Item	count
P	5 or more
O	5 or more
T	5 or more

All items are frequent as support count ≥ min-support count

2-itemset

Item	count
PO	4 or more
PT	4 or more
OT	4 or more

All 2 itemsets are frequent & closed

P(≥5) → PO (4 or more), PT (4 or more)

O(≥5) → PO (4 or more), OT (4 or more)

T(≥5) → PT (4 or more), OT (4 or more)

3 itemset

Item	count
POT	3

All 3 itemsets are frequent & closed

PO (4 or more) → POT (3)

PT (4 or more) → POT (3)

OT (4 or more) → POT (3)

Since 3-itemset is closed frequent itemset

So, subset count must be greater than its immediate superset.

i.e; PO, PT & OT support count ≥ 4

Similarly, since 2-itemsets are closed frequent itemset

So, P, O & T support count ≥ support count of PO, PT, OT

∴ P, O & T support count ≥ 5

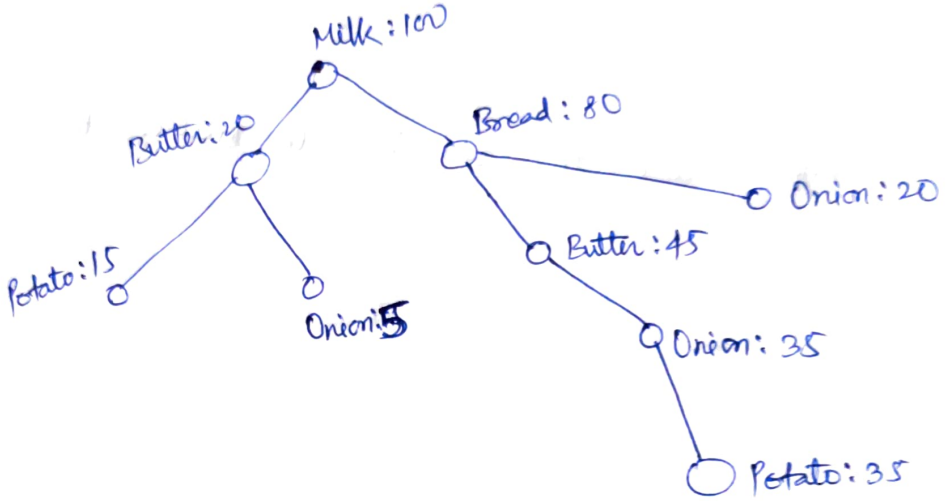
Also, POT is most frequent itemset.

5. The following FP-Tree is generated by a data scientist from daily transactions in a grocery store. The minsup = 50 and min conf = 80%. Using the tree perform the following:

a) Show the projected database for Onion

b) Find all the frequent k-itemsets for the largest k

c) Find strong association rules containing the k-items (for the largest k found in b)



Milk – M
Bread – Brd
Butter – But
Potato – P
Onion – O

Sol^n

Item	Conditional pattern base	Conditional fp-tree	frequent patterns
Potato	M Brd But O : 35 M But : 15	M But : 50	M But P : 50 (and all its subsets)
Onion	M Brd But : 35 M Brd : 20 M But : 5	M : 60	MO : 60
Butter	M Brd : 45 M : 20	M : 65	M But : 65
Bread	M : 80	M : 80	M Brd : 80
Milk	—	—	—

Item	frequency
Potato	15+35 = 50
Onion	35+20+15 = 70
Butter	45+20 = 65
Bread	80
Milk	100

strong association rule : {milk, butter, potato}

6. Cluster the following points {A[2,3], B[2,4], C[4,4], D[7,5], E[5,8], F[13,7]} using complete linkage hierarchical clustering algorithm. Assume Manhattan distance measure. Plot dendogram after performing all intermediate steps.

Soln:

	x	y
A	2	3
B	2	4
C	4	4
D	7	5
E	5	8
F	13	7

As per Manhattan distance,

d(distance) = $|x_1 - x_2| + |y_1 - y_2|$

$d(A,B) = |2-2| + |3-4| = 1$

$d(A,C) = |2-4| + |3-4| = 3$

$d(A,D) = |2-7| + |3-5| = 7$

$d(A,E) = |2-5| + |3-8| = 8$

$d(A,F) = |2-13| + |3-7| = 15$

$d(B,C) = |2-4| + |4-4| = 2$

$d(B,D) = |2-7| + |4-5| = 6$

$d(B,E) = |2-5| + |4-8| = 7$

$d(B,F) = |2-13| + |4-7| = 14$

$d(C,D) = |4-7| + |4-5| = 4$

$d(C,E) = |4-5| + |4-8| = 5$

$d(C,F) = |4-13| + |4-7| = 12$

$d(D,E) = |7-5| + |5-8| = 5$

$d(D,F) = |7-13| + |5-7| = 8$

$d(E,F) = |5-13| + |8-7| = 9$

	A	B	C	D	E	F
A	0					
B	1	0				
C	3	2	0			
D	7	6	4	0		
E	8	7	5	5	0	
F	15	14	12	8	9	0

(Distance matrix)

Here, min^m distance exists b/w A & B, so we will merge A & B in one cluster.

After that, we will create new distance matrix.

Iteration-1

	A, B	C	D	E	F
A, B	0				
C	(3)	0			
D	7	4	0		
E	8	5	5	0	
F	15	12	8	9	0

For complete linkage, we have to consider maxm value of distance

$\max\{d((A,C),d(B,C))\} = \max\{3,2\} = 3$

$\max\{d((A,D),d(B,D))\} = \max\{7,6\} = 7$

$\max\{d((A,E),d(B,E))\} = \max\{8,7\} = 8$

$\max\{d((A,F),d(B,F))\} = \max\{15,14\} = 15$

Iteration-2 Now, minm value in distance matrix is 3, so we will merge (A,B) with C.

	A,B,C	D	E	F
A,B,C	0			
D	7	0		
E	8	(5)	0	
F	15	8	9	0

$\max(d\{(A,B),D\},d(C,D))$
$= \max(7,4) = 7$

$\max(d\{(A,B),E\},d(C,E))$
$= \max(8,5) = 8$

$\max(d\{(A,B),F\},d(C,F))$
$= \max(15,12) = 15$

Iteration 3, Min^m distance in distance matrix is 5, so we will merge D & E.

$\max(d\{(A,B,C),D\}, d\{(A,B,C),E\})$

$= \max(7, 8) = 8$

$\max(d\{D,E\}, d\{E,F\})$

$= \max(8, 9) = 9$

	A,B,C	D,E	F
A,B,C	0		
D,E	8	0	
F	15	9	0

Iteration-4 Min^m value in distance matrix is 8, so we will merge A,B,C with D,E

$\max(d\{(A,B,C),F\}, d\{(D,E),F\})$

$= \max(15, 9)$

$= 15$

	A,B,C,D,E	F
A,B,C,D,E	0	
F	15	0

Nested clusters diagram

7. Classify the following points as core points, border points and noise points (DBSCAN). The minimum density requirement for a core point is 3 points (in addition to it) in its neighborhood radius of 1.5 units. Use Euclidean distance as the distance measure. The points are (1,2), (3,3), (4,3), (5,2), (3,4).

Sol:

Step-1: Distance matrix

$d(P_1,P_2) = \sqrt{(1-3)^2 + (2-3)^2}$

$= 2.24$

$d(P_1,P_3) = \sqrt{(1-4)^2 + (2-3)^2}$

$= 3.16$

$d(P_1,P_4) = \sqrt{(1-5)^2 + (2-2)^2} = 4$

$d(P_1,P_5) = \sqrt{(1-3)^2 + (2-4)^2} = 2.83$

$d(P_2,P_3) = \sqrt{(3-4)^2 + (3-3)^2} = 1$

$d(P_2,P_4) = \sqrt{(3-5)^2 + (3-2)^2} = 2.24$

$d(P_2,P_5) = \sqrt{(3-3)^2 + (3-4)^2} = 1$

$d(P_3,P_4) = \sqrt{(4-5)^2 + (3-2)^2} = 1.41$

$d(P_3,P_5) = \sqrt{(4-3)^2 + (3-4)^2} = 1.41$

$d(P_4,P_5) = \sqrt{(5-3)^2 + (2-4)^2} = 2.83$

	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	2.24	0			
P_3	3.16	1	0		
P_4	4	2.24	1.41	0	
P_5	2.83	1	1.41	2.83	0

Step-2: Find points within $\epsilon = 1.5$

P_1: None

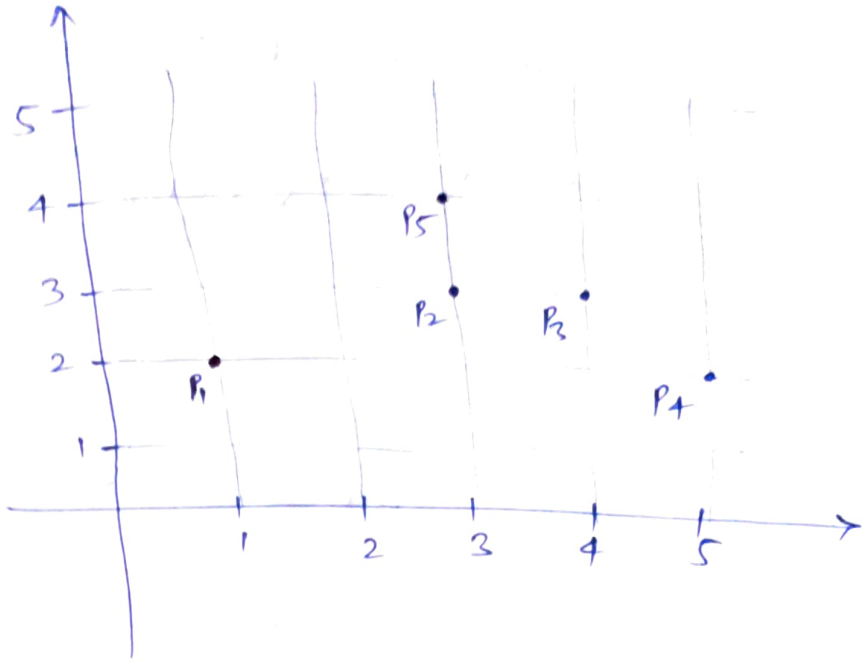
P_2: P_3, P_5

P_3: P_2, P_4, P_5

P_4: P_3

P_5: P_2, P_3

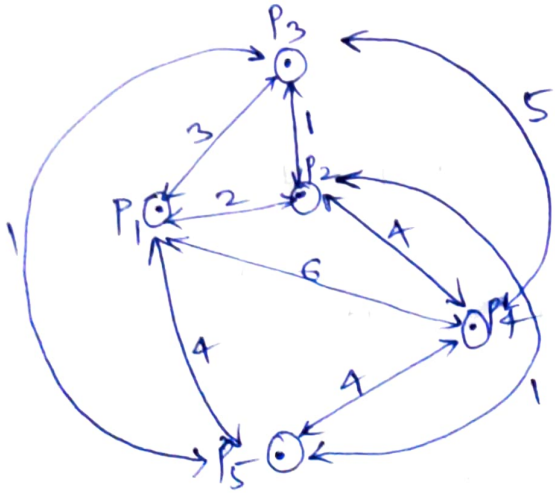
Point	Status
P_1	Noise
P_2	Core
P_3	Core
P_4	Border
P_5	Core



8. Given the following distance matrix, compute LOF (local outlier factor) of all points, based on 2-distance neighborhood consideration ($K=2$). Show intermediate steps.

	P_1	P_2	P_3	P_4	P_5
P_1	0	2	3	6	4
P_2	2	0	1	4	1
P_3	3	1	0	5	1
P_4	6	4	5	0	4
P_5	4	1	1	4	0

Sol^n:



✓ $dist_K(o)$ = distance b/w o and K^{th} NN

✓ $N_K(o) = \{o' \mid o' \text{ in } D, dist(o,o') \le dist_K(o)\}$

$dist_2(P_1) = 3$, $|N_2(P_1)| = 2$

$dist_2(P_2) = 1$, $|N_2(P_2)| = 2$

$dist_2(P_3) = 1$, $|N_2(P_3)| = 2$

$dist_2(P_4) = 4$, $|N_2(P_4)| = 2$

$dist_2(P_5) = 1$, $|N_2(P_5)| = 2$

$reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$

local reachability density, $lrd_k(o) = \dfrac{||N_k(o)||}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}$

$$lrd_2(P_1) = \frac{||N_2(P_1)||}{reachdist_2(P_3 \leftarrow P_1) + reachdist_2(P_2 \leftarrow P_1)}$$

$$= \frac{2}{2+3} = \frac{2}{5} = 0.4$$

$$lrd_2(P_2) = \frac{||N_2(P_2)||}{reachdist_2(P_1 \leftarrow P_2) + reachdist_2(P_5 \leftarrow P_2)}$$

$$= \frac{2}{\max\{1,1\} + \max\{1,1\}} = \frac{2}{1+1} = 1$$

$$lrd_2(P_3) = \frac{||N_2(P_3)||}{reachdist_2(P_2 \leftarrow P_3) + reachdist_2(P_5 \leftarrow P_3)}$$

$$= \frac{2}{\max\{1,1\} + \max\{1,1\}} = \frac{2}{1+1} = 1$$

$$lrd_2(P_4) = \frac{||N_2(P_4)||}{reachdist_2(P_2 \leftarrow P_4) + reachdist_2(P_5 \leftarrow P_4)}$$

$$= \frac{2}{\max\{1,4\} + \max\{1,4\}} = \frac{2}{4+4} = 2/8 = 1/4 = 0.25$$

$$lrd_2(P_5) = \frac{||N_2(P_5)||}{reachdist_2(P_2 \leftarrow P_5) + reachdist_2(P_3 \leftarrow P_5)}$$

$$= \frac{2}{\max\{1,1\} + \max\{1,1\}} = \frac{2}{1+1} = 1$$

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{||N_k(o)||}$$

$$LOF_2(P_1) = \frac{\frac{lrd_2(P_2) + lrd_2(P_3)}{lrd_2(P_1)}}{||N_2(P_1)||} = \frac{\frac{1+1}{0.4}}{2} = \frac{2}{0.8} = 2.5$$

$$LOF_2(P_2) = \frac{\frac{lrd_2(P_3) + lrd_2(P_5)}{lrd_2(P_2)}}{||N_2(P_2)||} = \frac{\frac{1+1}{1}}{2} = 1$$

$$LOF_2(P_3) = \frac{\frac{lrd_2(P_2) + lrd_2(P_5)}{lrd_2(P_3)}}{||N_2(P_3)||} = \frac{\frac{1+1}{1}}{2} = 1$$

$$LOF_2(P_4) = \frac{\frac{lrd_2(P_2) + lrd_2(P_5)}{lrd_2(P_4)}}{||N_2(P_4)||} = \frac{\frac{1+1}{0.25}}{2} = \frac{2}{0.5} = 4$$

$$LOF_2(P_5) = \frac{\frac{lrd_2(P_2) + lrd_2(P_3)}{lrd_2(P_5)}}{||N_2(P_5)||} = \frac{\frac{1+1}{1}}{2} = 1$$

9. Term frequency matrix for the five articles (A1 to A5) is shown below. Answer the following questions:

i) Find the cosine similarity between articles? Identify the two articles that are most similar.

ii) What are the TF-IDF values for (A4, War)?

Articles / Terms	Putin	Crude	NATO	War	Sanctions
A1	14	1	0	6	3
A2	0	21	5	0	0
A3	0	15	18	0	5
A4	5	2	0	12	0
A5	0	0	5	0	10

i) Sol^n Similarity b/w A & B : $Sim(A,B) = Cos(r_A, r_B) = \frac{r_A \cdot r_B}{\|r_A\| \|r_B\|}$

$\|A_1\| = \sqrt{14^2+1^2+0+6^2+3^2} = 15.56$

$\|A_2\| = \sqrt{0+21^2+5^2+0+0} = 21.59$

$\|A_3\| = \sqrt{0+15^2+18^2+0+5^2} = 23.96$

$\|A_4\| = \sqrt{5^2+2^2+0+12^2+0} = 13.15$

$\|A_5\| = \sqrt{0+0+5^2+0+10^2} = 11.18$

$$Sim(A_1, A_2) = \frac{14\times0 + 1\times21 + 0\times5 + 6\times0 + 3\times0}{\|A_1\| \|A_2\|}$$

$$= \frac{21}{15.56 \times 21.59} = 0.0625$$

$$Sim(A_1, A_3) = \frac{14\times0 + 1\times15 + 0\times18 + 6\times0 + 3\times5}{\|A_1\| \|A_3\|}$$

$$= \frac{15+15}{15.56 \times 23.96} = 0.080$$

$$Sim(A_1, A_4) = \frac{14\times5 + 1\times2 + 0 + 6\times12 + 0}{15.56 \times 13.15}$$

$$= \frac{70+2+72}{15.56 \times 13.15} = 0.704$$

$$Sim(A_1, A_5) = \frac{14\times0 + 1\times0 + 0\times5 + 6\times0 + 3\times10}{15.56 \times 11.18}$$

$$= \frac{30}{15.56 \times 11.18} = 0.172$$

$$Sim(A_2, A_3) = \frac{0 + 21\times15 + 5\times18 + 0 + 0}{21.59 \times 23.96} = 0.783$$ ✓

$$Sim(A_2, A_4) = \frac{0 + 21\times2 + 5\times0 + 0 + 0}{21.59 \times 13.15} = 0.148$$

$$Sim(A_2, A_5) = \frac{0 + 21\times0 + 5\times5 + 0 + 0\times10}{21.59 \times 11.18} = 0.104$$

$$Sim(A_3, A_4) = \frac{0\times5 + 15\times2 + 18\times0 + 0 + 5\times0}{23.96 \times 13.15} = 0.095$$

$$Sim(A_3, A_5) = \frac{0 + 15\times0 + 18\times5 + 0 + 5\times10}{23.96 \times 11.18} = 0.523$$

$$Sim(A_4, A_5) = \frac{5\times0 + 2\times0 + 0\times5 + 12\times0 + 0\times10}{13.15 \times 11.18} = 0$$

∴ Articles A_2 and A_3 are most similar.

ii) Ref.

Articles/ Terms	Putin	Crude	NATO	War	Sanctions
A_1	14	1	0	6	3
A_2	0	21	5	0	0
A_3	0	15	18	0	5
A_4	5	2	0	12	0
A_5	0	0	5	0	10

TF-IDF (A_4, War)

Using Cornell SMART TF-IDF model

$$\therefore TF(d,t) = \begin{cases} 0 & \text{if } freq(d,t)=0 \\ 1+\log(1+\log(freq(d,t))) & \text{otherwise} \end{cases}$$

$$\therefore TF(A_4, War) = 1+\log(1+\log(freq(A_4, War)))$$

$$= 1+\log(1+\log(12))$$

$$= 1+0.318 = 1.318$$

$$\therefore IDF(t) = \log\left(\frac{1+|d|}{|d_t|}\right)$$; $d \rightarrow$ document collection, $d_t \rightarrow$ set of documents containing term t

Here, $|d| = 5$, $|d_{War}| = 2$

$$\therefore IDF(war) = \log\left(\frac{1+|d|}{|d_{war}|}\right) = \log\left(\frac{1+5}{2}\right) = \log(3) = 0.477$$

$$\therefore TF\text{-}IDF(A_4, War) = TF(A_4, War) * IDF(War)$$

$$= 1.318 \times 0.477$$

$$= 0.629$$