Midterm 2019 Answers (Part A Only)

1. I. $P(x = \text{"so"} \mid y = \text{"Intent to Buy"}) = \dfrac{P(x=\text{"so"}, y=\text{"Intent to Buy"})}{P(y=\text{"Intent to Buy"})}$

$= \dfrac{1/6}{1/2} = \boxed{1/3}$

II. $P(x=\text{"see"}) = \dfrac{2}{6} = \boxed{1/3}$

III. $P(x_i=\text{"see"}, x_{i+1}=\text{"move"}) = \boxed{2/6}$ (2 comments where "see" and "move" appear in same sentence)

IV. $P(y=\text{"No Intent"} \mid x=\text{"bad"}) = \dfrac{P(y=\text{"No Intent"}, x=\text{"bad"})}{P(x=\text{"bad"})}$

$= \dfrac{1/6}{2/6} = \boxed{1/2}$

B. events A and B are independent if $P(A,B) = P(A)P(B)$

$P(x_i = \text{love}, x_j = \text{move}) = P(x=\text{love})\,P(x=\text{move})$

should equal this if independent → $4/6$ $= 1/6 \quad 4/6$

$P(x=\text{love}, x=\text{move}) = 1/6$ (only 1 review where "love" and "move" appear together).

C. priors:

$p(y=\text{No Intent}) = 3/6 = \boxed{1/2}$

$p(y=\text{Intent}) = \boxed{1/2}$

likelihood:

$p(x \mid y=\text{No Intent}) = p(x=\text{look} \mid y=\text{No}) \times p(x=\text{bad} \mid y=\text{No})$

$\boxed{1/9} = 1/3 \quad 1/3$

$$p(x \mid y=\text{Intent}) = p(x=\text{lookl} \mid y=\text{Intent}) \times p(x=\text{bad} \mid y=\text{Intent})$$

$$\boxed{1/9} = 1/3 \qquad 1/3$$

evidence:

$$p(x \mid y=\text{Intent})\,p(y=\text{Intent}) + p(x \mid y=\text{No})\,p(y=\text{No})$$

$$\boxed{1/9} = (1/9)\ (1/2) \qquad (1/9)\ (1/2)$$

posterior: $p(y=\text{Intent} \mid x) = \dfrac{p(x \mid y=\text{Intent})\,p(y=\text{Intent})}{p(x \mid y=\text{Intent})\,p(y=\text{intent}) + p(x \mid y=\text{No})\,p(y=\text{No})}$

$$\boxed{1/2} = \frac{(1/9)(1/2)}{(1/9)}$$

same for $P(y=\text{No Intent} \mid x)$

2. A.i. word count:

	old	blue	navy	sweater	trendy	wood	chair	dessert	jeans
A	0	0	0	0	1	1	1	0	0
B	2	1	1	0	0	0	0	1	1
C	1	1	0	1	1	0	0	1	0

B. one-hot:
same as A, except the 2 becomes a 1.

C.

	old	blue	navy	sweater	trendy	wood	chair	dessert	jeans
IDF	2	2	2.5	2.5	2	2.5	2.5	2	2.5
this is the same as A.i (word count) TF_A	0	0	0	0	1	1	1	0	0
TF_B	2	1	1	0	0	0	0	1	1
TF_C	1	1	0	1	1	0	0	1	0

$$IDF = 1 + \frac{N}{df(t)+1}$$

For words that appear 1 time:
(navy, sweater, wood, chair, jeans):

$$1 + \frac{3}{1+1} = 2.5$$

For words that appear 2 times:
(old, blue, discount, trendy)

$$1 + \frac{3}{3} = 2$$

	old	blue	navy	sweater	trendy	wood	chair	discount	jeans
$TFIDF_A$	0	0	0	0	2	2.5	2.5	0	0
$TFIDF_B$	4	2	2.5	0	0	0	0	2	2.5
$TFIDF_C$	2	2	0	2.5	2	0	0	2	0

B. (using word count)

	old	blue	navy	sweater	trendy	wood	chair	discount	jeans
wordcount$_q$	1	0	0	0	0	0	0	0	1
$TFIDF_q$	2	0	0	0	0	0	0	0	2.5

$$distance(Q, B) = \sqrt{(2-1)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 + (1-1)^2}$$

$$= \sqrt{1+1+1+1}$$

$$= \boxed{2}$$

TF_B	2	1	1	0	0	0	1	1	
TF_q	1	0	0	0	0	0	0	1	

query: OLD JEANS

distance(Q, C) =

$= \sqrt{(1-1)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2}$

$= \sqrt{1+1+1+1+1}$

$= \sqrt{5}$

$= \boxed{2.24}$

TF_C =	1	1	0	1	1	0	0	1	0
TF_Q =	1	0	0	0	0	0	0	1	1
diffs	0	1		1	1		1	1	

Since distance(Q, B) < distance(Q, C), recommend Product B.

(using TF-IDF)

C. query: OLD JEANS (changed from original query)

$$\cos(Q, A) = \frac{Q \cdot A}{\|Q\| \times \|A\|}$$

$Q \cdot A = 0$ dot product is 0, so similarity is also 0

$\cos_{sim}(A, Q) = 0$

$$\cos_{sim}(B, Q) = \frac{Q \cdot B}{\|Q\| \cdot \|B\|}$$

Norm: $\sqrt{x_1^2 + x_2^2 + \dots x_i^2}$ (definition)

$\|Q\| = \sqrt{2^2 + (2.5)^2}$

$= \sqrt{10.25} = 3.$

TFIDFB 4 2 2.5 0 0 0 0 2 2.5

from previous question 2

$$||B|| = \sqrt{4^2+2^2+2.5^2+2^2+2.5^2}$$

$$||B|| = \sqrt{16+4+6.25+4+6.25}$$

$$= \sqrt{36.5} = \sim 6$$

TFIDFQ 2 0 0 0 0 0 0 0 0.25

from above

$Q \cdot B = 4 \cdot 2 + 2.5 \cdot 2.5$ ← dot product of Q and B.

$= 8 + 6.25 = 14$

$$\cos_{sim}(Q, B) = \frac{14}{3 \times 6} = .875$$

$\cos_{sim}(Q, B) > \cos(Q, A)$

$\cos_{sim}(Q, B)$: 0.875 $\quad$ $\cos(Q, A)$: 0

Therefore, recommend product B.

d. Reasons why cosine similarity is used over Euclidean distance:

1. accounts for different document lengths
2. uses dot product, which handles high-dimensional data better.

3. a.

i. accuracy = $\boxed{11/15}$

ii. precision = $\boxed{2/4}$

iii. recall = $\boxed{2/4}$

iv. F1 score = $2 \cdot \frac{\text{Precision} \times \text{recall}}{\text{precision} + \text{recall}}$

$= 2 \cdot \frac{1/4}{1/2 + 1/2} = \boxed{1/2}$

b.

Actual \ Predicted	T	NT
T	2	2
NT	2	9

c. Recall: the model should be more accepting of false positives in order to catch more threats, since the cost of a false negative is extremely high.

d. Any model getting below 73.3% (11/15) accuracy is more or less useless; since you can achieve this accuracy by simply predicting "no threat" to every data point.