1. a.

i) Accuracy $= \frac{4+8}{4+2+1+8} = \frac{12}{15} = \frac{4}{5}$

ii) precision $= \frac{4}{4+1} = \frac{4}{5}$

iii) Recall $= \frac{4}{4+2} = \frac{4}{6} = \frac{2}{3}$

iv) F1 score $= 2 \cdot \frac{P \cdot R}{P+R} = 2 \cdot \frac{\frac{4}{5} \cdot \frac{2}{3}}{\frac{4}{5}+\frac{2}{3}} =$

b.

		predicted	
		W	M
Actual.	W	4	2
	M	1	8

c. p(gender_actual = "women") is a probability or proportion of actual women among all cases. On the other hand, p(gender_predicted = "women" | gender_actual = "women") is a conditional probability that only considers predicted women among actual women case.

d. Recall. Recall captures what proportion of actual positives was identified correctly.

2.

*
IB = Intent to Buy Tickets

NIB = Not intent to Buy Tickets

a. i) $P(X = \text{"so"} \mid y = IB)$

$= 1/3$

ii) $P(X = \text{"see"})$

$= P(X = \text{"see"} \mid y = IB) \cdot P(y = IB) +$
$\quad P(X = \text{"see"} \mid y = NIB) \cdot P(y = NIB)$

$= \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2} = \frac{2}{6} = \frac{1}{3}$

iii) $P(X_i = \text{"see"}, X_j = \text{"movie"}) = \frac{2}{6} = \frac{1}{3}$

iv) $P(y = NIB \mid X = \text{"bad"})$

$= \frac{P(X = \text{"bad"} \mid y = NIB) \cdot P(y = NIB)}{P(X = \text{"bad"})} =$

$P(X = \text{"bad"} \mid y = NIB) = \frac{1}{3}$

$P(X = \text{"bad"}) =$

$P(X = \text{"bad"} \mid y = IB) \cdot P(y = IB) +$
$P(X = \text{"bad"} \mid y = NIB) \cdot P(y = NIB)$

$= \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2} = \frac{2}{6} = \frac{1}{3}$

$p(y = NIB \mid x = \text{"bad"})$

$$= \frac{p(x=\text{"bad"} \mid y=NIB) \cdot p(y=NIB)}{p(x=\text{"bad"})} = \frac{\frac{1}{3} \cdot \frac{1}{2}}{\frac{1}{3}} = \frac{1}{2}$$

b. $p(x_i=\text{"love"}, x_j=\text{"movie"}) = \frac{1}{6}$

$p(x = \text{"love"}) = \frac{1}{6}$

$p(x = \text{"movie"}) = \frac{4}{6} = \frac{2}{3}$

$p(x = \text{"love"}) \cdot p(x = \text{"movie"}) =$

$= \frac{1}{6} \cdot \frac{2}{3}$

$\neq \frac{1}{6}$

$= p(x_i=\text{"love"}, x_j=\text{"movie"})$

$\therefore$ Not independent.

3.

a.

	Trendy	jeans	old	blue	red	wool
IDF	2	1.75	2	2	2.5	2.5
TF-A	1	1	0	0	0	0
TF-B	0	1	2	1	0	0
TF-C	1	1	1	1	1	1

b. Q : "old jeans"

TF-Q	0	1	1	0	0	0

Let $CS(A,B)$ be cosine similarity between A and B.

$CS(B,Q) \approx 0.867$

$CS(C,Q) \approx 0.506$.

Since $CS(B,Q)$ is larger, I will recommend product B.

c. product A document representation

$= [0.5 \quad 1.5 \quad -1 \quad 1 \quad 0.25]$

4. a.

	I	love	going	to	store	he	working	at	restaurant	is	closed	today	am	END
START	1/5				1/5	2/5						1/5		
I		1/2											1/2	
love			1/3				2/3							
going				2/2										
to					1/2				1/2					
store										1/2				1/2
he		1/2								1/2				
working								1/1						
at									1/1					
restaurant														2/2
is			1/2								1/2			
closed												1/1		
today	1/2													1/2
am							1/1							

b. I love working

$$= p(\text{"I"} \mid \text{START}) \cdot p(\text{"love"} \mid \text{"I"}) \cdot p(\text{"working"} \mid \text{"love"})$$

$$= \frac{1}{5} \cdot \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{15}$$

c. Since they are different in size, we need to use perplexity to make them comparable.