DS0560 HW4 answers

1. a. i. accuracy = $12/15$ = 80%

ii. precision = $4/5$ = 80%

iii. recall = $4/6$ = 67%

iv. F1 score = $2 \cdot \frac{P \times R}{P+R} = 2 \cdot \frac{.8 \times .67}{.8+.67}$ = 73%

b.

actual \ predicted	women	men	
women	4	2	$6/15$ = P(actual = women)
men	1	8	$9/15$ = P(actual = men)
	$5/15$	$10/15$	
	P(predicted = women)	P(predicted = men)	

c. P(predicted = women | actual = women) is the recall rate. Given an actual women's product, what is the probability it will be tagged as a women's product by the model? P(actual = women) is the marginal probability and represents the % positives in your dataset.

d. Recall since we want to capture all the actual women's products. However, a model could blindly tag every sample as positive and achieve a 100% recall rate, so you should note that some form of weighted average - like F1 score - is needed. However, F1 score in this instance would not weight recall heavily enough to be suitable.

2. A. i. $P(x=\text{so} \mid y=\text{Intent}) = \boxed{1/3}$

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

iii. $P(x_i=\text{see}, x_j=\text{move}) =$

$P(x=\text{see} \mid y=\text{intent}) \times P(x=\text{move} \mid y=\text{intent}) \times P(y=\text{intent}) + P(x=\text{see} \mid y=\text{no I}) \times P(x=\text{move} \mid y=\text{no intent}) \times P(y=\text{no I})$

$1/3 \cdot 1/1 \cdot 1/2 = 1/6 + \quad 1/3 \cdot 1/3 \cdot 1/2 = \boxed{2/9}$ ← this is if I ask for the evidence.

$\boxed{1/3}$ ← this is if I ask for prob of seeing both tokens in same document

iii. $P(y=\text{No intent} \mid x=\text{bad}) =$

$$\frac{P(x=\text{bad} \mid y=\text{No intent})\, P(y=\text{No intent})}{P(y=\text{No Intent})\, P(x=\text{bad} \mid y=\text{No intent}) + P(y=\text{intent})\, P(x=\text{bad} \mid y=\text{intent})}$$

$$\frac{1/3 \cdot 1/2}{1/2 \cdot 1/3 + 1/2 \cdot 1/3} = \boxed{1/2}$$

B. If events A and B are independent, then $P(A, B) = P(A) \times P(B)$.

$P(\text{love}, \text{movie}) = 1/6$

$P(\text{love}) = 1/6$

$P(\text{movie}) = 1/2$

$P(\text{love}) \times P(\text{movie}) \neq P(\text{love}, \text{movie})$

$1/6 \times 1/2 \neq 1/6,$

They are <u>not</u> independent.

3. a.

	trendy	jeans	old	blue	red	wool
TF(A)	1	1	0	0	0	0
TF(B)	0	1	2	1	0	0
TF(C)	1	1	1	1	1	1
IDF	2	1.75	2	2	2.5	2.5
TFIDF(A)	2	1.75	0	0	0	0
TFIDF(B)	0	1.75	4	2	0	0
TFIDF(C)	2	1.75	2	2	2.5	2.5

IDF for $df(t)=1 = 1+\frac{3}{1+1} = 2.5$

$df(t)=2 = 1+\frac{3}{2+1} = 2$

$df(t)=3 = 1+\frac{3}{3+1} = 1.75$

b.

	trendy	jeans	old	blue	red	wool
tf(query) =	0	1	1	0	0	0
tfidf(query) =	0	1.75	2	0	0	0

$\cos \text{similarity}(B, query) = \frac{B \cdot query}{\|B\| \times \|query\|} = \frac{1.75 \times 1.75 + 4 \times 2}{\|B\| \times \|query\|}$

$\|B\| = \sqrt{1.75^2 + 4^2 + 2^2} = 4.8$

$\|query\| = \sqrt{1.75^2 + 2^2} = 2.66$

$= \frac{11.06}{4.8 \cdot 2.66} = \boxed{0.867}$

$\cos \text{similarity}(C, query) = \frac{C \cdot query}{\|C\| \times \|query\|} = \frac{1.75 \times 1.75 + 2 \times 2}{5.25 \times 2.66} = \frac{7.06}{13.97} = \boxed{.505}$

$\|C\| = \sqrt{2^2 + 1.75^2 + 2^2 + 2^2 + 2.5^2 + 2.5^2} = 5.25$

Recommend product B, since $\cos \text{sim}(B, query) > \cos \text{sim}(C, query)$

c. Product A is **trendy jeans**:

trendy	-1	0	1.0	2	-2
jeans	2	3	-3	0	2.5
	1	3	-2	2	.5
divide by # words	0.5	1.5	-1	1	.25

→ trendy jeans vector

4. Docs after lemmatization / stopword removal:

1. I love go to store
2. He love work at restaurant
3. Store be close today
4. He be go to restaurant
5. Today I be work

* is/am → be (you can make these small assumptions, just note them!)

* going → go

	START	I	love	go	to	store	he	work	at	restaurant	be	close	today	END
START		0.2		0.2	0.4								0.2	
I			0.5										0.5	
love				0.5				0.5						
go					1									
to						0.5				0.5				
store											0.5			0.5
he								0.5				0.5		
work														
at									0.5					0.5
restaurant										1				
be														1
close														
today													1	
END														0.5

B. START → I ⇒ 0.2

I → LOVE ⇒ 0.5

LOVE → WORK ⇒ 0.5

WORK → END ⇒ 0.5

P("I love working") =

0.2 × 0.5 × 0.5 × 0.5 = 0.025

C. Longer sentences will inherently have smaller probabilities, so you should convert them to perplexity using $P = \frac{1}{\sqrt[n]{p(w)}}$, where n = # of tokens.