James Dolla

HW 4

1a. i. Accuray =>

$$\frac{12}{15} = \frac{4}{5} = 80\%$$

ii. Precision =>

$$\frac{4}{4+1} = 80\%$$

iii. Recall =>

$$\frac{4}{4+2} = \frac{2}{3} = 66.67\%$$

iv. F1 Score =>

$$= 2 * \left(\frac{.8 \times 2/3}{.8 + 2/3} \right)$$

$$= 72.72\%$$

1b.

		Predicted	
		0	1
Actual	0	8	1
	1	2	4

1c. P(gender_actual="women") is the probability that the actual gender is women.
P(gender_predicted ="women"| gender_actual="women") is the probability that the predicted gender is women given that the actual gender is women.

1d. Recall should be high, because we want to return as many potential women records at the expense of being wrong once in a while. F1 Score seems to balance recall and precision.

2ai. 1/3

ii. 2/6 = 1/3

iii. 1/3

iv. 1/2

2b. $P(\text{love, movie}) = 1/6$

if independent → $P(\text{love, movie}) = P(\text{love}) * P(\text{movie})$

$= 1/6 * 4/6 = \frac{4}{36} = \frac{1}{9}$

$\frac{1}{6} \neq 1/9$ ∴ not independent

3a.i.

	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
tfidf q →	0	1.75	2	0	0	0

3b.

$$\cos(b,q) = \frac{1.75^2 + 8}{\sqrt{1.75^2 + 4^2 + 2^2} * \sqrt{1.75^2 + 2^2}} = \frac{11.06}{4.8 * 2.66} = 0.87$$

$$\cos(c,e) = \frac{1.75^2+4}{\sqrt{2^2+1.75^2+2^2+2^2+2.5^2+2.5^2} * \sqrt{1.75^2+2^2}} = \frac{7.06}{5.25 * 2.66}$$

$$= .51$$

Since $\cos(b,e) > \cos(c,e) \rightarrow e$ is more similar to b

3c.

trendy	-1	0	1	2	-2
jeans	2	3	-3	0	2.5
Avg =>	0.5	1.5	-1	1	.25

4a.

not normalized

	I	love	go	to	store	he	work	at	rest.	is	close	today	am	end
Start	1				1	2						1		
I		1											1	
love			1				1							
go				2										
to					1				1					
store										1				1
he		1								1				
work								1						1
at									1					
rest.														2
is			1											
close											1			
today	1										1			
am							1							1

4b). $0.2 \times 0.5 \times 0.5 \times 0.5 = \frac{1}{2} \times \frac{1}{2} \Rightarrow \frac{1}{4} = 0.025$

4c. You would need remove the stopword "the" Since that's what we did when we constructed the confusion matrix. Then we need to use perplexity because it accounts for the different sizes of test corpuses.