

# DS0560 HW4 answers

I. a.

- accuracy =  $\frac{12}{15} = \boxed{80\%}$
- precision =  $\frac{4}{5} = \boxed{80\%}$
- recall =  $\frac{6}{6} = \boxed{100\%}$

IV.  $F1_{Soc} = 2 \cdot \frac{P \times R}{P+R} = 2 \cdot \frac{.8 \times .67}{.8 + .67} = \boxed{73\%}$

b.

		predicted	
		women	men
actual	women	4	2
	men	1	8
		$\frac{5}{15}$	$\frac{10}{15}$
		$P(\text{predicted}=\text{woman})$	$P(\text{predicted}=\text{men})$

c.  $P(\text{predicted}=\text{woman} | \text{actual}=\text{woman})$  is the recall rate. Given an actual women's product, what is the probability it will be tagged as a woman's product by the model?  $P(\text{actual}=\text{woman})$  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  $F1_{Soc}$  - is needed - however,  $F1_{Soc}$  in this instance would not weight recall heavily enough to be suitable.

$$2. A. i. P(x=so | y=Intent) = \boxed{\frac{1}{3}}$$

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

$$iii. P(x_i=see, x_j=movie) =$$

$$P(x=see | y=intent) \times P(x=movie | y=intent) \times P(y=intent) + P(x=see | y=no intent) \times P(x=movie | y=no intent) \times P(y=no intent)$$

$$\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6} \quad \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} = \boxed{\frac{2}{9}} \leftarrow \text{this is if I ask for } \overset{\text{the}}{\text{endings}}.$$

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

$$iii. P(y=No intent | x=bad) =$$

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

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

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

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

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

$$P(\text{movie}) = \frac{4}{6}$$

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

$$\frac{1}{6} \times \frac{4}{6} \neq \frac{1}{6},$$

They are not independent.

	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

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

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

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

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

b. trendy jeans old blue red wool

tf(query) = 0 1 1 0 0 0

tf\_idf(query) = 0 1.75 2 0 0 0

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

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

$$= \frac{11.06}{9.8 \times 2.66} = 0.867$$

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

$$\cos \text{ similarity } (C, \text{query}) = \frac{C \cdot \text{query}}{\|C\| \times \|\text{query}\|} = \frac{1.75 \times 1.75 + 2 \times 2}{\|C\| \times \|\text{query}\|} = \frac{7.06}{13.97} = 0.505$$

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

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

c. Product A is trendy jeans:

trendy	-1	0	1.0	2	-2
jeans	2	3	-3	0	2.5

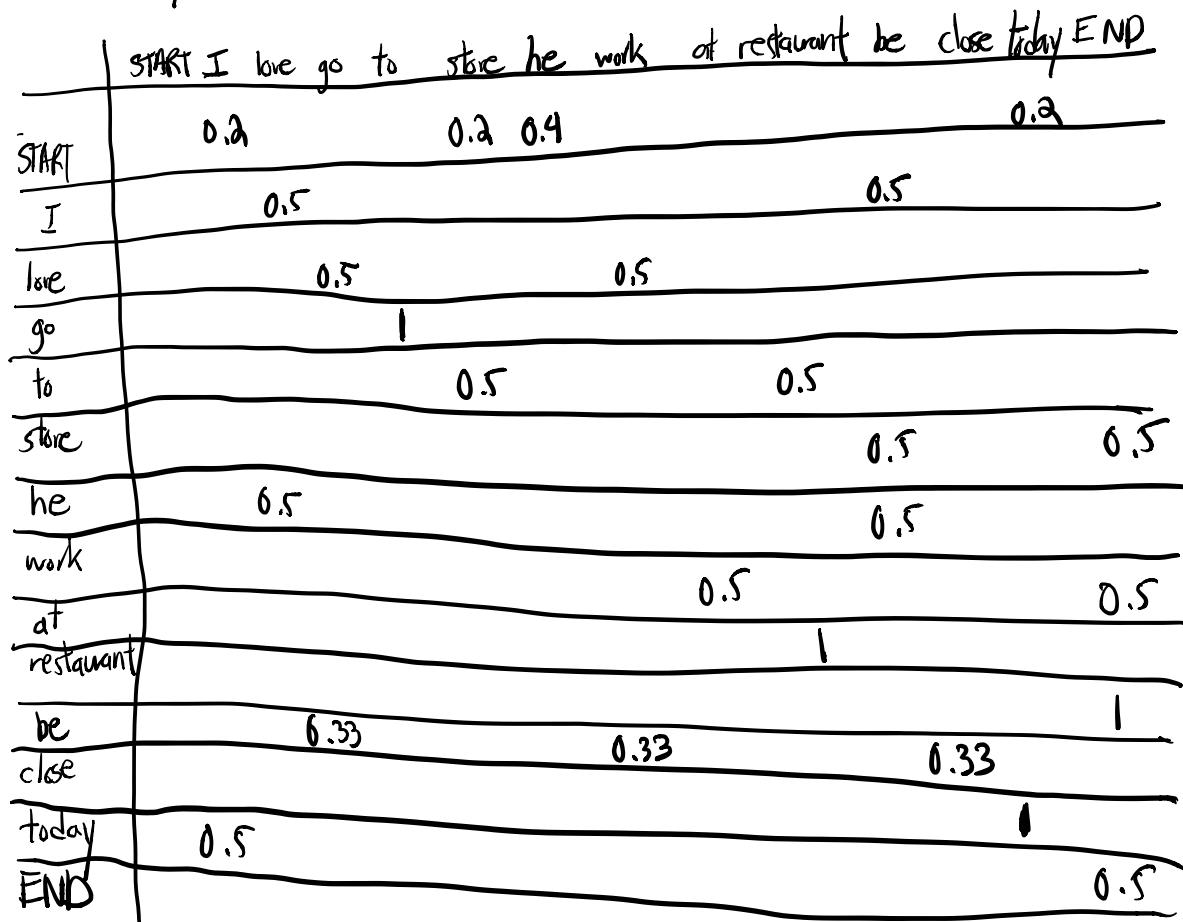
divide by  $1 \ 3 \ -2 \ 2 \ .5$

# words 0.5 1.5 -1 1 .25

trendy jeans vector

4. Docs after lemmatization/stemmed removal:

1. I love go to store
2. He love work at restaurant
3. store be close today \* is form → be (you can make these small assumptions, just note them!)
4. He be go to restuant \* going → go
5. Today I be work



B.

$$\text{START} \rightarrow I \Rightarrow 0.2 \quad P(\text{"I love working"}) =$$

$$I \rightarrow \text{LOVE} \Rightarrow 0.5 \quad 0.2 \times 0.5 \times 0.5 \times 0.5 = \underline{\underline{0.025}}$$

$$\text{LOVE} \rightarrow \text{WORK} \Rightarrow 0.5$$

$$\text{WORK} \rightarrow \text{END} \Rightarrow 0.5$$

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

