

1

a. i) Accuracy $A = 8/10$
 $B = 8/10$] Tie!

ii) Precision $A = 3/4 \rightarrow A$ is higher.
 $B = 4/6 = \frac{2}{3}$

iii) F1 score $A = 2 \cdot \frac{\frac{3}{4} \cdot \frac{3}{4}}{\frac{3}{4} + \frac{3}{4}} = \frac{2 \cdot \frac{9}{16}}{\frac{3}{2}} = \frac{18}{32} = \frac{9}{16}$
 $B = 2 \cdot \frac{\frac{2}{3} \cdot 1}{\frac{2}{3} + 1} = \frac{2 \cdot \frac{2}{3}}{\frac{5}{3}} = \frac{4}{5}$

b. $P(\text{Model A prediction} = \text{Model B prediction} \mid \text{actual rating} = \text{"negative"})$ B higher.

$$= 5/6$$

c.

A

B.

	P	N
P	3	1
Actual		
N	1	5

predicted

	P	N
P	4	0
Actual		
N	2	4

predicted

d. Recall since the cost of FN is a lot higher than FP, it should focus on FN in the expense of FP. Model B should be chosen as it has recall of 1.

2.

- a. • Structured Query Language → SQL
- Amazon Web Services → AWS
Amazon Web Services (AWS)
- MS Excel / Microsoft Excel → Excel

Why? → They are necessary to reduce the # of distinct words.

* SH = Strong Hire
PH = poor Hire

b.

$$\text{i)} P(Y = \text{SH}) = 4/10 = 2/5$$

$$\text{ii)} P(X = \text{"SQL"} | Y = \text{SH}) = 3/4$$

$$P(X = \text{"AWS"} | Y = \text{SH}) = 2/4$$

$$P(X = \text{"Python"} | Y = \text{SH}) = 3/4$$

$$\therefore P(X | Y = \text{SH}) = 3/4 \cdot 2/4 \cdot 3/4 = 18/64 = 9/32$$

1 2

64
3

$$\text{iii)} P(X = \text{"SQL", "AWS", "Python"})$$

$$= P(X = \text{"SQL", "AWS", "Python"} | Y = \text{SH}) \cdot P(Y = \text{SH})$$

$$+ P(X = \text{"SQL", "AWS", "Python"} | Y = \text{PH}) \cdot P(Y = \text{PH})$$

$$= 9/32 \cdot 2/5 + \cancel{1/32} \cdot 3/5$$

192

$$= \frac{18+3}{32+5} = \frac{21}{37} / 160. \approx 0.1181$$

(iv) $P(Y=SH | X = "SQL", "AWS", "Python")$

$$= \frac{P(X | Y=SH) \cdot P(Y=SH)}{P(X = "SQL", "AWS", "Python")}$$

$$= \frac{\frac{9}{32} \cdot \frac{2}{5}}{\cancel{21}/160} = \frac{\cancel{18}/21}{\cancel{6}/7} = 0.9526$$

C. $P(X_i = "Python" | X_j = "SQL") = \frac{3}{4}$.

~~Incorrect. Among 4 resumes that listed SQL, only two also have Python. Thus, not likely, but only the half of the time.~~

$$P(\text{Python}) = \frac{5}{10}$$

3.

a.

	nugget	shake	burger	fries	hamburger
IDF	2	2	2.5	2.5	2.5
TF-A	1	1	0	0	0
TF-B	0	0	1	1	0
TF-C	1	1	1	0	1
TF-Q	0	1	1	0	0

b. Let $ED(A, B)$ the Euclidean distance between A and B.

$$ED(Q, B) \approx 3.2$$

$$ED(Q, C) \approx 3.2$$

they are indifferent.

c.

d.

they class are \rightarrow

4. a.

class are today I am late he went they were home after END

START	$\frac{1}{5}$	$\frac{0}{1}$	$\frac{0}{1}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	
class		$\frac{0}{1}$	$\frac{1}{4}$				$\frac{3}{4}$
are			$\frac{1}{1}$				
today				$\frac{1}{2}$	$\frac{1}{2}$		$\frac{2}{2}$
I						$\frac{1}{1}$	
am				$\frac{1}{1}$			
late	$\frac{1}{1}$		$\frac{1}{2}$				
he					$\frac{1}{1}$		
went		$\frac{1}{1}$				$\frac{1}{1}$	
They						$\frac{1}{1}$	
were							$\frac{1}{1}$
home							
after	$\frac{1}{1}$						

Y.01 where it is blank.

-01

b. He went home

$$\begin{aligned} &= p(\text{"He"} \mid \text{START}) \cdot p(\text{"Went"} \mid \text{"He"}) \cdot p(\text{"Home"} \mid \text{"Went"}) \\ &\quad \cdot p(\text{END} \mid \text{"home"}) \\ &= 0.2 \times 1 \times 0.5 \times 0.01 = 0.001 \end{aligned}$$

c. They went to class.

$$\begin{aligned} &= p(\text{"They"}) \cdot p(\text{"went"}) \cdot p(\text{"class"}) \\ &= \frac{1}{19} \cdot \frac{2}{19} \cdot \frac{4}{19} \quad \text{Unique words of OUP.} \\ &\quad \text{Word Count!!!} \end{aligned}$$

5. a. False. For example, if the word "thorough" is given, $\setminus b\text{Thor}\setminus b$ will not search this whereas Thor will. It's a FP case. Thus $\setminus b\text{Thor}\setminus b$ will have lower FPR.

B,

c. False.

word2vec takes order into account. So they are different.

D. ~~True~~: False

Their cosine similarity ~~= 1.~~
Distance = 0

E. False. ? indicates it is not a capture group.

F. True. Since ASCII part is most common, ASCII part kept the same in UTF-8.
"a" is in ASCII.

G. True.

$$1 = \frac{2 \cdot P \cdot R}{P+R} \quad \text{where } P+R \neq 0,$$

is only possible when Precision and Recall
are 1 where there are no FP or FN.

H. False.

