1. Classification + Model Evaluation

	Actual	Model A		Model B		A and B agree
1.	N	P	X (FP)	P	X (FP)	T
2.	N	N	✓	P	X (FP)	F
3.	N	N	✓	N	✓	T
4.	N	N	✓	N	✓	T
5.	P	N	X (FN)	P	✓	F
6.	N	N	✓	N	✓	T
7.	N	N	✓	N	✓	T
8.	P	P	✓	P	✓	T
9.	P	P	✓	P	✓	T
10.	P	P	✓	P	✓	T

5/6 for 1b

a. i. accuracy: Model A: 8/10
Model B: 8/10
$\boxed{\text{neither}}$

ii. precision:

$$\frac{TP}{TP+FP} = \frac{3}{3+1} = \frac{3}{4} \text{ (Model A)} \leftarrow$$

$$= \frac{4}{4+2} = \frac{2}{3} \text{ (Model B)}$$

iii. recall:

$$\frac{TP}{TP+FN} = \frac{3}{3+1} = \frac{3}{4} \text{ (Model A)}$$

$$= \frac{4}{4+0} = 1 \text{ (Model B)} \leftarrow$$

F1 Score:

$$2\times\frac{P\times R}{P+R} = \frac{2\times .75\times .75}{.75+.75} = .75 \text{ (Model A)}$$

$$= \frac{2\times .67\times 1}{.67+1} = .80 \text{ (Model B)} \leftarrow$$

b. P(Model A pred = Model B pred | actual rating = negative) = 5/6 (see above chart)

c.

		predicted P	predicted N
actual	P	3	1
	N	1	5

d. Recall, trying to minimize # of FN (false negatives). Model B.
In practice, use a weighted average metric.

2 a. remove punctuation, group related terms → MS Excel becomes Excel. They are necessary to reduce the # of distinct vocab words and # of features in our model. Remove stopwords and perform stemming/lemmatization.

b. i. $P(y=\text{strong}) = \dfrac{c(\text{strong})}{c(\text{strong}) + c(\text{not strong})} = \dfrac{4}{4+6} = \dfrac{4}{10}$

ii. $P(x=\text{"SQL, AWS, Python"} \mid y=\text{strong}) = P(x=\text{SQL} \mid y=\text{strong}) \times P(x=\text{AWS} \mid y=\text{strong}) \times P(x=\text{Python} \mid y=\text{strong})$

$$= \frac{3}{4} \times \frac{2}{4} \times \frac{3}{4}$$

$$= \frac{18}{64} = \boxed{.28125}$$

iii. evidence:

$$\overbrace{P(x=\text{"SQL, AWS, Python"} \mid y=\text{strong})}^{0.28125} \times \overbrace{P(y=\text{strong})}^{0.4} +$$

$$\underbrace{P(x=\text{"SQL, AWS, Python"} \mid y=\text{weak})}_{\text{need to calculate}} \times \underbrace{P(y=\text{weak})}_{0.6}$$

$$\underbrace{P(x=\text{SQL} \mid y=\text{weak})}_{1/6} \times \underbrace{P(x=\text{AWS} \mid y=\text{weak})}_{1/6} \times \underbrace{P(x=\text{Python} \mid y=\text{weak})}_{1/3} = 1/108 = .0926$$

$$0.28125 \times 0.4 + 0.0926 \times 0.6 = \boxed{0.1181}$$

iv.

$$\overbrace{P(y=\text{strong} \mid X=\text{SQL, AWS, Python})}^{\text{posterior}} = \frac{\overbrace{P(x=\text{SQL, AWS, Python} \mid y=\text{strong})}^{\text{likelihood}} \overbrace{P(y=\text{strong})}^{\text{prior}}}{\underbrace{\sum_i P(x=\text{SQL, AWS, Python} \mid y=i) P(y=i)}_{\text{evidence}}}$$

$$\frac{0.28125 \times 0.4}{0.1181} = 0.9526 = \boxed{95.26\%}$$

c. Your colleague is asking if $P(Python|SQL) > P(Python)$.

$P(Python|SQL) = \frac{C(Python, SQL)}{C(SQL)} = \frac{3}{4}$

$P(Python) = 5/10$

$3/4 > 5/10$, so she is right.

If $P(Python|SQL) = P(Python)$, then the Python and SQL events would be independent.

3. Customer A: nugget shake

Customer B: burger fries

customer C: hamburger nugget shake burger

(it's your choice if you want to turn hamburger → burger) results will look different

a.

	nugget	shake	burger	hamburger	fries
TF(A)	1	1	0	0	0
TF(B)	0	0	1	0	1
TF(C)	1	1	1	1	0
IDF	2	2	2	2.5	2.5

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

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

TFIDF(A)	2	2	0	0	0
TFIDF(B)	0	0	2	0	2.5
TFIDF(C)	2	2	2	2.5	0
TF(Q)	0	1	1	0	0
TFIDF(Q)	0	2	2	0	0

Euclidean distance $(B, Q) =$

TFIDF(Q)	0	2	2	0	0
TFIDF(B)	0	0	2	0	2.5

$= \sqrt{(0-0)^2 (2-0)^2 (2-2)^2 (0-0)^2 (0-2.5)^2}$

$= \sqrt{4+6.25} = \boxed{3.20}$

Euclidean distance C(C,Q)=

TFIDF(Q)	0	2	2	0	0
TFIDF(C)	2	2	2	2.5	0

$$= \sqrt{(0-2)^2+(2-2)^2+(2-2)^2+(0-2.5)^2(0-0)^2}$$

$$= \sqrt{2^2+2.5^2} = \boxed{3.20}$$

The qrey is equally similar to B and C.

	nugget	shake	buger	fries
TF(A)	1	1	0	0
TF(B)	0	0	1	1
TF(C)	1	1	2	0
IDF	2	2	2	2.5

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

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

TFIDF(A)	2	2	0	0
TFIDF(B)	0	0	2	2.5
TFIDF(C)	2	2	4	0.
TF(Q)	0	1	1	0
TFIDF(Q)	0	2	2	0

Euclidean distance C(B,Q):

TFIDF(Q)	0	2	2	0
TFIDF(B)	0	0	2	2.5

$$= \sqrt{(0-0)^2+(2-0)^2+(2-2)^2+(0-2.5)^2}$$

$$= \sqrt{4+6.25} = 3.20$$

Euclidean distance C(C,Q):

TFIDF(Q)	0	2	2	0
TFIDF(C)	2	2	4	0

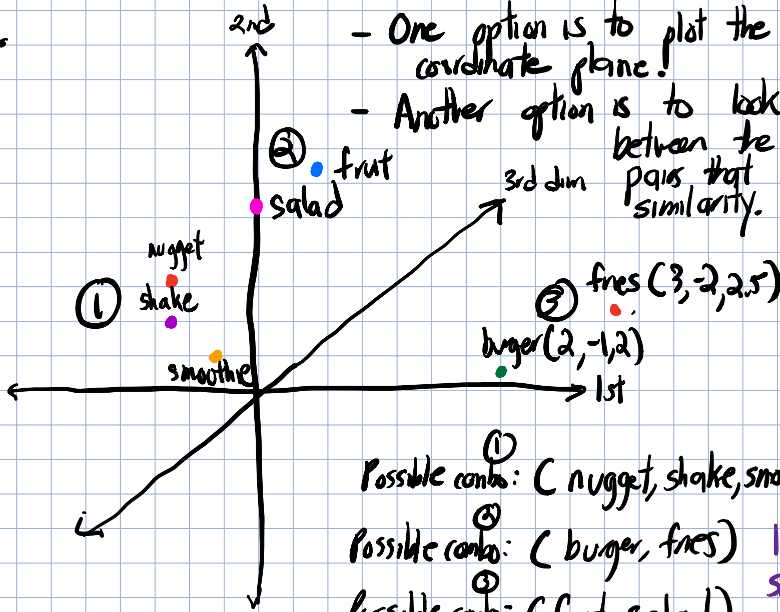
$$= \sqrt{(0-2)^2(2-2)^2(2-4)^2(0-0)^2}$$

$$= \sqrt{2^2+2^2} = 2.82$$

If you grap hamburger → burger, then $\boxed{\text{Customer C}}$ is a better match.

c.

- One option is to plot the points on a coordinate plane!
- Another option is to look at the distances between the vector and select pairs that have the highest similarity.



① Possible combo: (nugget, shake, smoothie)

② Possible combo: (burger, fries)

③ Possible combo: (fruit, salad)

Note: I graded this question very loosely. As long as you made some attempt to interpret the vectors, you received credit.

d.

nugget	-2	0	1
smoothie	-1	0	0.5
shake	-2	0.5	0

$\cos_{sim}(\text{nugget}, \text{smoothie})$:

$$= \frac{-2 \times -1 + 0 \times 0 + 1 \times 0.5}{\|\text{nugget}\| \times \|\text{smoothie}\|}$$

$$= \frac{2.5}{\sqrt{5} \cdot \sqrt{1.25}} = 1$$

$$\|\text{nugget}\| = \sqrt{(-2)^2 + (0)^2 + (1)^2} = \sqrt{5}$$

$$\|\text{smoothie}\| = \sqrt{(-1)^2 + (0)^2 + (0.5)^2} = \sqrt{1.25}$$

$$\|\text{shake}\| = \sqrt{(-2)^2 + (0.5)^2 + (0)^2}\ \sqrt{4.25}$$

$\cos_{sim}(\text{nugget}, \text{shake}) =$

$$= \frac{-2 \times -2 + 0 \times 0.5 + 1 \times 0}{\|\text{nugget}\| \times \|\text{shake}\|}$$

$$= \frac{4}{\sqrt{5} \times \sqrt{4.25}} = \frac{4}{4.61} = 0.87$$

$\cos_{sim}(\text{nugget}, \text{smoothie}) > \cos_{sim}(\text{nugget}, \text{shake})$, so nugget is closer to smoothie.

Note: word2vec similarities can actually range from -1 to 1.

Also, we typically normalize word2vec so that each vector is of unit length.

4.

A.

	class	be/is	today	I	late	go	he	they	home	after	END
START	1 0.2			2 0.4			1 0.2	1 0.2			
class		1 0.25									3 0.75
be/is			1 0.5		1 0.5						
today											2 1.0
I		1 0.5				1 0.5					
late	1 0.5		1 0.5								
go	1 0.5								1 0.5		
he						1 1.0					
they		1 1.0									
home										1 1.0	
after	1 1.0										
END											

all other transitions are 0.01 (Note, because the question was ambiguous, if you did not add in 0.01, I did not penalize)

B. P("HE WENT HOME") =

$= P(\text{"HE"} \mid \text{"START"}) \times P(\text{"WENT"} \mid \text{"HE"}) \times P(\text{"HOME"} \mid \text{"WENT"}) \times P(\text{END} \mid \text{"WENT"})$

$= 0.2 \times 1.0 \times 0.5 \times 0.01$

$= .001$

C. Unigram model is simply word count frequency $\rightarrow$ $P(\text{word}) = \dfrac{C(\text{word})}{\sum_i^V C(w_i)}$

THEY	WENT	TO	CLASS
$1/19$	$2/19$	(stopword)	$4/19 = \boxed{\dfrac{8}{6859}}$

V = set of distinct words