Q1 - Probability & Language models (3 points, 2 bonus)

You're part of a research team exploring a newly discovered planet. There are two species of aliens living here, Xaabs and Yaabs. At the entrance to a cave, you hear the following:

baaabaaaa

Your colleague has shared with you a simplified models, θ_X and θ_Y of the sounds these two species make.

Total things = 9

	θ_X	θ_Y
a	0.7	0.4
b	0.3	0.5
c	0.0	0.1

Q1a (3 points) - According to the two models, what is the likelihood of the sounds you heard, given it came from that species? Can you use this to make a prediction about what species you heard? What assumption(s) would you have to make?

Prob(X) = 0.07411
Prob(Y) = 0.0004096

It is more probable that the sound has come from Xaabs rather than yaabs. To confidently make this conclusion we would also have to assume that xaabs and yaabs are not bilingual. Also we would have to assume that the alphabet for each species is limited to these three sounds, for example if one species had a larger alphabet the sounds probability would be smaller.

Q1b (Bonus - 1 point) - For each species, what is the cross-entropy between the simplified model of the sounds they make and what you heard? What does cross-entropy mean here?

Xaab cross entropy = -sum(p(0.3)*log2(2/9), p(0.7)*log2(7/9)) = 0.904776556
Yaab cross entropy = -sum((0.5)*log2(2/9) + (0.4) * log2(7/9)) = 1.22999053248

Q1c (Bonus - 1 point) - For each species, what is the perplexity of the sound you heard? What does perplexity mean here?

Xaab pp = 1.8722544984245282

Yaab pp = 2.34565450532579

Perplexity, in this case, is giving us an idea on the efficiency of the language model. In this case we can say the xaab language shows a lower perplexity, this indicates that this is the more likely outcome as their is less uncertainty

Q2 - Smoothing (3 points, 3 bonus)

You ask you colleague to send you their raw data so you can compute a smoothed model:

	X counts	Y counts
a	70	40
b	30	50
c	0	10

Q2a (3 points) - Rewrite the unigram model (from Q1) using Laplacian smoothing.

	X Counts	Y counts
a	71 / 103 0.6	41 / 103 0.39
b	31 / 103 0.3	51 / 103 0.49
c	1 / 103 0.009	11 / 103 0.1

Total counts = 103

Q2b (1 point) - What problem would you run into if you tried to apply Good-Turing smoothing?

We need a count of one to efficiently apply Good-Turing, as we do not have any single counts this would not be possible

Q2c (Bonus - 2 points) - Compute the like likelihood of "cabaaabaaaa" for each species using the Laplace-smoothed model.

Xaab = 16.165515224881833 or 1.3604889599999995e-05

Yaab = 16.247852553506206 or 1.2850174234414884e-05

We can see that with the smoothing, the chances of the sound being from yaab speak is slightly greater.

Q3 - Sets (4 points)

Given the sets
A = { a , b , c }
B = { a , d , e , f }

and the universe
U = { a , b , c , d , e , f }

find the following:
Q3a (1 point) - A ∪ B = {a ,b ,c ,d ,e ,f }
Q3b (1 point) - A ∩ B = {a}
Q3c (1 point) - A ∖ B = {b,c}
Q3d (1 point) - A^{c} ∩ B = {d,e,f}

Key

- ∪ ,union, objects that belong to set A or set B
- ∩ , intersection, objects that belong to set A and set B
- \, relative complement, objects that belong to A and not to B
- Ac (A with hat), Complement all objects not in A

Q4 - Joint and conditional (4 points)

Now let X be a random variable defined by uniformly selecting elements of U (from Q3) and let p (A) = p (X = x ∈ A) (likewise for p (B)). Compute the following:

A = { a , b , c }

B = { a , d , e , f }
U = {a,b,c,d,e,f}

Q4a (1 point) - p (A) = 1/2

Q4b (1 point) - p (A , B) = 1/6
Q4c (2 points) - p (A | B) = $p(A|B) = \frac{count(A \cap B)}{count(B)}$ = 1/4

Q5 - Information theory (2 points, 1 bonus)

Q5a (2 points) - Compute the entropy of A : H (A)

Entropy of A when you pick an element from the universe.

A : H (A) = -sum(0.5 * log2*0.5) + (0.5 * log2*0.5) = 1

Q5b (Bonus - 1 point) - Compute the information gain of A given B : I G (A | B)

Work out entropy of A given B
IG of A|B is the entropy of A - entropy of A|B

Entropy of A|B = (((0.25)* log2(0.25))) + ((0.75) *(log2(0.75))) = 0.20751874963

B : I G (A | B) = 1 - 0.20751874963 = 0.79248125037

Q6 - Exponentiation & Logarithms (3 points, 1 bonus)

Q6a (2 points) - Simplify: $\log_N \frac{N}{2} + log_N 2$

Q6b (1 point) - Convert to an expression using only log base M: $\log_2 x$

Q6c (Bonus - 1 point) - Simplify: $\log_N M \cdot \log_M N$

Q6a = $log_b(b) = 1$
Q6b = $log_2(x) = \frac{log_m(x)}{log_m(2)}$

Q6c = 1