Problem 1

We must first create contingency tables for each of the terms in our query:

Contingency table for "obama":

Documents	Relevant	Nonrelevant	Total
"obama" Present $x_t = 1$	$s = 1.5$	2.5	$df_t = 4$
"obama" Absent $x_t = 0$	0.5	0.5	1
Total	$S = 2$	3	$N = 5$

$p_t = \frac{s}{S} = \frac{\frac{3}{2}}{2} = \frac{3}{4}$ $u_t = \frac{df_t - s}{N - S} = \frac{4 - \frac{3}{2}}{5 - 2} = \frac{5}{6}$

$c_t = log\frac{p_t}{1-p_t} - log\frac{u_t}{1-u_t} = log\left(\frac{\frac{3}{4}}{1-\frac{3}{4}}\right) - log\left(\frac{\frac{5}{6}}{1-\frac{5}{6}}\right) = log(3) - log(5) \approx -0.222$

Contingency table for "health":

Documents	Relevant	Nonrelevant	Total
"health" Present $x_t = 1$	$s = 1.5$	1.5	$df_t = 3$
"health" Absent $x_t = 0$	0.5	1.5	2
Total	$S = 2$	3	$N = 5$

$p_t = \frac{s}{S} = \frac{\frac{3}{2}}{2} = \frac{3}{4}$ $u_t = \frac{df_t - s}{N - S} = \frac{3 - \frac{3}{2}}{5 - 2} = \frac{1}{2}$

$c_t = log\frac{p_t}{1-p_t} - log\frac{u_t}{1-u_t} = log\left(\frac{\frac{3}{4}}{1-\frac{3}{4}}\right) - log\left(\frac{\frac{1}{2}}{1-\frac{1}{2}}\right) = log(3) - log(1) \approx 0.477$

Contingency table for "plan":

Documents	Relevant	Nonrelevant	Total
"plan" Present $x_t = 1$	$s = 1.5$	1.5	$df_t = 3$
"plan" Absent $x_t = 0$	0.5	1.5	2
Total	$S = 2$	3	$N = 5$

$p_t = \frac{s}{S} = \frac{\frac{3}{2}}{2} = \frac{3}{4}$ $u_t = \frac{df_t - s}{N - S} = \frac{3 - \frac{3}{2}}{5 - 2} = \frac{1}{2}$

$c_t = log\frac{p_t}{1-p_t} - log\frac{u_t}{1-u_t} = log\frac{\frac{3}{4}}{1-\frac{3}{4}} - log\frac{\frac{1}{2}}{1-\frac{1}{2}} = log(3) - log(1) \approx 0.477$

Now that we have calculated the c_t value for each of the terms in our query, we can now calculate the RSV value for each of the documents:

Doc1: $\sum_{x_t=q_t=1} c_t = -0.222 + 0.477 = 0.255$

Doc2: $\sum_{x_t=q_t=1} c_t = -0.222 + 0.477 = 0.255$

Doc3: $\sum_{x_t=q_t=1} c_t = -0.222 + 0.477 + 0.477 = 0.732$

Problem 2

Consider the following set of words and their English classification:

event	word	English?	probability
1	ozb	no	$\frac{4}{9}$
2	uzu	no	$\frac{4}{9}$
3	zoo	yes	$\frac{1}{18}$
4	bun	yes	$\frac{1}{18}$

Part 1

To compute the priors and conditionals for this series of documents, we will fill in the following table:

	prior	b	n	o	u	z
b	0	0	0	0	0	0
n	0	0	0	0	0	0
o	0	0	0	0	0	0
u	0	0	0	0	0	0
z	0	0	0	0	0	0

Part 2

Problem 3

Consider the following series of documents:

docID	Document Text
1	click go the shears boys click click click
2	click click
3	metal here
4	metal shears click here

Computing the probability for a query based upon a document will be done using the following formula:
$P(q|d) = \prod (\lambda P(t|M_d) + (1 - \lambda)P(t|M_c))$

In this formula, $\lambda = 0.5$, M_d is the language model for the document and M_c is the language model for the entire collection. To compute these, we need to generate language models across our collection for the terms in each query. Additionally, for each query, we'll need to create language models for each document in the collection.

First we will compute the M_c values for each of the terms in the three queries.

query term	M_c
click	$P(\text{click}\|c) = 0.44$
shears	$P(\text{shears}\|c) = 0.13$

Now we will compute the M_d values for each document for each of the terms in the three queries:

query term	M_{Doc1}	M_{Doc2}	M_{Doc3}	M_{Doc4}
click	0.5	1	0.00	0.25
shears	0.13	0.00	0.00	0.25

Here are the probabilities for three queries in this language model broken into two tables:

Query	λM_c	Doc1	Doc2
click	0.22	$0.22 + 0.50(0.50) = 0.47$	$0.22 + 0.50(1.00) = 0.72$
shears	0.07	$0.07 + 0.50(0.13) = 0.14$	$0.07 + 0.50(0.00) = 0.07$
click shears	0.03	$0.03 + 0.50(0.50 \cdot 0.13) = .06$	$0.03 + 0.50(1.00 * 0.00) = 0.03$

Query	λM_c	Doc3	Doc4
click	0.22	$0.22 + 0.50(0.00) = 0.22$	$0.22 + 0.50(0.25) = 0.35$
shears	0.07	$0.07 + 0.50(0.00) = 0.07$	$0.07 + 0.50(0.25) = 0.20$
click shears	0.03	$0.03 + 0.50(0.00 \cdot 0.00) = .03$	$0.03 + 0.50(0.25 * 0.025) = 0.06$

Using these values as $P(q|d)$ values, we can rank the documents for each query:

- "click" - Doc2, Doc1, Doc4, Doc3
- "shears" - Doc4, Doc1, Doc2, Doc3
- "click shears" - Doc1, Doc4, Doc2, Doc3

Problem 4

Part 1

Part 2

Part 3