Assignment 4 Solution

Problem 1

For all keywords in the query: only Doc 3 is relevant for the query, and this doc contains all keywords. Thus, N = 3, S = 1, and s = 1.

Obama:
df_Obama = 3
c_Obama = $log(\frac{(1+½)/(1-1+½)}{(3-1+½)/((3-3)-(1-1)+½)})$ = -0.22

health:
df_health = 2
c_health = $log(\frac{(1+½)/(1-1+½)}{(2-1+½)/((3-2)-(1-1)+½)})$ = 0.48
plan:
df_plan = 2
c_plan = c_health = 0.48

Thus, the RSV scores for the 3 documents are:

RSV_doc1 = -0.22 + 0.48 = 0.26
RSV_doc2 = -0.22 + 0.48 = 0.26
RSV_doc3 = -0.22 + 0.48 + 0.48 = 0.74

Problem 2

Parts 1 and 2

Priors:
P(e) = 1/9
P(not e) = 8/9

Conditional probabilities:
P(b|e) = 1/6
P(n|e) = 1/6
P(o|e) = 2/6
P(u|e) = 1/6
P(z|e) = 1/6

P(b|not e) = 1/6
P(n|not e) = 0/6, smoothed to: 0.01
P(o|not e) = 1/6
P(u|not e) = 2/6
P(z|not e) = 2/6

Document scores:
Score(e|zoo) = 1/9 * 1/6 * 2/6 * 2/6
Score(not e|zoo) = 8/9 * 2/6 * 1/6 * 1/6
The document is classified as not English.

Part 3

The vocabulary in this case is: o1, z2, b3, u1, u3, z1, o2, o3, b1, u2, n3.

Conditional probabilities:
P(z1|e) = 1/6
P(o2|e) = 1/6
P(o3|e) = 1/6

P(z1|not e) = 0/6, smoothed to: 0.01
P(o2|not e) = 0/6, smoothed to: 0.01
P(o3|not e) = 0/6, smoothed to: 0.01

Document scores:
Score(e|zoo) = 1/9 * 1/6 * 1/6 * 1/6 = 0.00051
Score(not e|zoo) = 8/9 * 0.01 * 0.01 * 0.01 = 0.00000088
The document is classified as English.

Problem 3

(best score per query in bold)

Query	Doc 1	Doc 2	Doc 3	Doc 4
click	0.46875	**0.71875**	0.21875	0.34375
shears	0.125	0.0625	0.0625	**0.1875**
click shears	0.0585	0.0449	0.0136	**0.0644**

P(click|doc1) = (4/8 + 7/16) / 2 = 0.46875
P(click|doc2) = (2/2 + 7/16) / 2 = 0.71875

P(click|doc3) = (0/2 + 7/16) / 2 = 0.21875
P(click|doc4) = (1/4 + 7/16) / 2 = 0.34375

P(shears|doc1) = (1/8 + 2/16) / 2 = 0.125
P(shears|doc2) = (0/2 + 2/16) / 2 = 0.0625
P(shears|doc3) = (0/2 + 2/16) / 2 = 0.0625
P(shears|doc4) = (1/4 + 2/16) / 2 = 0.1875

P(click shears|doc1) = 0.46875 * 0.125 = 0.0585
P(click shears|doc2) = 0.71875 * 0.0625 = 0.0449
P(click shears|doc3) = 0.21875 * 0.0625 = 0.0136
P(click shears|doc4) = 0.34375 * 0.1875 = 0.0644