Exercise 2 – Kullback-Leibler Divergence

a)

As a first step, we compute the term frequencies from the given table:

Profile	**A1**	**A2**	**A3**	**Q**
tf	**233**	**281**	**139**	**140**
term1	20	21	3	25
term2	36	100	23	40
term3	90	100	12	45
term4	75	3	67	10
term5	12	57	34	20

Table 1: Term frequencies **tf**

We can compute the author profile by computing the relative term frequency. We use Laplace smoothing with $\lambda = 1$.

Profile	**A1**	**A2**	**A3**	**Q**
term1	0.02702703	0.20903955	0.72661871	0.19047619
term2	0.71171171	0.25988701	0.04316547	0.14285714
term3	0.0990991	0.17514124	0.05035971	0.28571429
term4	0.10810811	0.18644068	0.17266187	0.23809524
term5	0.05405405	0.16949153	0.00719424	0.14285714]

The Kullback-Leibler Divergence for the query text Q and the author A_j is then given by

$$KLD(Q||A_j) = \sum_{i=1}^{m} q(t_i) \cdot \log_2 \left(\frac{q(t_i)}{a_j(t_i)} \right),$$

where m is the number of features, $q(t_i)$ and $a_j(t_i)$ are the occurrence probabilities for term t_i in Q or A_j, respectively. Therefore we have

$$KLD(Q||A_1) = 0.38845693$$
$$KLD(Q||A_2) = 0.19389402$$
$$KLD(Q||A_3) = 0.96495801$$

b)

Using the same computations as above, we get the following values for the Kullback-Leibler divergence for the given table:

$$KLD(Q||A_1) = 1.11360837$$
$$KLD(Q||A_2) = 0.10161531$$
$$KLD(Q||A_3) = 1.32055294$$