1

1.a

(a)
Let us note that due to the formula that defines each α_i the sum of all α_i is 1. Hence, by looking at each value as category we easily get that this is a categorical distribution.
(b)
this would happen if
1. there exists some k_j where $k_j^T q$ is large and
2. for all other k_i where $i \neq j$ $k_i^T q$ is small.
Let us remember that $a^T b = |a||b|cos(\theta)$ where θ is the angle between a and b. Lets us assume for simplicity sake that $|q| = 1$ and $|k_j| = 1$ for all j.
Then, for the two conditions above to hold we need that q and some single k_j are almost parallel, i.e their angle is small on the unit circle. Whereas all other k_i are almost in the opposite direction to q, i.e their angle with q is large on the unit circle.
(c)
Assuming that some k_j is parallel to q and all other k_i are in the opposite direction to q we get that $k_j^T q = 1$ and $k_i^T q = -1$ for all $i \neq j$.
We get that $e^{k_j^T q} = e^1 = e$ and $e^{k_i^T q} = e^{-1} = \frac{1}{e}$.
We get that $\alpha_j =$

(d)
This means that the output c will depend almost only on the the single v_j that corresponds to the k_j that is parallel to q. This is, obviously, not desirable since we want the output to be contextualized, i.e depend on the full sentence. Without this property the model will not be expressive enough.

2

2.d

The result of london baselinr is -
Correct: 25.0 out of 500.0: 5.0%
The result of the finetuned model is -
Correct: 10.0 out of 500.0: 2.0%

2.f

Correct: 115.0 out of 500.0: 23.0%

2.g

Let us note that the pretraining data contains the birthplace information about the individuals which are in the birth_dev dataset. It seems that the model is able to ”remember” the birthplace knowledge that it acquires during pretraining, and it is able to infer that knowledge during evaluation. Let us also note that the finetuning process does not make the model forget the knowledge it acquired during pretraining.

References