1

1.1

(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. Hence for the conditions above to hold we need that k_j is parallel to q and all other k_i are in the opposite direction to q.
Plus, we need that either q is very large (i.e large norm) or that k_j is very large or that the other k_i vectors norm is very large.

(c)

Assuming that the conditions above are correct, we get that $c \approx v_j$

(d)

This means that the output c will depend almost only on a 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.

1.2

(a)

Let us define the following matrix M:

$$M = c_1 \frac{a_1 a_1^T}{a_1^T a_1} + c_2 \frac{a_2 a_2^T}{a_2^T a_2} + \cdots + c_m \frac{a_m a_m^T}{a_m^T a_m}$$

If we multiply M by a_i we get that

$$M a_i = c_i a_i$$

i.e, a_i is an eigenvector of M with eigenvalue c_i. That means that

$$Mv_a = M(a_1 + \cdots + a_m) = Ma_1 + \cdots + Ma_m = c_1a_1 + \cdots + c_ma_m = v_a$$

Plus, when multiplying M by v_b we get that

$$Mv_b = c_1\frac{a_1a_1^Tv_b}{a_1^Ta_1} + c_2\frac{a_2a_2^Tv_b}{a_2^Ta_2} + \cdots + c_m\frac{a_ma_m^Tv_b}{a_m^Ta_m}$$

we know that $a_i^Tv_b = 0$ for all i since

$$a_i^Tv_b = a_i^Tb_1 + ... + a_i^Tb_m = 0$$

which is true because the two basis are supposed to be orthogonal. Hence

$$Mv_b = 0$$

Hence we finally get that

$$Ms = Mv_a + Mv_b = v_a$$

(b)

we define q as follows:

$$q = (k_a + k_b) * H$$

where H is some Huge number. It holds that

$$k_aq = k_bq = H$$

Plus, it holds that

$$k_iq = 0$$

if $i \neq a, b$. We get that $\alpha_a = \alpha_b = \frac{e^H}{n-2+2e^H}$ and $\alpha_i = \frac{1}{n-2+2e^H}$ for $i \neq a, b$.
For large enough H we get that $\alpha_a \approx \alpha_b \approx \frac{1}{2}$ and $\alpha_i \approx 0$ for $i \neq a, b$. Hence $c \approx \frac{1}{2}(v_a + v_b)$

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