Lucas Sipos

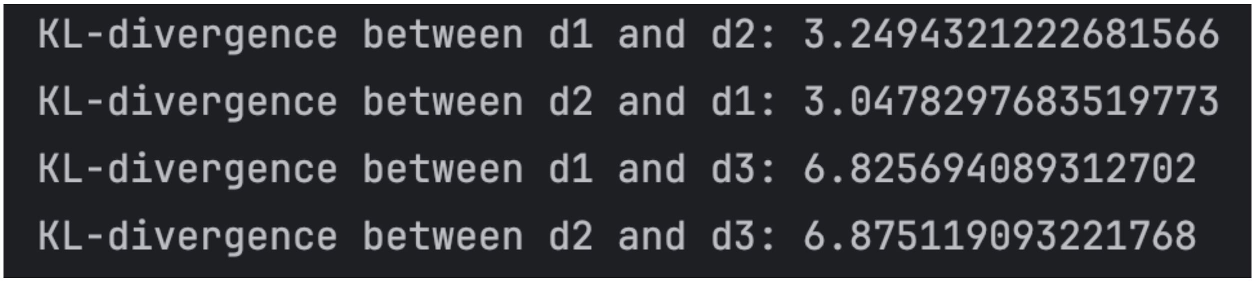
Input:

d1. “john fell down harry fell as-well down by the stream the sun shone before it went down mary was fine”

d2. "bill fell down jeff fell too down by the river the sun shone until it sunk down belinda was ill"

d3. "Clyde/Gucci (my cat) climbed the old oak tree, its green eyes sparkling in the sunlight as he was fearless of the height"

Results:



```
KL-divergence between d1 and d2: 3.2494321222681566
KL-divergence between d2 and d1: 3.0478297683519773
KL-divergence between d1 and d3: 6.825694089312702
KL-divergence between d2 and d3: 6.875119093221768
```

Comments:

As we can see, the first 2 inputs (d1 and d2) are somewhat like each other, and we get a pretty good KL-divergence score, but why do we get different scores? Simple, actually; 2 lists of strings can be different from each other by having different lengths of different words, meaning that the different words between d1 and d2 with respect to d1 are more than the different words between d2 and d1 with respect to d2. Let’s see how are they different:

1. d1 – d2: 'well', 'mary', 'went', 'fine', 'john', 'before', 'stream', 'harry' – length 8
2. d2 – d1: 'bill', 'sunk', 'jeff', 'ill', 'river', 'belinda', 'until' – length 7

Before talking about the last 2, I want to talk about epsilon and gamma, from my understanding. Epsilon is a variable that, even if there are missing values in the other set, they will be added with a very low probability and basically keep being almost 0, but the key difference is that now is computationally good, and the algorithm works without any error or 0 division. So because we have this epsilon, we need a gamma which balances the data, it's kind of like a normalisation so as to have the sum of all the data still be 1.

Now that we got that out of the way, I can explain why d1 and d3 have a smaller KL-divergence than d2 and d3, even if the number of different tokens is bigger for d1 and d3 (13) than d2 and d3 (12). Let’s get all the information about it:

1. d1 and d3: sizes – d1=14 and d3=18, differences - 13

2. d2 and d3: sizes – d2=13 and d3=18, differences – 12

So, because of that, they have different gammas, d1 and d3 gamma = 0.9992(7) and d2 and d3 gamma = 0.999(3). So, we already know that the gammas are different, and the bigger KL-divergence has a bigger gamma, which is obvious. But let’s get a difference between d1 and d2:

- In d1, the token down occurs three times, making it dominate the distribution
- In d2, the token down also occurs three times, but d2 has slightly fewer tokens overall (13 vs. 14) and a bigger gamma, so its relative weight is slightly larger in d2

In conclusion, the key is not only to count the missing tokens but also to see how they are distributed and what’s the weight of the penalty.