1. **Take the simple sentences we used in our word example Put these into the program and compute the K-L divergence scores for them, in both directions.**

Kullback–Leibler divergence (also called **relative entropy**) is a measure of how one probability distribution is different from a reference probability distribution. It quantifies how much one probability distribution differs from another probability distribution. It is the average number of extra bits needed to encode the data, due to the fact that we used distribution q to encode the data instead of the true distribution p.

Entropy :-
It is the number of bits required to transmit a randomly selected event from a probability distribution. With entropy we quantify how much information is in the data. It provides a lower bound on the number of bits needed to encode the data.

$$H(X) = E(I(X)) = \sum_{i=1}^{N} p_i \cdot \log_2\left(\frac{1}{p_i}\right)$$

Cross-entropy :-
It calculates the number of bits required to represent or transmit an average event from one distribution compared to another distribution.

In general the KL divergence can be computed as the **difference between cross entropy and entropy.** (Kullback-Leibler Divergence Explained, 2017)

$$D_{KL}(P||Q) = H(P,Q) - H(P)$$

The **KL divergence** between two distributions Q and P is given by the formula -

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

Here,
P(x) => True probability of one document
Q(x) => Approximated probability of another(relative) document

If two distributions perfectly match, DKL(p || q) = 0, otherwise it can take values between 0 and infinity. Lower the KL divergence value, the better we have matched the true distribution with our approximation.

Following are the stories taken =>

```
d1 = """ john fell down harry fellas-well mary was
fine down by the stream the sun shone before it went down"""

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

Following are the scores obtained after running the program =>
KL-divergence between d1 and d2: 3.704719577426546
KL-divergence between d2 and d1: 3.0973054936655595

The scores for Dkl(d1, d2) are low which indicates that both the documents have less divergence and their probabilities match to a great extent. Since the DKL(d1 || d2) != DKL(d2 || d1), it means that KL divergence score is asymmetric.

2. **Now create a third story that is very different to the other two, add it to the program and report how its score changes relative to the first two. Comment on whether the score makes sense.**

Third story =>

```
d3 = """If you're visiting this page,
you're likely here because you're searching for a random sentence.
Sometimes a random word just isn't enough,
and that is where the random sentence generator comes into play.
By inputting the desired number,
you can make a list of as many random sentences as you want or need."""
```

Following are the scores obtained after running the program =>
KL-divergence between d1 and d3: 7.622145574584418
KL-divergence between d2 and d3: 7.755080331918002

Yes, the scores do make sense. From the above results we can see that the divergence scores are **high** as compared to the results for Dkl(d1, d2) and Dkl(d2, d1). This is because the document d3 is different and dissimilar from d1 and d2. Since the document is different there will not be many terms that will be the same within documents.

3. **Explain what role epsilon and gamma play in the computation of K-L.**
Epsilon and gamma are hyperparameters that are useful in controlling the final results of KL divergence score. While calculating KL Divergence score in the above manner, there is a small problem.

Consider two documents as follows -
d1 = “henry, says”
d2 = “henry, has very strong hold over his region”

The vocabulary intersection of the above documents consists of terms **‘henry’.** The K-L divergence of d1 and 2 would result in 0 because **‘henry’** is the common term occurring in both the documents. This would mean that both documents have very low divergence score. However there are other terms in d2 and d1 that are not even considered while calculating K-L divergence score.

To avoid the above scenario, a simple **back-off** smoothing model is proposed in which term frequencies appearing in the document are discounted and all the terms which are not in the document are given weightage. (Using Kullback-Leibler Distance for Text Categorization, 2003)
Formula for back-off method :-

$$P(t_k, d_j) = \begin{cases} \beta P(t_k \mid d_j) & if\ t_k\ occurs\ in\ the\ document\ d_j \\ \varepsilon & else \end{cases}$$

where,

$$P(t_k \mid d_j) = \frac{tf(t_k, d_j)}{\sum_{x \in d_j} tf(t_x, d_j)}$$

This process of giving weightage is called as the **epsilon probability (ε)**. It is set to a small value instead of 0 to avoid the distance to be infinite.

Gamma (γ) is a normalization coefficient to account of epsilon, so the probability of a term in a category lies between 0 and 1 i.e. satisfies the properties of probability.

References:

Kurt, W. (2017). Kullback-Leibler Divergence Explained — Count Bayesie. [online] Count Bayesie. Available at: https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained [Accessed 23 Nov. 2019].

Bigi, B. (2003) Using Kullback-Leibler Distance for Text Categorization. In: Sebastiani F. (eds) Advances in Information Retrieval. ECIR 2003. Lecture Notes in Computer Science, vol 2633. Springer, Berlin, Heidelberg.

Shibuya, N. (2019). Demystifying KL Divergence. [online] Medium. Available at: https://medium.com/activating-robotic-minds/demystifying-kl-divergence-7ebe4317ee68 [Accessed 23 Nov. 2019].

Tsagkias, M. (2010). KL-divergence of two documents. Available at: https://web.archive.org/web/20130903010418/http://staff.science.uva.nl/~tsagias/?p=185