**SENTIMENT USE**

**Q1.** Take the simple sentences we used in our word example from the lecture.

Put these into the program and compute the K-L divergence scores for them, in both directions.

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.

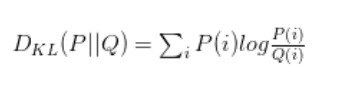
Comment on whether the scores make sense.

Explain what role *epsilon* and *gamma* play in the computation of K-L.

**Ans1:**

**KL Divergence**: ***[1]***

* The Kullback–Leibler divergence,KL Divergence is a measure of how one probability distribution is different from a second, reference probability distribution.
* For 2 documents, d1 and d2 let p and q be the probability distributions of documents 1 and 2.
* Two properties:
  + Summation of p and q equals 1
  + For each *i,* p(i) and q(i) should be greater than one.
* Then KL Divergence(Doc1||Doc2) is defined as summation of product of document 1 and logarithm of ratio of document1 to document 2.



where,Q(i) is the approximation and the P(i) is the true probability that is considered in matching Q(i).

It has 3 properties:

* KL Divergence(Doc1 || Doc2) is not same as KL Divergence(Doc2 || Doc1) due to asymmetry.
* If both documents have exactly same common words, KL Divergence is 0.
* It is additive for independent distributions.

The KL divergence infers that how well the probability distribution Q approximates the probability distribution P by calculating the cross-entropy minus the entropy.

**KL Divergence Score:** KL Divergence Score takes the values between 0 and infinity.

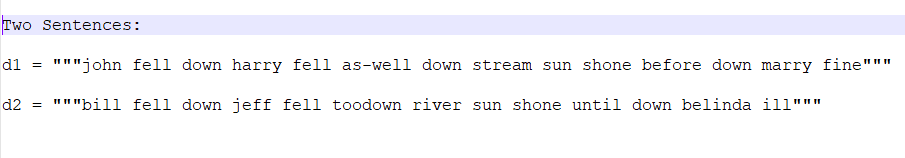
The lower the value of KL Divergence Score,the better true distributions are matched with the approximations. ***[2]***

**Entropy:** The measure of randomness in the information being processed.

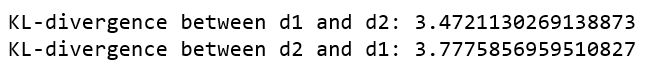
**Cross Entropy:** It is a measure from the field of information theory, building upon entropy, calculating the difference between two probability distributions.

KL Divergence calculates the relative entropy between two probability distributions, whereas cross-entropy can be thought to calculate the total entropy between the distributions.

The 2 sentences from the lecture notes itself are placed as d1 and d2. They are shown below:



KL Divergence Score among the two sentences is calculated and shown below:

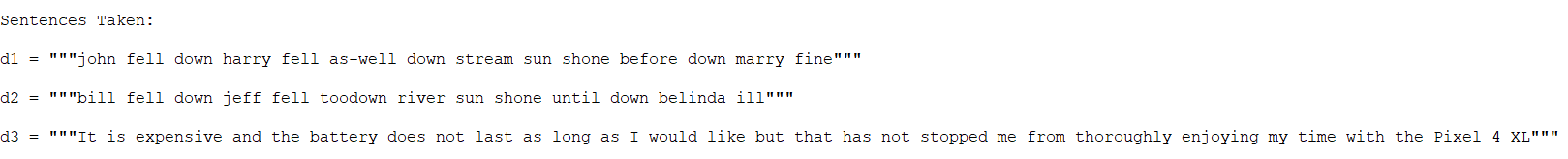


|  |  |
| --- | --- |
| K-L Divergence Among | Value |
| d1,d2 | 3.472 |
| d2,d1 | 3.777 |

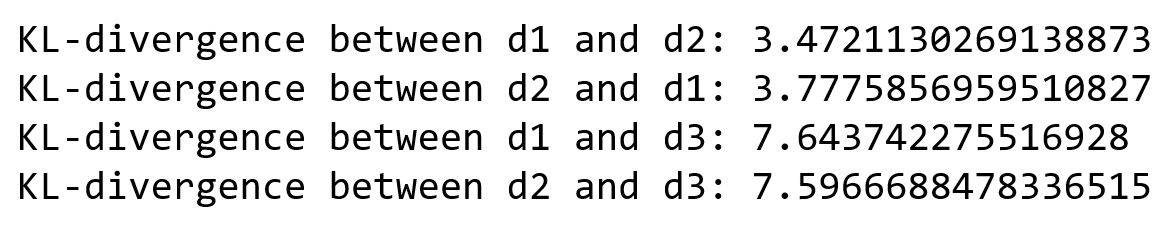
KL divergence, (relative entropy information gain or information divergence) of q(x) from p(x) measures how much information is lost when q(x) is used to approximate p(x). And it is different from the information lost when p(x) is used to approximate q(x). So the KL Divergence between documents 1 and 2 are different when calculated in both directions.

The documents 1 and 2 have more common words due to which their probability distribution is quite similar and thus the KL Divergence score is low.

The 3rd sentence different from the other two is shown below along with the two:



Calculated the KL Divergence Score among the three sentences is shown below:



|  |  |
| --- | --- |
| KL Divergence Among | Value |
| d1,d2 | 3.472 |
| d2,d1 | 3.777 |
| d1,d3 | 7.643 |
| d2,d3 | 7.596 |

The KL Divergence of third document differs from document 1 and 2 because the number of common terms in third one with rest of the documents is extremely low. Reason being the probability distribution is extremely diverge. Hence their KL Divergence score is quite highly comparative to that of document 1 with document 2 or vice-versa . Therefore, the KL Divergence score also makes sense.

**Hyper parameters**:

While calculating the KL Divergence among the 2 documents, we end up with 2 major problems:

1. KL Divergence is calculated in both directions due to the asymmetry.
2. Only the common terms among both documents are considered.

So, to avoid such drawbacks, Instead of using the common terms, whole vocabulary is to be considered.Thus smoothing comes but with Smoothing, the probability distribution doesn’t sum up to 1. So, back off smoothing that keeps the probability distribution sum to 1 and whole vocabulary is considered instead of common ones.

Example 1: There are two documents: **“Karishma likes mangoes”** and **“Mangoes are sweet”**. In both the documents, the common term is **“mangoes”** only. The rest of the words are not considered. Both the sentences are highly dissimilar but due to the common word “mangoes” and similar frequency(1), the KL Divergence(doc1,doc2) comes to be 0.

Example 2: There are two documents: **“Karishma likes mangoes”** and **“Apples are sweet”**. In both the documents, there is no common term. So, the KL Divergence(doc1,doc2) comes to be infinity.

To avoid such values of KL Divergence, hyper parameters are used:

* **Epsilon**: It is defined as the probability of a term which is not in a document. It is set to a small value instead of 0 to avoid the distance to be infinite. It is a value less than the minimum term probability occurring in either of the documents. Every word in the document is assigned a threshold probability instead of only to the common words.
* **Gamma**: It is defined as a normalization coefficient to account of epsilon, so a probability of a term in a category satisfies the properties of a probability (sum to 1).

**==================================================================================================**

**REFERENCES**

1. "Kullback–Leibler KL Divergence - Statistics How To", Statistics How To, 2019. [Online]. Available: <https://www.statisticshowto.datasciencecentral.com/kl-divergence/> [Accessed: 23- Nov- 2019]
2. M. Tsagkias, "KL-divergence of two documents | Manos Tsagkias", Web.archive.org, 2019. [Online]. Available: [https://web.archive.org/web/20130903010418/http://staff.science.uva.nl/~tsagias/?p=185](https://web.archive.org/web/20130903010418/http:/staff.science.uva.nl/~tsagias/?p=185) [Accessed: 23- Nov- 2019]