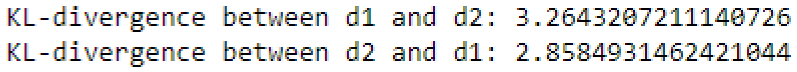
Practical 10

The first step to this practical was to take the simple sentences used in the lecture and plug them into the K-L divergence code provided. The K-L divergence between the two similar simple sentence stories (represented by d1 and d2) were calculated in both directions. These scores were as follows:

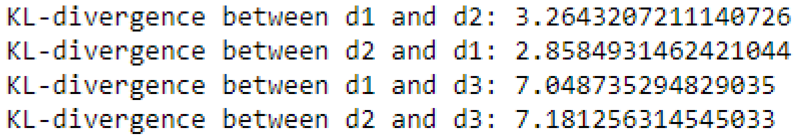


```
KL-divergence between d1 and d2: 3.2643207211140726
KL-divergence between d2 and d1: 2.8584931462421044
```

The K-L divergence between both simple sentence stories (represented here by d1 and d2) and a third completely different story (represented by d3) was also calculated. The third story was written in the same simple sentence style as the other stories, but with completely different words. The third story is as follows:

> *Katherine ate cake yesterday. Marian and Amie ate cake as well. The cake was a chocolate cake. All three girls ate cake until they were all so full they couldn't move. The next day, all three girls felt very sick.*

And the resulting K-L divergence scores were as such:



```
KL-divergence between d1 and d2: 3.2643207211140726
KL-divergence between d2 and d1: 2.8584931462421044
KL-divergence between d1 and d3: 7.048735294829035
KL-divergence between d2 and d3: 7.181256314545033
```

The scores between d1 and d2 remain the same and are much smaller than that of the completely different story. This makes perfect sense when considering the K-L divergence measurement. It measures the divergence between two probability distributions (in this case, the distribution of words in a story). It may be considered the measure of information loss between distribution P and distribution Q (in this program represented by the first and second texts – variable d1 and d2 – measured, respectively). It should be noted that the two same distributions might have different scores between them (e.g. the differing scores between measuring d1 and d2 - 3.26 – and d2 and d1 – 2.86). As a side note, this asymmetry, combined with the fact that it doesn't hold to triangle inequality, is why K-L distribution is not considered a "true" distance metric (Kurt, 2017). The closer the score is to 0, or the lesser the score, the more similar the two distributions. Considering this, the scores found above make perfect sense wherein the two similar stories have much lower and closer scores than the completely different story (although I had initially expected a score greater than seven, as very few of the words between the first two and the last story were similar (perhaps the few similar words such as "as-well" did lessen the score). In short, when considering the definition of K-L distribution, the scores found do indeed make sense.

There several parameters to this equation, but there are more beneath the surface of the initial equation, specifically in context of the probability distribution. One key parameter here is epsilon. Epsilon is a very small-valued probability which equals or represents the probability of unknown words. This probability value is given to terms which are not in the examined documents (Bigi, 2003). This is because if the probability of unknown words outside of the corpus were set at the expected probability value of zero (as in the likelihood of an unknown word outside of the corpus appearing in a sentence) then the calculations would be thrown off as they would be multiplying by zero. What the epsilon value does is nullify this problem by having a value so small it has an insignificant impact but also eliminates the problem of multiplying by zero. There are some constraints on the values. Firstly, this value is a probability given to terms not in the documents in the equation or terms not in the category in the equation. Therefore, epsilon's value must be smaller than the minimum probability of a term in the document or category for each possible term. Because of these constraints and that different values might work better on different documents, the epsilon value is obtained, not through an estimation equation, but experimentally. Epsilon, along with beta and gamma are chosen so that the corresponding probabilities sum to 1 (1 = 100% probability).

The parameter gamma depends on epsilon and is estimated by taking 1 (meaning 100% probability) and subtracting the difference in items between the two probability distributions, and multiplying by epsilon. This can be seen in the provided code where gamma = 1 – len of the differences between D1 and D2 times epsilon. Gamma is a normalization coefficient. In other words, it "smoothes" the probability distribution in a way that considers the context of the documents (as opposed to beta which is also a normalization coefficient but varies in terms of the size of the document) (Bigi, 2003). Changing gamma would result in changes to epsilon and beta (because, once again, these parameters must result in a total probability of 100% or rather all equal 1).

Works Cited

Bigi, B. (2003, April). Using Kullback-Leibler distance for text categorization. In European Conference on Information Retrieval (pp. 305-319). Springer, Berlin, Heidelberg.

Kurt, W. (2017, May 10). Kullback-Leibler Divergence Explained. Retrieved November 23, 2017, from https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained