1 Exercise 1

The three sentences which I have taken are:

1. From the lecture: " john fell down harry fellas-well mary was fine down by the stream the sun shone before it went down."
2. From the lecture: "bill fell down jeff fell too down by the river the sun shone until it sunk down belinda was ill"
3. My sentence: "We have just looked at tracking opinion but opinion should predict behaviour things I like are the things I buy so how does opinion get used to predict"

Calculcating the K-L score in all directions we have:

- kldiv(s1,s2) = 3.704
- kldiv(s2,s1) = 3.097
- kldiv(s1,s3) = 7.399
- kldiv(s2,s3)=7.559
- kldiv(s3,s1) = 7.165
- kldiv(s3,s2)=7.165

The K-L Divergence is a method to quantify of how two probability distributions differ; as it fails to satisfy both the triangle inequality and is assymetric (as we see in the score above), it fails to meet the criteria to be a metric (for ease, I'll refer to it as a measure going forward). The KL score can be considered as given two probability distributions P and Q, KL($P||Q$) is a measure of how much additional bits are required to be added to the probability distribution of P in order to fit the probability distribution Q. Using a loose example for intuition, if P was a uniform probability, and Q was a Gaussian, KL would be a measure of how many bits are required to 'make' the Uniform Distribution a Gaussian.

The KL Score for discrete random variables (for continuous, the integral is used) is defined as:

$$\text{KL}(P||Q) := \sum_{x\ inX} P(X)\log(\tfrac{P(X)}{Q(X)})$$

One note is that this definition implies that this measure is only defined if $Q(x) = 0 \implies P(x) = 0$ and in this case, the contribution of the term is considered as zero (rather than undefined) as $\lim_{x\to 0^+} log(x) = 0$. In the context of text classification, P is consiered to be the probability distribution followed by the document, while Q is the approximated probability distribution of the second document.

Based on the interpretation of this metric, if the distribution of P and Q are 'identical' it would imply a divergence score of zero and the lower this score is the more similar the distributions of P and Q.

Based upon this intereptation, the comparison scores between s1 and s2 make sense. As both scores are low, there is a bit of similarity between both documents. This is sensible as there are a number of words in both sentences which overlap, and therefore this low KL score is expected. We see the assymetry of this measure as the scores in both directions are different depending on which document is taken as the 'model' probability distribution (we see if s1 is taken as the model, less bits are required to for s2 to this distribution than the inverse).

The third sentence I have taken has very little overlap, both in regards to terms and sentiment, with the previous two sentences. We see for this sentence, comparing the other two sentences to it in both directions, that the KL score has increased. This implies the the number of bits required to fit s1 to the distribution of s3, and s2 to the distribution of s3, and vice versa, is more significant and hence there is a greater divergence between the probability distribution of the documents s1 and s2 to that of s3. This is a completely expected result because the sentence was explicitly used because there are very few terms in common between the three sentences and thus the higher KL score is exactly what we should have expected.

In the program provided, we see there is an implementation of smoothing backoff outlined in Kullback-Leibler Distance for Text Categorization. As two documents may have wildly different vocabularies (and not all terms will intersect, such as in the example sentences provided above) and there may be terms in one document which are not in the other, we want to account for this in the calculation of the KL score, and epsilon and gamme are used to account for this. The term epsilon is used to give weight for all of the terms in the model document that aren't used in the other document, and is set to be some small non-zero value as the KL score will tend to infinity as epsilon tends towards zero. Gamma is then introduced as a normalisation term to force the probability of a term being between zero and 1 (this need no longer be the case after introducing epsilon) and in order to ensure that it is still a probabilistic measure space gamma is required to force the probability of document terms being present to be between 0 and 1. Changing epsilon and gamma will result in different KL scores being produced. For example, if epsilon is instead set as

```
epsilon2 = min(min(_s.values())/ssum, min(_t.values())/tsum) * 0.00001
```

We see the following results are obtained instead of the ones listed above

- kldiv$_{\epsilon_2}$(s1,s2) = 6.146
- kldiv$_{\epsilon_2}$(s2,s1) = 5.114
- kldiv$_{\epsilon_2}$(s1,s3) =12.009
- kldiv$_{\epsilon_2}$(s2,s3) = 12.168
- kldiv$_{\epsilon_2}$(s3,s1) = 11.77
- kldiv$_{\epsilon_2}$(s3,s2)=11.77

which is exactly what we would expect; by setting epsilon as a smaller value we are giving a higher weighting to the non-intersecting terms and consequently pushing the KL score higher.