My understanding of KL Divergence

I couldn’t find any paper or document on KL Divergence for finding the divergence between two words. On the basis of all my readings about KL Divergence, this is what I tried to do to find the similarity between two words.

Example: (where two words are of same length)

Q: SILK R: MILK

KL(R,Q) = p(M)(log R(M)/Q(S)) + p(I)(log R(I)/Q(I)) + p(L)(log R(L)/Q(L)) + p(K)(log R(K)/Q(K))

= p(M)(log R(M)- log Q(S)) + 0 + 0 + 0

= (1/4)(log R(M)- log Q(S)) = 0.25 \* 3 = 0.75

Since divergence between M and S is 6 and to represent 6 we need ceil(log(6))= ceil(2.58)=3bits

p(x) is probability of occurrence of x in the final word

and [log R(M)- log Q(S)] this will give us the no of bits required to map the difference between the two letters

Hence the similarity is = 1 – divergence/max length of the either word

= 1 – 0.75/4 = 0.8125

Whereas similarity given by edit distance to us will be= 1- 1/5 = 0.800

Example: (where two words are of same length)

Q: SILK R: UILK

KL(R,Q) = p(U)(log R(U)/Q(S)) + p(I)(log R(I)/Q(I)) + p(L)(log R(L)/Q(L)) + p(K)(log R(K)/Q(K))

= p(U)(log R(U)- log Q(S)) + 0 + 0 + 0

= (1/4)(log R(U)- log Q(S)) = 0.25 \* 1 = 0.25

divergence between U and S is 2 and to represent 2 we need ceil(log(2))= ceil(1)=1bit

Hence the similarity is = 1 – divergence/max length of the either word

= 1 – 0.25/4 = 0.9375

Whereas similarity given by edit distance to us will be= 1- 1/5 = 0.8000

Example: (where two words are not of same length)

Q: SILK R: MILKS

Fill the left over spaces in the shorter length word by NULL (\0: assuming it is indicating the lowest possible character needed for its representation in the encoding scheme we are using) and then proceed in the similar manner.

Q: SILK\0 R: MILKS

KL(R,Q) = p(M)(log R(M)/Q(S)) + p(I)(log R(I)/Q(I)) + p(L)(log R(L)/Q(L)) + p(K)(log R(K)/Q(K)) + p(S)(log R(S)/Q(\0))

= p(M)(log R(M)- log Q(S)) + 0 + 0 + 0 + p(S)(log R(S)- log Q(\0))

Let us assume that S is represented by 4 bits in the encoding scheme used by us

= (1/5) (log R(M)- log Q(S)) + (1/5) (log 4) = 0.2 \*( 3 + 2 ) = 1

Hence the similarity is = 1 – divergence/max length of the either word

= 1 – 1/5 = 0.80

Whereas similarity given by edit distance to us will be= 1- 2/5 = 0.6

Observations

1. More accurate results are given by KL Divergence.
2. We can also see that KL Divergence differentiates between [SILK and MILK] & [SILK and UILK] similarities, whereas for edit distance both the pairs are equal similar.

Limitations

But in reality SILK should be replaced by AILK/WILK/DILK/XILK instead of UILK, because of fast typing nearby letters generally gets replaced by the actual letter