Classical Natural Language Processing

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Practice class session 3

Problem 3.1 Probabilities

a) Are X and Y as defined in the following table independently distributed, why?

x	a	a	b	b
y	a	b	a	b
$p(X = x, Y = y)$	0.32	0.08	0.48	0.12

b) Compute the entropies $H(X), H(Y), H(X|Y), H(Y|X), H(X, Y)$. Use the following log_2 values:

k	0.08	0.12	0.2	0.32	0.4	0.48	0.6	0.8
$log_2(k)$	-3.644	-3.059	-2.322	-1.644	-1.322	-1.059	-0.737	-0.322

c) Compute the Kullback-Leibler divergence $D(X||Y)$.

Problem 3.2 Language Models

Consider the following training text:

```
SALAMI IST IM SALAT. LAMA IST ALT.
```

a) Create an MLE n-gram model for letter unigrams. (Tip: Use fractions)

b) Create an MLE n-gram model for letter bigrams and draw the corresponding Markov Chain. Assume a start symbol.

Consider the following test text:

```
LAMA ISST
```

c) Compute the probabilities of the test text for each model.

d) Compute the perplexities of the models on the test text.

e) Use "add one" smoothing for both models and re-assess the probabilities and perplexities on the test set. Which LM does model the test text better?

f) Compute the probabilities and perplexities using a mixture model with $\lambda_1 = 0.1$ and $\lambda_2 = 0.9$ on the MLE estimates. What can you say about the quality of the model.