Statistical Methods of Language Technology

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

{ruppert,yimam}@informatik.uni-hamburg.de

Summerterm 2019

Assignment 3 Due date: 01.05.2019

Problem 3.1 Probability (5pts.)

a) A pattern C $\{yes, no\}$ to recognize a name event N $\{name, not\_name\}$ has the following properties:

$$P(C = yes|N = name) = 0.9$$
$$P(C = yes|N = not\_name) = 0.2$$

Assume the following:

- In newspaper text, around 5% of the words are names.
- In scientific text, around 1% of the words are names.

a) What is the probability to really see a name if C says so?

b) How low must the false positive rate P(C=yes|N=not-name) get so that this probability goes up to 50% for both kinds of text?

b) are X and Y as defined in the following table independently distributed?

x	0	0	1	1
y	a	b	a	b
$p(X = x, Y = y)$	0.3	0.1	0.2	0.4

c) Compute the entropies for:

a) $H(X), H(Y)$

b) $H(X, Y), H(X|Y), H(Y|X)$

c) $D(X||Y)$

Use the fact that $H(X|Y) = H(X, Y) - H(Y)$

Hint: You can use Google to compute expressions like "0.3 lg 0.3 + 0.25 lg 4". Check for correct bracketing.

Problem 3.2 Language Models (5pts.)

Download the homework data from Moodle. In the archive, you will find two files: Two German tokenized text with 50K lines each. Each line consists of a sentence; special tokens have been added at the beginning and at the end.

Example: `%^% %^% Leder : Vielleicht ringt Normann nur um Anerkennung . %$% %$%`

This sentence has 9 tokens, 10 bigrams and 11 trigrams: note that the special tokens %^% and %$% are only considered if needed. Tokens are separated by a space character.

a) List the 20 most frequent words from the training set.

b) Compute the percentage of tokens in the test data that have not been seen in the training data.

c) List the 20 most frequent bigrams from the training set.

d) Compute the percentage of bigrams in the test data that have not been seen in the training data.

e) Compute the percentage of trigrams in the test data that have not been seen in the training data.

f) How many sentences in the test data are estimated to have zero probability by an MLE bigram model from the training data?

g) Give the probabilities of the first 3 sentences from the test data, using a linear combination of 0-gram, unigram, bigram and trigram model with $\lambda_0 = 1.0 \times 10^{-10},\ \lambda_1 = 0.01,\ \lambda_2 = 0.2,\ \lambda_3 = 1 - (\lambda_0 + \lambda_1 + \lambda_2)$

You may use any programming language for this assignment. It is allowed to do parts of the exercise by hand. Please also submit your programs for error analysis.