Lehrstuhl für Informatik 6
Rheinisch–Westfälische Technische Hochschule Aachen
Prof. Dr.–Ing. H. Ney

May 5, 2015
Andreas Guta

3. Exercise Sheet **Statistical Methods in Natural Language Processing**

The solutions to the problems may be submitted until **Tuesday, May 12, 2015** before the exercise lesson. Either upload a digital version to the L2P, or bring your written/printed solution to the exercise lesson. Condition for obtaining the **Leistungsnachweis** (*Schein*) "Statistical Methods in Natural Language Processing" is the successful solution of 50% of the problems and the presentation of the solution of at least two problems in the exercise lessons.
The solutions to the problems can be submitted in groups of up to **two** students.
For programming exercises:

- The implementation has to be done in C/C++, but you can use additional standard Unix tools (tr, sed, awk...) for the preprocessing.
- Implementations have to be uploaded to L2P as well as printed out and submitted before the exercise lesson.
- Upload the programs as a `.tgz` or a `.zip` compressed directory to L2P. It must include a `Makefile` and must compile on Linux using simply the `make` command.
- Include a *short* description of the main data structures and algorithms used in your solution sheet.

1. [6 Points] In this exercise you will implement a multinomial text classifier ("count model") and test it on the corpus "20 newsgroups", consisting of mails sent to twenty different newsgroups on the internet. You can find this corpus (already preprocessed) on our webpage as a packed `.tgz` file. It contains a `README` file which describes the format of the corpus and gives some additional information.

   a) Implement a multinomial classifier (train it with relative frequencies). Consider the case where the vocabulary size can be restricted by giving a vocabulary (list of words).

   b) Plot the error rate as a function of the vocabulary size by selecting only the n first entries of the provided vocabulary. In order to get an idea of the evolution of the curve you have to use only small vocabulary sizes or plot using a logarithmic scale. Why does the error rate show this behaviour? (**Hint:** Look at the probabilities/scores for the different classes.)

   c) Propose a (simple) solution to the problem encountered in exercise 1b and plot the corresponding evolution of the error rate. (**Hint:** Alter slightly the parameter estimates, but make sure that the probability distributions are still normalized!)

   d) Generate the confusion matrices for your experiments and comment on them.

   e) Try your text classifier on the spam corpus available from our website (it has the same format as the 20 newsgroups corpus).

2. [1 Point] For a vocabulary of 4 words (A,B,C,D) draw or describe a finite state machine for a trigram language model with and without sentence end symbol.

3. [2 Points] Consider a language model where the sentence end is stochastically independent from the history, i.e., $p(\$|h) = p(\$)$ and $\sum_w p(w|h) = 1 - p(\$)$ for all h. Compute the probability function for the text lengths

$$p(N) = \sum_{w_1^N} p(w_1^N \$)$$

   as a function of $p(\$)$ and show that it is normalized, i.e.

$$\sum_N p(N) = 1\,.$$