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

April 18, 2018
Weiyue Wang, Parnia Bahar

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

- The solutions to the problems may be submitted until **April 25, 2018** before the exercise lesson. Upload a digital version to the L2P and 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 one problem in the exercise lessons.
- The solutions to the problems can be submitted in groups of up to **three** students.
- For programming exercises:
  - The implementation should be done in Python, but you can use additional standard Unix tools (tr, sed, awk...) for the preprocessing.
  - The implementation must be uploaded to L2P as a `.tgz` or a `.zip` compressed directory before deadline.
  - Include the results and a short description in your solution sheet.

1. [3 Points] Select **three** different tasks in statistical NLP, specify exactly the observations (input) and the output to be generated by the (Bayes) decision rule.

2. [2 Points] After examinig a collection of newspaper articles about politics, following *joint* probabilities have been estimated

	economy		¬ economy	
	crisis	¬ crisis	crisis	¬ crisis
oil	0.576	0.144	0.064	0.016
¬ oil	0.008	0.072	0.012	0.108

   where w denotes the presence and $\neg w$ the absence of word w. Compute the following:

   a) $p(\text{economy})$.
   b) $p(\text{oil})$ and $p(\neg\text{oil})$.
   c) $p(\text{economy}|\text{oil})$ and $p(\neg\text{economy}|\text{oil})$.
   d) $p(\text{oil}|\text{economy}, \text{crisis})$ and $p(\neg\text{oil}|\text{economy}, \text{crisis})$.

3. [1.5 Points] Suppose you are given a bag containing n unbiased coins. You are told that $n-1$ of these coins are normal, with heads on one side and tails on the other, whereas one coin is fake, with heads on both sides.

   a) Suppose you reach into the bag, pick out a coin uniformly at random, flip it, and get a head. What is the (conditional) probability that the coin you chose is the fake coin?
   b) Suppose you continue flipping the coin for a total of k times after picking it and see k heads. Now what is the conditional probability that you picked the fake coin?

c) Suppose you wanted to decide whether the chosen coin was fake by flipping it k times. The decision procedure returns FAKE if all k flips come up heads, otherwise it returns NORMAL. What is the (unconditional) probability that this procedure makes an error?

4. [1.5 Points] The members of the series $(p_k)_{k\in\mathbb{N}_0}$ obey following recursive equation

$$p_{k+1} = \frac{\lambda}{k+1} p_k, \quad k \in \mathbb{N}_0$$

for fixed $p_0 > 0$ and $\lambda > 0$.

a) Give an explicit representation of p_k, $k \in \mathbb{N}_0$.

b) For which values of p_0 is $(p_k)_{k\in\mathbb{N}_0}$ a probability distribution?

c) Repeat the exercise with the series $(p_k)_{k\in\mathbb{N}}$ with fixed $p_1 > 0$ and $\lambda > 0$.

5. [4 Points] you can find a tokenized text corpus "Europarl.txt" under:
`https://gigamove.rz.rwth-aachen.de/d/id/Wkgc87YB87MYj3?7&id=Wkgc87YB87MYj3`.
Collect statistics on the length of sentences (in words) and the length of words (in characters).

a) Compute the mean and variance for these samples.

b) If you plot the samples, are they normally distributed?

Hint:

- In order to open the link, you have to log in to `Gigamove` using your rwth account. You can find more information under `https://gigamove.rz.rwth-aachen.de/`
- Punctuations should be considered for the length of sentences, but **NOT** for the length of words. For example, for the sentence "`Thank you , Mr Segni .`", the sentence length is 6, and lengths of words are 5 3 2 5 respectively.
- You could use e.g. `matplotlib` for plotting the samples.