

# Natural Language Processing

## Assignment 1: Words

---

### 1 Introduction

This assignment involves material from classroom meetings 1 to 4. You should have watched the relevant videos, read the relevant chapters in the textbook and made a serious attempt at completing the relevant labs before you attempt this assignment. If you feel you have done that and still find the instructions unclear, you are welcome to email the course teachers and/or go to office hours to ask for help. The assignment is split into 2 sections, one about tokenization, one about language modelling and probability and statistics. Each section is worth 10 points. We expect between half a page and a page for each section. Please do not submit more than 4 pages overall. Your answers for each section should be self-contained.

### 2 Tokenization

In the videos and the book chapters, you learned about regular expressions (if you did not already know all about them before that ;)). In the labs, you got to play with regular expressions in the context of tokenization. You started with a very simple tokenizer and gradually refined it by looking at the errors it was making. If you have not done so already, try to improve it so that it reaches at least 95% precision and recall. Describe your approach to improving it. You may illustrate the problems you dealt with by adding snippets of code.

You will use a new test set `dev2-raw.txt` (available from `/local/kurs/nlp/basic2`) to run your tokenizer on. Write down for yourself your expectation of how good you think your tokenizer is going to be before running it. Compare the output of your system against `dev2-gold-sent.txt` and discuss the result. Are there still cases of under/oversplitting? If so, can you identify the reason? Can you identify difficult challenges?

### 3 Language Modelling

In Lab 3, you learned about how to use MLE for language modelling. Hopefully, in the lab, you saw its limitations. In Lab 4, you experimented with smoothing methods.

Define MLE and discuss its limitations. Explain the principles behind smoothing and how they help circumvent the limitations of MLE. Illustrate this with examples. (You may take them from the Sherlock Holmes text but use different examples than those we asked you to look at in the labs.)

Describe and compare different smoothing methods. Illustrate your answer with results from experiments you made in the lab. You may use the table you produced during the lab.

### 4 Grading Criteria

#### Basic Criteria

- Answers are given in understandable English.
- Answers are stated clearly and coherently.
- Answers are essentially correct.

#### Additional Criteria

- Answers are well motivated.
- Answers are well illustrated.
- Answers reveal extensive knowledge of the textbook chapter(s).

To pass the assignment, you must meet all the basic criteria on all subparts of the assignment. To get VG, you must in addition meet some of the additional criteria for most of subparts.

### 5 Submit the assignment

Submit your assignment as a pdf file named `firstname.lastname_assignment_1.pdf`. It should follow the style and margins given in the example submission even if not created with LaTeX. The submission is due on *studentportalen* before Friday November 16th at 20h00. Later submissions will be considered failed submissions and assessed after the final re-submission deadline on January 11th.