Choose a job you love, and you will never have to work a day in your life.

(Confucius)

Lab 2
Supervised Learning

In this lab, we apply supervised approaches to a dataset of hate speech[1]. The dataset consists of Wikipedia comments, labeled with the classes 'toxic, severe_toxic, obscene, threat, insult, identity_hate' or none.

The aim of the Lab is to develop a binary classification procedure that classifies text into hate speech / not hate speech. The minimal goal of the lab is to achieve a high F1-Score on the dataset. Note that the dataset is highly skewed regarding the classes: 89% of the texts are not hate speech. Hence, a simple baseline that classifies all texts as not hate speech has an accuracy of 89% already.

Possible applications of the classifier:

- Command line interface that takes a text file as an argument and outputs a hate speech score.
- Take tweets of politicians, calculate a hate speech score for each and compare the results.
- What are typical (non-)hate speech words?
- A tool that highlights/color codes/flags hate speech in a document on the sentence or word level.
- Whatever cool idea you might have :)

Technology:

Python – SciPy, sklearn, numpy, pandas, matplotlib

1https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data