# TNM098 - Lab 3

## Part 1: Image comparison

## Part 2: Text comparison

Ten different text files were given. Two of the ten files contained content from some of the other files. For one text, one sentence was copied, and for another a whole paragraph was copied. The task was to find out which texts contained the plagiarized content.

### Method

The problem was solved using Python. First, the text in the files had to be cleaned by removing punctuation and making all letters to lowercase. Each sentence was put on a single line. From the cleaned texts tables were made by letting each sentence be the key and setting the value to 1 in the Python data type dictionary. Now the dictionaries could be used for checking plagiarized content.

For all dictionaries, every entry in them was checked against all entries in the other dictionaries. If a match was found and it was more than 20 characters long, it was registered as plagiarism.

### Result

The plagiarized material was taken from file 2 and could be found in file 6 and 8. The computational time on Intel Core i5, Windows 8.1 was 1.62682950241881 seconds.

#### **In both text nr 6 and nr 2:**

i tried out the scales and found that my involuntary host weighed over 195 pounds a good deal of it around the middle

#### **In both text nr 8 and nr 2:**

rutherford was pacing with surgical precision up and down my den

#### **In both text nr 8 and nr 2:**

he looked slightly more self possessed than the day before and seemed to be in excellent physical condition

#### **In both text nr 8 and nr 2:**

i guessed at the contour beneath my wadded black silk dressing gown and re considered my original plan to throw him bodily out of the house for having come without my invitation