# Independent project: Preliminary description
# DAT450/DIT245 Machine learning for natural language processing

Ahmed Groshar, Gustav Molander, Noa Onoszko

December 10, 2020

## Introduction

In this project, we are going to perform text summarization with machine learning. The point is to shorten a text while retaining as much meaning as possible. There are many different approaches to this problem and our goal is to investigate a few different basic methods and if there is time, implement a state of the art method from the list at the nlpprogress.com website [1]. Below, there are some topics and approaches that we might consider:

- Extractive summarization with the TextRank algorithm
- Single sentence summarization
- Multi-sentence summarization
- Abstractive summarization with autoencoder LSTM and attention [2]
- Abstractive summarization with supervised learning and reinforcement learning (current SOTA) [3]
- How to improve cross-sentence coherence patterns

Ultimately, we aim to have at least 3 models with different characteristics. We might also compare our results with pretrained models.

## Dataset

For this project we are going to use the CNN / Daily Mail summarization dataset. It contains a test set of 11,490 instances, train set of 287,113 instances and a validation set of 13,368 instances. Each datapoint consists of a text article as predictor and a text summary as label.

There is also a Google Dataset with about 200 000 sentence-compression pairs for summarizing a single sentence. This is something we might find useful to use in our project.

## References

[1] http://nlpprogress.com/english/summarization.html.

[2] analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/.

[3] https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16838/16118.