# Check List

## Preliminary task

* ~~What type of documents does it contain~~
* ~~How many documents are there?~~
* ~~Calculate and visualize some simple statistics for the collection, e.g. the average document length, the average vocabulary size, etc.~~
* ~~Cluster the documents and visualize the clusters to see what types of groups are present (or whether the known classes can be found);~~
* ~~Index the documents so that you can perform keyword search over them;~~
* Train a Word2Vec embedding on the data and investigate the properties of the resulting embedding.

## Training models

* ~~Train a model to perform that task (by fine-tuning models on the training data).~~
* ~~Test pre-trained models on the task (if they already exist).~~
* Evaluate different models and compare their performance.

HINT: as a minimum here we would expect to see a linear classifier trained on the data (if an

appropriate for the task) and compare it with deep learning model, such as BERT.

## Possible extensions

* ~~Oftentimes we can improve performance of a model on a particular task by simply including additional data that is related to the task in its training set.~~
* There are many NLP challenges out there, so if you can’t find more data for the task you’re working on, look for another interesting challenge to work on.

## Task specific

* ~~Computing the cosine similarity between TF-IDF bag-of-words vectors for the sentences,~~
  + ~~Try BM25 instead~~
* ~~Extending the BOW vector to include n-grams (bigrams and trigrams)~~
  + ~~Maybe try more complex grams?~~
* ~~Consider the distance between average word embeddings~~
  + ~~Try more SOTA embeddings~~
* ~~Moreover, supervised models could be trained, in particular transformer-based models are often fine-tuned to classify pairs of sentences.~~
  + ~~Learning rate plot for better hyperparameter choice~~
* ~~LightGBM/XGBoost on the tabulare scores/data~~

## Currently working on:

Lars: Statistikk og visualisering av word embedding

Erik: Se litt nøyere på BERT fine tuning og andre modeller som LSTM og GRU

Andreas: n\_grams

Gjermund:

- 2 bokser fersken hermetikk

- laken

Sprit vodka, anasjuce

Tropisk nektar

Brus

6 l rødvin

0.5 l gin og vodka

1.5 l fanta exotic og 1.5l fanta lemon