NLP with Deep Learning and Advanced Machine Learning for Information Retrieval from Text Archives

The National Archives of Norway archives available from public sector in Norway. Almost all archives are written in Norwegian or Sami languages. Some of the archives are highly sensitive and should only be available for specific people, and some other archives can be public for everyone.

Natural Language Processing (NLP) is a very attractive research topic and has been growing dramatically in recent years, thanks to advances in data processing capabilities. Several companies (i.e. Google, Facebook) and research groups around the world (i.e. NLP group at Stanford and Harvard Universities) are actively working in this fields. There are several open sourced toolboxes within the field (i.e. TensorFlow from Google and PyTorch from Facebook).

The aim of this project is to apply NLP, deep learning and advanced machine learning techniques and technologies for information retrieval from raw documents. This information can be used to determine both the right context and metadata for the archives, as well as whether the archives contains sensitive information and could be made available for a given person.

Preliminary results on finding context information in some of analog, digitized and digital-born archives using classical machine learning algorithms (i.e. Logistic Regression, Random Forest) and pre-trained deep neural networks (i.e. BERT based models) are very promising and there are much more opportunities to explore.

To identify sensitive information, we believe that the following need to be identified:

- Names
- Institutions / organizations
- Email Address
- Date, Place
- Sensitive information (i.e. personal numbers etc.)
- Geographical information
- Context for the information

The National archives would like to make available different type of archives for students to use, in order to solve some of the above-mentioned challenges, including

- Handwritten text from 1600 century
- Machine written text
- Noark Archives (metadata) in version 3, 4 and 5 with public sector data
- Archives from “fagsystem” in SIARD format with json content information
- Archives from Social Media archives
- Digitally born material

All data are available in pdf, image and/or text format. As documents are in different formats, data extraction and cleaning and feature engineering are vital in the machine learning pipeline. Material with analog origin is scanned and interpreted (OCR or HCR).

Some of the archives will only be available from our offices (either in Trondheim or in Oslo) due to sensitive information, and the students must be prepared to sign “taushetserklæring”.

General sketch for the project thesis:

- Literature review on NLP and existing machine learning algorithms for NLP (i.e. text vectorization, classic machine leaning algorithms like Boosting Tree, Deep Neural Networks, etc.),
- Selection of the framework and creating an experimental setup (i.e. Python, OpenCV, Sci-Kit Learn, PyTorch, TensorFlow, Keras, etc),
- Implementation of the selected algorithms,
- Evaluation of the results and analysis.

General sketch for the follow-up master thesis:

- Application of deep learning techniques (i.e. Deep RNN, LSTM, CNN, Transformers etc) and comparing the results,
- Further development of the existing algorithms,
- Evaluation of the results.

Contact: Javad Rezaie <javrez@arkivverket.no>