Norsk regnesentral

De-identification of text documents for privacy protection

Many public & private organisations struggle to manage the personal data they gather or produce. This data may relate to patients, customers, welfare recipients, or even defendants in court cases. Such data must comply with privacy and data protection laws, such as the General Data Protection Regulation (GDPR) newly introduced in Europe. In particular, personal data cannot be distributed to third parties (or used for secondary purposes) without legal ground, such as the consent of the individuals to whom the data refers.

In many organisations, a large portion of this data takes the form of *text documents*. For instance, electronic health records often include a wide variety of clinical notes. The ubiquity of such text data is often problematic from a privacy perspective, as text documents are much harder to anonymise than structured databases.

In this thesis, you will develop new machine learning models to automatically de-identify text documents – that is, detect the occurrence of personal identifiers (such as names, addresses or telephone numbers) and mask them from the text. Although de-identification does not always amount to "full anonymisation" (as it may still be possible to identify persons through indirect clues), it is often a necessary first step and considerably reduce the disclosure risk.

More specifically, the student will develop and evaluate neural models for sequence labelling that take text documents as input and identify text spans including personal information. One particular challenge that will need to be addressed is how to handle semi-structured formats, such as clinical notes that are often divided in multiple sections.

This master thesis will be conducted as part of the CLEANUP (Ma**c**hine **le**arning for the **an**onymisation of **u**nstructured **p**ersonal data) project. The student will be expected to work on Norwegian data provided by one of the CLEANUP project partners and stored on TSD (see www.uio.no/tsd). The exact dataset will be determined at a later stage.

Contact person: Pierre Lison (plison@nr.no)