Aktivitetens navn

FAIR Embeddings for Textual Cultural Heritage (FAIR eTeCH)

Aktivitetens start-slut

March 1, 2019 - March 1, 2020.

Aktivitetens leder

Anne-Marie Pahuus & Kristoffer L Nielbo

Konsortium

- Aarhus University's Center for Humanities Computing Aarhus
- Danish Royal Library
- University of Southern Denmark

FAIR eTeCH participates in the Scandinavian collaboration *Nordic Digital Humanities Laboratory*, which has similar activities in Sweden, Finland and Norway.

Aktivitetens scope

Valid and scalable solutions to research questions for textual cultural heritage (TeCH) data will increasingly depend on access to so-called neural embeddings. Neural embeddings are abstract and distributed dense representations of language (characters, words, phrases) that are learned by data-intensive representation-learning algorithms implemented as deep neural network architecture. To ensure that DH researchers uses state-of-the-art technology for tackling complex TeCH problems, it is mandatory that they have access to pre-trained multi-level embeddings for their respective language, which follow the FAIR principles (findable, accessible, interoperable, and resuable). *FAIR Embeddings for Textual Cultural Heritage* (FAIR eTeCH) will pioneer FAIR embeddings for Scandinavian languages, which through a collaboration with national libraries and an innovative use of regulations pertaining to derived data can circumvent restriction on copyrighted and sensitive data.

One of the greatest challenges for large scale DH research is access to original or direct data (e.g., the content of newspaper article) because of copyright restrictions. In Denmark, for instance, a newspaper article in the Danish *MedieStream* has to be more than a century old in order to allow a researcher free data mining access. Embeddings however have status as derived data[1] that does not allow for a reconstruction of the original data source. Embeddings trained on large newspaper collections are more than adequate to solve problems related to semantic similarity and drift.

FAIR Danish Royal Library's current eScience infrastructure, the Cultural Heritage Cluster, does not contain GPU nodes necessary for deep learning and *Lamda* offers a robust solution

[1]The OECD Glossary of Statistical Terms defines a 'derived data element' as: A derived data element is a data element derived from other data elements using a mathematical, logical, or other type of transformation, e.g. arithmetic formula, composition, aggregation.

for piloting at a competitive price. In time NDHL will release embeddings for all Scandinavian languages hosted through national HPC service providers as part of the NDHL virtual laboratory.

Must-have leverancer

1. Acquisition of server and tests.
2. Implementation of pipeline.
3. Model training of neural embeddings for character, word and sentence-levels.
4. Release of neural embeddings and code/pipeline repository with associated DOIs through Zenodo.

Nice-to-have leverancer

Joint access to the server for all members of NDHL in order to ensure FAIR embeddings for all Scandinavian languages.

Økonomi

0. Hardware: Lambda Blade Server Premium m. Lambda Stack: 300.000 Dkr 1. Server acquisition: 100 hrs 2. Pipeline: 250 hrs 3. Model training: 250 hrs. 4. Release: 100 hrs.

Tidspunkt for leveranceplan

Ultimo March 2019