EGU22-10850
https://doi.org/10.5194/egusphere-egu22-10850
EGU General Assembly 2022

A Natural Language Processing-based Metadata Recommendation Tool for Earth Science Data

Armin Mehrabian, Irina Gerasimov, and Mohammad Khayat
NASA, Goddard Space Flight Center, Greenbelt MD, United States of America (armin.mehrabian@nasa.gov)

As one of NASA's Science Mission Directorate data centers, the Goddard Earth Sciences Data and Information Services Center (GES-DISC) provides Earth science data, information, and services to the public. One of the objectives of our mission is to facilitate data discovery for users and systems that utilize our data. Metadata plays a very important role in data discovery. As a result, if a dataset is to be used efficiently, it needs to be enhanced with rich and comprehensive metadata. For example, most search engines rely on matching the search query with the indexed metadata in order to find relevant results. Here we present a tool that supports data custodians in the process of creating metadata by utilizing natural language processing (NLP).

Our approach involves combining several text corpora and training a semantic embedding. An embedding is a numerical representation of linguistic features that is aware of the semantics and context. The text corpora we use to train our embedding model contains publication abstracts, our data collections metadata, and ontologies. Our recommendations are based on keywords selected from the Global Change Master Directory (GCMD) and a collection of ontologies including SWEET and ENVO. GCMD offers a comprehensive collection of Earth Science vocabulary terms. This data lexicon enables data curators to easily search metadata and retrieve the data, services, and variables associated with each term. When a query is matched against various keywords in the GCMD branch, the probability of the query matching these keywords is calculated. A similarity score is then assigned to each of the branches of the GCMD, and each branch is sorted according to this similarity metric. In addition to unsupervised training, our approach has the advantage of being able to search for keyword recommendations of different sizes, ranging from sub-words to sentences and longer texts.