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

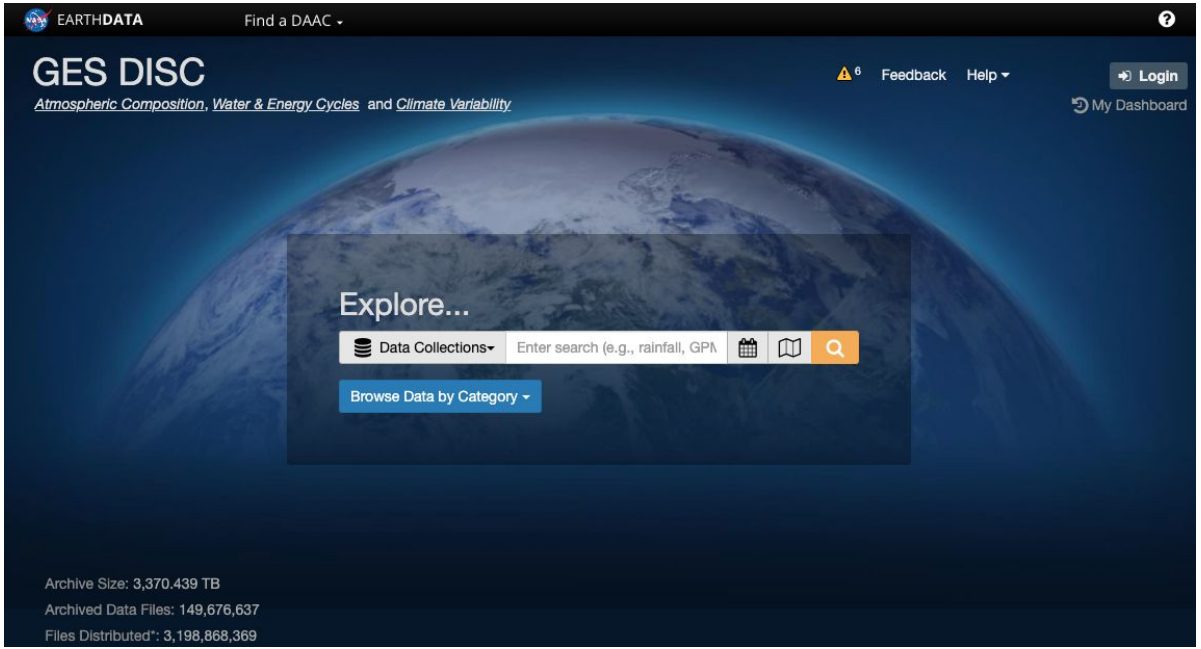
Armin Mehrabian, Irina Gerasimov, and Mohammad Khayat
NASA, Goddard Space Flight Center, Greenbelt MD, United States of America

Corresponding author's email: armin.mehrabian@nasa.gov

Who we are

- We are the NASA Goddard Earth Sciences (GES) Data and Information Services Center (DISC).
- Located at the Goddard Space Flight Center (GSFC) in Greenbelt, Maryland
- We are one of 12 NASA Science Mission Directorate Data Centers that provide **Earth science data**, **information**, and **services** to research scientists, applications scientists, applications users, and students.
- We archive and support data sets applicable to several NASA Earth Science Focus Areas including: Atmospheric Composition, Water & Energy Cycles, and Climate Variability.

https://disc.gsfc.nasa.gov/



Currently,

- We host more than 1.5k datasets
- 3 PetaBytes of data
- ~150M observation files

Data Discovery and Curation

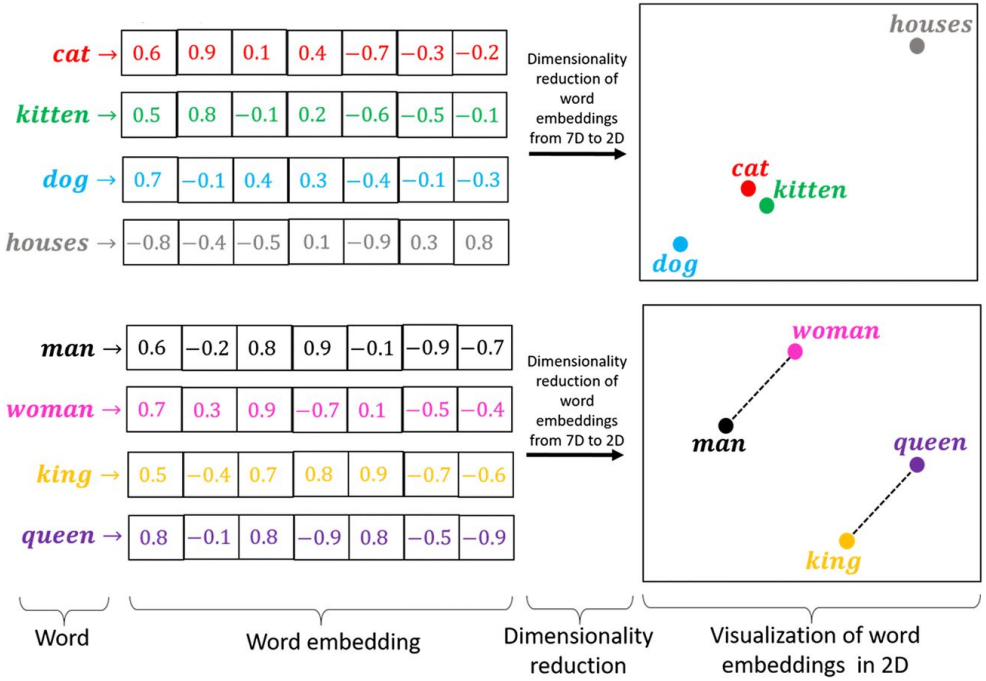
- NASA's Earth Observing System Data and Information System (EOSDIS) collection contains more than 65 petabytes of Earth science data which it opens up for open access
- A key element of our mission is data discovery and curation
- One aspect of our data curation process is to provide metadata with relevant scientific keywords
- We use Global Change Master Directory (GCMD) Keywords https://earthdata.nasa.gov/earth-observation-data/find-data/idn/gcmd-keywords
- The proposed tool here aims to assist data curators in finding proper science keywords for dataset metadata

Example science keyword in dataset metadata

```
"CollectionProgress" : "COMPLETE",
"ScienceKeywords" : [ {
  "Category" : "EARTH SCIENCE",
  "Topic" : "ATMOSPHERE",
  "Term" : "ALTITUDE",
  "VariableLevel1" : "BAROMETRIC ALTITUDE"
}, {
  "Category" : "EARTH SCIENCE",
  "Topic" : "LAND SURFACE",
  "Term" : "TOPOGRAPHY",
  "VariableLevel1" : "TERRAIN ELEVATION"
}, {
  "Category" : "EARTH SCIENCE",
  "Topic" : "LAND SURFACE",
  "Term" : "LAND USE/LAND COVER",
  "VariableLevel1" : "LAND USE/LAND COVER CLASSIFICATION"
} ],
```

Language Embeddings

- Language embeddings are representations of linguistic components using vectors
- Enables us to perform mathematical operations on words

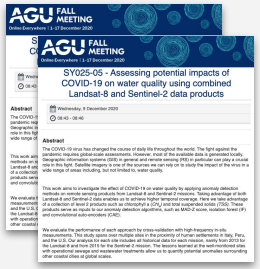


(cat ● kitten) = 0.9

(cat ● king) = 0.1

Source: Word Embedding: Basics. Create a vector from a word | by Hariom Gautam | Medium

Language Embedding Generation

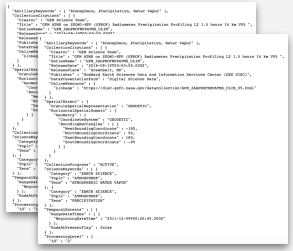


270K AGU Abstracts

NLP Pre-preprocessing

Train wordvectors
(Fastext, BERT, RoBERTa, etc.)

Model



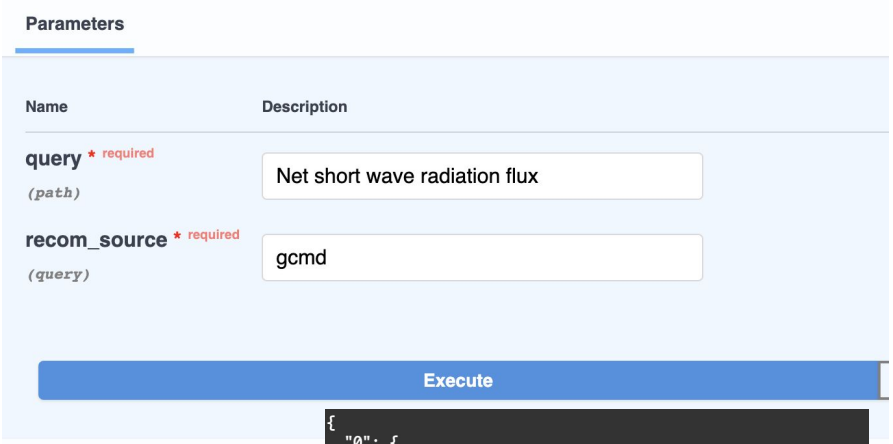
GES-DISC collections metadata

Flatten .json files into plain text
(select set of fields)

biodiversity = [0.23, -0.45, 0.89, ., ., -0.76]

m2tunpcld = [0.31, -0.30, 0.53, ., ., -0.49]

Science Keyword Recommender




```
{
  "0": {
    "Topic": "ATMOSPHERE",
    "Term": "ATMOSPHERIC RADIATION",
    "Variable_Level_1": "SHORTWAVE RADIATION",
    "Variable_Level_2": "DOWNWELLING SHORTWAVE RADIATION",
    "Variable_Level_3": null,
    "Detailed_Variable": null
  },
  "1": {
    "Topic": "ATMOSPHERE",
    "Term": "ATMOSPHERIC RADIATION",
    "Variable_Level_1": "LONGWAVE RADIATION",
    "Variable_Level_2": "DOWNWELLING LONGWAVE RADIATION",
    "Variable_Level_3": null,
    "Detailed_Variable": null
  },
  "2": {
    "Topic": "ATMOSPHERE",
    "Term": "ATMOSPHERIC RADIATION",
    "Variable_Level_1": "LONGWAVE RADIATION",
    "Variable_Level_2": "UPWELLING LONGWAVE RADIATION",
    "Variable_Level_3": null,
    "Detailed_Variable": null
```


- Embedding method
  Fasttext
  - slow training
  - supports sub-words
- Embedding trained on
  - **270k** AGU Abstracts
  - GES-DISC Data Collections Metadata
- Used embedding alone
  (Unsupervised)
- Selects keywords from all GCMD keywords and supported ontologies (ENVO, SWEET)

Thank You!

Feel free to contact me with any questions you may have.
armin.mehrabian@nasa.gov