CONSTRUCTOR UNIVERSITY

Data Structures and Processing

Dr. Carlos Henrique Brandt
cbrandt@constructor.university

CONSTRUCTOR UNIVERSITY

Hands-on / Homework

ETL pipeline with Github data

- We are going to query some (semi-structured) data from GitHub and organize that in a couple of (structured) datasets, files and directories.
  - GitHub API docs: https://docs.github.com/en/rest

Exercise:

- To query for the 100 most starred Python-based repositories (of this year), save the respective JSON content in individual files in their own directories.
- Create (related) tables to represented the (JSON) information.
- Download the corresponding *readme* files, and create an index (of words) relating the most overall frequent words to the repositories.

ETL pipeline with Github data

- Since we are not interesting in Github's API features, here is the URL to query for such information:

  https://api.github.com/search/repositories?q=created:>2023-01-01+language:python&sort=stars&order=desc&per_page=1

  - Copy-n-paste this URL into your browser to see the results for **1** repository

- The response is composed by some top-level attributes ('total_count', 'incomplete_results') that are not of our interest here, 'items' are.

ETL pipeline with Github data

- The contents of ‘items’ is what we want: those are the repositories' metadata we queried for.
- Save each ‘items’ metadata block in a JSON file, in a directory named after *owner* and/or *repository name*.
- Organize such information (i.e., inside items) into:
  - a "main" table for the items, each record represents a repository;
    - remove all unnecessary URLs: keep only home_url.
  - a "owners" table, with the information inside items' owner object;
    - same for license and topics.
- Download corresponding readme files, save them next to metadata's .json file.
  - Create an index of words for each readme file and compute words frequencies;
  - Merge them all into one "index" table of the (100) most frequent words;
  - Remember: an index (table) related word/term to source/location.

ETL pipeline with Github data

- Submit the Jupyter Notebook (.ipynb) file and corresponding ‘requirements.txt’ or ‘environment.yml’ file.