

Data Intake Report

Name: en-fr-translation-dataset

Report date: 12/06/2023

Internship Batch: LISUM21

Version: 1.0

Data intake by: Bogdan-Remus Pintilie

Data intake reviewer:

Data storage location: <https://www.kaggle.com/datasets/dhruvildave/en-fr-translation-dataset>

Tabular data details:

Total number of observations	number_of_rows.py – my PC couldn't do it
Total number of files	1
Total number of features	2
Base format of the file	.csv
Size of the data	8.2 GB

Proposed Approach:

- Set up an environment
- Upload csv file
- Read the file using different methods (pandas, dask, modin)
- Use a smaller subset of the data, reading the entire file at once can be memory-intensive
- Measure the computational efficiency of each method and document findings
- Perform basic validation on data columns (remove special characters and white spaces)
- Create a YAML file defining the separator and column names
- Read the YAML file and validate the number of columns and column names
- Write the file in pipe-separated (“,”) text format in gzipped format







