src folder

Inspect schema.org_properties and identifiers, investigate tree.ipynb

- Inspects schema.org_properties and identifiers in WDC data
- Inspects the Open Icecat catalog hierarchy

Get Icecat data, merge with WDC, investigate datasets.ipynb

- Gathers Icecat data from database
- Merges Icecat data with WDC data
- Calculates statistics for datasets

Build datasets.ipynb

- Constructs hierarchies
- Constructs final datasets used in the experiments

Analysis.ipynb

- Contains the basis for the error and ensemble analysis
- For hierarchical classification systems, calculates amount and fraction of total misclassifications that have their origin on each hierarchy level
- Calculates error overlap between certain classification systems
- Identifies types of misclassifications
- Calculates performance per first-level category (sub-trees)

Classification_Dict_and_Flat.ipynb

- Contains experiments for the dictionary-based classification system
- Contains experiments for the flat traditional classification system
- Plots data

Classification_LCPN.ipynb

- Contains experiments for the hierarchical traditional classification system

Classification_Fasttext.ipynb

- Contains experiments for the flat fastText classification system
- Contains experiments for the hierarchical fastText classification system
- Contains experiments for the hierarchical neural network + fastText classification system

Classification_Deep_Learning_Flat.ipynb

- Contains experiments for the flat neural network classification system

Classification_Deep_Learning_hierarchical.ipynb

- Contains experiments for the hierarchical neural network + traditional classification system

data folder

files_index.csv

- Index file for the Open Icecat
- Product_id is used to construct request URL for the actual product data

offers_english.json.gz

- WDC Training Dataset for Large-scale Product Matching
- Basis for constructing WDC dataset (a subset of this full set)

matched_WDC_data.json

- Matched data between offers_english.json.gz (WDC) and Icecat dataset based on gtin

rdc-catalog-train.tsv

- Raw Rakuten training set

rdc-catalog-gold.tsv

- Raw Rakuten testing set

training and testing data folder

- Contains final datasets used in the experiments
- Contains respective hierarchy objects
- Splits into training, validation, and testing sets are done in the code, but the resulting datasets after the split are contained in the folder “train, val, test splitted” as well

df_all_icecat.pkl

- Open Icecat products

json folder

- Contains Icecat training, validation, and testing set and WDC dataset in json format

icecat_category_counts_in_wdc_dataset.csv

- Contains counts of instances in WDC dataset per category of Icecat training dataset