DBLP Project

Identifying Duplicate Entries in a Bibliography

C. Costea P. Groenen P. Jagroep D. Pop J. van der Schoot

EDA

1 2 3

dblp-X.csv

pbooktitle.json
pbooktitlefull .json

ptype.json

train.csv
validation.csv
test.csv

pjournal.json
pjournalfull .json

Concatenate → Remove useless features → |year| → Swap author - title → Clean & sort author → Key decomposition (4)

ASCII → Remove whitespaces → Lowercase → Remove punctuation → Abbreviate → Group duplicates

ASCII → Group duplicates

(1) Years difference → Overlap → Same features file1 <-> file2 → (2) One-hot-encoding → Normalization → Modeling GBDT

Experiment 1: OpenRefine Clustering

Experiment 2: journal data

Experiment 3: Semantic Scholar API

Error Handling

1. NaN: <10 % -> fill in
   >10 % -> raise error
2. Missing features
3. Types
4. Outliers

Local Results