BIG DATA PROJECT - DBLP

UNIVERSITEIT VAN AMSTERDAM

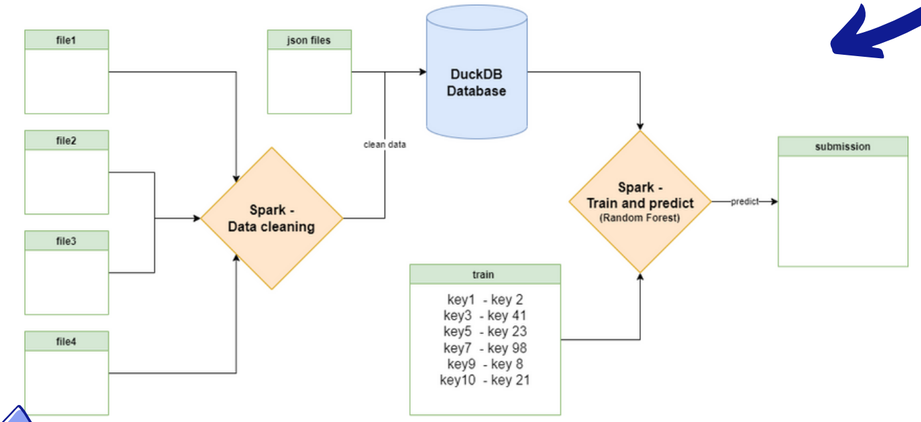
AUTHORS
Anna Doura, Katrin Grunert,
Jonathan A. Harris, Oscar Palafox, Lucas Prieto

Data Exploration

- Author and title were switched
- Negative years
- Duplicates in different languages
- Non-ascii characters in the json files

python™ OpenRefine

Our pipeline



Data Cleaning

- Author Title switch -> titles end with ".", most authors contain "|"
- Negative years -> **absolute values**
- Non-ascii characters encode as ascii characters Ä -> A
- Detect missing pkeys with OpenRefine and manually fix them
- **Capitalize first letter of every word** in the title to preserve starting letters when using letter-wise Jaccard

Our approach

- We focused on the **scalability** and **efficiency** of our method

DuckDB: OLAP DB read-only storage
PySpark: Parallelizable

- We use **letter-wise jaccard similarity** to avoid the need for translation

PySpark DuckDB

Results

79.58%
accuracy

A simple and efficient approach without translation and **reducing SQL storage costs by 25%**

Our insights

- Compare **sets** of **unique letters** instead of full text.
- **Techincal words** have **simlar letters** in different laguages: "Integrate", "Integrar", "Intégrer"
- This avoids translation, lemmatization, tokenization

Discussion

- Cleaning and prediction transformations are independent tasks that are easy to parallelize
- With Spark, our pipeline will be able to handle larger amounts of data