UFMG

DCC
DEPARTEMANETO DE
CIÊNCIA DA COMPUTAÇÃO

CS+X

CNPq

PPORTAL:

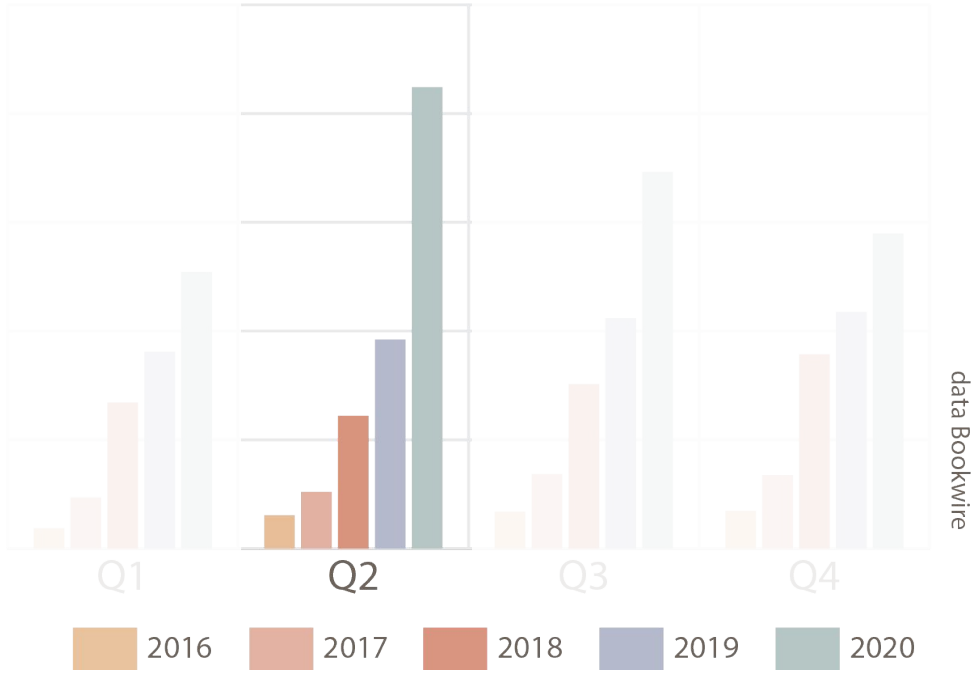
Public Domain Portuguese-language Literature Dataset

Mariana O. Silva mariana.santos@dcc.ufmg.br
Clarisse Scofield clarissescofield@dcc.ufmg.br
Mirella M. Moro mirella@dcc.ufmg.br

Book Industry:
Digital Transformation

#BRAZIL

Brazil Quarterly ebook download unit sales, 2016 to 2020



Amazon purchase of Goodreads stuns book industry

Alarm from Authors' Guild, and many Goodreads users, over 'shocking vertical integration' but at least one writer declares move 'cool'

Alison Flood

Tue 2 Apr 2013 14.15 BST

▲ Goodreads news, bad news? The Goodreads website

Americanas compra plataforma digital literária Skoob

"A Skoob soma conteúdo e conhecimento do universo literário para as nossas marcas, principalmente Americanas e Submarino", diz Marcio Cruz, CEO da plataforma digital da Americanas S.A..

Por **Marina Filippe**

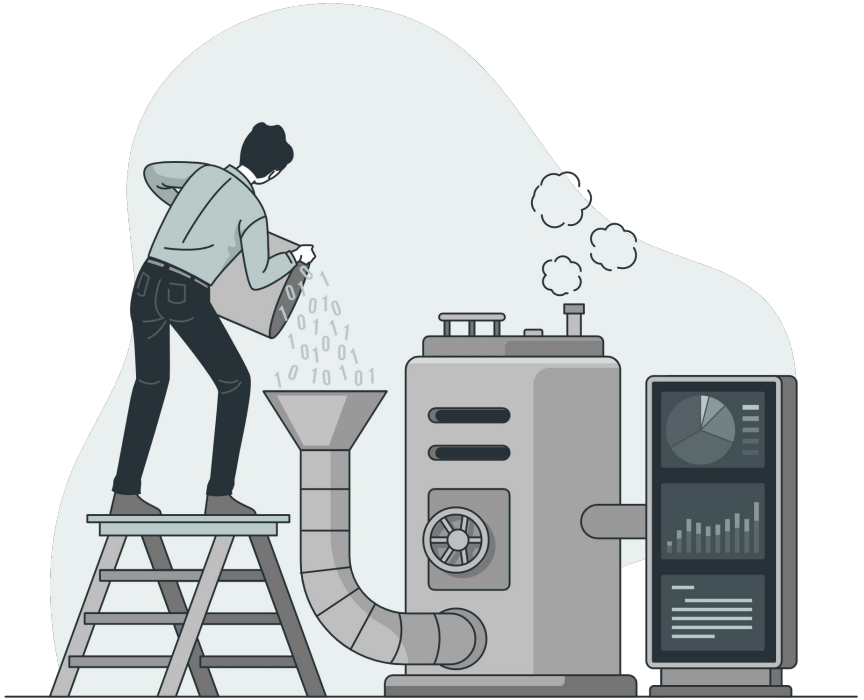
Publicado em: 16/09/2021 às 10h23
Alterado em: 16/09/2021 às 10h43

Tempo de leitura: **3 min**

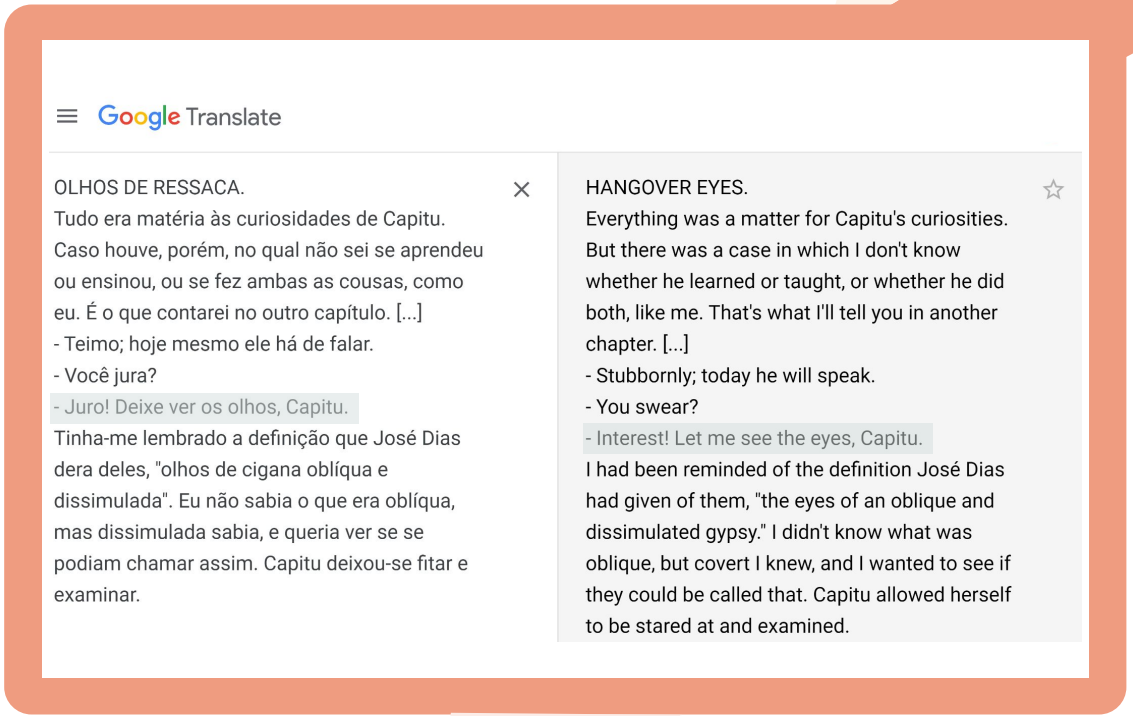
Solution → Data

Problem → Data

- Data extraction and preprocessing is **time-consuming**
- Current scenario → **few open enriched datasets**
  - Focuses on one dimension of the problem
  - Limited to English-written books



Portuguese in NLP



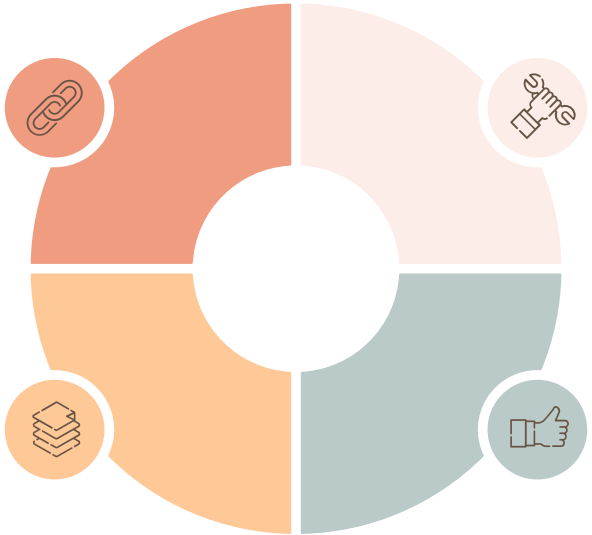
PPORTAL: a Public domain PORTuguese lAnguage Literature dataset

Data Integration

numerous public domain works from three digital libraries

Enriched metadata

for works, authors and online reviews extracted from Goodreads



Feature engineering

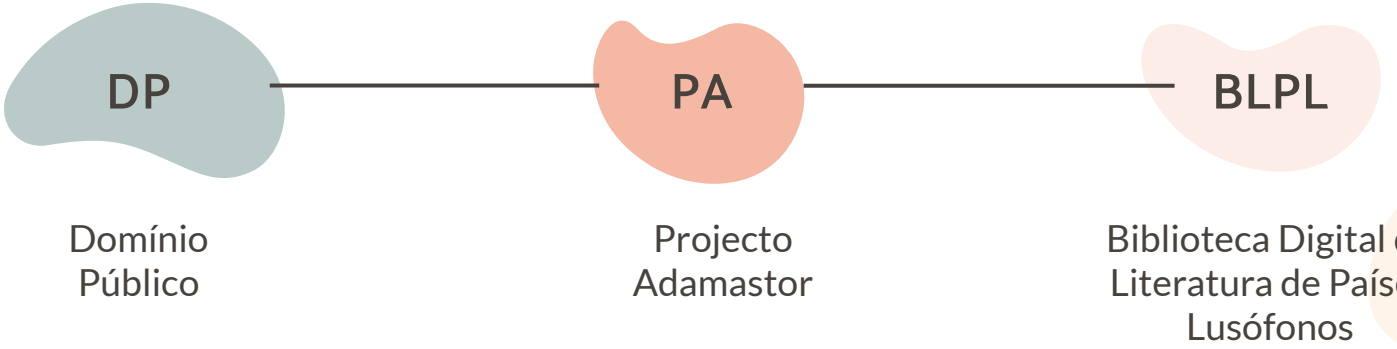
on the metadata to create meaningful additional features

Open access

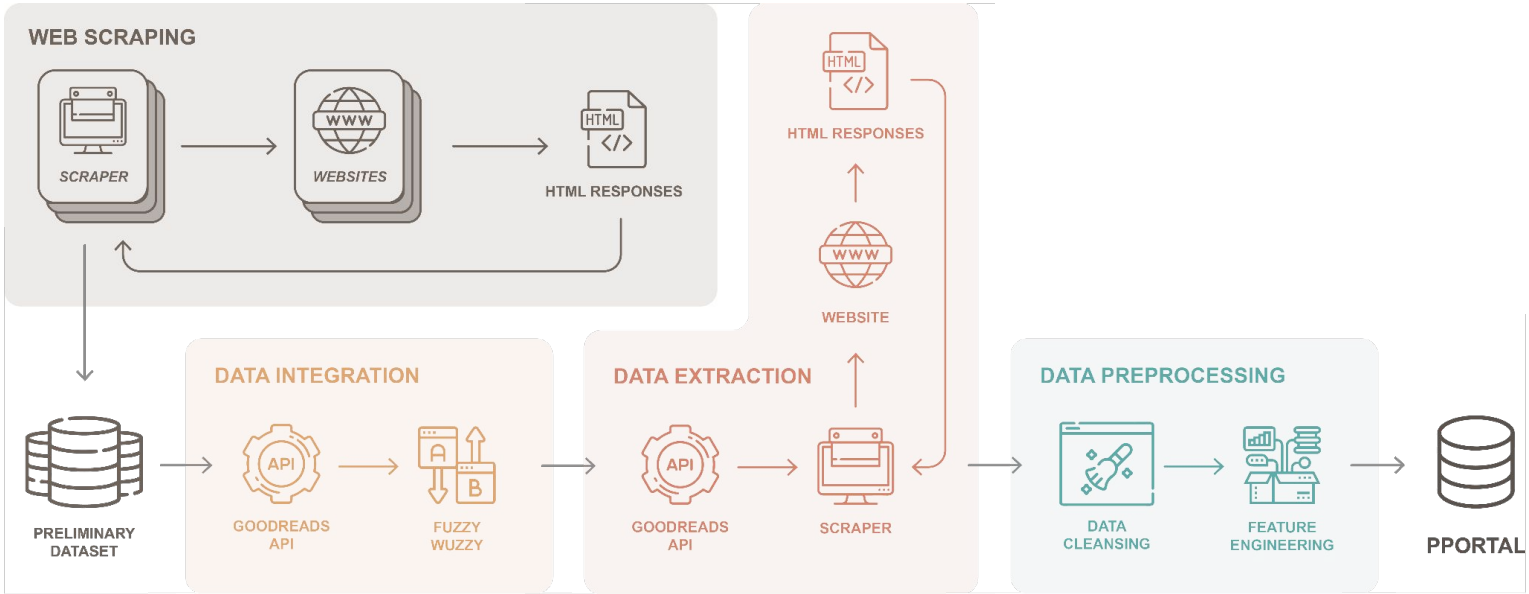
two formats

(Dump SQL and compressed .csv files)

Data Sources



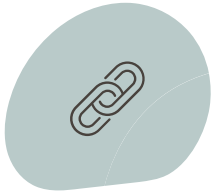
Dataset Building Process



PPORTAL Statistics

80

Literary genres

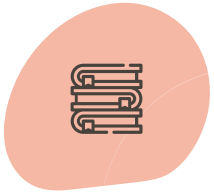


2,388

Integrated Records (Goodreads)

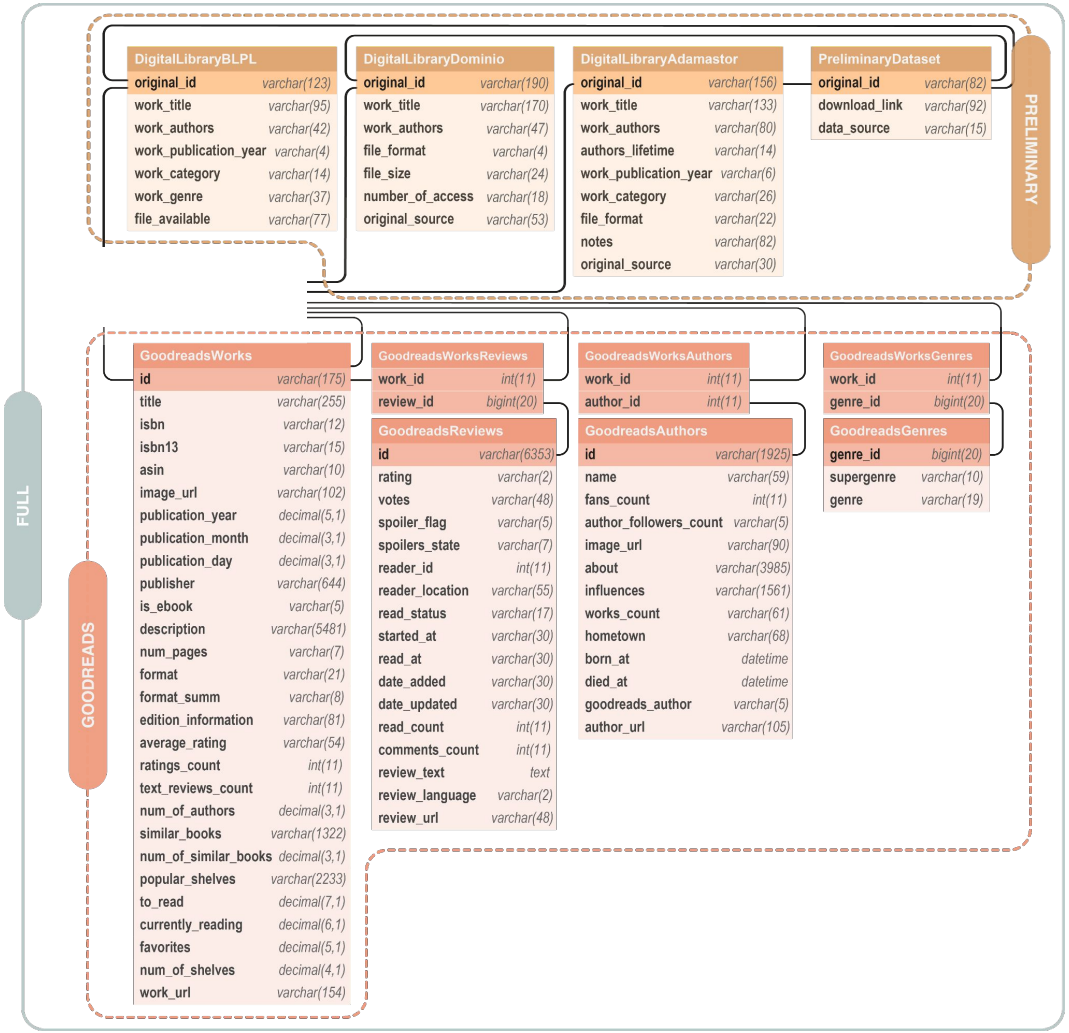
9,585

Records with downloads



82,313

Public domain works



Data Content

Applicability

- Text Classification
- Named Entity Recognition (NER)
- Sentiment Analysis

- Recommender System
- Success Prediction
- Social Network Analysis (SNA)

Challenges and Limitations

- **Data Integration** — Keeping them only in the preliminary set or manually search on Goodreads
- **Data Quality** — Consider an additional data source to try imputing the incomplete content
- **Distinct Genres** — Consider fuzzy matching approaches to find similar genres

Concluding Remarks

- Manually data collection is **time-consuming**
- Solution: **PPORTAL**
- Integrate and centralize public domain works
- Download links and valuable metadata
- Elements of the book industry ecosystem:
  - Works
  - Authors
  - Readers
  - Reviews

Future Directions

01 Additional data sources

Handling missing data

02 Fuzzy Matching

Mitigating distinct genres issue

03 Data growth

Update-oriented collecting phase

04 Expected projection

- Extracting text files
- Integrating BraCID

DOI 10.5281/zenodo.4890048

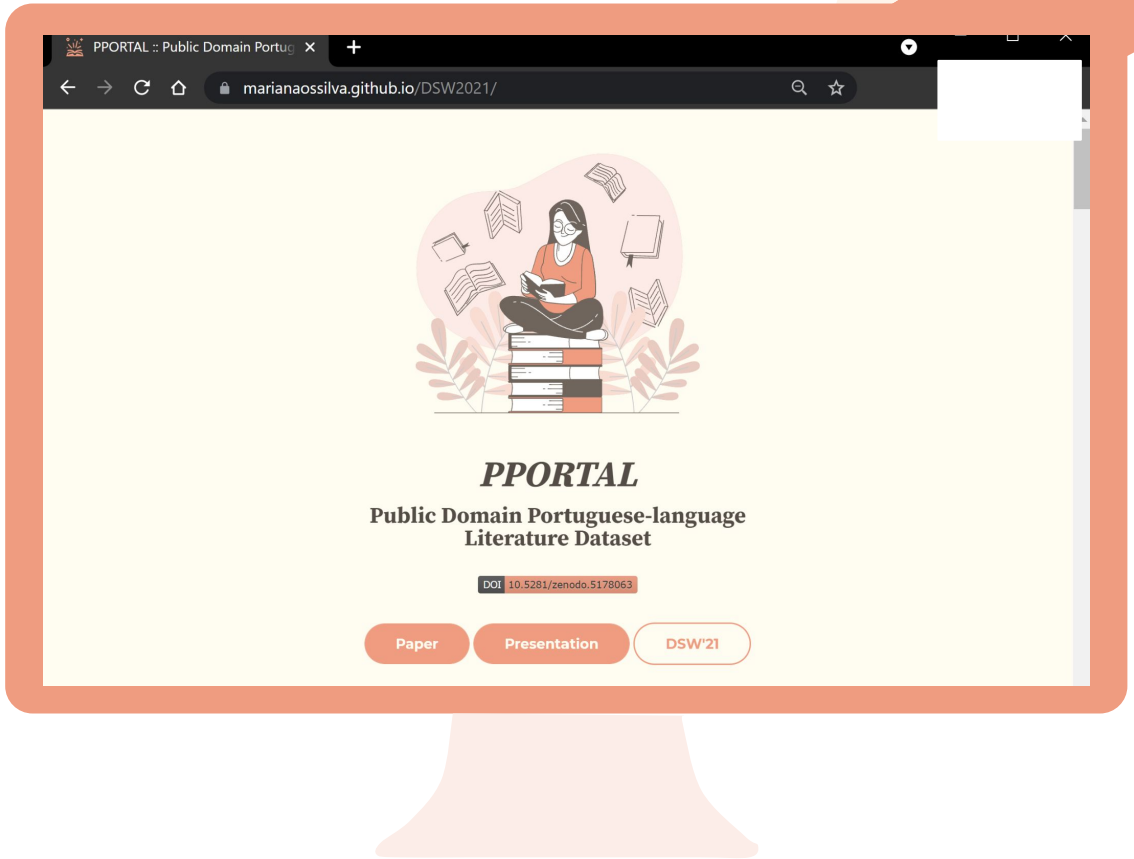
Format and Usage

DOI 10.5281/zenodo.5178063

Publicly available in an open-access Zenodo repository

Downloaded from its project webpage

- Dump SQL
- CSV files



SCAN ME

Thank you!

Any questions?

mariana.santos@dcc.ufmg.br

Download PPORTAL
Relational schema
SQL dump
CSV files

DOI: 10.5281/zenodo.5178063

UFMG

DCC
DEPARTEMANETO DE
CIÊNCIA DA COMPUTAÇÃO

CS+X

CNPq