

PPORTAL:

Public Domain
Portuguese-language
Literature Dataset

Mariana O. Silva
Clarisse Scofield
Mirella M. Moro

mariana.santos@dcc.ufmg.br
clarissescofield@dcc.ufmg.br
mirella@dcc.ufmg.br

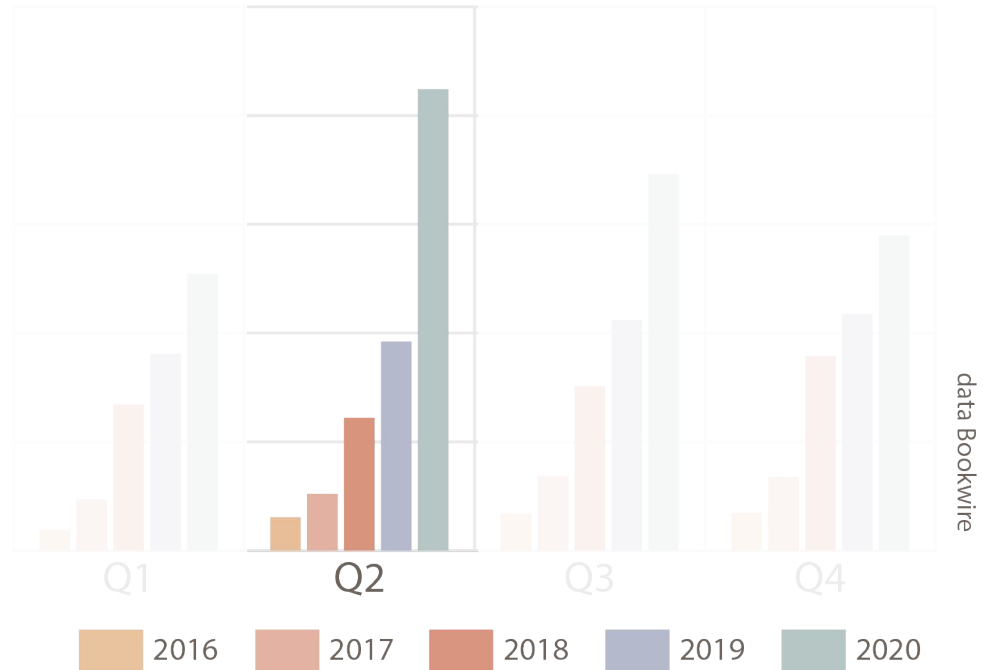


Book Industry: Digital Transformation



#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



91



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

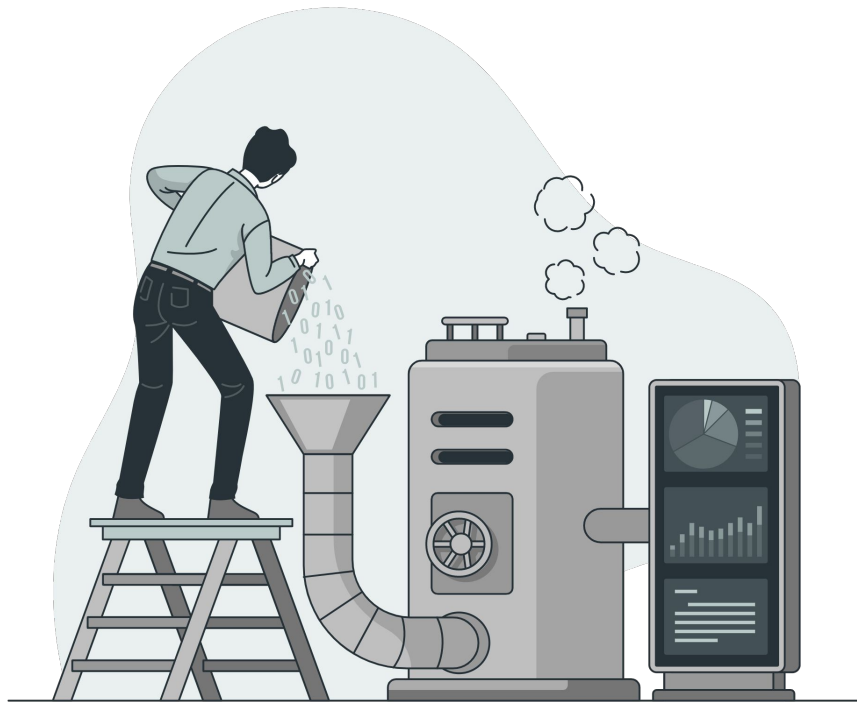
🕒 Tempo de leitura: 3 min



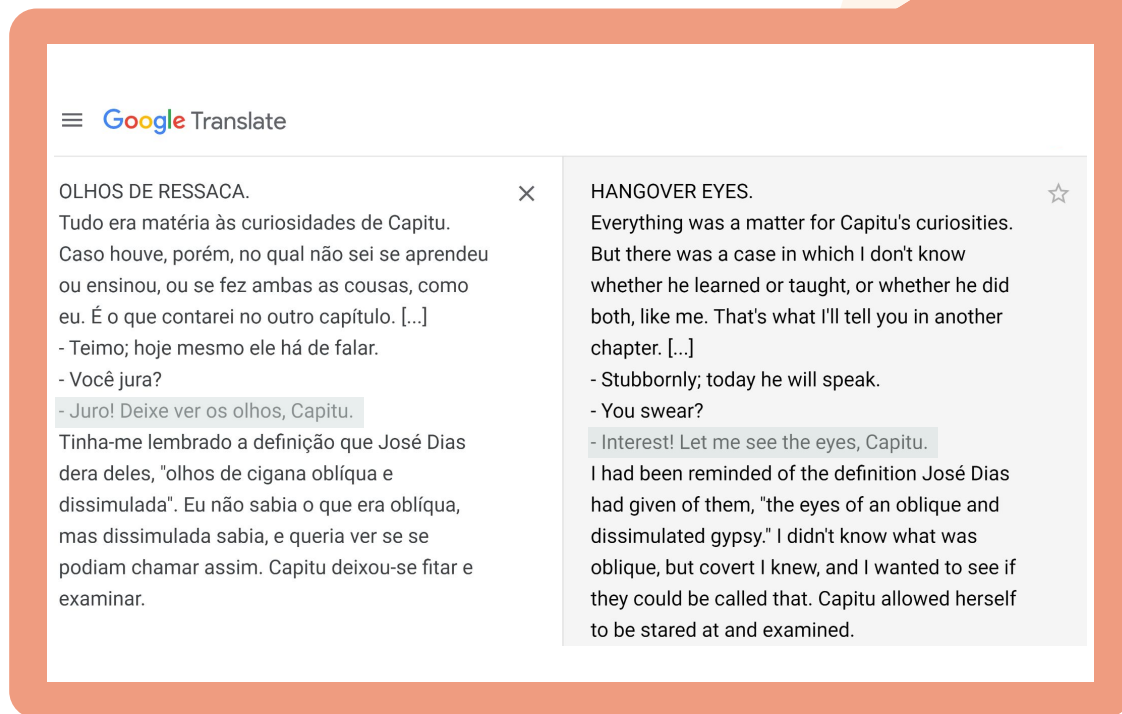
Solution → Data

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



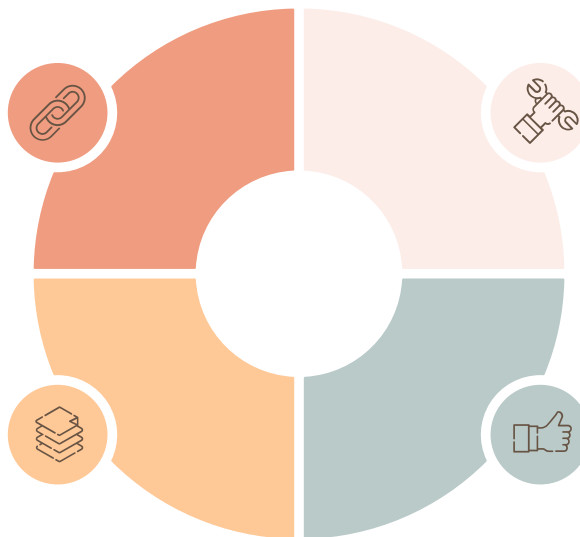
PPORTAL: a Public domain PORTuguese lAnguage Literature dataset

Data Integration

numerous public domain works from three digital libraries

Enriched metadata

for works, authors and online reviews extracted from Goodreads



Feature engineering

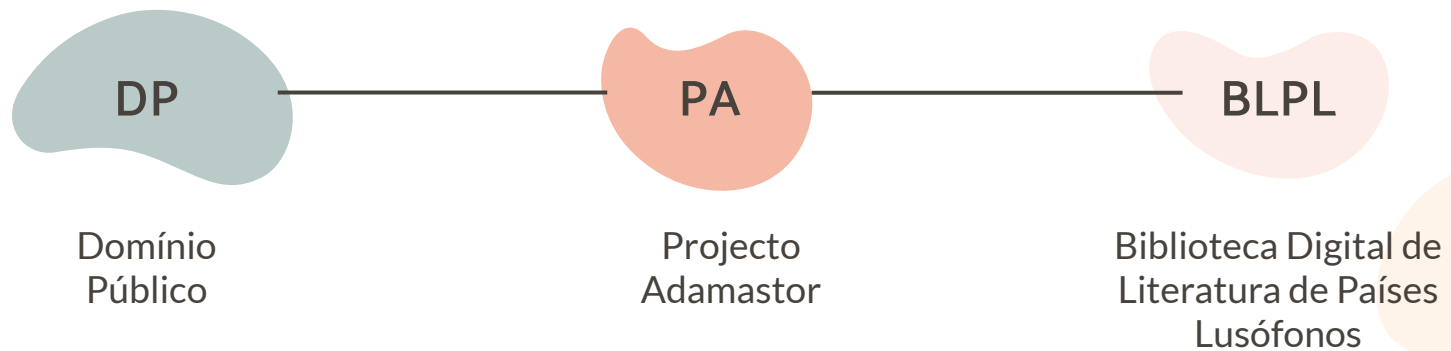
on the metadata to create meaningful additional features

Open access

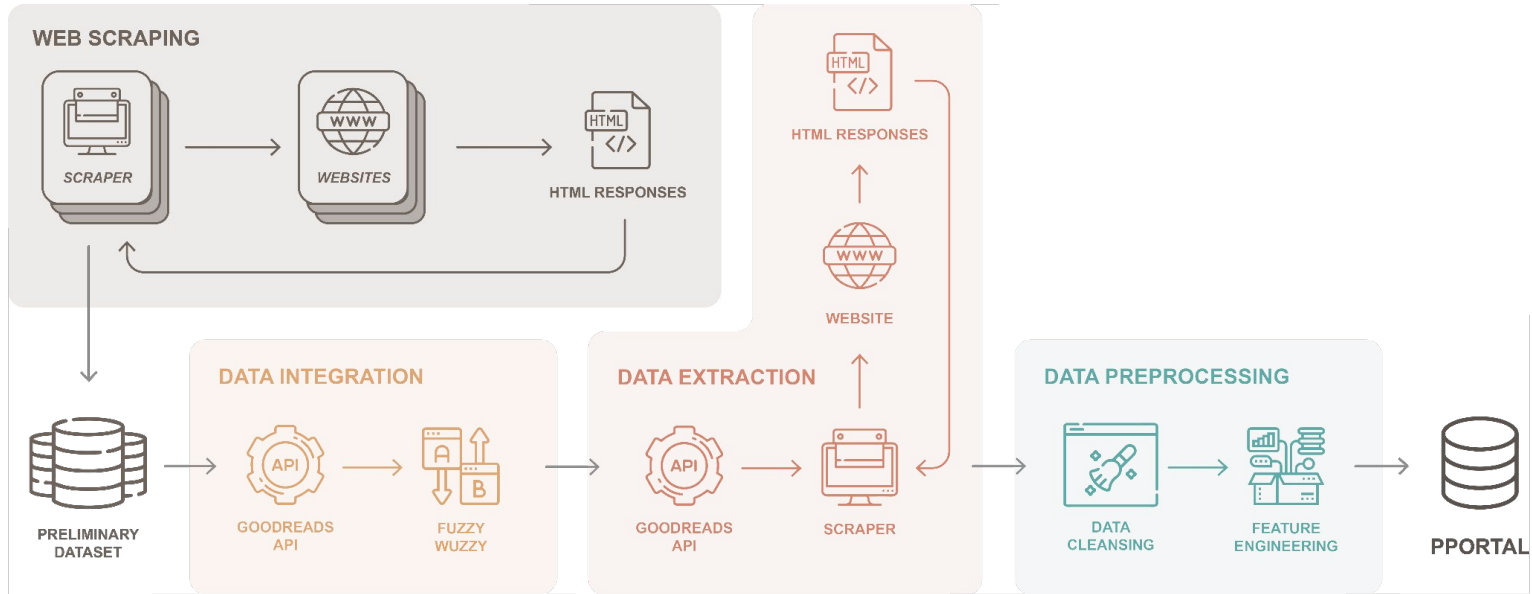
two formats

(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

DigitalLibraryBLPL

original_id	varchar(123)
work_title	varchar(95)
work_authors	varchar(42)
work_publication_year	varchar(4)
work_category	varchar(14)
work_genre	varchar(37)
file_available	varchar(77)

DigitalLibraryDominio

original_id	varchar(190)
work_title	varchar(170)
work_authors	varchar(47)
file_format	varchar(4)
file_size	varchar(24)
number_of_access	varchar(18)
original_source	varchar(53)

DigitalLibraryAdamastor

original_id	varchar(156)
work_title	varchar(133)
work_authors	varchar(80)
authors_lifetime	varchar(14)
work_publication_year	varchar(6)
work_category	varchar(26)
file_format	varchar(22)
notes	varchar(82)
original_source	varchar(30)

PreliminaryDataset

original_id	varchar(62)
download_link	varchar(92)
data_source	varchar(15)

GoodreadsWorks

id	varchar(175)
title	varchar(255)
isbn	varchar(12)
isbn13	varchar(15)
asin	varchar(10)
image_url	varchar(102)
publication_year	decimal(5,1)
publication_month	decimal(3,1)
publication_day	decimal(3,1)
publisher	varchar(644)
is_ebook	varchar(5)
description	varchar(5481)
num_pages	varchar(7)
format	varchar(21)
format_summ	varchar(8)
edition_information	varchar(81)
average_rating	varchar(54)
ratings_count	int(11)
text_reviews_count	int(11)
num_of_authors	decimal(3,1)
similar_books	varchar(1322)
num_of_similar_books	decimal(3,1)
popular_shelves	varchar(2233)
to_read	decimal(7,1)
currently_reading	decimal(6,1)
favorites	decimal(5,1)
num_of_shelves	decimal(4,1)
work_url	varchar(154)

GoodreadsWorksReviews

work_id	int(11)
review_id	bigint(20)

GoodreadsReviews

id	varchar(6353)
rating	varchar(2)
votes	varchar(48)
spoiler_flag	varchar(5)
spoilers_state	varchar(7)
reader_id	int(11)
reader_location	varchar(55)
read_status	varchar(17)
started_at	varchar(30)
read_at	varchar(30)
date_added	varchar(30)
date_updated	varchar(30)
read_count	int(11)
comments_count	int(11)
review_text	text
review_language	varchar(2)
review_url	varchar(48)

GoodreadsWorksAuthors

work_id	int(11)
author_id	int(11)

GoodreadsAuthors

id	varchar(1925)
name	varchar(59)
fans_count	int(11)
author_followers_count	varchar(5)
image_url	varchar(90)
about	varchar(3985)
influences	varchar(1561)
works_count	varchar(61)
hometown	varchar(68)
born_at	datetime
died_at	datetime
goodreads_author	varchar(5)
author_url	varchar(105)

GoodreadsWorksGenres

work_id	int(11)
genre_id	bigint(20)

GoodreadsGenres

genre_id	bigint(20)
supergenre	varchar(10)
genre	varchar(19)

Data Content

Applicability

Text
Classification



Named Entity
Recognition
(NER)



Sentiment
Analysis



Recommender
System



Success
Prediction

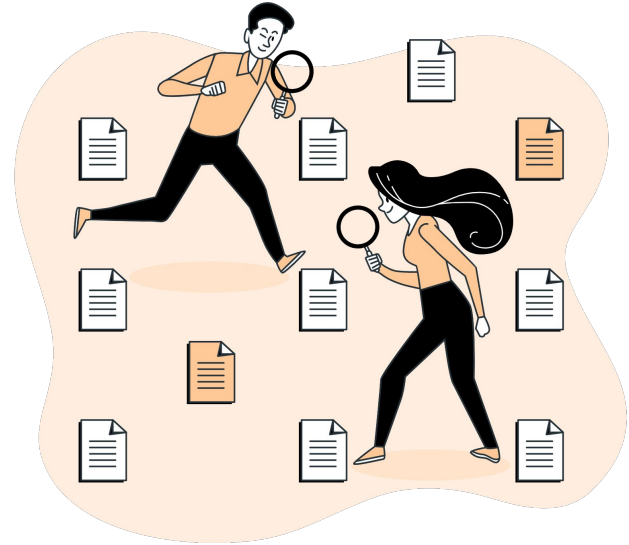


Social Network
Analysis (SNA)



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

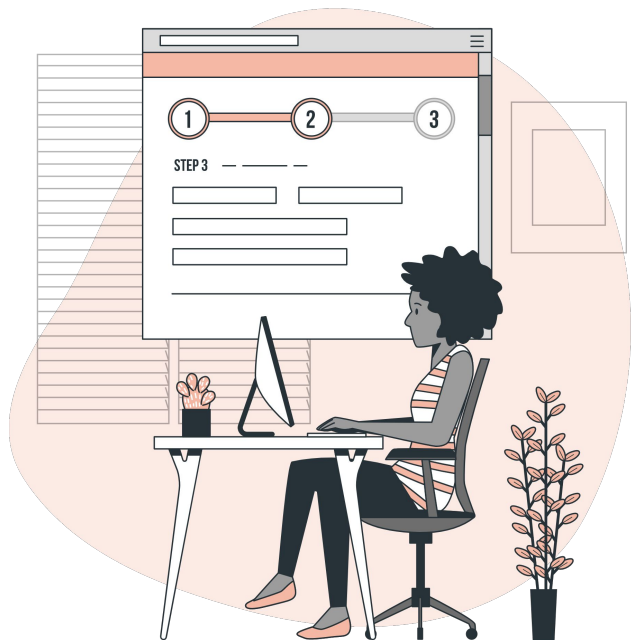
Future Directions

01 Additional data sources

Handling missing data

03 Data growth

Update-oriented collecting phase



02 Fuzzy Matching

Mitigating distinct genres issue

04 Expected projection

- Extracting text files
- Integrating BraCID

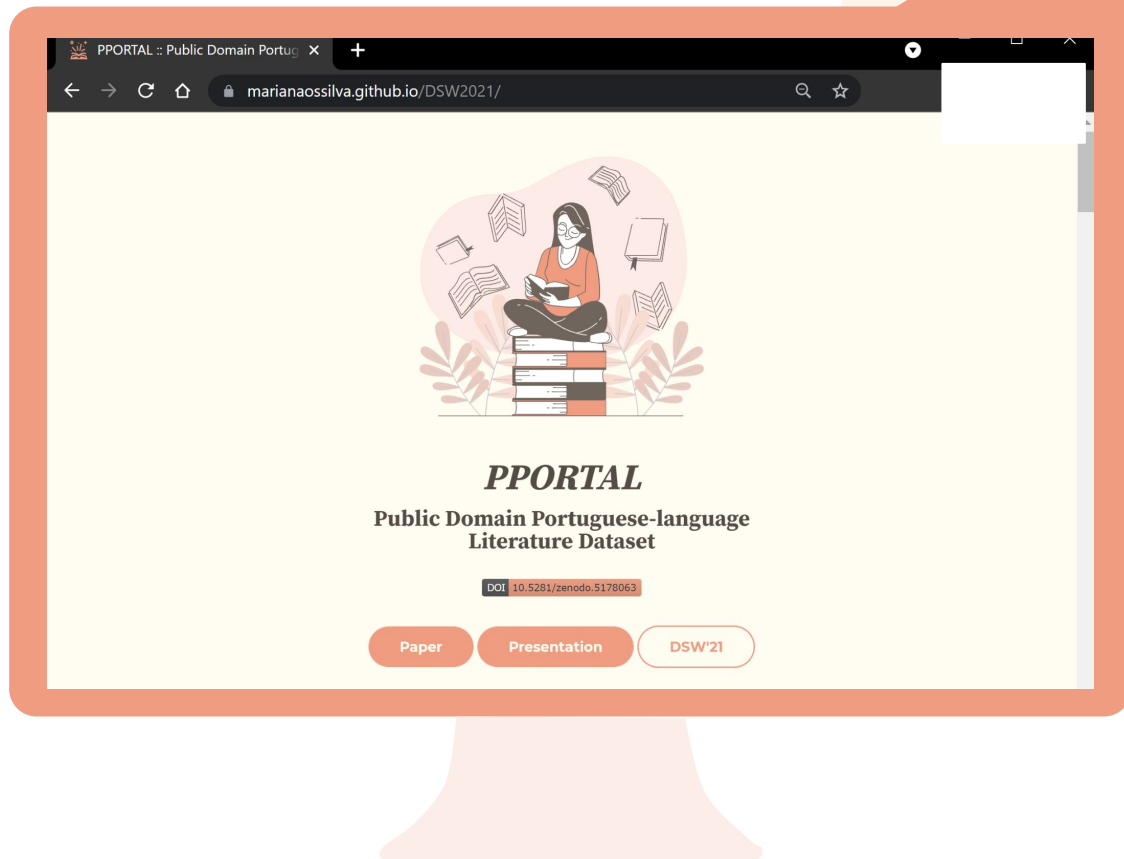
Format and Usage

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

UF m G

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

CS *x*

 **CNPq**