U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

NEWS ARTICLES

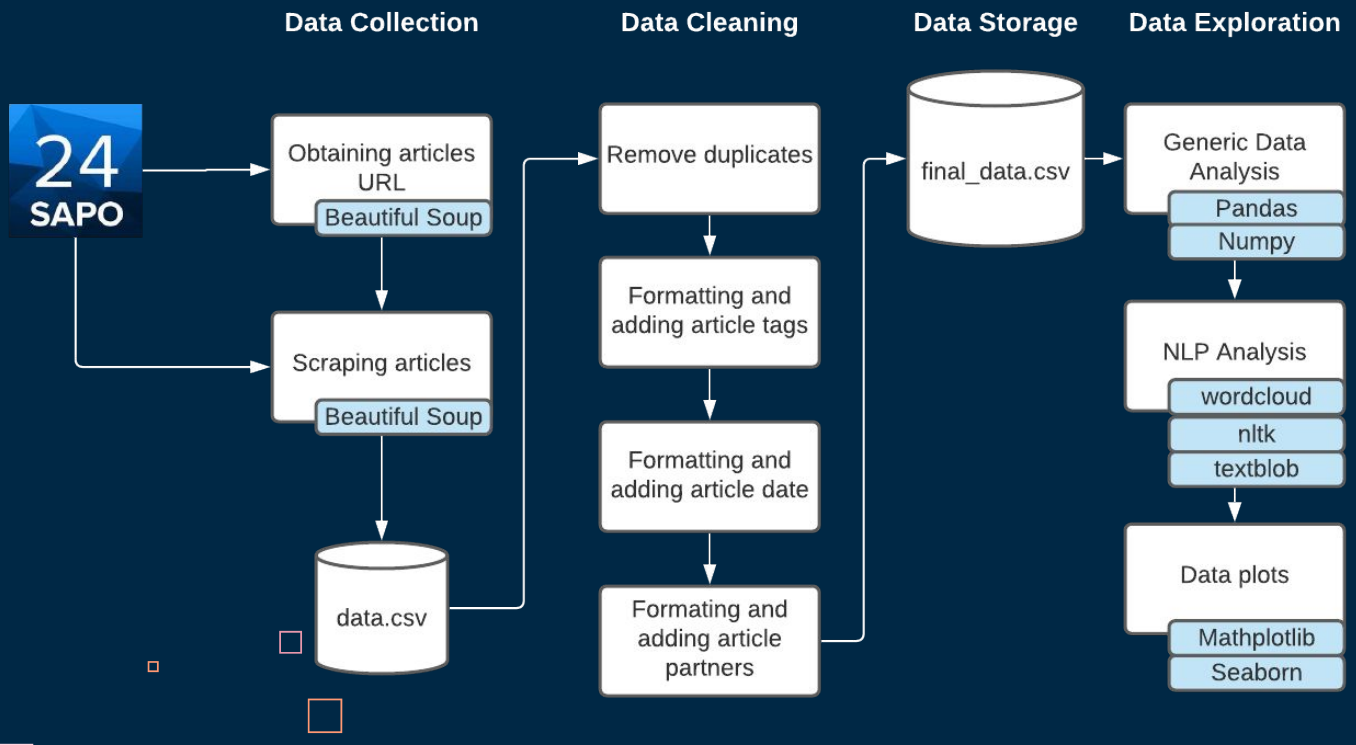
Information Processing and Retrieval

Milestone #1 | Data Preparation

Master in Informatics and Computing Engineering

Pedro Galvão up201700488
Rodrigo Reis up201806534
Tiago Alves up201603820

Data Preparation Pipeline



Data Collection

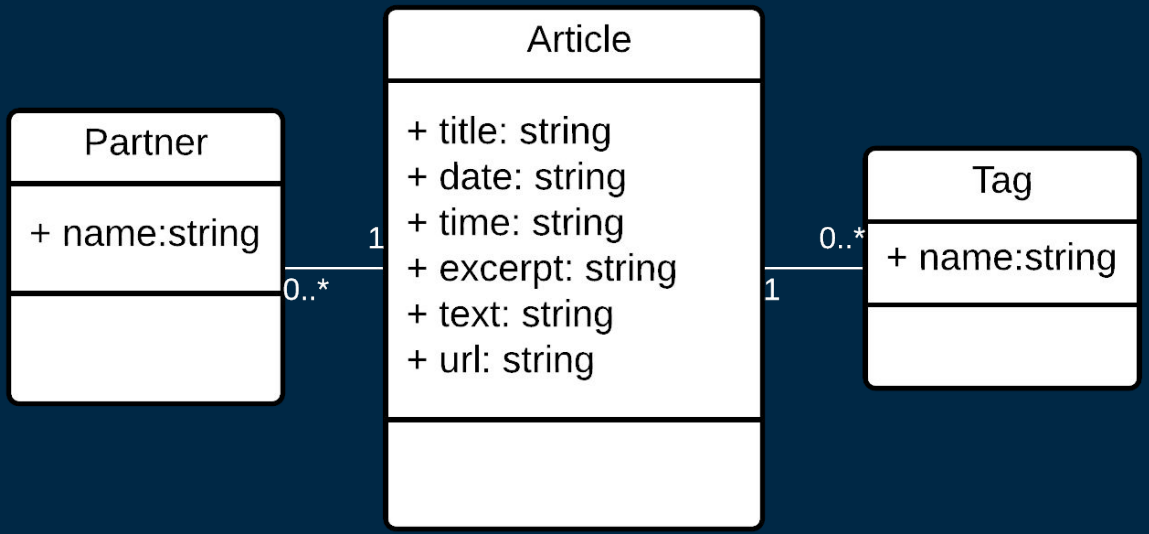
Scrapping page

GET REQUEST
www.sapo24.pt/atualidade?página=<page-number>

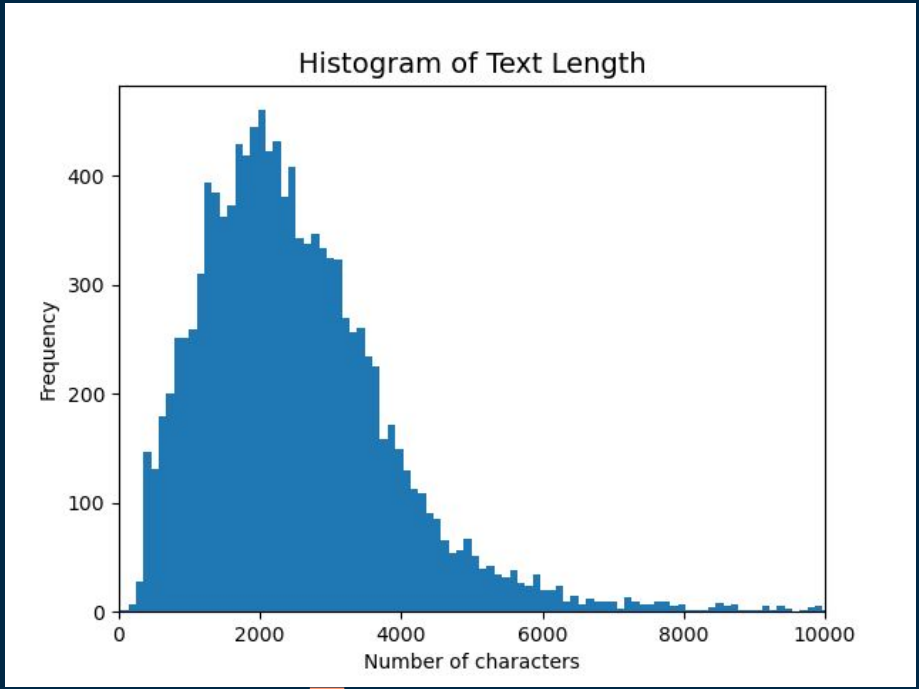
Scrapping article by article

GET REQUEST
www.sapo24.pt/atualidade/<article-title>

Data Model

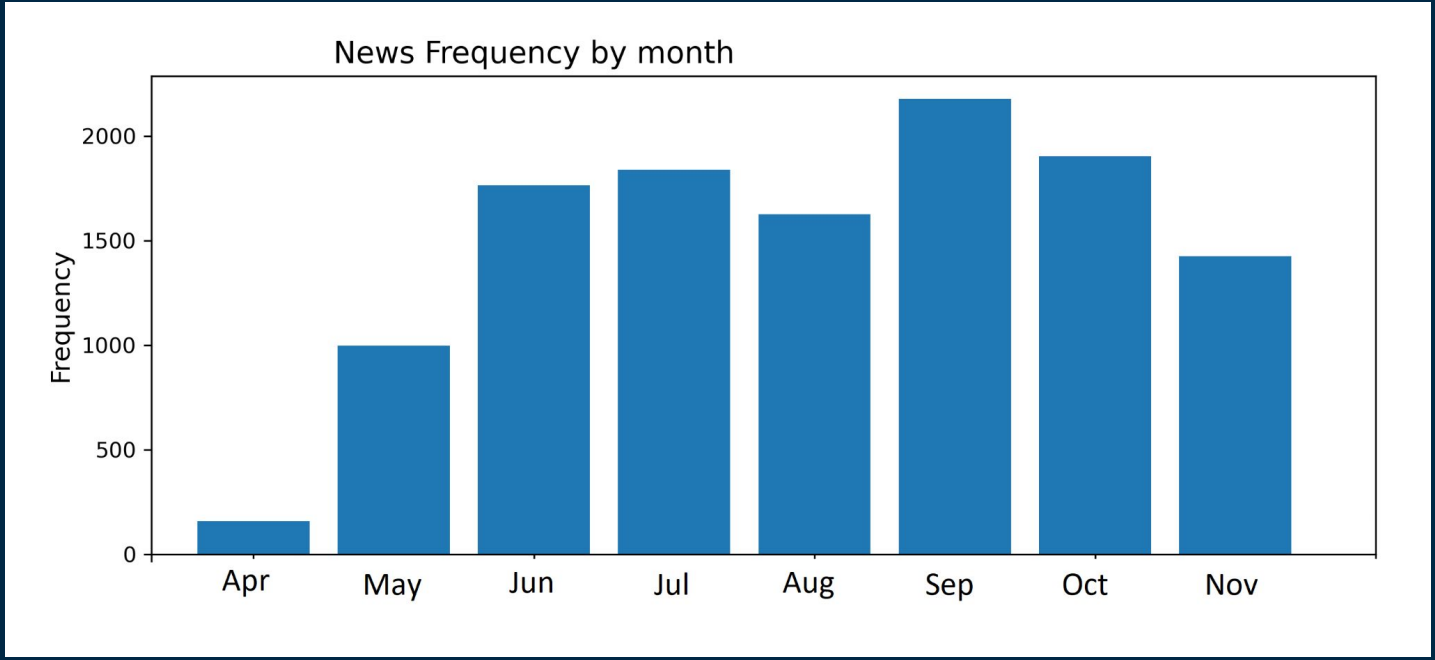

Partner
+ name:string
1
0..*
Article
+ title: string
+ date: string
+ time: string
+ excerpt: string
+ text: string
+ url: string
0..*
1
Tag
+ name:string

Data Characterization Articles Text Length

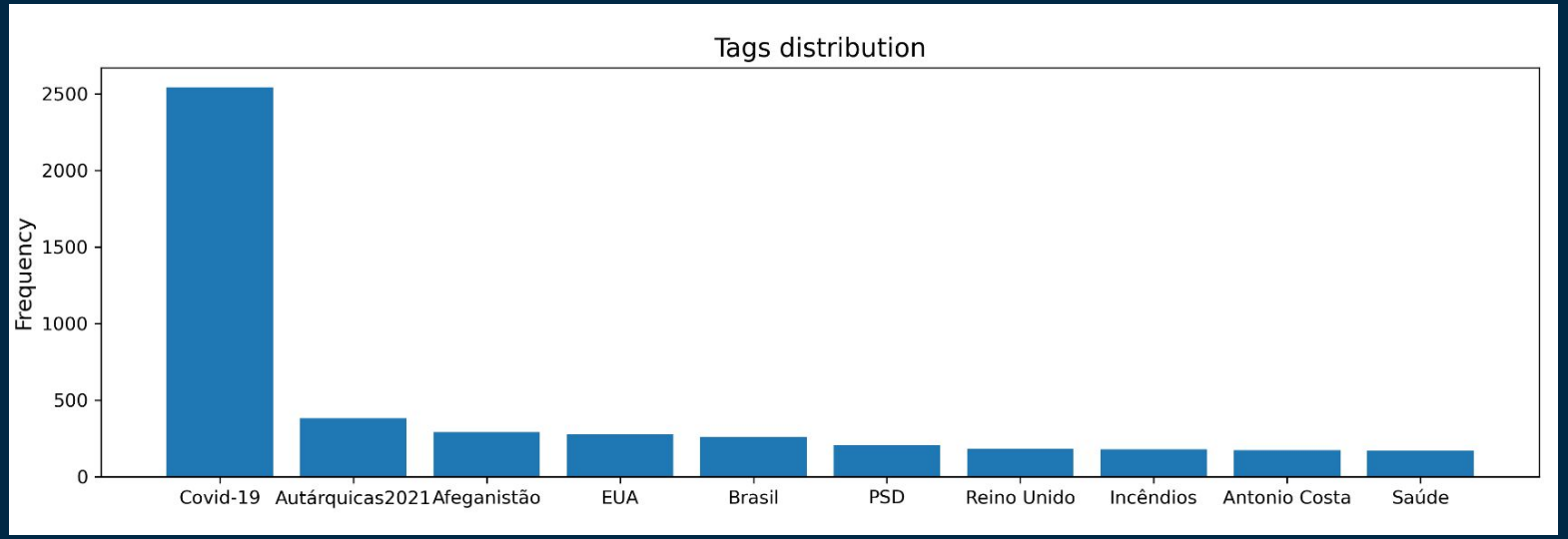


The text length of the article can be approximated to a normal distribution.

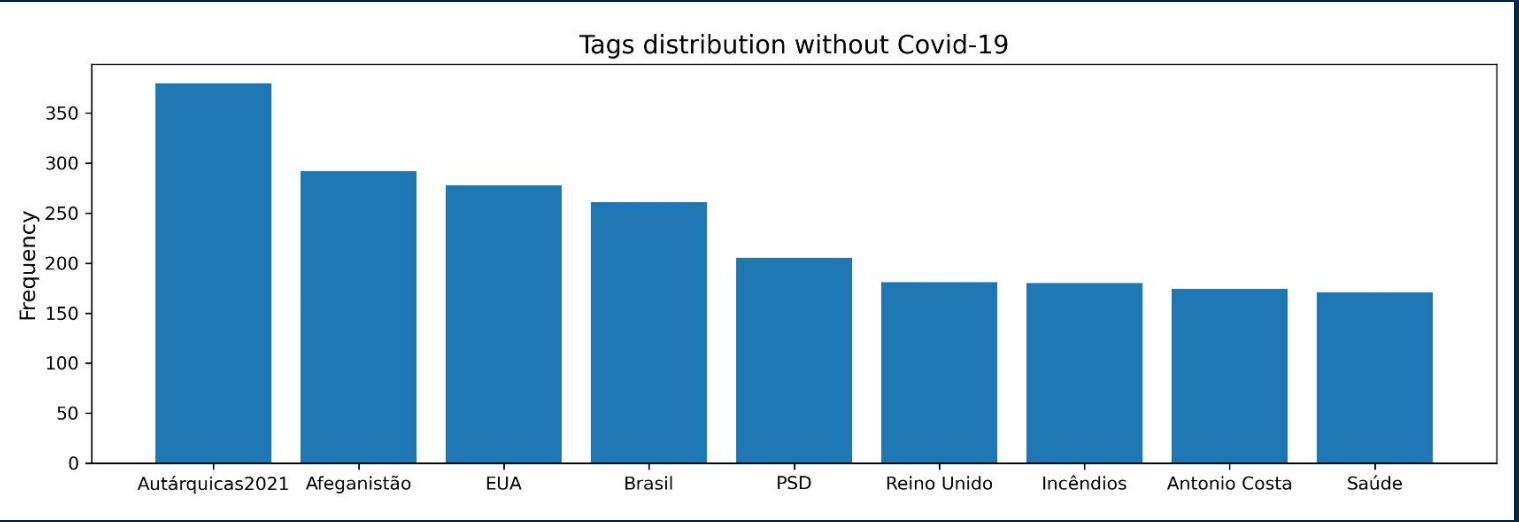
Data Characterization News by month



Data Characterization Most used tags



Data Characterization Most used tags without Covid-19



Data Characterization Most used words

Conclusion

✅ We accomplish all the proposed goals for Milestone #1

✅ We are ready for next Milestone

THANK YOU!