

NEWS ARTICLES

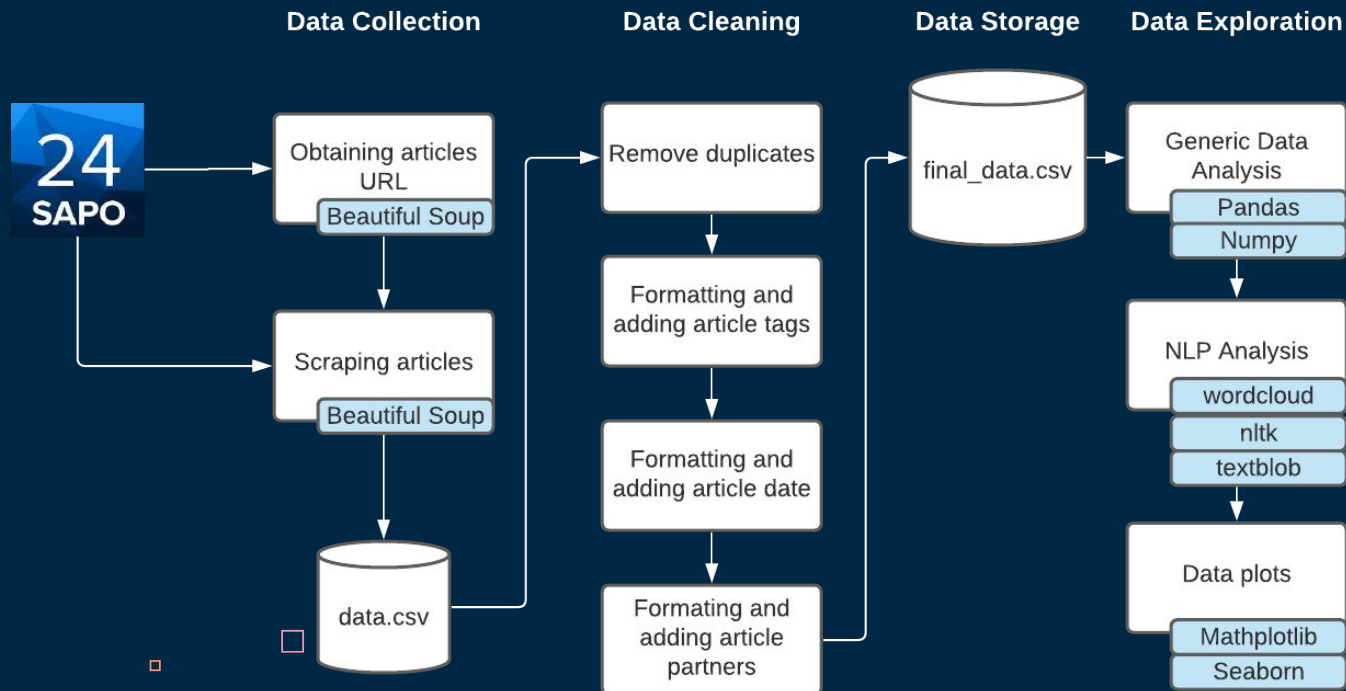
Information Processing and Retrieval

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?pagina=<page-number>

Scrapping article by article

GET REQUEST

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

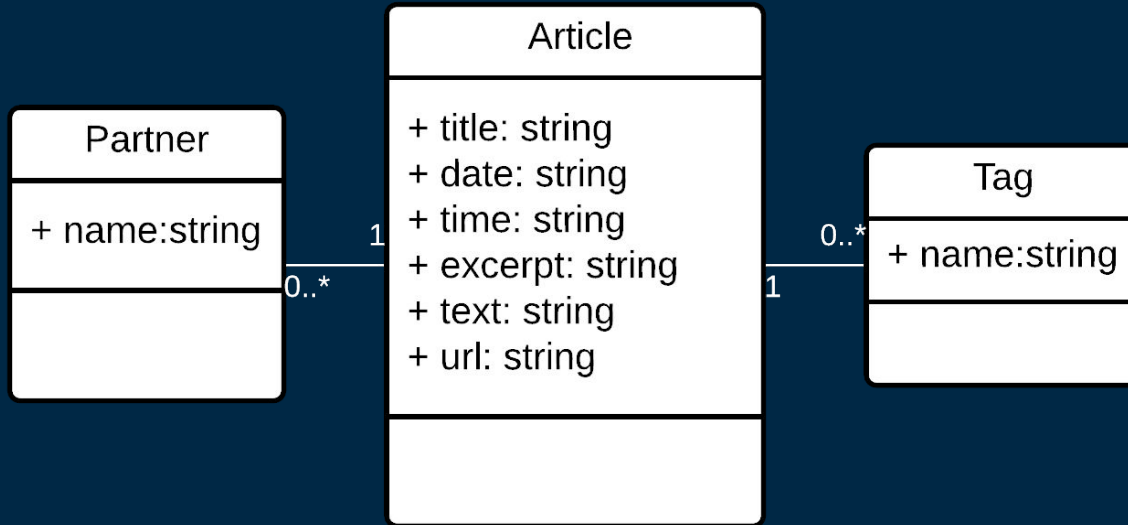


The screenshot shows the SAPO24 website interface. At the top is a navigation bar with the SAPO logo and links for MAIL, JORNAIS, CARROS, CASAS, EMPREGO, BLOGS, PROMOS, WOMANLIFE, and TUDO. Below this is a secondary bar with categories: 24, Atualidade, Economia, Desporto, Vida, Tecnologia, Local, Opinião, Jornais, and Arquivo Lusa. A third bar contains the text: 'Hoje o dia foi assim', 'Acho que vais gostar disto', 'É desta que leio isto', 'Para ouvir no SAPO24', and 'Covid-19. Regras editoriais'.

The main section is titled 'ATUALIDADE'. It features several article thumbnails:

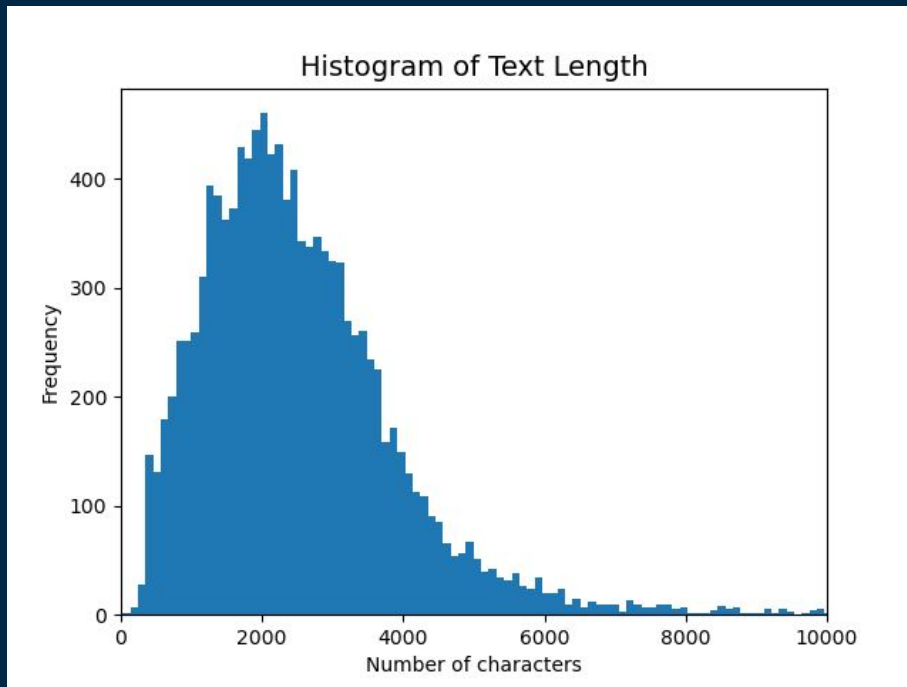
- Top Left:** A photo of a city street with a large building. The headline is 'Porto e Braga entre as 95 cidades mundiais na "lista A" de líderes ambientais'. The date is '18 NOV 2021 06:27'.
- Top Right:** A photo of a man wearing a face mask. The headline is 'Família de homem atropelado por carro de Eduardo Cabrita recebe pensão de sobrevivência de 246 euros por mês'. The date is '17 NOV 2021 19:34'.
- Bottom Left:** A photo of two men in suits. The headline is 'Dois homens condenados pelo assassinio de Malcolm X vão ser absolvidos'. The date is '18 nov 2021 07:59'. The article text states: 'Dois homens condenados pelo assassinio do ativista norte-americano Malcolm X, um dos maiores símbolos da luta contra o racismo, devem ser absolvidos hoje, quinta-feira, mais de 50 anos depois, anunciou hoje o gabinete do procurador de Manhattan, Cyrus Vance.'
- Bottom Middle:** A thumbnail of a newspaper page. The headline is 'Revista de Imprensa: Faltam professores e Marcelo afasta novo estado de emergência'. The date is '18 nov 2021 07:49'. The article text states: 'Os principais destaques nos jornais e revistas desta quinta-feira, dia 18 de novembro.'
- Bottom Right:** A photo of three recycling bins (green, yellow, and blue). The headline is 'Portugal já reciclou mais 400 mil toneladas de embalagens face a 2020'. The date is '18 nov 2021 07:32'. The article text states: 'Portugal reciclou este ano, até outubro, mais 8% de embalagens do que em igual período do ano passado, num total de 400 mil toneladas, indicam números hoje divulgados pela Sociedade Ponto Verde (SPV).'

Data Model



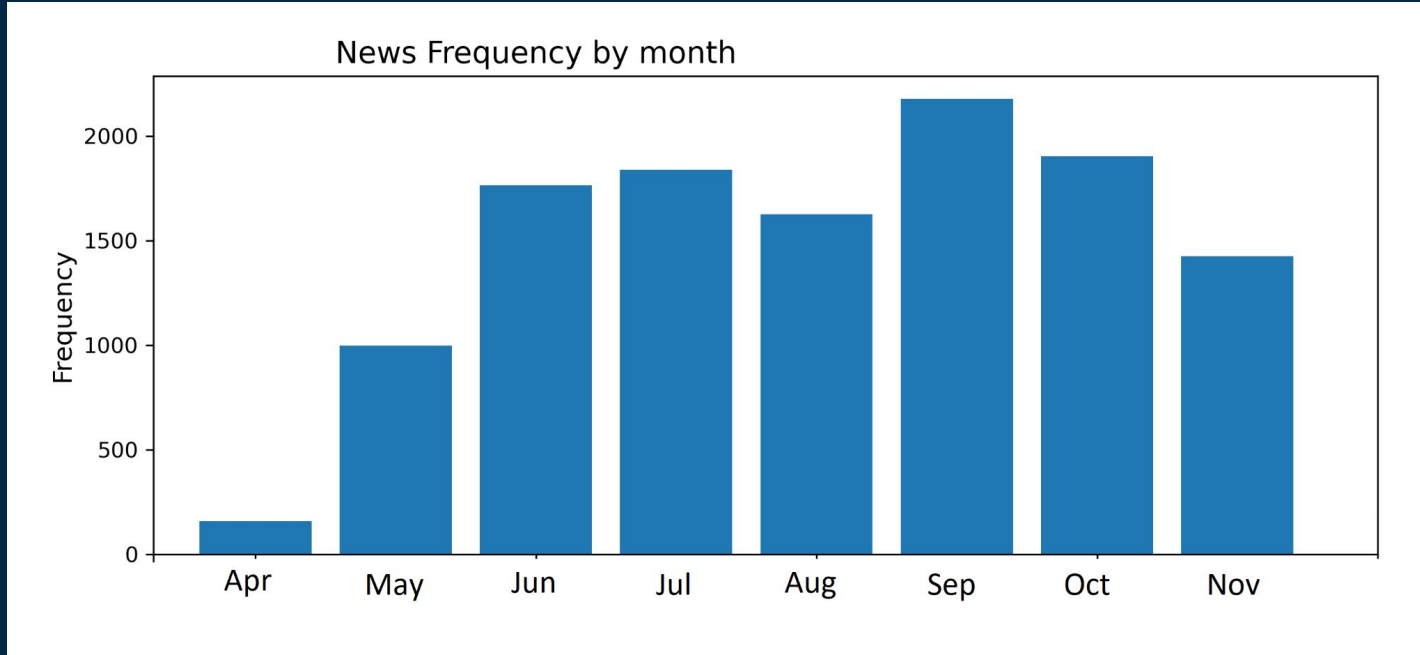
Data Characterization

Articles Text Length



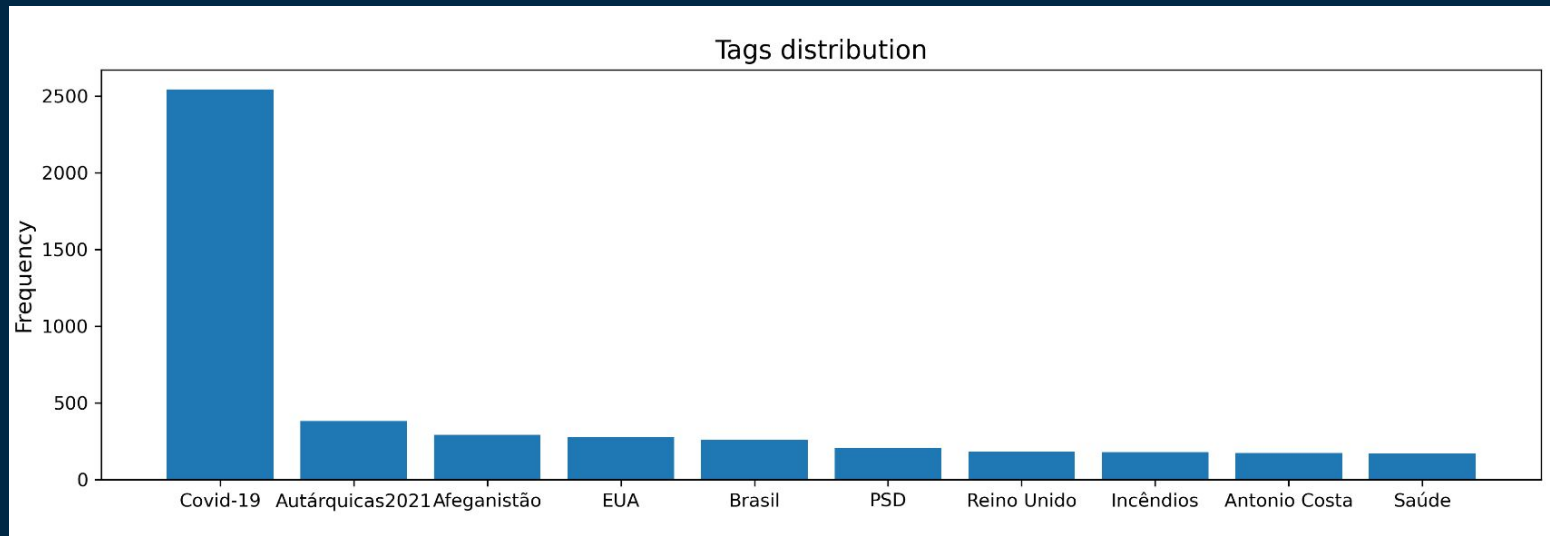
The text length of the article can be approximated to a normal distribution.

Data Characterization News by month



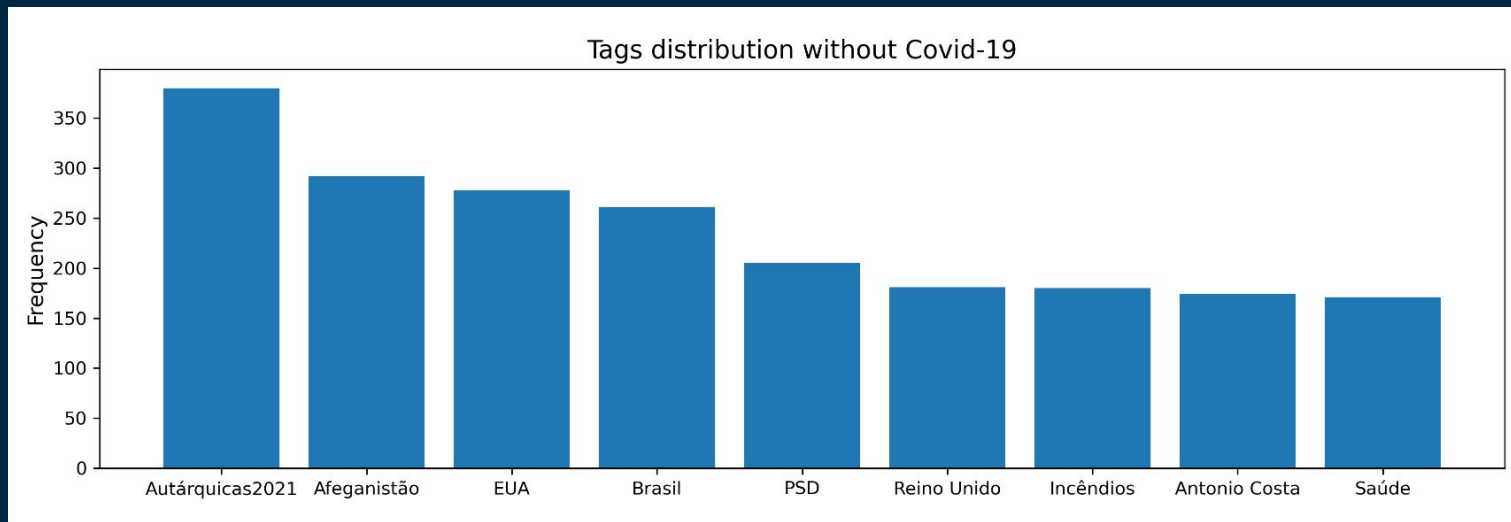
Data Characterization

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!

