

**U.**PORTO

**FEUP** FACULDADE DE ENGENHARIA  
UNIVERSIDADE DO PORTO

# Information Processing and Retrieval

Master in Informatics Engineering and Computation at FEUP, U.Porto

Diogo Fonte  
Rodrigo Figueiredo  
Sofia Rodrigo  
Vitor Cavaleiro





# Dataset choice



ML Resources

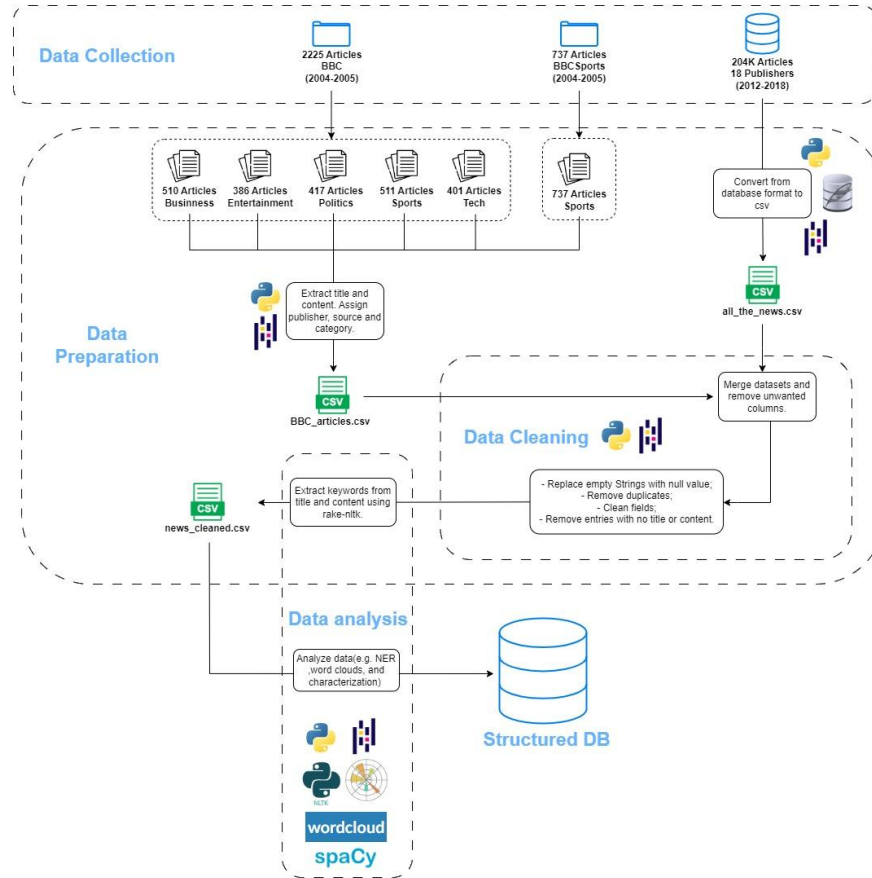


## Dataset content

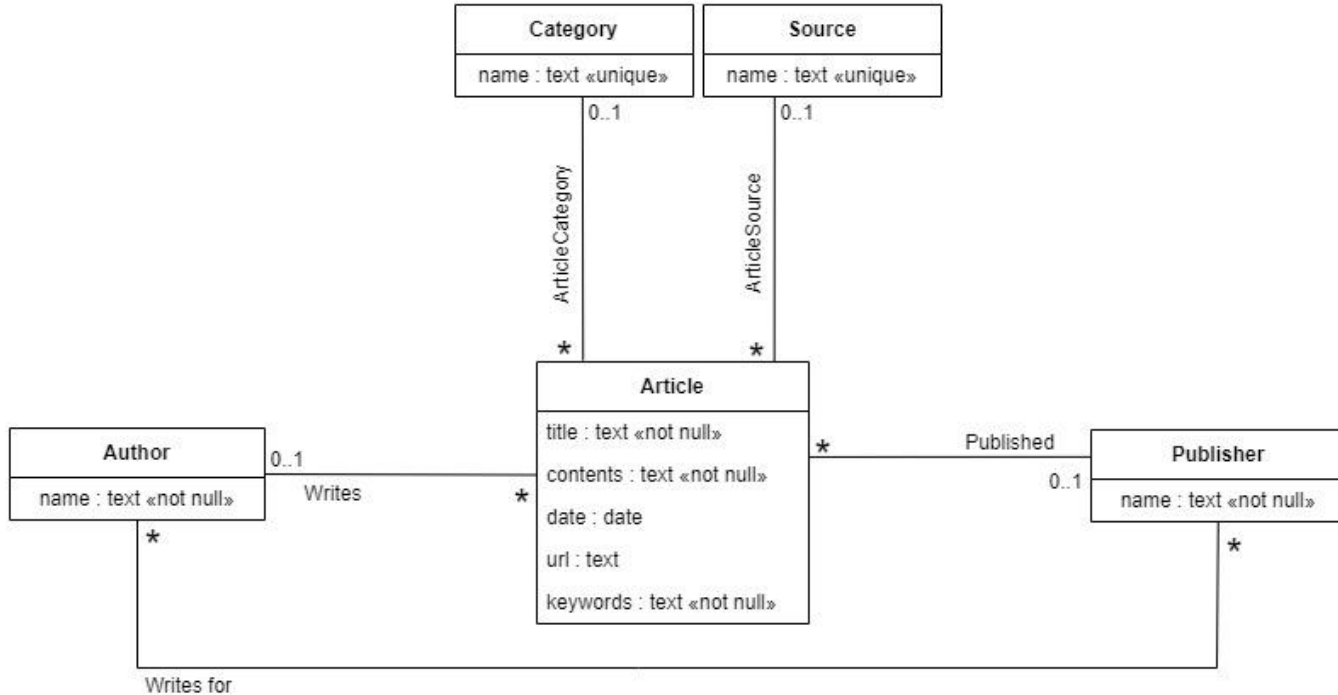


The New York Times

# Data processing pipeline



# Domain data model





# Data sample

	title	author	date	content	publisher	source	category	url	keyphrases
0	Agent Cooper in Twin Peaks is the audience: on...	Tasha Robinson	2017-05-31	And never more so than in Showtime's new...	Verge	Longform	NaN	NaN	defining murder victim laura palmer told fbi s...
1	AI, the humanity!	Sam Byford	2017-05-30	AlphaGo's victory isn't a defeat for hum...	Verge	Longform	NaN	NaN	amazing master like ke jie ."" alphago master;...
2	The Viral Machine	Kaitlyn Tiffany	2017-05-25	Super Deluxe built a weird internet empi...	Verge	Longform	NaN	NaN	youtube star miranda sings ' disappointing net...
3	How Anker is beating Apple and Samsung at thei...	Nick Statt	2017-05-22	Steven Yang quit his job at Google in th...	Verge	Longform	NaN	NaN	entire consulting industry around helping star...
4	Tour Black Panther's reimagined homeland with ...	Kwame Opam	2017-05-15	Ahead of Black Panther's 2018 theatrical...	Verge	Longform	NaN	NaN	dollar marvel cinematic universe;would give bl...

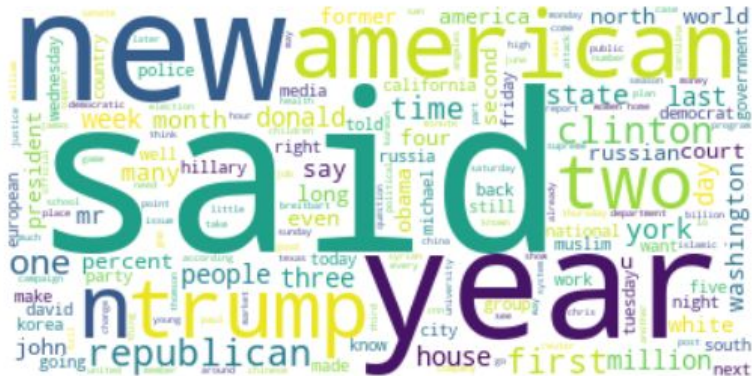
# Characterization



## Keywords



## Entities



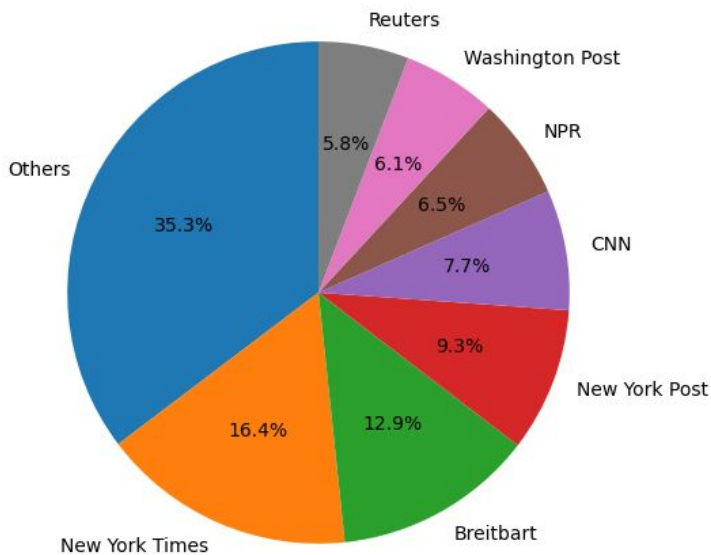
## Labels



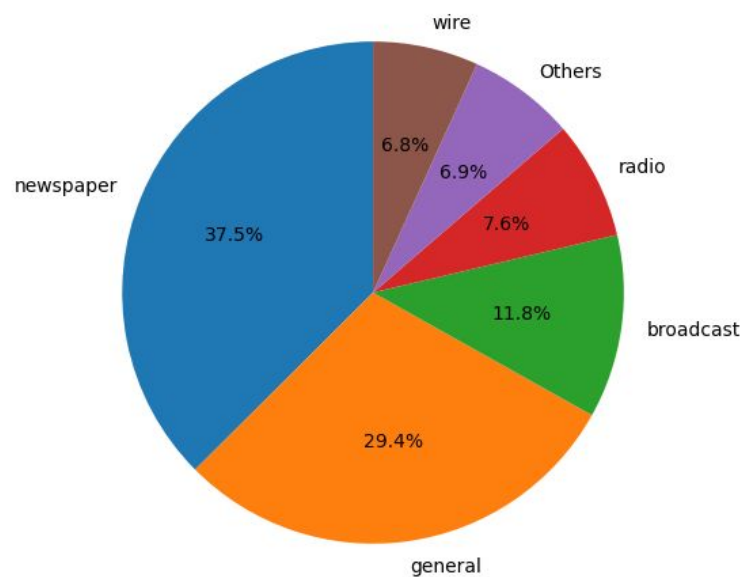


# Characterization

Publisher Distribution



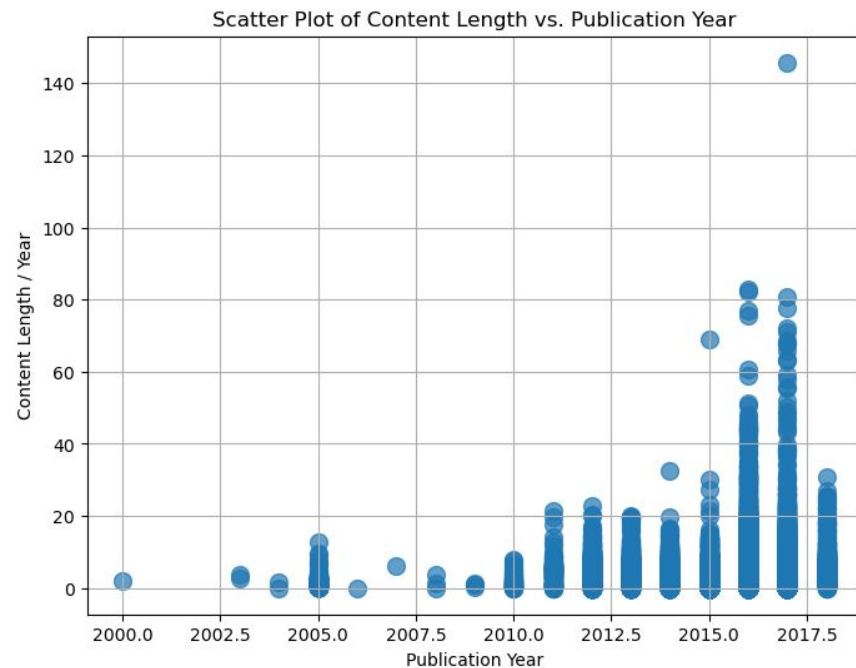
Source Distribution







# Characterization





## Search tasks

- Find news articles where Trump spoke on the immigration crisis
- Find news about LeBron's best performance in game
- Find articles related to latest crimes investigated by the police
- Find news articles regarding the conflicts between republicans and democrats