Web content mining and information extraction to review Covid-19 rapid diagnostic test evaluations

Dègninou Yehadji
Technological University Dublin
Ireland
B00108474[at]mytudublin.ie

Introduction

- It is being hard for academics and health professionals to summarize or keep up to date with the information, particularly in the context of fast moving COVID-19 research
- Advanced computational approaches such as web mining and natural language processing can be used to review scientific literature
- Project to used a set of web mining, information extraction and text analytics strategies to extract and analyze abstracts from the WHO Global literature on coronavirus disease

Objective

- Generate a structured database from article abstracts on COVID-19 rapid diagnostic test evaluations
- Identify top journals in which such studies are published,
- Identify and visualize top keywords used in article titles and summaries
- Summarize diagnostic test performances.

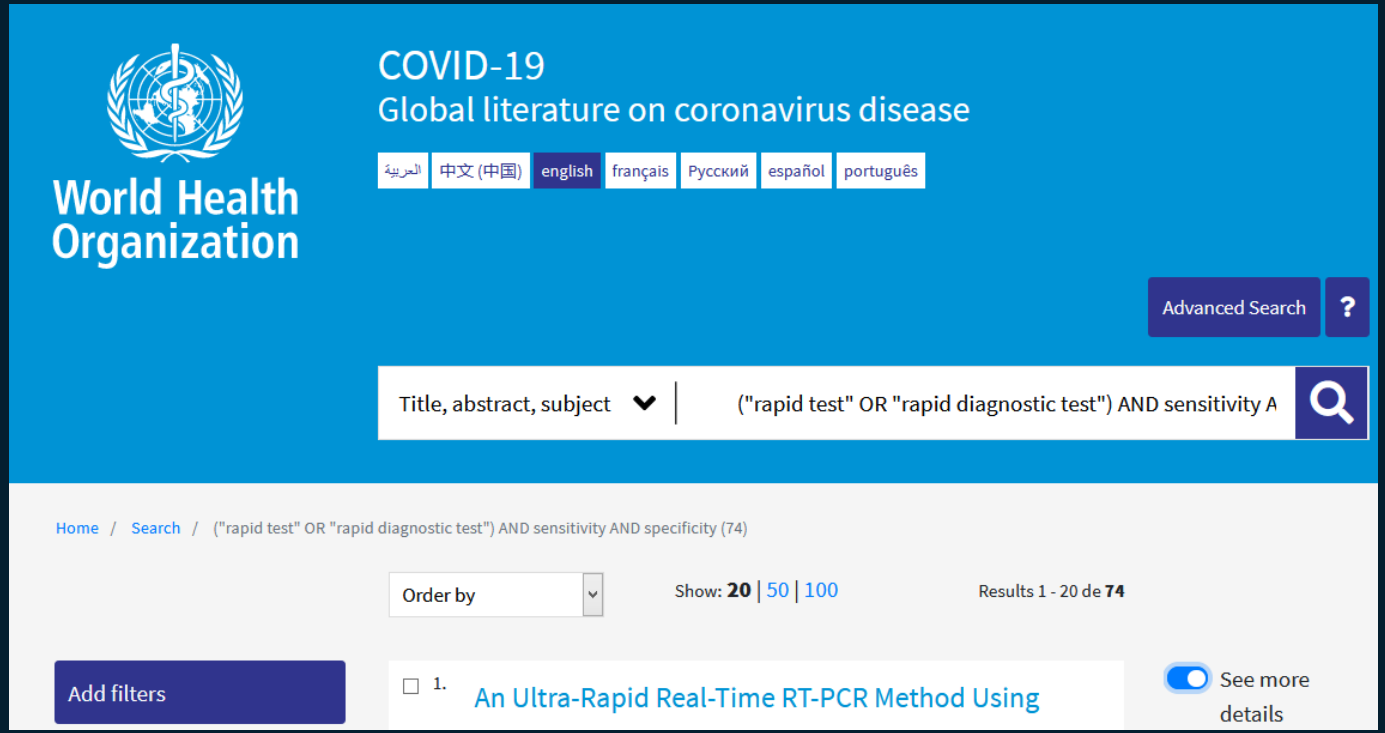
Methods – Overall approach

Search on the WHO Global literature on COVID-19 database to retrieve abstracts of publications related to the COVID-19 rapid diagnostic tests

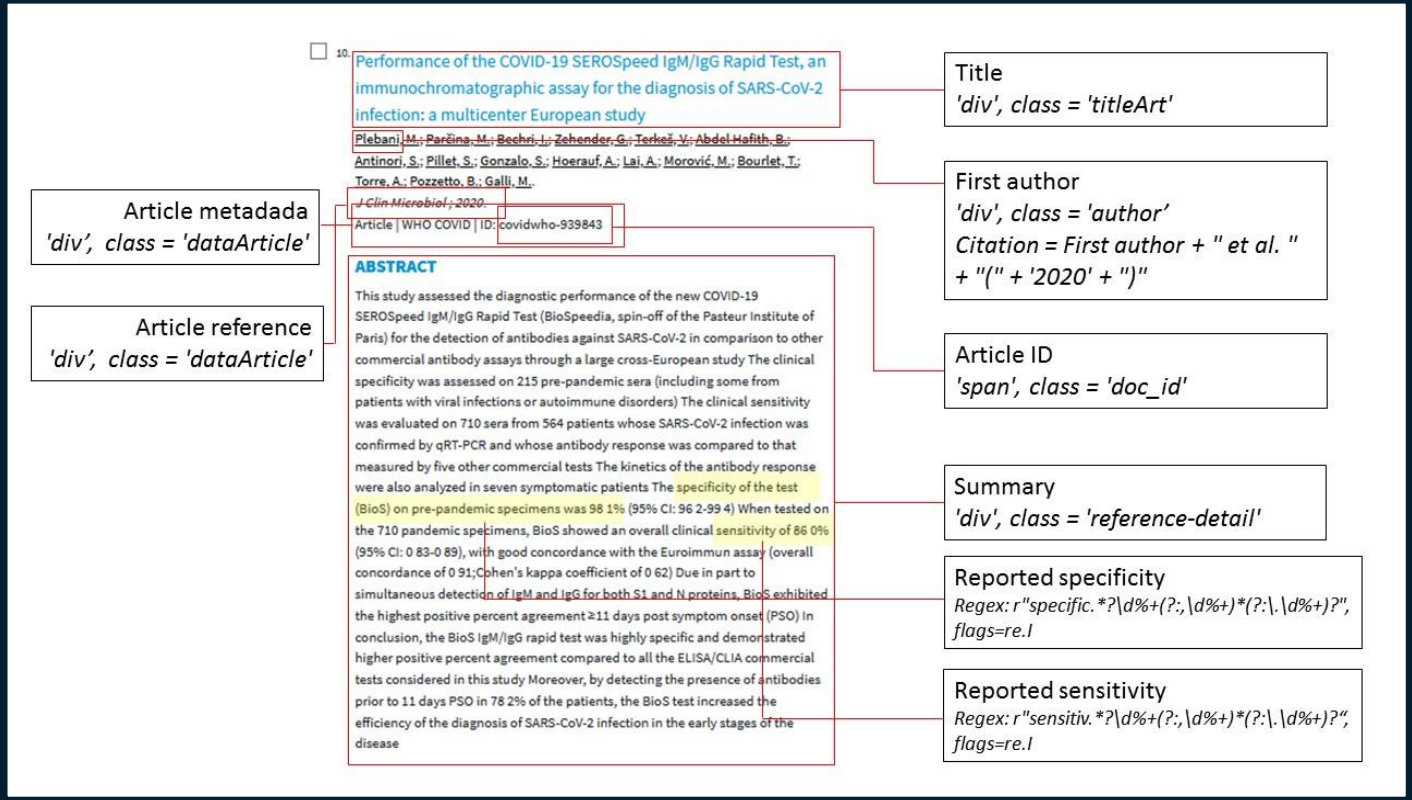
Web mining and information extraction was performed with Python

- Requests: HTTP for Humans™ library used to send http requests allowing query to the URL
- Beautifulsoup4 library to scrap the webpage content
- The re module to extract test performances (sensitivities and specificities) from abstracts

Methods – Search strategie



Methods – Web scraping



Methods – Data analysis

- Text preprocessing and analysis on titles and abstracts using NLTK
- Term frequencies in titles and abstracts on the preprocessed tokens
- WordCloud module to draw word clouds on preprocessed strings
- Frequencies of journal titles to identify top journals in which studies are published

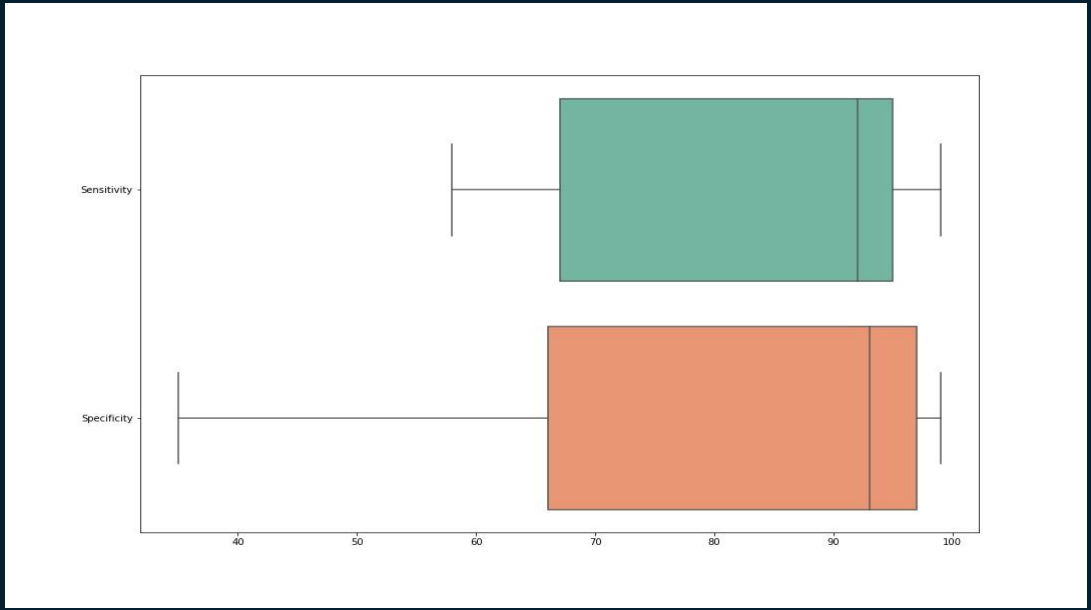
Results – The dataset

#	Column	Valid record	Data type
0	ID	73	object
1	Metadata	73	object
2	Citation	73	object
3	Reference	73	object
4	Journal title	73	object
5	Title	73	object
6	Url	73	object
7	Raw sensitivity extract	73	object
8	Raw specificity extract	73	object
9	Sensitivity (Clean)	13	float64
10	Specificity (Clean)	17	float64
11	Summary	73	object

Results – Top journals

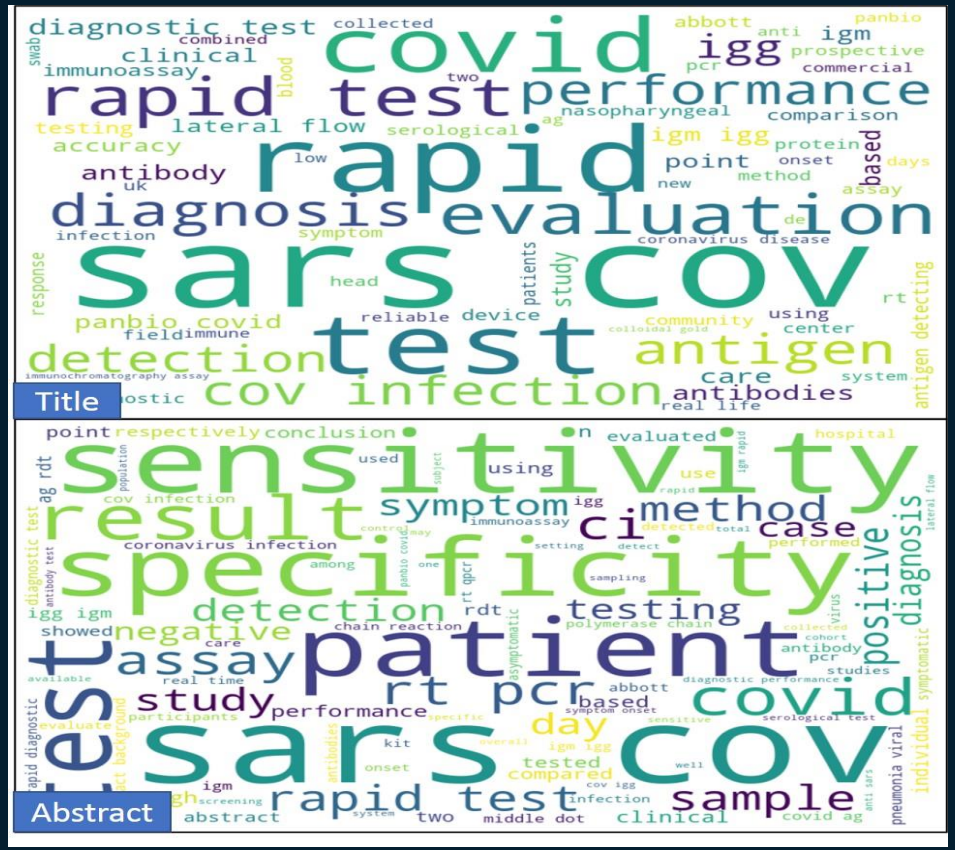
	Frequency	
Journal	n	%
Preprint	33	45
J Clin Virol	8	11
Clin Microbiol Infect	3	4
J Clin Microbiol	2	3
Am J Clin Pathol	2	3
Indian J Med Res	1	1
Journal of Medical Virology	1	1
Int J Infect Dis	1	1
Complex & Intelligent Systems	1	1
Clin Transl Med	1	1
Other journals	20	27
Total	73	100

Results – Test performances



The average sensitivity was 83% (Sd = 15) for sensitivity and 80% (Sd = 20) for specificity, with medians of 92% and 93% respectively

Results – Keywords



Conclusion

- The approaches adopted in this project allowed extraction of relevant information from the WHO Global literature on COVID-19 website containing abstracts of research articles.
- The dataset generated was analysed to summarize COVID-19 rapid diagnostic test performances, and to determine and visualize keywords in titles and summaries
- Limitation: the project does not meet the quality criteria of a systematic review. Improvement: performing a manual check on a sample the diagnostic performances retrieved.

THANKS!