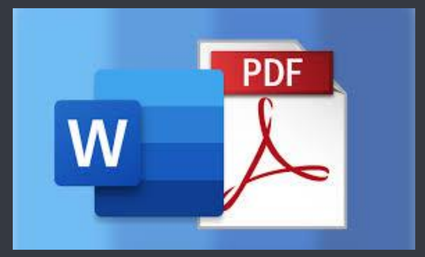
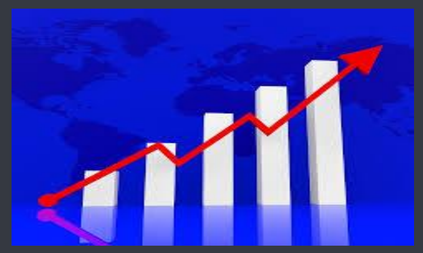
QUICK CANCER REPORT STARTING FROM A RAW DATABASE





REGISTRO POBLACIONAL DE CÁNCER
General Pueyrredon

rpc@ine.gov.ar

Introduction

This is a useful tool for registry practice and it has several important utilities.

It can be used to automatically convert a set of Word or PDF files into an ordered database, so you don't have to read each file.
Can also be used to quickly identify existing cancer cases in a database, to start registering in CanReg.

But above all, it can be used to prepare a quick report to the institutions that provide us with their information and thus strengthen the exchange between both.

Methods

1°. Scrapping
Input: Files (histopathological reports) .docx .pdf
Process: Read the files, scrap the files and obtain a data table
https://github.com/cballejo/basic_cancer_report/tree/main/report_scraping
file: Convert docx files to csv templates.R (RStudio®) - output: raw data table (csv)

2°. Data processing
Input: raw data table (csv)
Process: Read the .csv file, detect diagnostic keywords, obtain a cancer data table
https://github.com/cballejo/basic_cancer_report/tree/main/data_processing
file: Database leaker.R (RStudio®) - auxiliary file: keyword detection function - output: processed data table (csv)

3° Report
Input: processed data table (csv)
Report structure:• Written introduction• Table of incident cases per year• Table of cancer cases per year and according to location site• Interannual frequency comparison graph found in the institution (top 10 primary sites)• Summary measures of cases• Frequencies of cancer cases by gender and location•Final comments
https://github.com/cballejo/basic_cancer_report/tree/main/final_report
file: Report.Rmd (downgrade R) - auxiliary files: templates and header banner - output: final word file (docx)

Results

Reports prepared using this methodology were compared with reports from the records in the CanReg system. The results obtained were analyzed in terms of diagnostic quality and the number of cases detected in the 10 most incident topographical sites.
The sources chosen were 3 different institutions of the health system of the district of General Pueyrredon, in the province of Buenos Aires, Argentina.
No differences were observed in terms of the number of cases detected. However, there were 3% of cases that, with the new proposed methodology, were included in another category called "other tumor sites".

In conclusion, the tool that is shared is very useful for those cancer registries that want to process thousands of files with reports, for those that want to detect cancer cases in a database, and for those that want to immediately deliver a statistical report to the institution providing the data.

Bibliography

https://www.rstudio.com/resources/books/

Frequent questions

https://github.com/cballejo/basic_cancer_report/tree/main/frecuent_questions