RESEARCH EXCELLENCE FRAMEWORK
2021 IMPACT DATA ANALYSIS

Authors
Miguel Vieira, Arianna Ciula, Geoffroy Noël, Tiffany Ong
{jose.m.vieira, arianna.ciula, geoffroy.noel, tiffany.ong}@kcl.ac.uk

Institution
King's Digital Lab, King's College London

01 Introduction

The REF 2021 Impact Data Analysis was a small project, between King's Digital Lab (KDL) and colleagues in the King's College London (KCL) Research Management & Innovation Directorate (RMID), to support the analysis of the college's REF 2021 impact case studies and environment statements.

The data used during the development of this project includes 153 impact case studies and environment statements, in PDF (5-10 pages of text each), which follow standard templates but are expressed with heterogeneous descriptions and language.

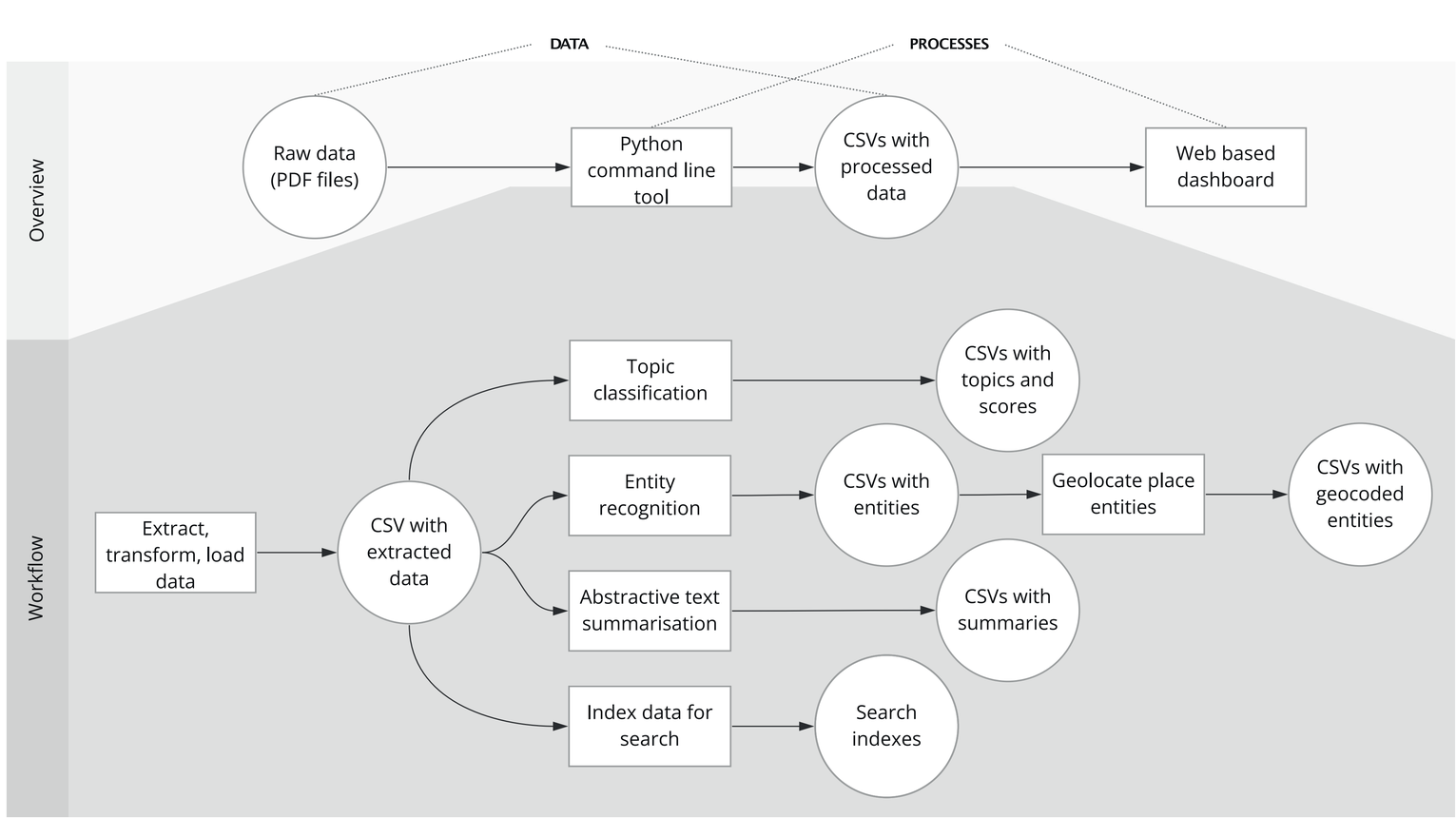
The project was set up to help RMID and research impact leads to address the questions:

- What are the main types of impact KCL has delivered? Which pathways have been used to deliver those impacts?
- Who are our key partners and beneficiaries of our impacts?
- Where are they - local (London), national or global?
- Is there a correlation between discipline and types of impact or pathways to impact used?
- What are the areas identified as strengths, areas for development and future plans?

02 Methodology

The project has two main components, a Python command line tool to do the data processing including running the machine learning processes, and a web-dashboard to present the results of the data processing:

- The process starts with extracting data from relevant sections (mainly the summary of the impact, details of the impact, sources to corroborate the impact) of the documents into a single CSV file, which is then used by the different machine learning processes;
- Zero shot topic classification is applied to extract impact categories, fields of research and impact pathways' outputs;
- A transformers-based language model is used to extract entities (mainly organisations and locations) from the data. The location data is further enriched by applying geocoding to gather coordinates and place geometries;
- Abstractive text summarisation is used to create summaries of the documents;
- And an indexing process indexes all the text both to perform keyword and semantic searches.



Architecture and workflow diagram, King's Digital Lab 2022

03 Processing

Topic classification has been applied to the documents using different authority lists to classify the data according to different perspectives:

- Impact categories, extracted from the whole text of the documents based on the nine REF-defined areas of impact;
- Fields of research (FoR), extracted from the section Underpinning research based on the Australian and New Zealand Standard Research Classification FoR classification;
- Pathways' outputs, extracted from the sections Summary, Details of the impact, based on the list of outcomes/outputs used by Researchfish, the impact data collection tool adopted by UKRI.

Entity extraction has been applied to different sections of the case studies to extract mentions of Organisations, Places and Products.

- GPE: Geo-political entities, countries, cities, states
- LOC: Non-GPE locations, mountain ranges, bodies of water
- NORP: Nationalities or religious or political groups
- ORG: Companies, agencies, institutions, etc.
- PRODUCT: Objects, vehicles, foods, etc. (not services).

In the dashboard, entities extracted from the sections Summary, Sources to corroborate the impact are grouped together in the Partners view. Entities extracted from the section Details of the impact appear in the Beneficiaries view.

Locations: entity extraction has been applied to the documents to extract Places mentions (GPE, LOC). The extracted entities were geocoded and classified according to the categories local (to London), national (UK) and global (rest of the world).

Search is possible in the dashboard using either a lexical search, that matches documents that contain the exact search terms, or by semantic search, that matches documents that contain words with meaning supposedly related to the search term.

04 Technologies

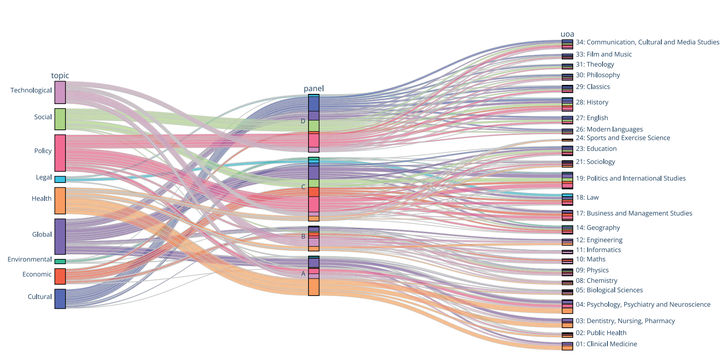
Python packages:

- Typer - to build the command line interface application;
- pandas - to load and manipulate the data;
- txtai - a library to build AI applications, it is used to extract text from the PDF documents, for topic classification, for abstractive text summarisation, and for semantic and lexical search;
- spaCy - natural language processing library used for entity extraction;
- GeoPy - to geolocate the extracted places using the OpenStreetMap Nominatin service;
- Streamlit - to build the dashboard;
- Plotly - to create the charts and visualisations.

github.com/kingsdigitallab/refida

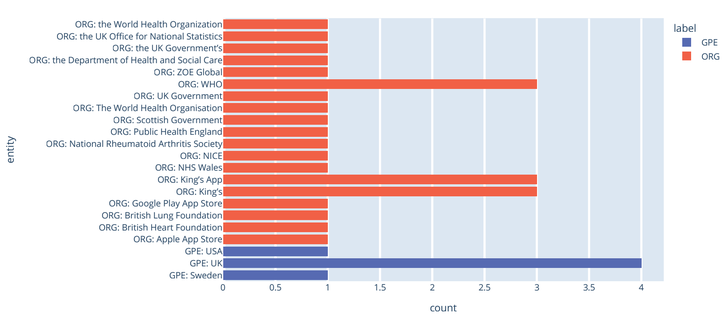
05 Results

The results of the data processing can be explored and visualised via a web-based interactive dashboard. The landing page of the dashboard displays a table and summary about the data. The rest of the dashboard is divided into sections according to the research questions. Search and filters can be tuned to get insights for single or multiple documents.



Alluvial diagram with impact categories extraction

Scatter plot with pathways' outputs extraction for panel D



Bar chart with entity extraction from a COVID-19 impact case study

Map with outputs of entity extraction and geolocation

06 Conclusion

The dashboard remains a tool with an exploratory function with the caveats that:

- All of the insights provided in the dashboard should not be accepted as final answers and should be reviewed; needless to say, the output requires further interpretation and analysis outside the system;
- Due to the time-frame and amount of data, there was no provision to train/fine tune the algorithms/models used; this means that some results may be more accurate than expected while others will be worse than expected and could even be useless.

The topic classification model used has an out of the box F1 score of 0.68 - 0.72 (for unseen and seen labels). In our evaluation the F1 score for the impact categories was 0.61 (0.74 precision, 0.52 recall) for topics assigned with a minimum confidence value of 0.5 or higher.

RMID organised a series of workshops to discuss the analysis of the REF impact documents with the support of the dashboard. Final analysis is not available to KDL yet but feedback has been very positive with intention to build on and possibly expand functionalities in the future.

kdl-info@kcl.ac.uk

kdl.kcl.ac.uk

@kingsdigitallab

KING'S College LONDON