RADAR PROJECT

HORIZON SCAN AUTOMATION

Data Science Internship

Autumn 2024

Github Repo

Team Members

Anh Le - 14195743

Ashley Wang - 14121545

Sagar Bhagwatka - 24613616

Table of Content

Project Background

The Centre for Work Health and Safety has launched the National Work Health and Safety Radar, an initiative that will deliver the latest insights on work health and safety (WHS) in Australian workplaces twice a year. The National WHS Radar will empower Australian regulators, academics, and leaders to take a proactive approach to WHS, informing existing and future policies, practices, and research projects.

Every six months, the Centre's Research Team will delve into a range of new WHS data and evidence, and deliver regular and actionable insights about WHS in an Australian context. From a list of trusted public data sources (websites), they will implement a horizontal scan through the latest:

- Literature and grey literature
- News articles
- Media release
- Relevant policies and legislations

As per the current practice, the Research Team will need to manually review the materials to filter the latest and most relevant insights, then input the following information into an Excel spreadsheet:

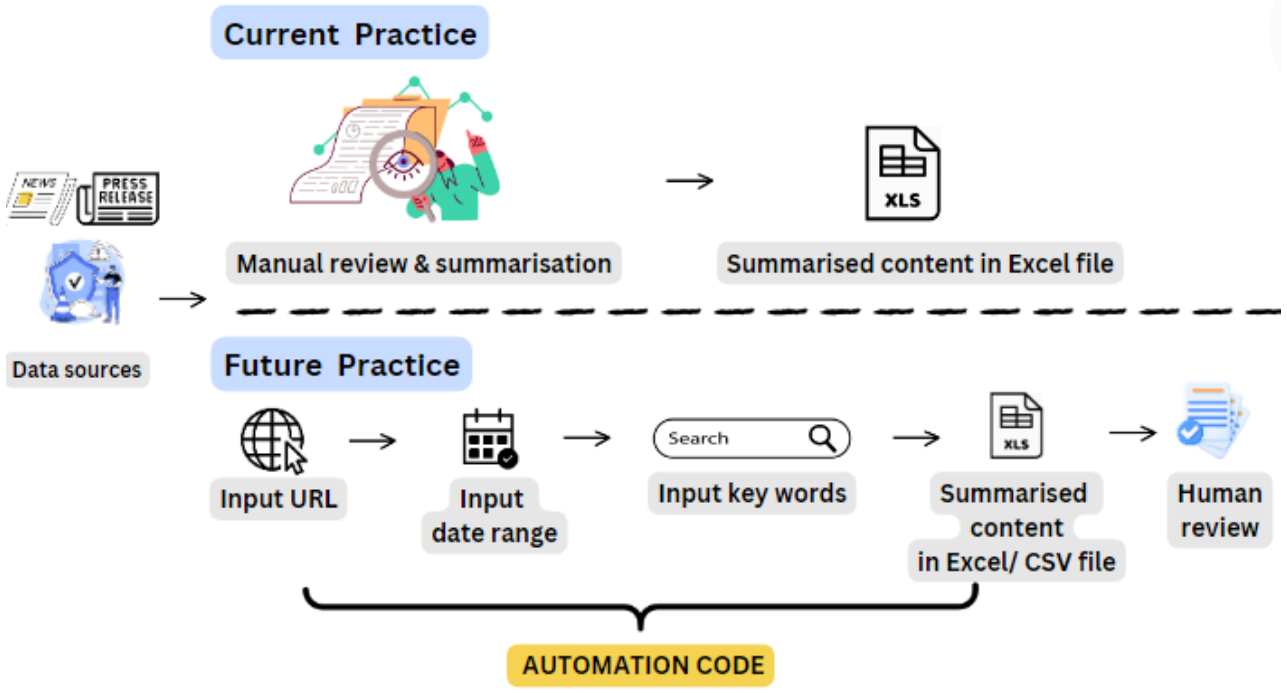
- Industry/ Report Section
- Short Headline
- Insight Description
- Importance
- Rating (5 = high, 1 = low)
- Link/ Reference
- Comment

Project Aims

Due to an extensive manual article review process in place, the project aims to **automate**:

- the extraction of up-to-date and pertinent content from the data sources
- cleaning and analysing the information to ensure the materials are relevant to the search keywords and within an expected date range
- and finally delivering valuable summaries and insights from the cleaned data

The below figure illustrates the current practice of Radar Horizon Scan process and the future practice after adopting the automation code.



Methodologies

The project entails four key steps:

1. Generating all links from Media/News pages.
2. Scraping data, including Title, Content, Publish Date, and Link of Article, using the newspaper3k library.

3. Cleaning data to filter articles within a chosen date range and applying threshold cosine similarity scores for different keywords.

4. Utilizing Large Language Models to summarize each article, employing tools including LLAMA2, Gemini Pro, and LLAMA Chat.

Results

We developed a code base that is able to achieve most of the project aims mentioned above. Particularly, using the user input of the News or Media Releases page of the website, the code can automatically extract all the links from that URL. However, depending on the number of links contained in that URL, the amount of time to extract them might vary. From the list of those links, these pieces of information will be scraped:

- Headline
- Content
- Date
- Reference URL

Next, based on user input of the date range and keyword, the data will be filtered accordingly to ensure its relevance to the reporting purpose. At this step, a similarity score of each article will also be shown. This score ranges from 0 - 1. The higher the score is, the more relevant the article is as compared to the keyword input, and vice versa. As per our observation, the threshold for the similarity score will range at approximately 10-15%. Nevertheless, this threshold might change depending on each keyword. In this case, the user need to use their domain knowledge to decide a suitable similarity threshold.

The below image illustrates an example of results obtained after the data from the Safework SA website has been cleaned and filtered.

	Headline	Content	Date	Reference URL	Similarity Score
18	Asbestos removal licence suspended after safet...	The contractor also failed to give SafeWork SA...	2024-01-24	https://www.safework.sa.gov.au/news-and-alerts...	0.145463
20	Locking in our major public holidays	The Public Holidays Act 2023 was finalised in ...	2023-11-16	https://www.safework.sa.gov.au/news-and-alerts...	0.058703
37	2023 Augusta Zadow Awards announced	The SafeWork SA awards were presented by the G...	2023-10-20	https://www.safework.sa.gov.au/news-and-alerts...	0.206498
46	Carbon Monoxide: A threat in your workplace	Carbon monoxide: a threat in your workplace\n\...	2024-03-25	https://www.safework.sa.gov.au/news-and-alerts...	0.085564
57	Changes to regulations for Amusement devices	On 25 December 2023 amendments relating to amu...	2023-11-30	https://www.safework.sa.gov.au/news-and-alerts...	0.174004
58	Holiday rostering reminder	Section 13B of the Shop Trading Hours Act 1977...	2023-12-15	https://www.safework.sa.gov.au/news-and-alerts...	0.019332
61	Retail work hazards under the microscope	From March, SafeWork SA inspectors will visit ...	2024-03-15	https://www.safework.sa.gov.au/news-and-alerts...	0.180482
62	Building site blitz reaches new heights	Since 3 July, SafeWork SA inspectors have perf...	2023-11-13	https://www.safework.sa.gov.au/news-and-alerts...	0.097568
65	Dangers of bypassing safety devices	Inspectors recently attended a worksite where ...	2023-12-12	https://www.safework.sa.gov.au/news-and-alerts...	0.304078

After the articles have been filtered, we will perform data summarisation using Large Language Models. In this project, we used LLAMA2, Gemini Pro and LLAMA Chat. While LLAMA2 and LLAMA Chat are free of charge and could generate detailed summaries, Gemini Pro is able to provide more concise text summarisation. During our code development, we have tried different prompt input for these three models to ensure the output is accurate. Up to date, no hallucination has been detected in the text summarisation results.

Challenges and Mitigation Strategies

Encountered Problems

Challenges impeded the project's progress include:

1. **Authentication Requirements**: Internal links necessitating authentication hindered data scraping efforts. Without proper credentials, accessing such content was unfeasible, leading to incomplete data collection.
2. **Website Restrictions**: Approximately 20% of the targeted websites blocked data scraping attempts. Additionally, some websites demanded subscriptions for access, further complicating data collection.
3. **Non-Textual Content**: Certain articles within the scraped links were not in textual format, such as PDF or image files, making it arduous to

extract meaningful information. This issue hampered the analysis and summarization processes.

4. **Time and Computational Demands**: Executing the project's code demanded substantial time and computational resources. The process of scraping, cleaning, and analyzing vast volumes of data necessitated robust computational capacity, resulting in prolonged execution times and potential resource constraints.

5. **Dependency on Manual Selection**: Post data cleaning, the number of results depended on manual selection based on the cosine similarity threshold. This manual intervention introduces subjectivity and may affect the consistency of results.

6. **Selection of NLP Tool**: Comparison of three NLP tools revealed that Gemini Pro is recommended if computational resources are limited, although additional costs may apply. Conversely, LLAMA Chat provides better statistical details and more concise summaries, making it preferable if computational resources are not a constraint and it is free of charge.

Mitigation Strategies

To overcome these challenges, several mitigation strategies are proposed:

1. **Alternative Authentication Methods**: Exploring alternative authentication methods or obtaining necessary credentials could facilitate data scraping from authenticated sources.

2. **Content Transformation**: Developing algorithms or tools to convert non-textual content, such as images or videos, into textual format could enhance the project's ability to extract information from diverse sources.

3. **Optimization of Code and Resources**: Optimizing code efficiency and utilizing scalable computational resources, such as cloud computing services, can alleviate the time and resource constraints associated with data processing.

Conclusion

In conclusion, the Horizon Scan Automation for the WHS Radar Initiative has been a fruitful project with certain major achievements. Using different techniques including web scraping, semantic search from Natural Language Processing and Large Language Models, we have been able to automatically extract latest and most relevant insights from diverse sources of data. However, it is worth noting that the project encountered a number of challenges related to authentication requirements, website restrictions, content format, resource demands, manual selection dependency, and NLP tool selection. By implementing appropriate mitigation strategies, such as alternative authentication methods, content transformation techniques, optimization of code and resources, these challenges can be addressed, facilitating the successful execution of the project objectives.