TAIC project proposal

Project proposal for 159333 Massey University

James Thompson - 21011195

14/07/2023

Background

This is a project that a familial contact has been wanting to do for the last year. Recently he was inspired by LLMs abilities to read and interpret natural language. He wanted to use this new technology to analyse all of the reports TAIC have released in the past decade. These reports can be found here Investigations | TAIC.

Goal

The final goal for this project would be to build a system that is as flexible as possible so that it could be setup and used across marine, rail and air accident investigation reports.

It would give insights that could normally only be achieved by manually reading lots and lots of reports. This way of doing it is very manpower intensive and so it is not often done. With a system to do the reading and understanding it can leave the manpower usable elsewhere.

Scope

As this is a very large project only so much can be a achieved in this singular semester. As such the stated goal for the 159333 project would be to build a proof of concept/value which would study the maritime reports from the last 10 years.

The end result would be two fold. A python project that did the heavy lifting and all of the computer science work and a 2 page report summarising the findings (this is excluding any reports/requirements of 159333 itself). This python project could be the groundwork for a more generalised engine that can do extraction and analysis across many different types of reports. The 2 page report will be the evidence used for the proof of value and can be given to non technical people.

How

Tools

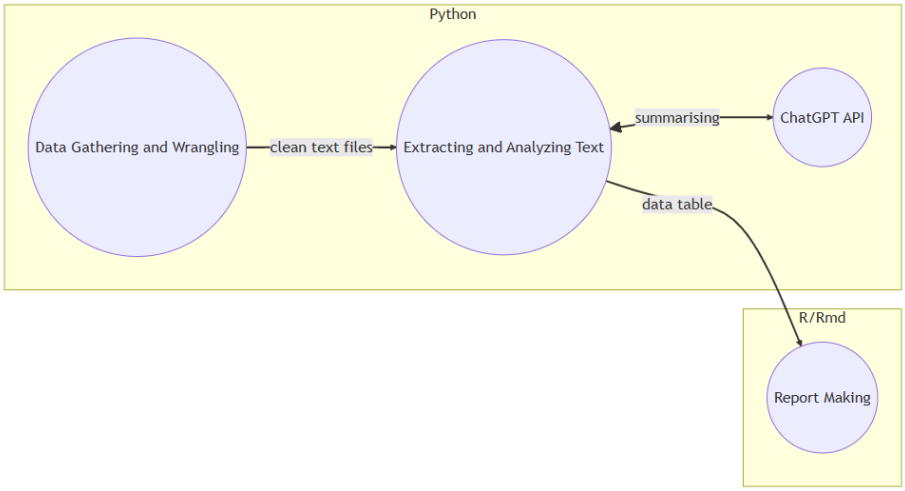
Code I intend to complete this using python as the main driver of the data science work. As the initial inspiration came from the recent advancements of LLMs (particularly gpt-3) there will be as much utilisation of LLMs through the openAI api as possible. With R used to do the final statistical analysis and report making.

Management VCS will be GitHub with the whole project being in an open repository. Feedback from stakeholders will be managed using GitHub Issues. The tracking of milestones both big and small can be done with GitHub projects.

Solution plan

I expect throughout the semester the solution design could change drastically.

However on outset I am working towards a pipeline like below:



Longer term it would be good to reduce the amount of tedious NLP python code and increase the utilisation of openAI models and potentially training our own on the reports themselves.

Milestones

There will be 3 milestones.

By the time of reaching the breadth milestone we will have enough substance of data to start doing some deep statistical analysis. The basic milestone can only have summary statistics.

Basic To achieve this there will be an extraction function. It will take the pdfs and determine the causes of the accidents.
These causes will be put togather into a data table which can then be sent to R for some basic correlation and relationship analysis.
Once this milestone has been completed the framework for extracting information into the data tables would be built.

Breadth Now that the framework of turning all the pdfs into a datatable has been completed there should be work on increasing what we look at.

For example we could start extracting time of day, outcomes (i.e damages, lives lost etc) and others that seem appropriate at this point.

Depth Going back to the causes we can start to turn these from simple themes like maintenance, fatigue, rushed etc and extend these causes into groups and subgroups. For example maintenance could be expanded to frequency and quality.