# Proposal: Job Title: Web Content Analysis & Data Retrieval using Language Models

Objective: To undertake a comprehensive web content analysis and data retrieval of 1,500 links supplied in a CSV format. The primary aim of the project involves extracting and categorizing specific details from the articles linked therein.

## Procedure:

1. The first task involves the analysis of mentions of the specified subjects (Paul Phua or Wei Seng Phua). Implement a program to extract all these relevant mentions and identify unrelated contexts.
2. Utilize a natural language processing (NLP) Algorithm to identify specific categories like judicial cases mentioned, any ties to criminal groups (14k, Triads, gang), and mentions of the acquittal of cases.
3. Using Web scraping tools, extract necessary information such as the language of the article, the date of publication, the publisher's name and specialty, the publisher's country, and global contact email.

## Technologies and Tools: A combination of the following tools would be employed;

1. Bard API (Or Gpt API): It would facilitate deep contextual understanding of the articles.
2. Python: This is suitable for web scraping and NLP tasks.
3. Beautiful Soup & Selenium: These would be used for web scraping to gather necessary meta information from the articles.

Estimated Time: The entire operation is estimated to take approximately 10 working hours. This is inclusive of coding, testing, data extraction, and refining the data acquired.

Alternative Solutions: While Bard API has excellent capacities, OpenAI's GPT-3 could also be an alternative option for the language model required due to its exceptional performance in contextual understanding and information extraction. The decision between the two depends on project requisites, specifics, and budget.

The key to any successful project lies in clarity and precision. This description encompasses all the aspects of the project and its execution strategy.

## Detailed procedure and schedule:

Step1:

1. Examination of the data and preparation of a strategic approach.

2. Implementation of a Python program to determine mentions of specified subjects (Paul Phua or Wei Seng Phua) and identifying nonrelevant mentions.

Step2:

1. Implementation of python script using web scraping tools (Beautiful Soup & Selenium) and Bard API (Or Gpt API) to extract required details like:

o Identify and categorize judicial cases mentioned

o Detect mentions of Paul Phua belonging to a criminal group (14k, Triads, gang)

o Track mentions of the acquittal of cases.

o Language of the article

o Date of publication

o Publisher's name

o Publisher's specialty, focusing on relations to gambling.

o Publisher's country

o Publisher's global contact email

Step3:

1. Testing, refining and finalization of the data acquired.

2. Submission of the final product and all associated files.