King Saud University
College of Computer and Information Sciences
Department of Information Technology
IT 362: Principles of Data Science
Project – Phase#1

Saudi Law project

LOGBOOK

Phase#1

First Semester 2025

Project Overview

Saudi Labor Law Q&A Dataset
This project aims to build a structured dataset of the Saudi Labor Law, including its official articles and frequently asked questions (FAQs), in order to make the information more accessible for analysis, visualization, and potential applications such as question - answering systems.

The dataset captures:
• Articles of the law (raw legal text).
• FAQs (question - answer pairs).
• Metadata linking each entry to its source URL or official PDF.

This ensures that the dataset can be used in future machine learning and data science projects that require legal text processing in Arabic.

Data Collecting

Data Collection Period: [10 September 2025] To [24 September 2025]

Data sources :

- Bureau of Experts at the Council of Ministers(As pdf) : The official text of the Saudi Labor Law.

Link :http://laws.boe.gov.sa/BoeLaws/Laws/LawDetails/08381293-6388-48e2-8ad2-a9a700f2aa94/1

- Istitlaa Platform (By web scrapping) : Draft amendments to the Saudi Labor Law regulations by MHRSD on the Istitlaa platform.

Link:https://istitlaa.ncc.gov.sa/ar/Labor/Hrsd/Regulations5/Pages/default.aspx

- Qiwa Platform (By web scrapping) : Qiwa is an MHRSD digital platform offering labor law content and electronic services for private-sector employers and workers.

Link :https://qiwa.sa/ar/labor-law

- Master’s Theses from King Saud University Digital Library(As pdf): Master’s theses containing analytical studies on Saudi labor law.

Link:https://drive.google.com/drive/folders/1wLQwAG_AKiWEzeCQFQtK6wy6IhCuZFrG

These sources were selected because they represent official and reliable references that ensure the accuracy and validity of the legal data, which enhances the credibility of the project and the quality of its outcomes. In addition, these sources supported by explanatory materials and simplified FAQs that help bridge the gap between complex legal wording and practical understanding. This combination of precise legal texts and clarifying explanations enables us to obtain a comprehensive and balanced dataset, making it easier to process and analyze in future applications such as question–answer systems.

Challenges and Actions taken:

Website Structure :

- Qiwa presents the Saudi Labor Law in separate sections, each under a unique URL. This made scraping more complex because the data was not centralized on a single page.
- Solution: A CSV file was created with: Column 1 = URL of the section Column 2 = Extracted text content This ensured the data was organized and traceable to its source.

Different Data Formats :

- The law text and the FAQ section had different structures. The law text was straightforward, while FAQs required a question–answer format.
- Solution: FAQ data was stored in a separate CSV file (labor_law_faq.csv) with: Column 1 = Question Column 2 = Answer This separation ensured consistency and ease of later processing.

WebDriver Configuration :

- Setting up Selenium WebDriver for headless scraping posed a challenge.
- Solution: Proper configuration was required to ensure compatibility with dynamic content loading and to allow scraping without opening a browser window. This involved fine-tuning browser options for performance and stability.

Prevent web scraping :

- Some websites prevent or hinder automated data collection by using dynamic JavaScript loading, protections like CAPTCHA, restrictions in the robots.txt file, or automated behavior detection systems (rate-limiting, WAF). This caused data collection interruptions, incomplete results, repeated errors during execution, and even the risk of account suspension or IP blocking.
- Solution: Shifted to sources that allow web scraping and ensured proper permissions and legal compliance.

Text Encoding :

- Arabic text initially appeared corrupted in Excel.
- Solution: Exported the files with UTF-8-SIG encoding to ensure proper display.

Web Scraping

Web Scraping Tools Used

- Selenium : Interacting with dynamic web pages and handling JavaScript-based content.
- webdriver_manager : Automatically installs and manages ChromeDriver.
- pandas : Organizing and saving the scraped data in tabular format.
- csv : Writing extracted results into CSV files.
- Requests (common tool, not used in our script) : Sending HTTP requests to fetch static web content.
- BeautifulSoup (common tool, not used in our script) : Parsing and extracting information from HTML documents.

Libraries, Modules, and Methods Used in the Web Scraping

Libraries Used

- selenium : Web scraping and interacting with dynamic web pages.
- webdriver_manager : Automatically installs and manages ChromeDriver.
- pandas : Data processing and saving results into CSV.
- csv : Writing data directly into CSV files.
- os : File and path management.
- time : Adding delays (e.g., sleep) during page loading.

Selenium Modules Used

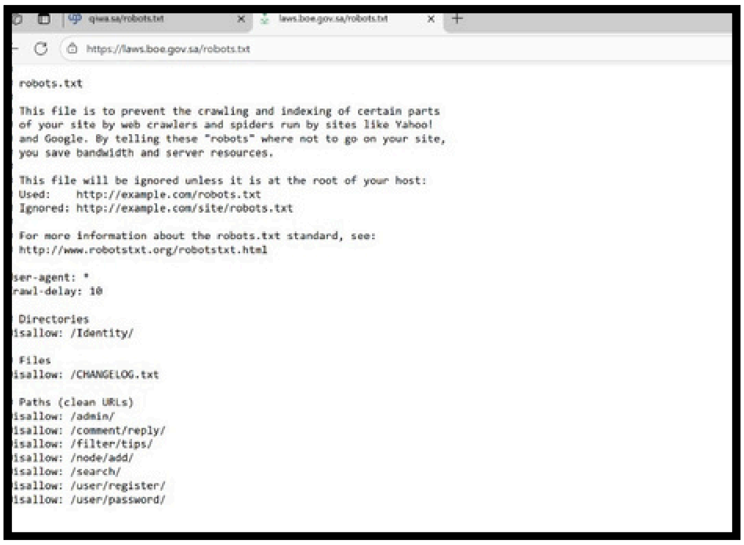
- webdriver : Launch and control the browser.
- By : Locate elements using CSS selector, XPath, ID, etc.
- WebDriverWait : Wait until a specific condition is met.(the page is fully loaded)
- expected_conditions (EC) : Predefined conditions used with WebDriverWait (e.g., presence, visibility).

Methods Used

- driver.get(url) : Open the target web page.
- driver.find_element(...) / driver.find_elements(...) :Find element(s) on the page.
- element.click() : Click on an element (e.g., expand accordion for answers).
- element.text : Extract text content from an element.
- wait.until(EC....) :Wait until a condition is met (e.g., element is present).
- EC.presence_of_element_located(locator) : Check that the element exists in the DOM.
- EC.visibility_of_element_located(locator) : Check that the element is visible before interacting.
- list.append(item) : Store extracted results in a list.
- csv.writer.writerow(row) : Write one row into a CSV file.
- pandas.DataFrame.to_csv(path, ...) : Save results into a CSV file.

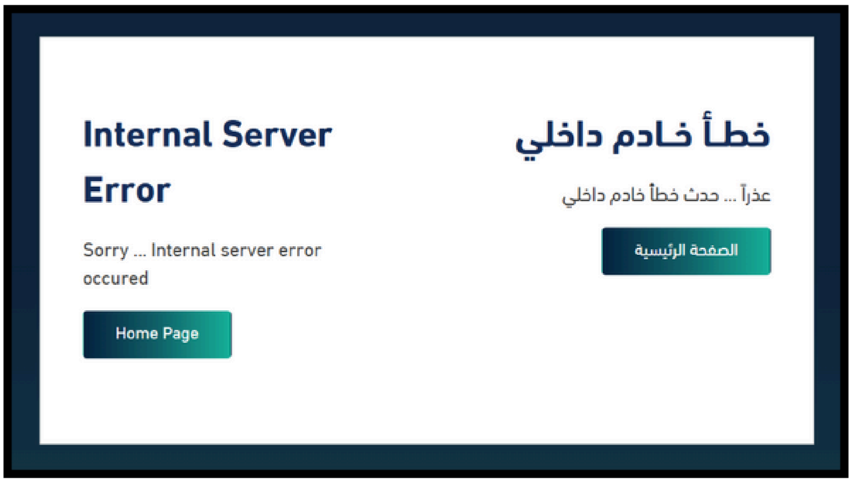
Robots.txt:

1- Bureau of Experts at the Council of Ministers


qiwa.sa/robots.txt
laws.boe.gov.sa/robots.txt
https://laws.boe.gov.sa/robots.txt

```
robots.txt

This file is to prevent the crawling and indexing of certain parts
of your site by web crawlers and spiders run by sites like Yahoo!
and Google. By telling these "robots" where not to go on your site,
you save bandwidth and server resources.

This file will be ignored unless it is at the root of your host:
Used:    http://example.com/robots.txt
Ignored: http://example.com/site/robots.txt

For more information about the robots.txt standard, see:
http://www.robotstxt.org/robotstxt.html

ser-agent: *
rawl-delay: 10

Directories
isallow: /Identity/

Files
isallow: /CHANGELOG.txt

Paths (clean URLs)
isallow: /admin/
isallow: /comment/reply/
isallow: /filter/tips/
isallow: /node/add/
isallow: /search/
isallow: /user/register/
isallow: /user/password/
```

2- Istitlaa Platform


Internal Server Error

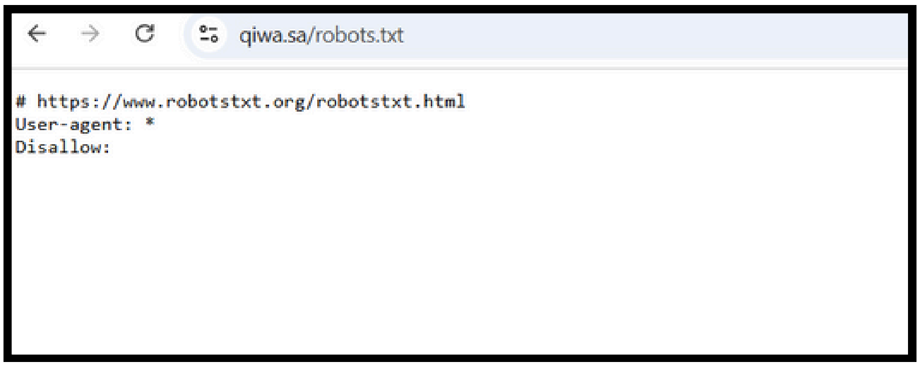
Sorry ... Internal server error occured

خطأ خادم داخلي

عذراً ... حدث خطأ خادم داخلي

The website didn't allow to enter the page.

3- Qiwa Platform



```
# https://www.robotstxt.org/robotstxt.html
User-agent: *
Disallow:
```