**Python Requirements Document**

**Overview**

The main goal is to extract specific information such as **titles, author names, emails, departments, countries, and phone numbers** from both websites and PDF files, and to store the extracted data in an Excel sheet.

**Requirement 1: Website Data Extraction**

**Objective**:  
Extract structured information from a given website URL.

**Details**:

* Input: Any valid website URL. (Ex: <https://chemrxiv.org>)
* Output:
  + **Title** of the page
  + **Author names**
  + **Email addresses**
  + **Department names**
  + **Country**
  + **Phone numbers**

**Requirement 2: PDFs Document Processing**

**Objective**:  
Download PDFs from websites and extract data into an Excel sheet.

**Details**:

* Input: URL of a PDFs file. (Ex:- <https://www.preprints.org/> )
* Users are given a page limit.
* Actions:
  + Download the PDF.
  + Extract the following:
    - **Title**
    - **Author names**
    - **Email addresses**
    - **Department names**
    - **Country**
    - **Phone numbers**

**Requirement 3: Web Profile Information Extraction**

**Objective**:  
Extract profile information from any website, such as titles, author names, emails, departments, countries, and phone numbers.

**Details**:

* Input: URL of a profile page. (Ex: <https://chemrxiv.org>)
* Output:
  + **Title**
  + **Author/Person name**
  + **Email addresses**
  + **Department names**
  + **Country**
  + **Phone numbers**

**Summary**

* Performed web scraping to extract data from HTML content and stored the results in an Excel sheet.
* Given a website link and a page limit, download all PDFs available within the specified pages.
* Structured data extraction of names, contact information, and affiliations from various formats on a given website.