

Tabula

Tabula is a powerful tool for extracting tabular data from PDF documents. It allows you to specify the region of the PDF containing the table and convert it into structured data, such as CSV or JSON format.

You might find Tabula useful for several reasons:

1. **Extracting Data from Government Reports:** Tabula can help journalists extract data from government reports published in PDF format. Many official reports in Iran, such as economic indicators, demographic statistics, or government budgets, are often released in PDF form. Tabula enables journalists to extract tabular data from these reports, making it easier to analyze and interpret the information.
2. **Analyzing Financial Data:** Journalists covering financial news or investigating economic issues may encounter financial statements, balance sheets, or corporate reports presented in PDFs. With Tabula, journalists can extract financial data from these PDFs and analyze it in spreadsheet software or other data analysis tools. This allows for deeper analysis, identification of patterns or anomalies, and better reporting on financial matters.
3. **Scrutinizing Election Results:** During elections, Tabula can assist journalists in analyzing and processing election results that are often published as scanned PDF documents. By extracting the tabular data using Tabula, journalists can quickly calculate and verify vote counts, analyze trends, and compare results across different regions or electoral districts.
4. **Data-driven Investigative Journalism:** Tabula empowers journalists engaged in data-driven investigative journalism. By extracting data from PDFs, journalists can uncover hidden patterns, identify relationships, or detect irregularities in datasets related to corruption, public spending, or social issues. Tabula enables them to transform unstructured data into a structured format for in-depth analysis and reporting.
5. **Local News Reporting:** Tabula can be invaluable for local journalists covering community news in Iran. It allows them to extract data from local government reports, public records, or community surveys that are often published in PDFs. By analyzing and visualizing the extracted data, journalists can better understand and report on local issues such as public services, infrastructure development, or social welfare.

Here's a general guide on how to use Tabula to extract tabular data from PDFs:

1. **Install Tabula:** Start by installing Tabula on your machine. You can download Tabula from the official website (<https://tabula.technology/>) or use the Tabula command-line tool. Note that Tabula is a Java-based tool, so you'll need to have Java installed.
2. **Load the PDF:** Launch Tabula and open the PDF file that contains the table you want to extract. You can do this by either uploading the PDF file through the Tabula interface or specifying the file path in the command-line tool.
3. **Select the Table Region:** Use the mouse cursor to draw a rectangle around the table region in the PDF document. Ensure that the rectangle covers the entire table and doesn't include any other surrounding elements.

4. **Preview the Table:** Once you've selected the table region, Tabula will display a preview of the extracted data. Verify that the preview accurately captures the tabular structure and content of the table.
5. **Adjust Table Selection (if needed):** If the table selection is not accurate or if the table spans multiple pages, you can manually adjust the selection or add additional table regions to cover the entire table.
6. **Choose the Output Format:** Tabula allows you to export the extracted table data in various formats, such as CSV, JSON, or TSV. Select the desired output format that suits your needs.
7. **Export the Table Data:** Click on the "Export" or "Extract" button to extract the selected table region and save it as structured data in the chosen output format. Specify the destination file path and file name for the extracted data.
8. **Validate and Clean the Extracted Data:** After extracting the table data, review the output file to ensure the extracted data is accurate and complete. Clean the data if necessary, removing any irrelevant rows or columns, and perform any additional data processing steps.

Remember to comply with the legal and copyright restrictions when using Tabula or any other tool for extracting data from PDF documents. Ensure you have the necessary rights and permissions to extract and use the data from the PDFs you are working with.

Advanced Topics

Copying a Template from One Page to Another:

In some cases, PDF documents may have similar table structures repeated across different pages. Tabula allows you to copy the template of a table from one page and apply it to another page with a similar structure. This feature saves time and effort by avoiding the need to manually recreate the table structure on each page. By copying the template, you can ensure consistency and easily extract the desired data from multiple pages.

OCR

When you take a picture of a document or scan it, the computer sees it as an image with pixels, not as actual words or text. OCR software analyzes these images and uses advanced algorithms to recognize the shapes of letters, numbers, and other characters in the image. It then converts those shapes into actual text that a computer can understand and work with. Understanding OCR is relevant to knowing how to use Tabula properly. Tabula is a tool designed specifically to extract tabular data from PDF files. However, it has a limitation. If a PDF is image-based, meaning the text in the PDF is in the form of images and not actual selectable text, Tabula won't be able to directly extract the data. Tabula relies on the text being recognized and selectable within the PDF.

In such cases, before using Tabula, you would need to use an OCR software or tool to convert the image-based PDF into a searchable and selectable text-based PDF. Once the PDF has

undergone OCR and the text is recognized, Tabula can then be used to extract the tables from the converted text.

There are a few quick ways to determine if you need OCR (Optical Character Recognition) for a particular document or image:

1. **Searchable Text:** If you're unable to search for specific words or phrases within a document or image using the standard search function, it indicates that the content is not recognized as text. OCR can convert the document into searchable text, enabling you to locate and extract information more efficiently.
2. **Copy and Paste:** Attempt to select and copy text from the document or image. If you can't copy the text or if it pastes as nonsensical characters or images, it suggests that the content is not in a text-based format. OCR can convert the content into editable text that can be copied and pasted.

Example:

Table That Can Be Parsed:

The table on page 1470 of the paper [Climate change impact on cultivated and wild cacao in Peru and the search of climate change-tolerant genotypes](#) can be parsed—notice that you are able to select text in the table.

Table That Cannot Be Parsed (Image-Based):

On the other hand, Figure 3 on page 1469 of the paper [Climate change impact on cultivated and wild cacao in Peru and the search of climate change-tolerant genotypes](#) would require OCR to extract the text, since the text is in image form. Note that much of the figure could not be converted to text at all, since much of the figure is a map of land.

When you do need OCR, there are many OCR tools that could suit your needs:

1. **Tesseract:** An open-source OCR engine developed by Google. It supports multiple languages and has extensive community support. [Link: <https://github.com/tesseract-ocr/>]
2. **OCRopus:** An open-source OCR system developed by Google, OCRopus includes the Tesseract OCR engine along with additional components for layout analysis, post-processing, and more. [Link: <https://github.com/tmbdev/ocropy/>]
3. **CuneiForm:** CuneiForm is an open-source OCR tool developed by Cognitive Technologies. It offers multilingual support and can handle complex document layouts. [Link: <https://github.com/Cognitive-Technologies/CuneiForm>]
4. **Abbyy FineReader:** A comprehensive OCR tool known for its high accuracy and advanced features. It offers layout analysis, language support, and image enhancement options. [Link: <https://www.abbyy.com/en-us/finereader/>]

5. Adobe Acrobat Pro: A widely-used PDF software with built-in OCR capabilities. It provides text recognition, document conversion, and post-processing options. [Link: <https://acrobat.adobe.com/>]
6. Google Cloud Vision OCR: A cloud-based OCR service by Google. It offers accurate text extraction, language support, and integration with other Google Cloud services. [Link: <https://cloud.google.com/vision>]

These OCR tools differ in terms of their features, accuracy, language support, and integration options. Selecting the right tool depends on specific requirements and preferences.