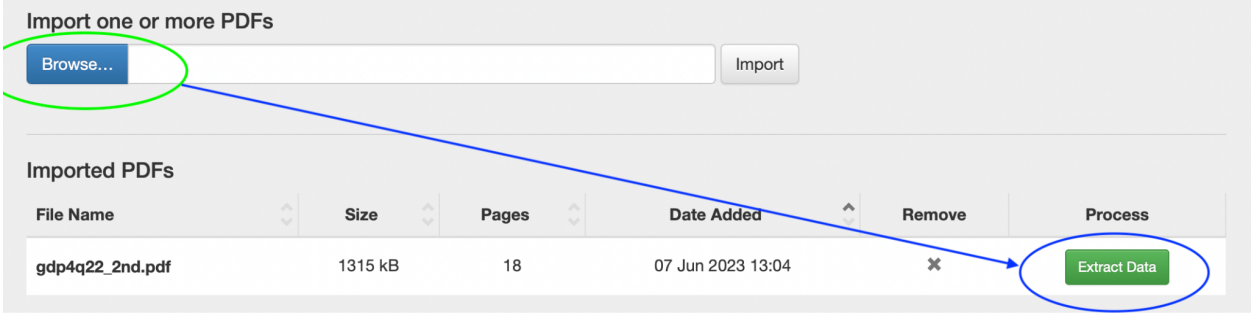
تبولا (Tabula)

تبولا (Tabula) یک ابزار قدرتمند برای استخراج داده‌های جدولی از فایل‌های پی‌دی‌اف است. این ابزار به شما امکان می‌دهد تا بخشی از فایل پی‌دی‌اف را که حاوی جدول است مشخص کنید و آن را به داده‌های ساختاریافته نظیر فرمت سی‌اس‌وی (CSV) یا جیسون (JSON) تبدیل کنید.

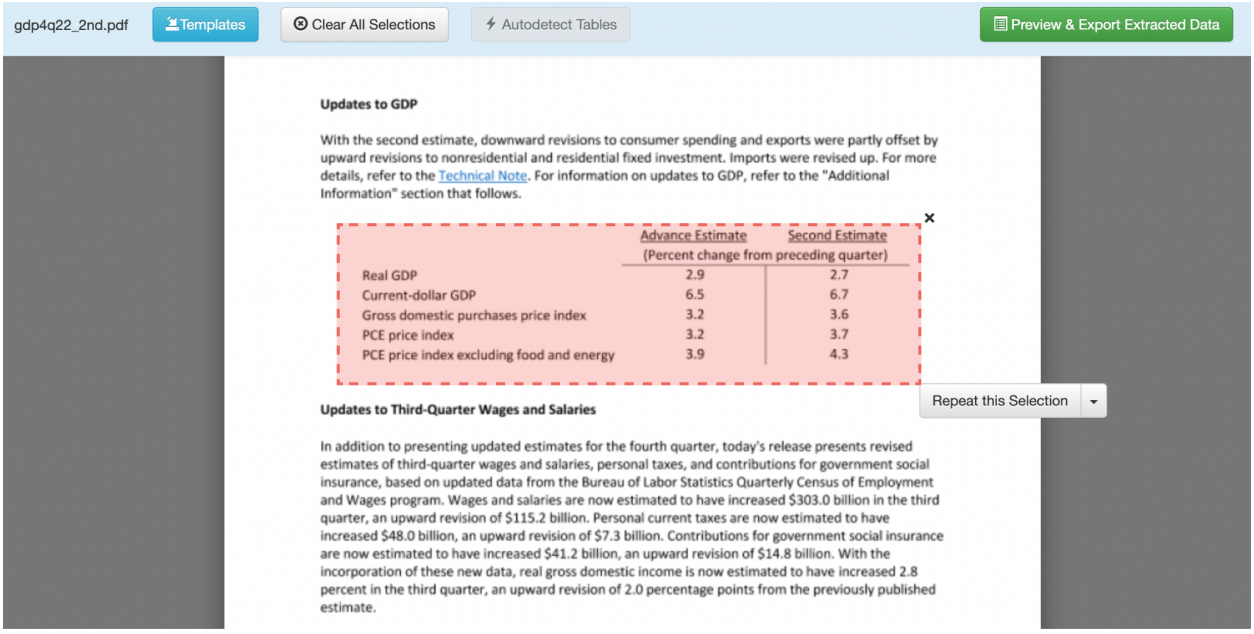
این یک راهنمای کلی برای نحوه استفاده از تبولاست:

۱. **نصب تبولا:** ابتدا تبولا را روی کامپیوتر خود نصب کنید. این کار را می‌توانید با رفتن به وبسایت تبولا و یا استفاده از ابزار کامندلاین (command-line) تبولا انجام دهید. توجه داشته باشید که تبولا ابزاری جاوا- محور است و به همین دلیل شما نیاز به نصب جاوا دارید.

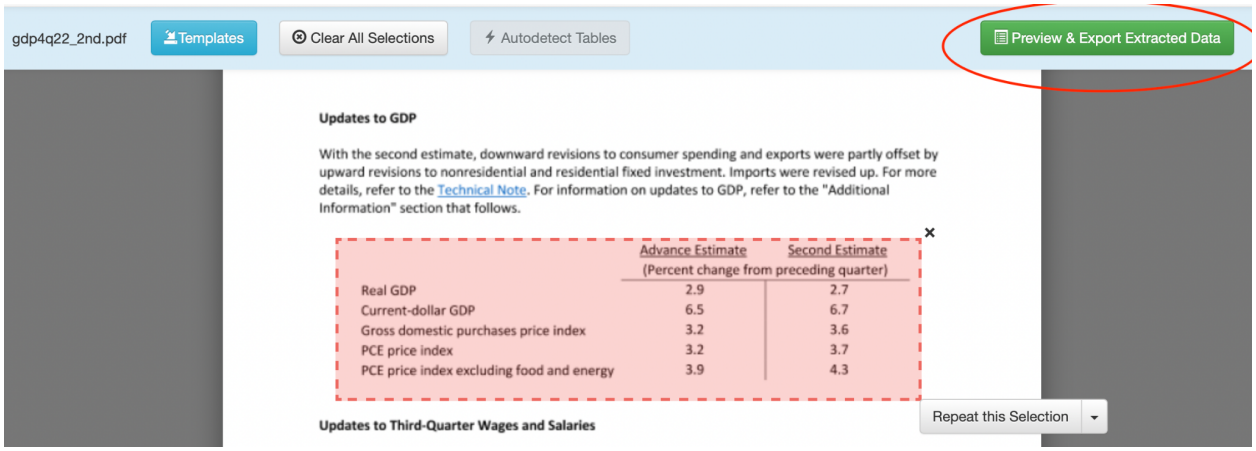
۲. **بارگذاری فایل‌های پی‌دی‌اف:** تبولا را باز کنید و فایل‌های پی‌دی‌اف خود را در آن بارگذاری کنید.



۳. **انتخاب بخش حاوی جدول:** با کمک نشانگر ماوس، جدول داخل فایل پی‌دی‌اف خود را انتخاب کنید. محل انتخاب شده باید فقط شامل جدول مد نظر باشد.



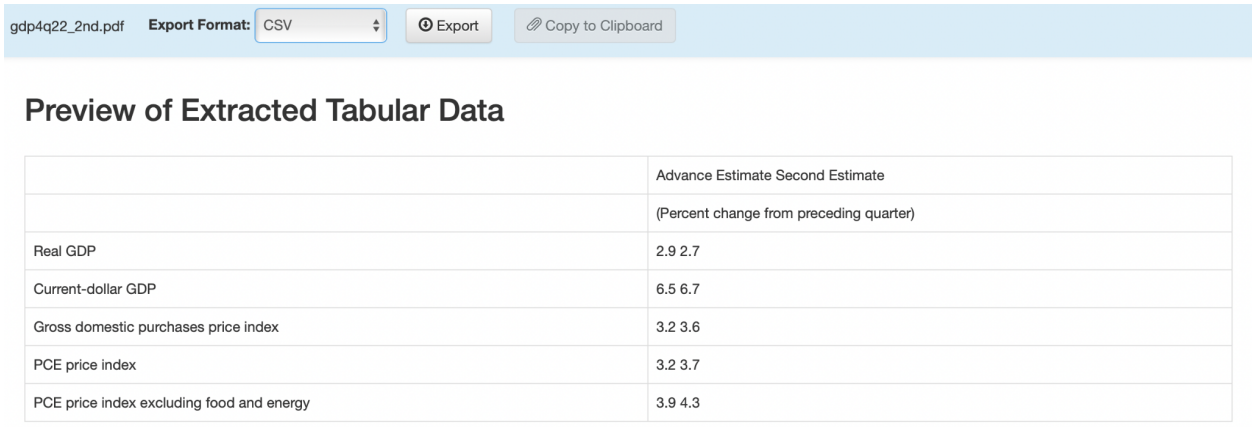
۴. بررسی جدول: پس از انتخاب جدول مدنظر، تبولا داده‌های استخراج شده را به صورت پیش‌نمایش به شما نشان می‌دهد. مطمئن شوید که این داده‌ها حاوی تمامی اطلاعاتی باشد که مد نظرتان است.



۵. تغییر محل تعیین شده جدول در صورت نیاز: اگر جدول مد نظرتان را به درستی انتخاب نکرده‌اید و بخش‌هایی از داده استخراج نشده‌اند، و یا جدول شما در چند صفحه قرار دارد، می‌توانید کار انتخاب جدول و تغییرات مد نظرتان را به صورت دستی انجام دهید.

۶. تعیین فرمت خروجی داده: تبولا به شما این امکان را می‌دهد تا برای فایل خروجی خود فرمت‌های مختلفی انتخاب کنید؛ فرمت‌هایی نظیر سی‌اس‌وی، جیسون یا تی‌اس‌وی (TSV).

۷. استخراج داده: حال می‌توانید با کلیک روی گزینه Export یا Extract، داده خود را استخراج و ذخیره کنید. در این‌جا می‌توانید محل ذخیره را نیز تعیین کنید.

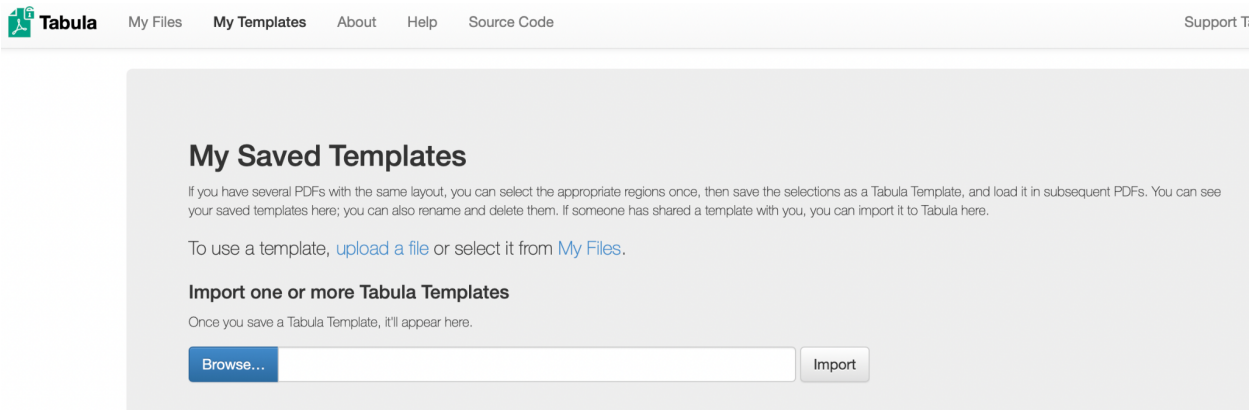


۸. راستی‌آزمایی و پاکسازی داده‌های استخراج شده: پس از استخراج داده‌ها، آنها را بررسی کنید تا از دقیق و کامل بودن آن مطمئن شوید. در صورت لزوم داده‌ها را پاکسازی کرده و سطر و ستون‌های بلااستفاده را حذف و یا داده‌های بیشتری را به این مجموعه داده اضافه کنید.

هنگام استفاده از تبولا، حواستان به موارد قانونی و محدودیت‌های کپی‌رایت مرتبط با فایل‌های پی‌دی‌اف مد نظرتان باشد و مطمئن شوید که اجازه استخراج داده از فایل‌ها را دارید.

موضوعات پیشرفته

استفاده از یک قالب (Template) مشخص در صفحات مختلف
در مواردی، فایل‌های پی‌دی‌اف از یک ساختار جدولی یکسان در تمامی صفحات خود استفاده می‌کنند. در این موارد، تبولا به شما این امکان را می‌دهد تا قالب جدول را در یک صفحه تعیین و آن را روی صفحات دیگر اعمال کنید. با این‌کار می‌توانید در زمان صرفه‌جویی کنید.



نویسه‌خانوی نوری یا اُسی‌آر (OCR)

زمانی که از یک سند عکس می گیرید یا آن را اسکن می کنید، کامپیوتر آن را نه به صورت کلمات و متن، که به صورت تصویری حاوی پیکسل می‌بیند. اُسی‌آر، این تصاویر را تجزیه و تحلیل می کند و از الگوریتم های پیشرفته برای تشخیص شکل حروف، اعداد و سایر کاراکترهای تصویر استفاده می کند. اُسی‌آر سپس این اشکال را به متن واقعی تبدیل می کند تا کامپیوتر بتواند آن را درک و با آن کار کند.

درک اُسی‌آر با دانستن نحوه استفاده صحیح از تبولا مرتبط است. تبولا ابزاری است که به طور خاص برای استخراج داده‌های جدولی از فایل‌های پی‌دی‌اف طراحی شده است. با این حال، تبولا محدودیت‌هایی هم دارد. اگر یک فایل پی‌دی‌اف مبتنی بر تصویر باشد، به این معنی که متن در پی‌دی‌اف به شکل تصویر ذخیره شده باشد و قابل انتخاب نباشد، تبولا نمی‌تواند مستقیماً داده‌ها را استخراج کند. به عبارتی تبولا تنها با فایل‌های پی‌دی‌افی کار می‌کند که قابلیت انتخاب متن دارند.

در چنین مواردی، قبل از استفاده از تبولا، باید از یک نرم افزار یا ابزار اُسی‌آر برای تبدیل فایل پی‌دی‌اف مبتنی بر تصویر به فایل پی‌دی‌اف مبتنی بر متن قابل جستجو و انتخاب استفاده کنید. پس از آن که فایل پی‌دی‌اف مورد اُسی‌آر قرار گرفت و متن آن شناسایی و تشخیص داده شد، می توان از تبولا برای استخراج جداول از متن تبدیل شده استفاده کرد.

برای این که تشخیص دهید آیا برای استخراج متن به اُسی‌آر نیاز دارید یا نه، چند راه سریع وجود دارد:

۱. **متن قابل جستجو**: اگر نمی توانید کلمات یا عبارات خاصی را در یک سند یا تصویر با استفاده از عملکرد جستجوی استاندارد جستجو کنید، این بدان معناست که محتوا به عنوان متن شناخته نمی شود. اُسی‌آر می تواند این سند یا تصویر را به متن قابل جستجو تبدیل کند و به شما این امکان را بدهد تا اطلاعات خود را پیدا و استخراج کنید.

۲. **کپی و پیست کردن**: سعی کنید متن مورد نظر خود را انتخاب و کپی کنید. اگر نمی‌توانید متن را کپی کنید یا پس از کپی کردن متن، هنگام پیست کردن آن، متن، بی‌معنا و حاوی تصاویری بی‌معنی است، معنایش آن است که محتوای شما در قالب متنی نیست. در این حالت نیز اُسی‌آر می تواند محتوا را به متن قابل ویرایش و استخراج تبدیل کند.

مثال:
جدولی که قابل تجزیه باشد: جدول صفحه ۱۴۷۰ این متن قابلیت تجزیه دارد. توجه کنید که شما می‌توانید این جدول را داخل متن انتخاب کنید.

جدولی که قابل تجزیه نباشد (مبتنی بر تصویر است): تصویر شماره سه صفحه ۱۴۶۹ همان متن اما، قابل انتخاب نیست و برای استخراج محتوای آن به اُسی‌آر نیاز است. توجه داشته باشید که بخش زیادی از این محتوا به طور کلی قابلیت تبدیل شدن به متن را ندارند و تصویر و نقشه هستند.

برای اوسی‌آر ابزارهای بسیاری وجود دارد که این موارد نمونه‌هایی از آن هستند:
۱. Tesseract: یک سیستم متن‌باز تهیه شده از سوی گوگل است که با چندین زبان کار می‌کند و خدمات همگانی گسترده‌ای دارد.
۲. OCRopus: یک سیستم متن‌باز تهیه شده از سوی گوگل است که شامل Tesseract و اجزای دیگری برای تجزیه و تحلیل طرح، پس پردازش و موارد دیگر است.
۳. CuneiForm: یک سیستم متن‌باز تهیه شده از سوی Cognitive Technologies است که قابلیت کار با زبان‌های مختلف را دارد و می‌تواند با متونی که طرح‌بندی پیچیده دارند، کار کند.
۴. Abbyy FineReader: یک ابزار جامع اُسی‌آر است که به دلیل دقت بالا و ویژگی های پیشرفته‌اش شناخته شده است. این ابزار دارای خدماتی نظیر آنالیز طرح بندی، پشتیبانی از زبان و گزینه‌های بهبود کیفیت تصویر است.
۵. Adobe Acrobat Pro: یک نرم افزار پی‌دی‌اف پرکاربرد با قابلیت اُسی‌آر داخلی است. این نرم‌افزار دارای قابلیت تشخیص متن، تبدیل سند و گزینه‌های پس پردازش است.
۶. Google Cloud Vision OCR: یک سرویس اُسی‌آر مبتنی بر فضای ابری (کلاود) گوگل است که ضمن داشتن قابلیت استخراج دقیق متن، با زبان‌های مختلف کار می‌کند و قابلیت ادغام با سایر خدمات فضای ابری گوگل را دارد.

این ابزارهای اُسی‌آر از نظر ویژگی، دقت، پشتیبانی زبان و ادغام با هم تفاوت دارند و قیمت آنها نیز متفاوت است. انتخاب هر یک از این اُسی‌آرها به نیازها و اولویت‌های شما بستگی دارد.