فاز اول پروژه مبانی بازیابی

اعضای گروه: نرگس کریمیان، الهه رضاپناه، زهرا رستمی

کد ما برای استخراج اطلاعات ساختاریافته از یک سند PDF طراحی شده است و هدف آن سازماندهی اطلاعات به صورت سلسله‌مراتبی از عناوین اصلی، زیرعناوین، آیتم‌های لیست و محتوای متنی معمولی است و خروجی کد، داده‌ها را به شکل سلسله‌مراتبی و در قالب JSON ذخیره می‌کند.

مثلا در سند ما:

– عناوین اصلی مانند "Depressive Disorders" و "Major Depressive Disorder" به عنوان کلیدهای اصلی در JSON قرار می‌گیرند.

– زیرعناوین مانند "Diagnostic Criteria" و "Development and Course" به عنوان کلیدهای داخلی در هر عنوان اصلی قرار می‌گیرند.

– آیتم‌های لیست مانند لیست نشانه‌ها و معیارهای تشخیصی در قالب لیست‌های متنی زیر کلید مربوطه ذخیره می‌شوند.

– متن عادی در قالب رشته‌های متنی در کلیدهای مناسب ذخیره می‌شود.

ما در کد خود دو تابع اصلی داریم:

1. extract_text_elements_from_pdf(pdf_path) : این تابع PDF را می‌خواند و همه عناصر متنی و اندازه فونت آنها را استخراج می‌کند. داده‌ها در قالب یک لیست از دیکشنری‌ها ذخیره می‌شوند.

2. parse_text_elements(elements) : عناصر متنی را تجزیه و تحلیل می‌کند تا محتوا را بر اساس سلسله‌مراتب تعیین‌شده توسط اندازه فونت‌ها سازماندهی کند.

به طور کلی می‌توان گفت :

این برنامه با استفاده از اندازه فونت و علائم شماره‌گذاری موارد زیر را شناسایی می‌کند:

– عناوین اصلی و زیرعناوین: بر اساس اندازه فونت و نوع قالب‌بندی.

– آیتم‌های لیست: با تشخیص شماره‌گذاری‌های معمول مانند "۱."، "الف." و سایر علائم مشابه.

– محتوای معمولی: متنی که به عنوان عنوان یا آیتم لیست شناسایی نمی‌شود.

روند کد به این صورت است که:

1. شناسایی عناوین اصلی و زیرعناوین :

– عناوین اصلی توسط اندازه فونتی که به طور قابل‌توجهی بزرگ‌تر از میانگین است شناسایی می‌شوند و زیرعناوین نیز فونتی کمی بزرگ‌تر از میانگین دارند.

– عناوین و زیرعناوین زمانی نهایی می‌شوند که خطی با قالب متفاوت شناسایی شود.

2. شناسایی آیتم‌های لیست :

– آیتم‌های لیست از طریق شماره‌گذاری (مثلاً "۱." یا "الف.") یا حروف الفبا شناسایی می‌شوند. زمانی که یک آیتم جدید شناسایی شود، آیتم قبلی ذخیره شده و خط فعلی به لیست افزوده می‌شود.

3. پردازش محتوای عادی:

– متنی که به عنوان عنوان یا زیرعنوان شناسایی نمی‌شود، به عنوان محتوای معمولی مرتبط با بخش در نظر گرفته می‌شود و تا رسیدن به عنوان، زیرعنوان یا آیتم لیست بعدی در بافر جمع‌آوری می‌شود و سپس در داده‌های JSON خلاصه‌سازی می‌شود.

همچنین علت استفاده ما از این الگوریتم در کدمان این است که ، روش سلسله‌مراتبی امکان تجزیه و تحلیل انعطاف‌پذیر اسناد ساختاریافته را فراهم می‌کند که در آن اندازه‌های فونت و شماره‌گذاری نشانگر عناصر ساختاری هستند. این روش برای اسنادی که ساختار عناوین و زیرعناوین دارند کاربردی است.