OCR
فناوری تبدیل تصویر متن به خود متن
(به زبان ساده)

● **فرزاد دادرس***
کارشناس ارشد کتابداری و اطلاع‌رسانی و
مدّرس گروه ICT دانشگاه صنعتی مالک اشتر

چکیده
این مقاله دربردارندهٔ تجربه‌ای نوین در عرصهٔ فناوری‌های رقومی‌سازی مدارک[1] است. نویسنده به همراه گروهی از همکاران خود در عرصه استفاده کاربردی و مؤثر از این فناوری، تجربه‌ای چندساله را پشت‌سرگذاشته است امید می‌رود که بیان این تجربه مورد استفادهٔ پژوهشگران و دانشجویان نیز باشد. در این مقالهٔ داستانی، نخست مشکلاتی که سازمان‌ها و مراکز اطلاع‌رسانی با آن روبه‌رو هستند، مورد بحث قرار گرفته تا نیاز به استفادهٔ بهینه از این فناوری برای همهٔ خوانندگان به بهترین نحو آشکار شود. وقتی مشخص شد که چرا به فناوری OCR نیازمندیم، به توضیح در مورد آن پرداخته‌ام. در این مسیر، علاوه بر ارائهٔ اطلاعات تخصصی به زبان ساده، فرایندهای مدیریتی این‌گونه مسائل را در قالب مدیریت پروژه‌های درون‌سازمانی، توضیح داده‌ام و سعی کرده‌ام تا بخشی از تجربیات کاربردی گذشته نیز در متن داستان به خوانندگان تقدیم شود.
کلیدواژه‌ها: او. سی. آر، تشخیص نوری، کارکترها، شناسایی نوری حروف

مقدمه

این مقاله را در دوره‌ای می‌نویسم که دیگر داد سخن دادن از برخی فناوری‌های نوین مانند دیجیتال‌سازی، می‌رود که به تدریج از مد و تب و تاب مقاله‌نویسان بیفتد. شاید یکی از دلایل از مد افتادن این موضوعات ـ در عرصه ارائه اطلاعات و دانش جدید ـ آن باشد که بسیاری از سخن‌پراکنان این نوشته‌ها، پس از برق‌انداختن واژه‌های قرضی، پر زرق و برق فناوری، قادر به حرکت با نیازهای عملیاتی سازمان‌ها و مراکزی که خواهان استفاده از این فناوری‌ها هستند نبوده و در عرصهٔ ارائه تجربه‌های کاربردی، حرفی برای گفتن و تجربه‌ای برای به‌کارانداختن ندارند.

به هرصورت، زمانی‌که یورش دانش جدید که نه، یورش واژگان جدید توسط مترجمان غیرحرفه‌ای و گاه ناشی اما شیک‌پسند، انبوه این واژگان را به حوزه‌های خاص سرازیر می‌کند، موضوعات زخمی‌شده و به حال‌خود رها می‌شوند؛ در چنین شرایطی، تولد حلقه‌های معیوب تولید اطلاعات و دانش جدید، چندان دور از انتظار نیست. هرچند ممکن است برخی بگویند که مسئولیتی برعهدهٔ مترجمان به لحاظ اجرایی‌کردن و تسلط عملی و کاربردی به مطالبی که ترجمه می‌کنند نیست، اما داشتن تسلط موضوعی وظیفهٔ اصلی مترجم است که در بسیاری از موارد دیده نمی‌شود. می‌خواهم بگویم که مقالهٔ علمی و پژوهشی، الزاماً نشانه‌ای از دانش و تسلط موضوعی پدیدآورنده نیست.

ناامیدانه هم سخن نمی‌گویم، چرا که در این میان، پژوهشگران و تولیدکنندگان حقیقی دانش هم بسیارند و باری را که مصرف‌کنندگان و نمایش‌دهندگان واژه‌ها بر زمین می‌گذارند، بر دوش گذاشته و به دست اهلش می‌سپارند؛ مسئولیتی که بر عهدهٔ سازمان‌ها و انجمن‌های تخصصی است تا دست کم با تشخیص این دو گروه از یکدیگر ـ و نه با دادن امتیازی بیشتر ـ توان‌شناختی خود را به رخ هر دو دسته بکشند!

حاصل آنکه این نوشته دربارهٔ کارها و واژه‌های شیک، که فرنگی‌ها تولید کرده‌اند داد سخن نداده است، بلکه تنها یک معرفی ساده، کوتاه و داستان‌گونه از یک فناوری پیشرفته است به پشتوانهٔ تجربیات کاربردی نویسنده و تنی چند از همکاران او بویژه در حوزهٔ OCR فارسی.

اما داستان...

من یک کتابدار هستم. هفته‌ای چند روز بعدازظهرها وقتم را در کتابخانه‌ای که اتفاقاً دوستم کتابدار آنجاست به مطالعه سپری می‌کنم. داستان مربوط می‌شود به ماجرایی که دوست‌کتابدارم برایم تعریف‌کرد. او یکی از تکنیکی‌ترین کتابدارانی است که تاکنون دیده‌ام و البته یکی از دلسوزترین آنها. او سرش برای رفع نیاز اطلاعاتی مراجعه‌کنندگان درد می‌کند، و اما آغاز داستان از زبان او:

همان‌طور که درکتابخانه نشسته بودم و برای رفع‌خستگی، گاهی

ما به شکلی از فرمت الکترونیکی منابع، مخصوصاً کتاب‌ها نیاز داریم که به سادگی قابل جست‌وجو باشد تا کاربر به سادگی و خیلی سریع بتواند موضوع یا موضوع‌های مورد نظر خود را در کتاب پیدا کند و حتی بتواند آن را برای خود چاپ‌کند یا به صورت فایل مورد استفاده قراردهد

به سقف قشنگ کتابخانه و گاهی به رفتار مراجعه‌کننده‌هایی که برای پیدا کردن کتاب‌ها، مقالات یا عکس‌های مورد نظرشان به کتابخانه آمده بودند نگاه می‌کردم، چشمم به قفسه‌ای افتاد که عده‌ای را دور خود جمع کرده بود. گویا موضوع مربوط می‌شد به چند جلد کتاب ارزشمند و قدیمی که تعداد استفاده‌کننده‌های آنها به طرز چشمگیری بیشتر از تعداد جلدهای موجود در کتابخانه بود. عجیب بود که روزهای بعد این صحنه را بارها و بارها در آن کتابخانه و چند جای دیگر هم دیدم. دوست‌کتابدار من هم به این جمع ملحق شد تا مسئله را حل و فصل کند. مسئله دقیقاً همان بود: کمبود منابع، تعداد انبوه استفاده‌کننده‌ها و مشکل اساسی‌تر اینکه: تهیه این منابع به دلیل نایاب بودن امکان‌پذیر نبود.

کتابدار چه باید می‌کرد؟ در آن لحظه به کدام یک از استفاده‌کننده‌ها باید کتاب را امانت می‌داد؟ مسئلهٔ حسّاس و کلیدی دیگری نیز وجود داشت: بر سر کتاب بعد از آن همه استفاده چه می‌آمد؟ طبیعی است که بعد از مدتی چیزی به شکل و هویت کتاب از آنها باقی نمی‌ماند. شاید به این دلیل که ماهیت کتاب هم مثل شمع است؛ نور می‌افشاند و می‌میرد.

بعدها دوست‌کتابدار جریان را این‌طور برایم تعریف کرد:

پیش از هر تصمیمی، سعی‌کردم تا صورت مسئله را تا حد امکان به صورت روشن و بدون ابهام در ذهنم مرتب‌کنم و بعد هم آن را روی کاغذ بنویسم. این کار را به خاطر این کردم که کلک همیشگی ذهن را بلد بودم و چند بار این کلک را از ذهن خودم خورده بودم:

وقتی فکر می‌کنی چیزی را می‌دانی یا فهمیده‌ای، بهترین راه، پیش از منتقل کردن آن به دیگران، نوشتن آن مطلب است. مخصوصاً اگر پیچیدگی یا اهمیت خاصی داشته باشند. تازه در مرحلهٔ نوشتن است که متوجه می‌شوی کجاهای مسئله را نفهمیده‌ای یا برایت مبهم است و تلاش دوباره‌ات برای تجزیه و تحلیل مسئله باعث می‌شود که آن را بهتر درک‌کنی و در نتیجه بهتر بتوانی به دیگران منتقل‌کنی. بعد هم خدا را شکر می‌کنی که زودتر در مورد آن مطلب در جایی صحبت نکرده‌ای و حالا خوشحالی که می‌توانی حقّ مطلب را به خوبی ادا کنی.

پس صورت مسئله را نوشتم و برای درک کامل ابعاد مختلف آن به قدرکافی وقت گذاشتم و آن را چندین بار اصلاح و تکمیل‌کردم تا مطمئن شدم تمام زوایای آن برایم روشن شده و هیچ نقطهٔ ابهامی وجود ندارد. بعد به خودم گفتم که حالا می‌توانم مسئله را حل‌کنم یا آن را برای مشاوره با دیگران مطرح‌کنم. در نتیجه ضروری دیدم در مورد راه‌حل‌هایی که به ذهنم رسیده بود فکر کنم، سپس در صورت نیاز مشورت کنم.

راه‌حل‌هایی که به ذهن کتابدار رسید

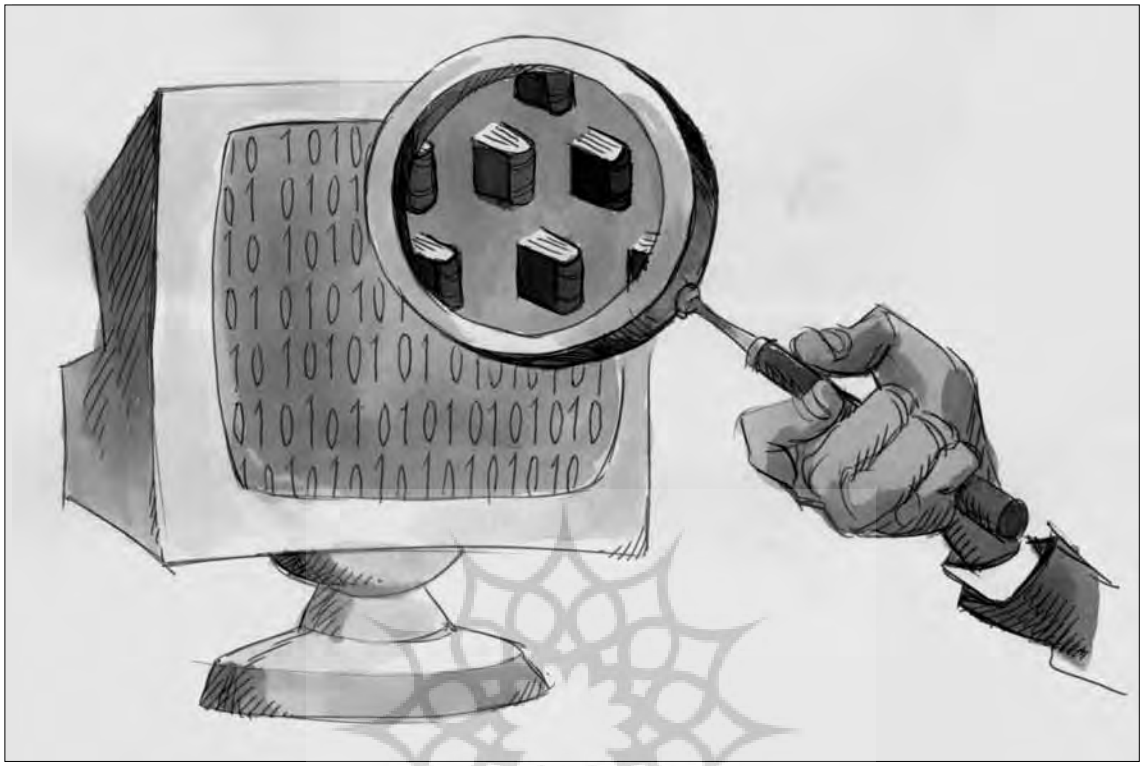
بدیهی بود که اولین راه‌حل باید می‌توانست سرعت چشمگیری در حل‌مشکل کتابخانه داشته باشد، چرا که به هیچ دلیلی نمی‌شود نیاز مراجعه‌کننده را معطل‌گذاشت. همچنین این راه‌حل باید همزمان مشکل فرسودگی کتاب‌ها به دلیل استفاده بیش از حد و مشکل ناکافی بودن تعداد کتاب‌ها را نسبت به تعداد زیاد مراجعه‌کننده‌ها حل می‌کرد.

در نتیجه جامع‌ترین راه‌حلی که در نگاه اول به ذهن او رسید، تهیهٔ شکل یا فرمت الکترونیکی کتاب‌ها بود. به این ترتیب، هم کتاب‌های نفیس از فرسودگی در امان می‌ماندند و هم می‌شد به تعداد کافی و با همان‌کیفیت و به سرعت تعداد مناسبی از آنها را تکثیرکرد. اما مشکل اینجا بود که این‌کار یک فرایند طولانی و پرهزینه بود. طولانی‌بودن آن استفاده‌کننده‌ها را مدتی از استفاده محروم می‌کرد و پرهزینه بودن آن (تهیه سخت‌افزار مناسب و ...) برای کتابخانه امکان‌پذیر نبود. پس باید از روش دیگری برای رسیدن به هدف اقدام می‌کرد و همان‌طوری که خودش بعدها تعریف‌کرد، گفت می‌خواهد تصور کند که کسی پیش از او این مشکل را داشته و این کار را انجام داده است و من نباید آن را دوباره انجام دهم و برای آن هزینه کنم.

در نتیجه دوست کتابدارم فکر کرد که شاااااااید – بخوانید شاید– کتابخانهٔ ملی بتواند در این زمینه به من کمک کند یا فهرستی از مراکزی که اقدام به ساخت الکترونیکی منابع مکتوب می‌کنند و یا منابعی را الکترونیکی کرده‌اند داشته باشد و شاید کتاب‌های مورد نظرشان در این فهرست‌ها موجود باشد و ما بتوانیم آنها را بخریم. بنابراین به سرعت با کتابخانهٔ ملی تماس‌گرفت و گفت که با مسئول بخش اشتراک منابع کاردارد.

مسئول بخش اشتراک منابع گفت که نسخه‌های الکترونیکی کتاب‌های مورد نظر ما در کتابخانهٔ ملی موجود نیست، اما کتابخانهٔ ملی می‌تواند با استفاده از تجهیزات سخت‌افزاری حرفه‌ای (اسکنر یا پویشگر)، کتاب‌های مورد نظر را به شکل الکترونیکی درآورده و در مقابل یکی از نسخه‌های الکترونیکی کتاب را در قبال این خدمت برای استفاده در آرشیو منابع الکترونیکی کتابخانهٔ ملی نگهداری کند.

در نتیجه، کتابدار زیرک ما توانست با هماهنگی کتابخانهٔ ملی و



در طـول چند روز تعطیلی که پیش‌آمده بود کتاب‌های مورد نظر را به کتابخانه ملی منتقل کرده و آنها را اسکن کند و شکل الکترونیکی آنها را به دسـت آورد؛ لازم به توضیح اسـت که کتابخانه ملی هیچ وقت تعطیل نیست .

مشکلات بعدی کتابخانه

مشکل اوّل حل شد. حالا هم مسئلهٔ فرسودگی منابع حل شده بود و هم مراجعه‌کنندگان به راحتی و بدون هیچ محدودیتی می‌توانسـتند به هر تعداد از نسـخه‌های کتاب‌های مورد نظر خود دسترسـی داشته باشند. حتی اگر ده‌ها و صدها مراجعه به این منابع صورت بگیرد، هیچ محدودیتی در تکثیر ثانیه‌ای این منابع وجود ندارد[۲].

اما مشـکل جدید، زمانی خود را نشـان داد کـه مراجعه‌کنندگان می‌خواسـتند به صفحات خاصی از کتاب خـود مراجعه کنند. به همین دلیـل، ناچار بودند که با اسـتفاده از صفحهٔ فهرسـت مندرجات کتاب و برداشـت شـمارهٔ صفحهٔ مورد نظر خود، صفحات کتاب را ـ که حالا تبدیل به عکس شـده بودند ـ تک تـک ورق بزنند تا به مطلب خود برسند؛ چه کار طولانی و پر دردسری!

این بار که اسـتفاده‌کننده‌ها متوجه حضور یک کتابدار کارشـناس و باهوش در کتابخانه شـده‌بودند، خودشان سراغ او رفتند و مشکل را توضیح دادند. حل این مسـئله دیگر در حوزه تجربه و تخصص کتابدار نبـود. او حـالا لازم می‌دید که باب مشـورت را با دیگر کارشناسـان بازکند. او این طور ادامه داد:

مسـئله بسـیار روشـن بود و ما تجربه‌ای برای حل این مسـئله نداشـتیم و تنها چیزی که می‌دانسـتیم ـ و البته مهم‌ترین نکته بود ـ این بود که می‌دانسـتیم به چه خدماتی در مورد این مسئلهٔ خاص نیاز داریم، اما نمی‌دانستیم که چگونه می‌توانیم این خدمات را فراهم‌کنیم و ارائه بدهیم. نیاز دقیق ما این بود:

«ما به شکلی از فرمت الکترونیکی منابع، مخصوصاً کتاب‌ها نیاز داریم که به سـادگی قابل جست‌وجو باشد تا کاربر به سادگی و خیلی سـریع بتواند موضوع یا موضوع‌های مورد نظر خـود را در کتاب پیدا کنـد و حتی بتواند آن را برای خود و چاپ‌کند یا به صورت فایل مورد استفاده قراردهد، اما چگونه؟»

گوشی را برداشتم و شمارهٔ تلفن مشاور کتابخانه در فناوری‌های‌نوین را شماره‌گیری کردم.

مشاوران کتابخانه‌ها

در بسـیاری از مواقع، مشـاوران کتابخانه‌ها و مراکز اطلاع‌رسانی و بقیه سـازمان‌هایی که به نوعی درگیر استفاده از فناوری‌های جدید برای بهبود مسـائل و مشکلات خود می‌باشند، افراد جوان، تندذهن و کارآزموده‌ای هسـتند که تجربه‌های کارآمد و اثربخشـی را در حوزه دانش مربوط به فناوری و پیاده‌سـازی آن کسب‌کرده‌اند. پس طبیعی بود که صدای جوانی از آن طرف خط به کتابدار کتابخانه پاسـخ داده باشد. دوست کتابدارم گفت که مسئله را به دقت و کاملاً روشن برای مشاور تعریف‌کردم و قرار شد تا در اولین فرصت در کتابخانه جلسه‌ای را برای تبادل نظر و بررسی راهکارها برگزارکنیم.

راهکارهای دیگر فناوری از زبان مشاور کتابخانه

مشـاور کتابخانه هم یک کتابدار مسـلط به مسـائل فناوریهای نویـن و مرتبط به کتابخانه و مراکز اطلاع‌رسـانی و نیز یک تحلیلگر کارآزمودهٔ سیستم بود. مشاورجوان، مطالب خود را این‌گونه آغاز کرد:

با توجه به اینکه کتابداردقیق و موشکاف کتابخانه، صورت مسئله

بخش رقومی‌سازی هر مرکز یا کتابخانه قبل از اقدام به شروع پویش و رقومی‌سازی‌مدارک خود، باید بر حسب شرایط، امکانات و توانایی‌های خود، فرایند هدفمند و قابل پیاده‌سازی را برای سازمان خود طراحی و بومی‌سازی کند. در این فرایند، مراحل و چگونگی انجام کارها، فرایندهای کنترلی و نیروی انسانی مورد نیاز به دقت تدوین شده‌اند و تازه بعد از این مراحل، آغاز کار منطقی است

را بسیارحرفه‌ای، جامع و روشن تعریف‌کردند، ما می‌توانیم بدون اتلاف وقت، برای ایجاد یک برداشت یکسان از مسئله، به سرعت وارد بحث اصلی یعنی راهکارهایی بشویم که فناوری پیش روی ما گذاشته است. اما پیش از شروع بحث مایل هستم تا دیگر مزایا و معایب استفاده از منابع پویش شده را توضیح دهم تا ضرورت نیاز به راه حل جدید، بیشتر برای شما روشن و قطعی شود و بعد برویم سراغ راه حل اصلی.

همان‌طور که خودتان هم به این نتیجهٔ درست رسیده بودید، فناوری‌های رایجی که برای تولید خروجی‌های الکترونیکی وجود دارند قادرند توسط پویشگرها صفحات چاپی را به تصاویری از متن تبدیل‌کنند. بایدگفت که این حرکت از اولین قدم‌ها برای ایجاد فرایند رقومی‌سازی منابع در کتابخانه است. گفتم از اولین قدم‌ها و نه اولین قدم، زیرا اولین قدم، طراحی و تدوین خود فرایند در یک محیط انتزاعی است؛ حال می‌خواهد این فرایند مربوط به رقومی‌سازی مدارک باشد و یا هر فرایند دیگری، به هر حال طراحی آن قبل از اجرا ضروری است. بخش رقومی‌سازی هر مرکز یا کتابخانه قبل از اقدام به شروع پویش و رقومی‌سازی مدارک خود، باید بر حسب شرایط، امکانات و توانایی‌های خود، فرایند هدفمند و قابل پیاده‌سازی را برای سازمان خود طراحی و بومی‌سازی کند. در این فرایند، مراحل و چگونگی انجام کارها، فرایندهای کنترلی و نیروی انسانی مورد نیاز به دقت تدوین شده‌اند و تازه بعد از این مراحل، آغاز کار منطقی است.

طبیعی است که مجریان و ناظران طرح، باید دقیقاً مطابق فرایندهای طراحی شده، اجرای کار را دنبال کنند. در این صورت، ضریب اطمینان رسیدن به اهداف از پیش تعیین شده بسیار بالا رفته و احتمال انحراف و یا خطا در اجرا بسیار کاهش می‌یابد. این ضریب افزاینده در کیفیت مدیریت و اجرا، موجب می‌شود تا سازمان در هزینه‌های خود صرفه‌جویی کرده و حتی یک شیب‌کاهنده ایجاد کند و مهم‌تر اینکه به دلیل بدانجام شدن کارها ناچار به صرف هزینه‌های اضافی یا پذیرش نقص در خدمات‌رسانی نباشد. این، درحالی است که در بیشتر سازمان‌ها و مراکز و کتابخانه‌ها، درست عکس این داستان اتفاق می‌افتد: اول کار انجام می‌شود، بعد روی آن فکر می‌کنند! در این‌گونه سازمان‌ها، پروژه‌هایی که با فناوری ارتباط دارند، با اتکا به دانش ذهنی و تجربیات افرادی معدود انجام می‌شوند و هیچ رویکرد روشمندی که نگاهی نظام‌مند به مسائل داشته باشد و مدیران را ملزم به مکتوب‌کردن استراتژی‌های انجام کارها کند وجود ندارد. با این توضیحات، مروری خواهیم داشت بر دیگر مزایا و معایب روشِ انتخاب شده:

مزایا:

- امکان ایجاد دسترسی کامل کاربران به منابع مورد نیاز(بدون محدودشدن به متغیرهایی مانند تعداد درخواست‌کننده‌ها)
- امکان حفظ و نگهداری بهتر مدارک اصلی
- تکثیر آسان
- امکان راه‌اندازی خدمات اشتراک منابع الکترونیکی
- امکان راه‌اندازی خدمات سفارش و دریافت الکترونیکی

معایب:

- محدودیت چشمگیر در تهیه رسانه‌های ذخیره‌سازی مدارک پویش‌شده
- عدم امکان جست‌وجوی متن، بویژه در منابع فارسی

در مورد مزایا به طور مشخص، می‌توان گفت که می‌توانید خدمات جدیدی را برای کاربران به فهرست خدمات پیشین خود بیفزایید. اما در مورد معایب و مشکلات این روش:

مشکل اول:

حجم فایل‌های‌تصویری از حجم فایل‌های متنی بیشتر است. این مسئله موجب می‌شود تا سازمان شما برای ذخیره‌سازی مدارک در حجم‌انبوه، هزینه‌های سنگینی بپردازد و این هزینه‌ها برای بسیاری از کتابخانه‌ها، مراکز اسناد و سازمان‌هایی که دارای حجم بالایی از مدارک هستند، مانعی جدی بر سر راه رقومی‌سازی و الکترونیکی کردن مدارک تلقی می‌شود.

مشکل دوم:

مشکل دوم را شما خودتان به‌خوبی و با دقت‌کافی توصیف کردید. یک مشکل بسیارجدی که حتی در صورتی که کتابخانه شما یا هر سازمان دیگری تصمیم به پرداخت هزینه های ذخیره‌سازی بگیرد، در خدمات‌رسانی نهایی به کاربر با آن روبه‌روست و مشکلات بسیاری را برای کاربران ایجاد می‌کند: مسئله جست‌وجو در متون‌الکترونیکی.

باید بگویم که درحال حاضر، فناوری، یک راه حل مطلوب برای حل این مسئله ارائه کرده است؛ راه‌حلی که شما را قادر می‌سازد تا خروجی‌های الکترونیکی متنی و کم‌حجمی تولیدکنید که به‌راحتی مانند یک فایل متنی ساده قابل جست‌وجو و تکثیر باشد و حجم بسیار پایین‌تری نسبت به فایل‌های تصویری داشته باشد: فناوری OCR.

فناوری [۳] OCR

OCR[۴] فناوری تبدیل تصاویر گرفته شـده از متن به خود متن است. خروجی حاصل از این فناوری، دو اشکال بزرگ ذخیره و بازیابی مدارک را که در قسمت قبل به آن اشاره شد حل می‌کند.

نرم‌افزارهـای OCR قادرنـد تصاویر پویش‌شـده از متـن را به میلیون‌ها نقطهٔ نوری تبدیل کنند. به این ترتیب، نرم‌افزار می‌تواند یک درک محاسـباتی از صفحهٔ پویش‌شـده برای خود ایجـاد کند و نقاط سـفید صفحـه را صفر و نقاط تاریـک یا همان نوشـته‌های متنی را یـک فرض کند. حال، نرم‌افزار می‌تواند نقاط تاریک یا نوشـته‌ها را با الگوهایی از حروف که از قبل در نرم‌افزار ذخیره‌شده یا به آن آموزش داده شده، مقایسه کرده و حدس بزند که هر کدام از این سلسله نقاط تاریک شـبیه کدام حرف از حروف فارسـی یا انگلیسی است و آنگاه آن حـرف را در یـک فایل متنی قرار داده و بـرای تمام نقاط تاریک عکس، همین فرایند را ادامه می‌دهد تا متن کامل از تصویر به دسـت آیـد. طبیعتـاً این نرم‌افزار از نقاط سـفید یا صفر صفحـه برای ایجاد حاشیه‌هایی که در صفحات وجود دارند و همچنین برای فاصله‌گذاری میان کلمات استفاده می‌کند.

به این ترتیب، نرم افزار تصاویر رقومی‌شده را به یک تصویر قابل پردازش تبدیل می‌کند. حال، ما می‌توانیم هر چیزی را به سـادگی در متن جدید تولید شده جست‌وجو کنیم و مثلاً ببینیم که کلمه مورد نظر ما در چه جاهایی از متن به کار رفته و چند بار؛ کاری که به هیچ وجه نمی‌توانستیم با تصاویر ایجاد شده از متن انجام دهیم.

اگـر بخواهیم مروری بـر دیگر کاربردهای این فناوری داشـته باشیم، باید به کاربردهای زیر اشاره کنیم:

- آرشیو الکترونیکی متون مجلات
- آرشیو الکترونیکی پایان‌نامه‌ها و مقالات
- آرشیو اسناد و تصاویرمتنی

ویژگی همهٔ این آرشـیوها وجود قابلیت جست‌وجو و حجم پایین به واسطهٔ استفادهٔ از فایل‌های متنی است.

انواع OCR

OCR یـا روی متون‌تایپـی انجـام می‌شـود یـا روی متـون دسـت‌نویس. یعنـی یا باید متنـی را که از قبل تایپ و چاپ شـده و فایل‌متنی آن موجود نیست و ما فقط نسخه چاپی آن را داریم OCR کنیم یا متن‌های دست‌نویس را. متن‌های دست‌نویس هم دو دسته‌اند؛ یا گسسـته‌اند یا پیوسته. متن گسسته دست‌نویس، مانند فرم‌هایی که مثلاً برای نام‌نویسـی کنکور اسـتفاده می‌کنیم و شما باید حرف‌های نام و نام‌خانوادگی‌تان را در مربع‌های کوچک بنویسـید. منظور از متن پیوسـته دست‌نویس هم همان دست نوشته‌های خودمان است که هر روز در کلاس دانشگاه یا دفتر کارمان روی کاغذ می‌نویسیم.

OCR انگلیسی و فارسی

OCR انگلیسـی سال‌هاسـت کـه به بهره‌بـرداری رسـیده و درحال‌حاضر اسـتفاده‌های زیادی دارد. OCR فارسی در ایران دو سه سـالی اسـت که مورد توجه قرار گرفته و چند گروه تحقیقاتی و چند شـرکت کوشیده‌اند تا نرم‌افزار مستقلی برای این کار تولیدکنند، اما به دلایل متعدد، تاکنون نتوانسـته‌اند تا در حد مطلوبی از پس مشکلات

OCR فناوری تبدیل تصاویر گرفته شده از متن به خود متن است. خروجی حاصل از این فناوری، دو اشکال بزرگ ذخیره و بازیابی‌مدارک را که در قسمت قبل به آن اشاره شد حل می‌کند

OCR فارسی برآیند.

البته کارهایی که روی OCR گسسـته فارسـی انجام شده، به ظاهر در مراحل پایانی خود قرار دارد، ولی همان‌طور که اشـاره شـد، OCR پیوسـته فارسی روی متون تایپی هنوز سال های بسیاری کار دارد. البته در مورد این مسـئله اسـتثنای بسیار مهمی وجود دارد، چرا که یکی از شرکت‌های کم سـروصدایی که سال‌هاست روی OCR پیوسته متون تایپی فارسی کار کرده است به نتایج بسیار مهمی رسیده و در آیندهٔ نزدیک، خدمت بسـیار جدیدی را در عرصه رقومی‌سـازی مدارک به سـازمان‌ها ارائه خواهدکرد و من در قسمت های بعد به آن اشاره خواهم‌کرد.

مشکلات OCR فارسی

بسـیاری از مشکلات OCR فارسی به ویژگی‌های زبان فارسی مربوط می‌شـود. مثلاً ما در فارسی حروفی داریم که شباهت بسیاری بـه هم دارنـد مثل : «د» «ذ» «ر» «ز» «ژ» «ت» «ث». اعداد نیز دارای چنین شباهت‌هایی با یکدیگر هستند. مثلاً اعداد «۱» «۲» «۳» کـه تنها در یک دندانـه با هم تفاوت دارند و صفر در سیسـتم نوشتاری ما شبیه یک نقطه است. همچنین هر یک از حروف فارسی را می‌توان به اشکال مختلف نوشت. مثلاً حرف س در کلمات مختلف، یک‌جا س را با دندانه می‌نویسـند، یک جا بدون دندانه و دیگری س را می‌کشـد. حال، وقتـی این حروف متفاوت به یکدیگر می‌چسـبند، مشـکلات بیشتر می‌شوند. تشـخیص صحیح و دقیق حروف و اعداد فارسی و در نتیجه کلمات را دشوارکرده و ضریب خطای خروجی‌های نرم‌افزار OCR فارسی را افزایش می‌دهد.

مرور کلی

حـال می‌خواهیم مـروری کلی به فرایند OCR داشته‌باشـیم و مراحل آن را دقیق‌تر به‌خاطر بسپاریم. مراحل انجام OCR به شرح زیر است:

- بازشناسی‌حروف
- بازشناسی‌الگو
- بازشناسی‌دنباله‌ها
- مدل‌سازی و پردازش‌زبانی

- **بازشناسی‌حروف**

در این مرحله، نرم‌افزار مکان قرارگرفتن حروف را از دیگر اجزای

نرم‌افزارهای OCR قادرند تصاویر پویش‌شده از متن را به میلیون‌ها نقطهٔ نوری تبدیل‌کنند. به این ترتیب، نرم‌افزار می‌تواند یک درک محاسباتی از صفحهٔ پویش‌شده برای خود ایجاد کند و نقاط سفید صفحه را صفر و نقاط تاریک یا همان نوشته‌های‌متنی را یک فرض کند

صفحه (مثلاً خطوط کادر) تشخیص می‌دهد و حروف را از یکدیگرجدا می‌کند.

• بازشناسی‌الگو

در ایـن مرحله نرم‌افزار با تعدادی، شـرط می‌فهمد که حرف جدا شـده مثلاً «گ» هسـت یا نه؟ در حقیقت، در این مرحله «گ» جدا شـده توسـط نرم‌افزار با «گ» هـای نمونهٔ خود نرم‌افزار مقایسـه و بازشناسی می‌شود.

• بازشناسی‌دنباله‌ها

کار با تشـخیص صحیح حروف تمام نمی‌شـود. در این مرحله، نرم‌افزار باید بفهمد که این حرف فقط «گ»است یا به حروف دیگری نیز چسبیده‌اسـت؟ به این ترتیب نرم‌افـزار دنباله‌های‌حروف را نیز در کلمات تشخیص می‌دهد.

• مدل‌سازی یا پردازش‌زبانی

حروف به هم چسبیده که کلمات را درست می‌کنند، باید معنی‌دار شـناخته شوند. در این مرحله، بررسی می‌شود که چه کلماتی در زبان وجـود دارند و چه ترکیبی از کلمات مجاز هسـتند. برای تشـخیص ترکیب‌های مجاز یک کلمه یا معنی‌دار بودن کلمه، به تهیه بانک‌های اطلاعاتی نشـان اسـتاندارد نیاز است. در مراحل پیشـرفته‌تر ایده‌آل مدل‌سـازی‌گرامری و مدل‌سـازی معنایی که نشـان می‌دهد جمله‌ها از نظر دسـتوری و معنایی درسـت هسـتند یا نه انجام می‌شود. البته فرایندهای مرحلهٔ اخیر، بیشتر درمرحلهٔ آزمایشگاهی قراردارند.

فرایند های کنترل در نرم افزارهای OCR

برخـی از نرم‌افزارهـای OCR از امکانات خاصـی برای کنترل خروجی‌های خود اسـتفاده می‌کنند، زیرا در این سیسـتم‌ها همیشـه احتمال خطا هست.

• کنترل Reject

در این فرایندکنترلی نرم‌افزار آن دسـته از حروفی را که نتوانسته تشـخیص‌دهد، علامت‌گذاری و اعلام می‌کند. در این صورت، اپراتور نرم‌افـزار یا به صورت دسـتی و یا با اسـتفاده از نرم‌افزارهای کنترل املایی[5] اقدام به تصحیح خطاها می‌کند.

• کنترل Spell Checker

در برخـی دیگر از سیسـتم‌های OCR نرم‌افزارهـای کنترل و تصحیح املای‌کلمات قرار داده شـده‌اند. این نرم‌افزارها که درحقیقت باید یک بانک اطلاعات اسـتاندارد از کلمات یک زبان خاص، باشـند کلمات غلط را با نمونه‌های صحیح آنها در بانک اطلاعات تطبیق داده و شـکل صحیح کلمه را جایگزین نموده یا به اپراتور سیستم پیشنهاد می‌دهند. با قراردادن این کنترل‌ها اکنون نرم‌افزارهای OCR انگلیسی خروجی‌های مناسبی را ارائه می‌دهند، اما نرم‌افزارهایOCR فارسی که در داخل کشور طراحی و پیاده‌سازی شده‌اند، همچنان راه پیشرفت را طی می‌کنند و امید اسـت که تا چند سـال آینده بـه تولید انبوه و صنعتی برسند.

وضع کنونی OCR فارسی در ایران

با اینکه چند گروه تخصصی در صدد ایجاد نرم‌افزارهای سـاخت داخل هسـتند و هنوز راه طولانی تا رسـیدن به کیفیت مطلوب برای متون فارسـی تایپی پیوسـته وجود دارد، اما این مسئله موجب نشده اسـت تا گروهی از کارشناسان یکی از شرکت‌های داخلی برای دادن خدمات OCR فارسـی در انتظار تکمیـل نرم‌افزارهای داخلی برای آغاز سـرویس‌دهی بمانند. این گروه به نتایج چشمگیری در این زمینه رسـیده‌اند و به زودی اطلاع‌رسـانی عمومی خود را آغـاز می‌کنند و شـما هم می‌توانید از خدمـات OCR فارسـی در کتابخانه و مرکز اطلاع‌رسانی خود استفاده کنید.

نتیجه‌گیری مشاور

پس از تمام شدن صحبت‌های مشاور که بیشتر شبیه یک کارگاه آموزشـی بود، همه راضی به نظر می‌رسـیدند، چرا کـه تمام مطالب گفته شـده در جلسـه را آموخته بودند و علاوه بر آن، بینش و نگرش آنهـا در مورد فرایندهای رقومی‌سـازی و طراحی این فرایندها به‌کلی تغییر کرده بود. این مطالب را می‌شـد هم در مدیریت پروژه‌های فنی کتابخانـه و هم در کارهای اجرایی به کار برد. ضمن‌اینکه مطالب تازه اشتیاق خاصی را در مخاطبان ایجاد کرده بود.

سرانجام، قرار شد تا مشاورکتابخانه با حضور شرکتی که خدمات OCR فارسـی ارائـه می‌کرد، مدیـر کتابخانه و کتابـدار متخصص جلسـه‌ای بگـذارد و در آن راهکارهای اجرایی کار را بررسـی کنند، چرا که نیاز شـدیدی در این قسـمت از خدمات کتابخانه به کاربران و استفاده‌کنندگان آن احساس می‌شد.

شاید شـما هم به زودی تصمیم بگیرید که از خدمات ارائهٔ متن کامل با قابلیت جسـت‌وجوی متنی و با کمک فناوری OCR فارسی بهره‌مند شوید.

پی‌نوشت‌ها

* Dadrass_f@yahoo.com

1. Document Digitalization

۲. البته ممکن است مشـکلات مالکیت معنوی (کپی رایت) محدودیت‌هایی را در تکثیر منابع ایجاد کند.

3. Optical Character Recognition

4. Optical Character Reader

5. Spell Checker

منبع

۱. نشانی اینترنتی OCR : آموزش الفبای فارسی به رایانه.
http://www.iritn.com/index.php?action=show&type=news&id=6080