بسمه تعالی

وزارت ارتباطات و فناوری اطلاعات
پژوهشگاه ارتباطات و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)
ITRC

درخواست ارائه پیشنهاد (RFP)

پروژهٔ توسعه وب‌سرویس نویسه‌خوان نوری (OCR) برای زبان فارسی

زیرمجموعه: طرح جویشگر

نسخه: ۱,۰

آذرماه ۹۵

سطح دسترسی: عمومی

فهرست مطالب

مقدمه

یکی از پردازش‌های مهم برروی تصاویر دیجیتال، پردازش تصاویر متنی می‌باشد. متون موجود در تصاویر ممکن است به صورت نویسه‌ها یا کاراکترهای چاپی (با فونت‌های مختلف) و یا به صورت دست‌نوشته یا دست‌خط (handwriting) از افراد مختلف باشند. هدف از این پردازش‌ها تشخیص این متون در تصویر و تبدیل خودکار آنها به فایل‌های متنی قابل ویرایش می‌باشد. اگر تصویر ورودی شامل نویسه‌های چاپی باشد به این فرایند Optical Character Recognition (OCR) یا بازشناسی نویسه‌های نوری گفته می‌شود ولی اگر تصویر ورودی شامل یک دست‌نوشته باشد به این فرایند تشخیص دست‌نوشته یا تشخیص دست‌خط (Handwriting Recognition) می‌گویند.

توسعه سامانه‌های OCR در زبان فارسی با پیچیدگی‌ها و مشکلاتی نسبت به زبان‌های لاتین همراه است. به عنوان مثال از آنجا که در زبان‌های لاتین نویسه‌ها به صورت مجزا از هم نوشته می شوند کار شناسایی آنها بسیار ساده است اما در زبان فارسی ابتدا باید کلمات (که دارای کاراکترهای متصل هستند) به کاراکترهای مجزا تبدیل شوند (که خود فرایند مشکلی است) و سپس عمل شناسایی انجام گیرد. مشکل دیگر خط فارسی این است که حروف فارسی بسیار به هم شبیه بوده و گاه تفاوت آنها در حد یک نقطه است.

مراحل مختلف پردازش تصاویر متنی به صورت زیر است:

- دریافت تصویر سند و بهینه‌سازی آن
- تحلیل محتوای سند و ناحیه‌بندی خودکار آن
- شناسایی نواحی مختلف سند با توجه به نوع اطلاعات آن
- در صورت لزوم ویرایش متن شناسایی‌شده و غلط‌یابی آن
- تولید فایل خروجی به فرمت دلخواه کاربر (مثلا فایل MS-word)

در این پروژه سعی می‌شود که مشکلات مختلف پردازش زبان فارسی بر طرف گردیده و یک وب‌سرویس OCR با دقت و کارایی قابل قبول برای زبان فارسی توسعه داده شود.

۱ تعریف پروژه

در پروژه حاضر، یک سامانهٔ OCR با کیفیت بالا برای متون فارسی و به صورت وب‌سرویس مدنظر است. معمولاً ورودی سامانه‌های OCR به صورت تصویری از یک سند متنی است که با یک رزولوشن خاص اسکن شده است. هرچه این رزولوشن بیشتر باشد، تصویر سند باکیفیت‌تر بوده و جزئیات نوشته‌ها بیشتر مشخص است، در نتیجه فرایند تشخیص را می‌توان با دقت بیشتری انجام داد. لیکن ازآنجاکه معمولاً اسکن کردن یا عکس‌برداری از متون، با رزولوشن بالا، موجب افزایش حجم آنها، اشغال بیشتر حافظه و کندشدن ارسال آنها می‌شود، کاربران ترجیح می‌دهند که تصاویر را با رزولوشن پایین‌تری ذخیره نمایند.

در سامانه‌های OCR، مهم این است که سامانه بتواند فونت‌های مختلف را با سایزها و حالت‌های مختلف (تیره، ایتالیک و ...) شناسایی نماید. همچنین عملکرد سامانه در مواجهه با تصاویر با رزولوشن‌های پایین هم مهم است. تصاویر گرفته‌شده از متون ممکن است در شرایط نامناسب از لحاظ نور محیط و یا زاویه نسبت به صفحه گرفته شده باشند (مثلاً با دوربین گوشی‌های هوشمند) که عملکرد سامانهٔ OCR در این شرایط نیز مهم است.

برای زبان فارسی تاکنون چند سیستم OCR (البته برای فونت‌های خاص) توسعه داده شده است که در این پروژه توسعهٔ آنها مدنظر می‌باشد. منظور از توسعهٔ سامانه‌های OCR، افزایش دقت و کیفیت سامانه برای پوشش فونت‌های بیشتر و با سایزهای مختلف و همچنین تشخیص تصاویر با کیفیت پایین (از لحاظ رزولوشن، زاویه تصویر، شرایط نوری مختلف و ...) می‌باشد.

در این پروژه، ارائهٔ سامانهٔ OCR به صورت سرویس تحت وب مدنظر است که با استفاده از آن کاربر بتواند تصاویر متون چاپی موجود در حافظهٔ کامپیوتر خود و یا تصاویری را که با گوشی هوشمند خود از متون مختلف گرفته است، به راحتی آپلود کرده و متن قابل ویرایش معادل آن را در مدت زمان اندکی دریافت نماید.

۲ هدف پروژه

هدف از این پروژه توسعهٔ یک سرویس OCR تحت وب با دقت و کیفیت بالا برای زبان فارسی می‌باشد. با استفاده از این سامانه کاربر قادر خواهد بود تصاویر اسکن‌شده از متون چاپی و یا تصاویری که با گوشی هوشمند خود از متون مختلف گرفته است را ارسال کرده و متن قابل ویرایش معادل آن را دریافت نماید. سرویس OCR می‌تواند به عنوان یکی سرویس پرکاربرد، توسط موتورهای جستجوی بومی ارائه گردد. در نتیجه این سامانه باید بتواند نیاز کاربران را با تمرکز بر زبان فارسی پوشش دهد. با در نظر گرفتن نیازمندی‌های کاربران ایرانی، این سامانه باید بتواند دستاوردهای زیر را محقق نماید:

- پاسخگویی به نیازهای کاربران ایرانی برای دو زبان فارسی و انگلیسی
- پوشش فونت‌های رایج با سایزهای مختلف
- تشخیص تصاویر متنی حاصل از اسکنرها و دوربین‌های مختلف مانند دوربین گوشی‌های هوشمند
- تشخیص تصاویر متون چاپی با رزولوشن‌ها و کیفیت‌های مختلف
- پاسخگویی سریع و برخط به کاربران

۳ مراحل اجرا و شرح خدمات پروژه

پیشنهاددهنده می‌بایست مراحل و فازبندی انجام کار را با جزئیات کامل ارائه کند. توجه به موارد زیر در پیشنهاد ارسالی ضروری است:

موارد کلی

- درصورتی‌که پیشنهاددهنده بخشی یا تمام سرویس OCR را به صورت محصول در اختیار دارد، باید وضعیت محصول به طور دقیق ذکر شود. بیان جزئیات محصول باید شامل مشخصات کلیدی، دقت، کیفیت، تعداد کاربران، سرعت پاسخ‌گویی و معماری سخت‌افزار و نرم‌افزار مورد استفاده باشد.
- پیشنهاد ارسالی محدود به موارد مشخص شده در این مستند نمی‌باشد و در صورت داشتن ایده جدید و یا ویژگی کاربردی در صورت ارائه، ارزیابی و لحاظ خواهد شد.
- در پیشنهاد ارسالی می بایست نحوه بهینه‌سازی و ارتقاء کیفیت سامانه در حین اجرای پروژه ذکر گردد.
- نیازمندی‌های سامانه در تعامل با جویشگرهای بومی باید در پیشنهاد ارسالی مشخص شود.
- در پیشنهاد ارسالی می‌بایست حتی‌الامکان از منابع موجود کشور در بخش سخت‌افزار و نرم‌افزار استفاده شود.
- در پیشنهاد ارسالی می‌بایست برنامه توسعهٔ بازار و طرح تجاری نیز ارسال شود. در طرح تجاری نحوه همکاری با موتورهای جستجو، تسهیم درآمد و جذب کاربران ارائه گردد.

ویژگی‌های فنی

هدف اصلی این پروژه، توسعه سامانه OCR برای زبان فارسی با قابلیت شناسایی انواع فونت‌ها و شناسایی تصاویری متنی با کیفیت پایین می‌باشد. در این راستا پیشنهاد دهنده باید برنامه و استراتژی خود را در زمینه‌های زیر مشخص نماید:

- تهیه دادگان جامع از تصاویر متنی
- معماری سامانه OCR
- استراتژی سامانه در برخورد با نویزهای تصویر و تصاویر با کیفیت پایین
- ارائه سامانه به صورت وب‌سرویس

همچنین با توجه به اهمیت منابع و دادگان تولید شده در حین اجرای پروژه لازم است پیشنهاد دهنده در مدل کسب و کار سیاست خود را در خصوص امکان فروش، به اشتراک‌گذاری یا ارائه رایگان محصولات و خدمات منتج از پروژه دیده باشد.

در ادامه مشخصه‌ها و الزامات فنی مدنظر برای وب‌سرویس OCR مشخص شده است.

- الزامات فنی:
  - قابلیت تشخیص متون چاپی از حداقل ۱۰ فونت رایج
  - قابلیت تشخیص متون چاپی با سایزها و حالت‌های مختلف (تیره، ایتالیک)
  - قابلیت تشخیص تصاویر متنی با رزولوشن پایین (۱۵۰ DPI و پایین‌تر)
  - قابلیت تشخیص تصاویر متنی با کیفیت‌های پایین، در شرایط نوری مختلف و دارای زاویه و انحنا
  - قابلیت تشخیص تصاویر متونی حاصل از اسکنرها و دوربین‌های مختلف (مانند دوربین گوشی‌های هوشمند)
  - دقت متوسط ۹۵٪ برای بازشناسی نویسه‌های نوری برای فونت‌های مختلف و شرایط مختلف تصویر
  - قابلیت بازشناسی نویسه‌های فارسی و انگلیسی، اعداد و علائم نگارشی
  - طراحی و پیاده‌سازی وب سرویس و API برای کلیه بخش‌های سیستم
  - استفاده از سیستم‌های موازی و توزیع شده برای بازشناسی نویسه‌های نوری
  - قابلیت پاسخگویی به تعداد بالای کاربر همزمان
  - ذخیره‌سازی تصاویر ارسالی کاربران
  - ذخیره‌سازی و بایگانی متون بازشناسی‌شده کاربران
  - پوشش نیازهای غیرکارکردی نظیر سرعت پاسخ‌گویی، زمان پاسخ‌گویی به کاربران همزمان، سرویس‌دهی همزمان به کاربران، دسترس‌پذیری، تحمل‌پذیری در برابر خطا، واسط کاربری، امنیت سیستم در برابر نفوذهای احتمالی
  - امکان مانیتورینگ و مدیریت سیستم به منظور تنظیم، کنترل، و گزارش‌گیری از بخش‌های مختلف سیستم

- الزامات نگهداری:
  - ارائه خدمات پشتیبانی رایگان به مدت دو سال
  - نگهداری از سامانه به صورت رایگان به مدت دو سال

- واسط کاربری:
  - طراحی واسط کاربری مناسب برای ارسال تصویر و دریافت متن به فرمت دلخواه

مراحل اجرای پروژه

اجرای پروژه به مدت دوازده ماه در سه مرحله، فاز اول (سه ماه)، فاز دوم (سه ماه) و فاز سوم (شش ماه) انجام می‌گردد.

- فاز اول: سرویس اولیه شامل امکان ارسال تصاویر متنی با کیفیت خوب از ۵ فونت رایج و دریافت متن بازشناسی‌شده.

- فاز دوم: ارتقاء کیفی سرویس نویسه‌خوان نوری برای بازشناسی ۵ فونت جدید و با تصاویر دارای رزولوشن پایین (۱۵۰ DPI و پایین‌تر).
- فاز سوم: توسعه سرویس نویسه‌خوان نوری برای بازشناسی تصاویر متنی با کیفیت‌های پایین و با شرایط مندرج در بخش ویژگی‌های فنی.

خروجی‌های پروژه

- ارائه وب‌سرویس نویسه‌خوان نوری با دقت و شرایط ذکرشده در بند ویژگی‌های فنی.
- ارائه دادگان تصاویر متنی تهیه شده در حین انجام پروژه به همراه متن معادل هر تصویر (حداقل ۲۰۰۰ صفحه از ۱۰ فونت رایج).
- ارائه ابزارهای جانبی تهیه شده در حین انجام پروژه شامل مواردی مانند:
  - ابزار ناحیه‌بندی تصویر
  - ابزار بهسازی کیفیت تصویر

۴ روش آزمون کیفیت و صحت خروجی

معیارهای ارزیابی نیازهای کارکردی

برای سنجش کارایی سامانه‌های OCR معمولاً از دو معیار دقت (accuracy) و صحت (correctness) بازشناسی نویسه‌ها استفاده می‌شود. معیار دقت، نسبت نویسه‌های بازشناسی‌شدهٔ درست به کل نویسه‌ها را بیان می‌کند. معیار صحت نیز میزان درستی کل رشتهٔ بازشناسی‌شده را نسبت به رشته نویسهٔ صحیح بیان می‌دارد. فرمول محاسبه معیارهای دقت و صحت (که برحسب درصد بیان می‌شوند) به صورت زیر است:

دقت = (N-D-S) / N*۱۰۰

صحت = (N-I-D-S) / N*۱۰۰

که در روابط بالا N تعداد کل نویسه‌ها، D تعداد نویسه‌های حذف‌شده در بازشناسی، S تهداد نویسه‌های جایگزین‌شدهٔ اشتباه در بازشناسی و I تعداد نویسه‌های اضافه‌شده در بازشناسی می‌باشد.

معیار دیگر، نرخ خطای نویسه‌ها (CER[۱]) می‌باشد که نسبت نویسه‌های بازشناسی‌شدهٔ نادرست به کل نویسه‌ها می‌باشد و به صورت زیر محاسبه می‌شود:

CER(%) = ۱۰۰ – Accuracy(%)

معیارهای ارزیابی نیازهای غیرکارکردی

- دسترس‌پذیری
- کارایی
  - زمان پاسخ‌گویی
  - توانایی در پاسخ‌گویی به درخواست‌های همزمان
  - زمان پاسخ‌گویی به درخواست‌های همزمان
- آمار بازدید
  - متوسط تعداد کل کاربران در روز
  - متوسط تعداد کاربران یکتا در روز
  - متوسط تعداد درخواست‌ها در روز
- تحمل‌پذیری در برابر خطا
- امنیت

[۱] Character Error Rate