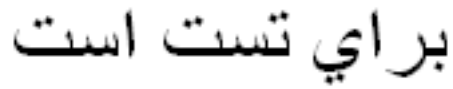
عنوان: **tesseract-ocr و پشتیبانی از زبان عربی**
نویسنده: وحید نصیری
تاریخ: ۱۳۹۰/۱۲/۲۷ ۱۰:۲۵:۰۰
آدرس: www.dotnettips.info
برچسب‌ها: OCR

tesseract-ocr
، یک OCR سورس باز توسعه یافته توسط شرکت HP در بین سال‌های 1985 تا 1995 است و اکنون شرکت گوگل کار نگهداری و توسعه آن‌را به عهده دارد. کیفیت نویسه خوانی انگلیسی آن فوق‌العاده بالا است. در آخرین نگارش آن پشتیبانی از زبان عربی هم را اضافه کرده است.
برای نصب آن ابتدا نگارش قابل حمل آن‌را دریافت و سپس فایل‌های مرتبط با زبان عربی را نیز باید دریافت کنید. پس از دریافت این‌دو، فایل‌های زبان عربی را در پوشه tessdata کپی کنید.

کار کردن با آن هم به سادگی اجرای فرمان زیر است:

```
tesseract.exe image.tif file -l ara
```

پارامتر اول نام تصویر، پارامتر دوم نام فایل متنی خروجی است (خودش یک txt را به صورت خودکار به فایل تولیدی اضافه می‌کند) و در آخر زبان عربی مشخص شده است.
برای نمونه تصویر زیر را



به صورت متن زیر نویسه خوانی کرد:

«برا ی ای ذست است»
فعلا ابزاری را برای ویرایش فایل‌های مرتبط با تشخیص زبان عربی ارائه نداده‌اند. بنابراین برای استفاده از آن جهت تشخیص متون فارسی مشکل وجود دارد چون «گچ پژ» را نمی‌تواند تشخیص دهد و به اینجا که می‌رسد کلا سیستمش به هم می‌ریزد.
انجمن پرسش و پاسخ آن هم در اینجا قرار دارد.

فایل‌های اجرایی و زبان عربی این برنامه را از آدرس‌های زیر هم می‌توان دریافت کرد:
Mirror: tesseract-ocr-3.01-win32-portable.zip & tesseract-ocr-3.01.ara.tar.gz

نظرات خوانندگان

نویسنده: حسین مرادی نیا
تاریخ: ۱۳۹۰/۱۲/۲۷ ۱۰:۳۴:۴۰

این طور که پیداست عربی رو هم خوب تشخیص نداده و خروجی به دست آمده جالب نیست!!!

نویسنده: Sarvari Dariush
تاریخ: ۱۳۹۰/۱۲/۲۷ ۱۱:۱۱:۴۰

.You are accessing this page from a forbidden country

.That’s all we know

نویسنده: وحید نصیری
تاریخ: ۱۳۹۰/۱۲/۲۷ ۱۱:۲۶:۴۸

Mirror اضافه کردم به انتهای متن اصلی.

نویسنده: رضا
تاریخ: ۱۳۹۱/۰۷/۲۶ ۱۹:۴۶

سلام پروژه ساخت باکس فارسی راه افتاده است اگر فرصت کردید به آنجا سر بزنید
https://github.com/reza1615/PersianOcr/wiki