

هووالعلیم

روش های بازشناسی سترون فارسی

دکتر مهدی خلیلی
مهندس مهسا نظری

عضو هیأت علمی دانشگاه پیام نور
عضو هیأت علمی دانشگاه پیام نور

سرشناسه	: خلیلی، مهدی، ۱۳۵۴ -
عنوان و نام پدیدآور	: روش‌های بازشناسی متون فارسی/ مهدی خلیلی، مهسا نظری.
مشخصات نشر	: تهران: تایماز، ۱۳۹۵.
مشخصات ظاهری	: ۱۰۴ص.: مصور، جدول، نمودار.
شابک	: 978-600-403-154-7
وضعیت فهرست نویسی	: فیبا
یادداشت	: کتابنامه: ص. ۹۲-۹۴.
موضوع	: نویسه‌خوانی نوری
موضوع	: الگوشناسی نوری
موضوع	: فارسی -- داده‌پردازی
شناسه افزوده	: نظری، مهسا، ۱۳۵۵ -
رده بنیاد کنگره	: ۱۳۹۵ ۸۹/خ/۸۱۶۴۰ TA
رده بندی دیویی	: ۰۰۶/۴۲۴
شماره کتابشناس ملی	: ۴۲۰۰۷۵۶

انتشارات تایماز

روش‌های بازشناسی متون فارسی

ناشر:	تایماز
مؤلف:	دکتر مهدی خلیلی - مهندس مهسا نظری
طراحی روی جلد:	مهندس اعظم ستایش
مدیر اجرایی:	مجید باشعور
چاپ و صحافی:	تایماز
نوبت چاپ:	اول ۱۳۹۵
شمارگان:	۳۰۰
قیمت:	۸۰۰۰۰ ریال
شابک:	۹۷۸-۶۰۰-۴۰۳-۱۵۴-۷

حق چاپ محفوظ و متعلق به مولفین می باشد.



آدرس: تهران - خیابان کارگر جنوبی - خیابان روانمهر - پلاک ۱۵۸ - طبقه ۳ صندوق پستی: تهران ۶۴۸-۱۳۱۴۵
تلفن: ۰۲۱-۶۶۴۰۲۷۴۴ موبایل: ۰۹۱۲۷۲۴۴۵۳۵ ایمیل: taymazpub@yahoo.com

فهرست مطالب

فصل اول: شناسایی الگو و بازشناسی متن..... ۱

مقدمه..... ۲

۱-۱) شناسایی الگو..... ۲

۲-۱) بازشناسی متن..... ۴

۱-۲-۱) روش سی نگ..... ۴

۲-۲-۱) روش قطعه بندی..... ۶

۳-۱) استخراج ویژگی..... ۷

۴-۱) قطعه بندی متن..... ۸

فصل دوم: بازشناسی حروف..... ۱۱

مقدمه..... ۱۲

۱-۲) بازشناسی نوری حروف..... ۱۳

۲-۲) انواع سیستم‌های بازشناسی نوری حروف..... ۱۷

فصل سوم: بازشناسی متن فارسی..... ۲۱

۱-۳) بازشناسی متن فارسی..... ۲۲

۲-۳) ویژگی‌های متون فارسی..... ۲۲

۳-۳) روش‌های جداسازی متون فارسی..... ۲۵

فصل چهارم: تحلیل آماری متون فارسی.....	۲۷
مقدمه.....	۲۸
۱-۴) بازشناسی برون خط متون فارسی.....	۲۸
۱-۱-۴) یک روش دو مرحله ای برای بازشناسی زیرکلمات چاپی.....	۳۱
۲-۱-۴) شناسایی دنباله زیر حروف متون چاپی فارسی با استفاده از قابلیت پردازش زمانی-مکانی شبکه عصبی با انتشار هدایت یافته.....	۳۵
۳-۱-۴) بازشناسی متن چاپی فارسی بر مبنای جداسازی هوشمند.....	۳۸
۴-۱-۴) بازشناسی کلمات دست نویس فارسی به کمک تبدیل قطبی - لگاریتمی و مدل جمع مارکوف.....	۴۰
۵-۱-۴) بازشناسی متون فارسی با استفاده از مدل زبانی <i>n-gram</i> و پالایش گرامری.....	۴۴
۶-۱-۴) استفاده از مناطق شش ریز کلمات چاپی فارسی برای کاهش فضای جستجو در بازشناسی آنها.....	۴۷
۷-۱-۴) خوشه یابی تصویر زیر کلمات در حروف قدیمی و حجیم چاپی با استفاده از معیار مقایسه تصویری.....	۴۹
۸-۱-۴) قطعه بندی عبارات متون فارسی با استفاده از شبکه های عصبی.....	۵۱
۲-۴) بازشناسی بر خط متون فارسی.....	۵۴
۱-۲-۴) روشی ساده برای بازشناسی برخط حروف مجزای فارسی.....	۵۶
۲-۲-۴) روشی ساده برای بازشناسی برخط زیر کلمات فارسی.....	۵۸
۳-۲-۴) بازشناسی برخط حروف فارسی بر پایه مدل مخفی مارکوف.....	۶۱

۴-۲-۴)	بازشناسی حروف برخط فارسی با استفاده از ویژگی‌های ساختاری.....	۶۳
۴-۲-۵)	بازشناسی برخط حروف مجزای دست نویس فارسی بر اساس تشخیص گروه بدنه اصلی با استفاده از ماشین بردار پشتیبان.....	۶۶
۴-۳)	بازشناسی ارقام فارسی.....	۶۹
۴-۳-۱)	کاربرد تطابق شکل در بازشناسی ارقام دست‌نویس.....	۶۹
۴-۳-۲)	تشخیص فازی اعداد دست‌نویس فارسی.....	۷۰
۴-۳-۳)	روش ترکیبی توسط توده ذرات دودویی برای بازشناسی ارقام دست‌نویس فارسی.....	۷۲
۴-۳-۴)	بازشناسی ارقام فارسی/عربی با استفاده از الگوی دودویی محلی.....	۷۳
۷۷	فصل پنجم: مقایسه روش‌ها.....	
۷۸	مقدمه.....	
۷۸	۱-۵) بررسی بازشناسی برون خط متون فارسی.....	
۸۶	۲-۵) بررسی بازشناسی برخط متون و حروف فارسی.....	
۸۸	۳-۵) بررسی بازشناسی ارقام فارسی.....	
۹۱	منابع.....	

مقدمه

نگهداری اسناد به شکل کاغذی در گذشته نه چندان دور در همه سطوح دولتی و شخصی رایج بوده است. امروزه با رشد فناوری‌های دیجیتال و استفاده از فناوری‌ها در نگهداری اسناد نیاز به تبدیل متن‌های دست نویس احساس شد. حتی اسنادی که به صورت تصویری نگهداری می‌گردند نیز نیاز به به روز رسانی پیدا می‌کنند که این امر نیز یک نیاز دیگر برای تبدیل این تصاویر به متن می‌باشد. به طور کلی، شناسایی متن از تصاویر دست نوشته ایی یا چاپی و تبدیل آن‌ها به قالب متنی را بازشناسی متن می‌نامند. بازشناسی متن شاخه ایی از شناسایی الگو در علم هوش مصنوعی می‌باشد. در اصل شناسایی الگو عملیات دسته بندی داده‌ها را که الگوهای ما هستند انجام می‌دهد. شناختن الگو با بهره گرفتن از اطلاعات آماری الگوها یا دانش قبلی آن‌ها را دسته بندی می‌کند. کاری که در بازشناسی متن نیز انجام می‌شود.

در بازشناسی متن با توجه به ویژگی حروف، دسته بندی صورت می‌گیرد. ورودی با این دسته‌ها مقایسه شده و نزدیک‌ترین دسته ورودی به عنوان پاسخ پذیرفته می‌شود. روش‌های مختلفی برای بازشناسی متن پیشنهاد داده شده‌اند که به طور کلی می‌توان آنها را به دو بخش کلی نگر و تحلیلی تقسیم نمود. در این کتاب سعی شده است روش‌های بازشناسی متون فارسی در پنج فصل، بصورت کامل مورد بررسی قرار گیرد. در فصل اول به موضوع «شناسایی الگو و بازشناسی متن» پرداخته شده و مباحث شناسایی الگو، بازشناسی متن، استخراج ویژگی و قطعه بندی متن شرح داده می‌شود. فصل دوم این کتاب در قالب «بازشناسی حروف» به بیان بازشناسی نوری حروف و انواع سیستم‌های بازشناسی نوری حروف پرداخته می‌شود. فصل سوم، در پوشش «بازشناسی متن فارسی»، موضوعات بازشناسی متن فارسی، ویژگی‌های متون فارسی و روش‌های جداسازی متون فارسی را مورد بررسی قرار می‌دهد. تحلیل آماری متون فارسی، شامل بازشناسی بیرون خط متون فارسی، بازشناسی بر خط متون فارسی و بازشناسی ارقام فارسی در فصل چهارم شرح داده

می شود. فصل پنجم نیز به مقایسه روش ها در قالب های بررسی بازشناسی برون خط متون فارسی، بررسی بازشناسی برخط متون و حروف فارسی و بررسی بازشناسی ارقام پرداخته می شود.

باید توجه داشت که بی شک، این کتاب خالی از کمبودها و نقائص نبوده و مولف از تمامی پیشنهادات و انتقادات خوانندگان عزیز که به نشانی پست الکترونیکی m.khalili@pniu.ac.ir ارسال شود، استقبال می نماید.

مهدی خلیلی - مهسا نظری

زمستان ۹۴