

سامانه استاندارد ساز و خطایاب

متون علمی فارسی

مؤلف

میرزا السادات حسینی بهشتی

عضو هیئت علمی پژوهش علوم و فناوری اطلاعات ایران (ایرانداک)

beheshti@iandoc.ac.ir



تهران ۱۳۹۷



کتابخانه ملی جمهوری اسلامی ایران

سرشناسه	: حسینی بهشتی، ملوک السادات، ۱۳۳۳ -
عنوان و نام پدیدآور	: سامانه استاندارد ساز و خطایاب متون علمی فارسی / مولف ملوک السادات حسینی بهشتی.
مشخصات نشر	: تهران: چاپار، ۱۳۹۷.
مشخصات ظاهری	: ۱۹۲ ص.: جدول، نمودار.
شابک	: 978-600-276-116-3
وضعیت فهرست نویسی	: فیبا
یادداشت	: کتابنامه.
یادداشت	: نمایه.
موضوع	: فارسی -- داده پردازی
موضوع	: Persian language -- Data processing
موضوع	: ویرایشگرهای متنی (برنامه کامپیوتر)
موضوع	: Text editors (Computer programs)
موضوع	: خطایابی املائی
موضوع	: Spelling errors
رده بندی کنگره	: ۹۷۰ اس ۵/ح ۲۶۴۳ PIR
رده بندی دیویی	: ۹۴۳
شماره کتابشناسی ملی	: ۹۴۳



تهران: ۸۸۸۹۹۶۸۰ - ۰۲۱۹ - ۰۹۰۲۱۹
 تهران - خیابان ولی عصر (عج) - بالاتر از زنجیرک
 کوچه نوریخس - پلاک ۴۰ - نشر چاپار
 سامانه استاندارد ساز و خطایاب متون علمی فارسی
 مؤلف: ملوک السادات حسینی بهشتی
 عضو هیئت علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایفای)
 صفحه آرا: رحیم کبیرصابر
 طراح جلد: سرور محمدی بزنج
 ناظر فنی چاپ: مریم تقضلی
 چاپ اول ۱۳۹۷
 شمارگان: ۱۰۰۰ نسخه
 شابک: ۳-۱۱۶-۲۷۶-۶۰۰-۹۷۸ ISBN: 978-600-276-116-3
 حق هرگونه چاپ و انتشار محفوظ است.
 قیمت: ۲۹.۰۰۰ تومان

چکیده

روزانه هزاران مستند متنی متنوع در حوزه‌های مختلف علمی بر روی وب جهان گستر قرار می‌گیرد. این مستندات می‌تواند شامل پایان‌نامه‌ها، مقاله‌ها، گزارش‌های علمی و مواردی از این قبیل باشد. نگارش متن این مستندات علمی جهت حفظ کتباختی باید بر اساس اصول ثابت انجام گیرد، اما همواره به‌طور غیر عمدی در خوش سلیقه‌های مختلفی در طول تاریخ می‌شود. اگرچه این تغییرات ناشی از پویابودن زبان و سلاقیات ذهن بشری است، اما این پویایی و خلاقیت پردازش ماشینی متن را با چالش‌های متعددی روبرو می‌کند و دقت پردازش داده‌ها را به میزان چشمگیری پایین می‌آورد. علاوه بر تنوع نگارشی، غلط‌های سهوی املائی نیز وجود دارد که فحوای گفته‌های متن را منحرف کرده و درک آن را با مشکل مواجه می‌کند. بنابراین، کلیه نویسه‌های متن باید به حالت استاندارد تبدیل شوند و عاری از هرگونه خطاهای املائی گردند. پژوهشگران طرح حاضر سامانه‌ای برای استانداردسازی و خطایابی متون علمی فارسی طراحی کرده‌اند که این سامانه متون نوشتاری علمی و تخصصی فارسی را به لحاظ صحت نگارشی و املائی بررسی می‌کند و متن را به شکل استاندارد در می‌آورد. این سامانه با معیار ۹۰.۰۸ و معیار بلوی ۹۱.۶۶ به استانداردسازی می‌پردازد و با معیار اف ۹۲ درصد را با رتبه‌بندی ۸۲ درصد به اصلاح خطای املائی می‌پردازد.

کلیدواژه‌ها: چالش‌های زبان فارسی، طبقه‌بندی روش‌ها، استانداردسازی نویسه‌ها، خطاهای املائی، پردازش ماشینی

فهرست مطالب

۱۷	 ۱. مقدمه
۱۹	 ۲. مطالعات نظری
۱۹	 ۱-۲. ویرایش
۲۰	 ۱-۱-۲. ویژگی‌های ویراستار
۲۰	 ۲-۱-۲. وظایف ویراستار
۲۰	 ۳-۱-۲. مراحل ویرایش
۲۴	 ۲-۲. بخش‌های ویرایش فنی
۲۴	 ۱-۲-۲. نشانه‌گذاری
۳۷	 ۳-۲. خط فارسی
۳۷	 ۱-۳-۲. تاریخچه خط فارسی
۳۸	 ۲-۳-۲. مشکلات خط فارسی
۳۹	 ۳-۳-۲. راهنمای شیوه خط فارسی
۵۳	 ۴-۲. آوانگاری
۵۵	 ۵-۲. عددنویسی
۵۵	 ۱-۵-۲. عددنویسی رقمی
۵۶	 ۲-۵-۲. عددنویسی حروفی
۵۶	 ۳-۵-۲. به هر دو شکل (حروفی-رقمی)
۵۶	 ۴-۵-۲. مختلط (ترکیب رقم و حروف)
۵۷	 ۶-۲. حروف ابجد و ابجدی در شماره‌گذاری
۵۸	 ۷-۲. فرمول‌نویسی
۵۸	 ۸-۲. سال‌نگاری
۵۹	 ۹-۲. کوتاه‌نوشت‌ها (نشانه‌های اختصاری)
۶۰	 ۱۰-۲. نویسه‌های زبان فارسی (دستور خط فارسی، فرهنگ املائی زبان فارسی)
۶۸	 ۱۱-۲. قواعد کلی دستور خط فرهنگستان زبان و ادب فارسی درباره کلمات مرکب

۶۹	۱۲-۲. مشکلات دستور خط فرهنگستان زبان و ادب فارسی.....
۷۳	۱۳-۲. جمع‌بندی.....
۷۵	۳. مطالعات عملی و کاربردی.....
۷۵	۱-۳. دادگان مورد نیاز برای برنامه‌های استاندارد سازی و خطایابی.....
۷۵	۱-۱-۳. پیکره.....
۷۷	۲-۱-۳. فرهنگ لغت.....
۸۰	۲-۳. روش‌های خطایابی و استاندارد سازی.....
۸۰	۱-۲-۳. روش فاصله حرف.....
۸۰	۲-۲-۳. ساماندهی کلمات.....
۸۱	۳-۱-۳. فاصله ویرایشی کمینه.....
۸۸	۴-۲-۳. آوا ده سازی.....
۸۹	۵-۲-۳. مدل انتقال نویزی.....
۸۹	۶-۲-۳. مدل‌های زبانی.....
۸۹	۷-۲-۳. مدل مبتنی بر یادگیری.....
۹۰	۸-۲-۳. مدل‌های آوایی.....
۹۶	۳-۳. روش‌های ارزیابی سامانه‌های خطایابی و استاندارد ساز.....
۹۷	۴-۳. معرفی سامانه‌های تجاری غیر فارسی.....
۱۰۸	۵-۳. معرفی پژوهش‌های انجام گرفته بر روی زبان فارسی.....
۱۰۸	۱-۵-۳. چالش‌های پردازش املائی زبان فارسی.....
۱۱۰	۲-۵-۳. مروری بر برجسته‌ترین پژوهش‌های خطایابی متون عمومی زبان فارسی.....
۱۲۵	۴. دادگان.....
۱۲۵	۱-۴. معرفی دادگان.....
۱۲۵	۲-۴. آماده سازی دادگان.....
۱۲۶	۱-۲-۴. پروتکل.....
۱۲۶	۲-۲-۴. شیوه آماده سازی.....
۱۲۷	۳-۲-۴. دقت آماده سازی.....
۱۲۹	۵. الگوریتم.....
۱۳۰	۱-۵. معرفی الگوریتم استاندارد سازی (ویرایش فنی).....
۱۳۰	۱-۱-۵. داده.....
۱۳۱	۲-۱-۵. ویرایش.....
۱۳۴	۲-۵. معرفی الگوریتم خطایابی املائی.....
۱۳۴	۱-۲-۵. داده.....

۱۳۴.....	۲-۲-۵. آماده‌سازی ساختار داده‌ها.....
۱۳۵.....	۳-۲-۵. اصلاح خطای املائی.....
۱۳۷.....	۶. فرایند کلی پردازش.....
۱۳۹.....	۷. ارزیابی.....
۱۳۹.....	۱-۷. استانداردسازی.....
۱۴۱.....	۲-۷. خطایابی.....
۱۴۳.....	۸. ورودی و خروجی هر بخش از سامانه.....
۱۴۳.....	۱-۸. زبان پیاده‌سازی.....
۱۴۳.....	۲-۸. پیش‌پردازش متن.....
۱۴۳.....	۳-۸. یادگیری ماشین و الحاق قواعد.....
۱۴۵.....	۴-۸. خطایابی املائی.....
۱۴۸.....	۵-۸. سامانه.....
۱۴۹.....	۹. نرم‌افزار پیرجور.....
۱۴۹.....	۱-۹. راهنمای تحریر متن.....
۱۴۹.....	۲-۹. بارگذاری متن.....
۱۵۱.....	۱۰. بحث و نتیجه‌گیری.....
۱۵۳.....	منابع.....
۱۵۷.....	الحاقیات: پروتکل.....
۱۸۵.....	نمایه.....