

عنوان مقاله:

طراحی و تولید مجموعه دادگان اخبار فارسی IHU-PersianNewsDataSet-Javadzade-et-
al دانشگاه جامع امام حسین (ع)

محل انتشار:

فصلنامه دستاوردهای نوین در برق، کامپیوتر و فناوری، دوره 2، شماره 3 (سال: 1401)

تعداد صفحات اصل مقاله: 19

نویسندگان:

حسین حسینی - دانشجوی کارشناسی ارشد دانشگاه جامع امام حسین (ع)

محمد قلعه نوئی - دانشجوی کارشناسی ارشد دانشگاه جامع امام حسین (ع)

محمد مهدی مختاری - دانشجوی کارشناسی ارشد دانشگاه جامع امام حسین (ع)

محمدعلی جوادزاده - استادیار دانشگاه جامع امام حسین (ع)

خلاصه مقاله:

اگر چه کمبود داده برای تحقیقات در حوزه پردازش زبان طبیعی یکی از چالش ها مهم است لیکن این چالش در خصوص زبان فارسی حادتر جلوه می کند، برای همین یافتن مجموعه دادگان باکیفیت و جامع در زبان فارسی کار دشواری است. علاوه بر آن دارا بودن برخی مشکلات از قبیل قابلیت دسته بندی و عدم رعایت استاندارد ذخیره سازی از نمونه مشکلات مجموعه دادگان موجود می باشد که هر کدام از این موارد می تواند بر میزان یادگیری مدل، نتایج و میزان خطا در آزمایش ها تاثیر بگذارد. به همین منظور تمامی این دلایل سبب شد که به دنبال جمع آوری و تهیه مجموعه دادگانی باشیم که تمام این گونه مشکلات را پوشش و میزان خطا هنگام به کارگیری داده ها در مدل های مختلف را کاهش دهد. ما در این پژوهش خزشگری را در جهت جمع آوری دادگان متنی طراحی و استفاده نموده ایم که با خزش بر روی یکی از پایگاه های خبری توانسته است مجموعه ای از دادگان را در پنج ستون عنوان، خلاصه، متن، برچسب و تاریخ انتشار خبر جمع آوری نماید. داده های متنی به کمک یکی از کتابخانه های مخصوص زبان فارسی در زبان برنامه نویسی پایتون، نرمال سازی شده و در دو فرمت CSV و XML ذخیره سازی شده و در اختیار پژوهشگران همکار قرار گرفته است. برچسب ها در این مجموعه داده شامل ۱۳ برچسب اصلی ورزشی، هنر و رسانه، فرهنگ، علم و پیشرفت، سیاسی، سیاست خارجی، زندگی، خانواده، جامعه، تعلیم و تربیت، بین الملل، اقتصادی و استان ها می باشد. از جمله کارهایی که بر روی این مجموعه داده قابل انجام است می توان به دسته بندی متن، استخراج متن، خلاصه سازی متن و تشخیص عنوان اشاره کرد. همچنین از ویژگی های بارز این مجموعه داده می توان به جامعیت، تعداد داده های مناسب، وجود ویژگی های مفید، دارا بودن ویژگی های منحصر به فرد و همچنین ذخیره سازی در قالب استاندارد اشاره کرد. این مجموعه داده محصول گروه پردازش زبان دانشگاه جامع امام حسین (ع) می باشد و از طریق لینک مذکور در پانویس صفحه بعد و با رعایت حق کپی رایت قابل دریافت و استفاده می باشد.

کلمات کلیدی:

مجموعه داده، اخبار فارسی، پردازش زبان طبیعی، مجموعه داده اخبار فارسی، یادگیری ماشین، دسته بندی متن، استخراج متن، خلاصه سازی متن، تشخیص عنوان

لینک ثابت مقاله در پایگاه سیویلیکا:

<https://civilica.com/doc/1521900>



