عنوان مقاله:
دسته بندی متون کتب فارسی در موضوعات مختلف با تکنیکهای متن کاوی و یادگیری ماشین

محل انتشار:
پانزدهمین کنفرانس ملی پژوهش های کاربردی در علوم برق،کامپیوتر و مهندسی پزشکی (سال: 1402)

تعداد صفحات اصل مقاله: 11

نویسنده:
فرانک خونساریان - *دانش آموخته کارشناسی ارشد مهندسی فناوری اطلاعات دانشگاه تربیت مدرس تهران*

خلاصه مقاله:
امروزه حجم زیادی از داده هایی که به صورت روزانه در جهان تولید میشود مربوط به داده های متنی یا همان داده های غیرساختیافته میباشد و سازمانها میتوانند از طریق تحلیل و پردازش این داده ها به اطلاعات با ارزش و دانش مفیدی برای بهبود فرآیند کسب و کار خود دست یابند. در واقع با استفاده از تکنیکهای متن کاوی یا پردازش متن میتوان با کشف الگوهای نهان موجود در داده ها و تبدیل آن به اطلاعات با معنا به سادگی مجموعه داده های بزرگ را تحلیل نمود.این پژوهش که با هدف دستهبندی متون موجود در کتب فارسی از منظر محتوای آنها به دسته های مختلف اقتصادی، تاریخی، روانشناسی، سلامتی، فنی و مهندسی و هنری که خود میتواند نمونه ای از داده هایی که در زمینه های مختلف تولید میشود میباشد انجام شده است. به این منظور ابتدا قسمتی از خلاصه مربوط به کتب در این دسته ها را در یک مجموعه گردآوری نمودیم سپس با استفاده از برخی از تکنیکهای متنکاوی متون فارسی به پیش پردازش و آماده سازی متون پرداختیم و پس از آن با استفاده از الگوریتم های یادگیری ماشین مانند ماشین بردار پشتیبان و گرادیان تقویتی به پیشبینی دسته بندی متون پرداختیم. در پایان دقت پیش بینی این روش ها را با یکدیگر مقایسه نمودیم که مشخص گردید که ماشین بردار پشتیبان عملکرد بهتری نسبت به الگوریتم گرادیان تقویتی دارد پس از آن ابر کلمات که نشاندهنده کلمات پرتکرار و مهم در هر دسته میباشد را نیز ترسیم نمودیم. میتوان دسته بندی متون فارسی را از طریق ساخت مدلهای یادگیری ماشین شناسایی نمود و کلمات مهم و پرتکرار موجود در آن ها را نیز تشخیص داد.

کلمات کلیدی:
داده های متنی، متنکاوی، متون فارسی، الگوریتمهای یادگیری ماشین، ابر کلمات