



مطالعه و بررسی یک متدلوژی ترکیبی مبتنی بر الگوریتم تشابه کسینوسی و مفاهیم کلاسه بندی برای دسته بندی اسناد متنی

سیروس شیخی^۱، اصغر درویشی^۲

۱- دانشجوی دکتری تخصصی نرم افزار دانشگاه آزاد اسلامی واحد تهران جنوب

SiroosSheikhi@gmail.com

۲- استادیار و عضو هیئت علمی گروه مهندسی نرم افزار دانشگاه آزاد اسلامی واحد تهران جنوب

A_Darvishy@azad.ac.ir

خلاصه

کلاسه بندی اسناد متنی یکی از پایه ای ترین و پرکاربردترین مفاهیم در علوم کامپیوتر می باشد. از این دست از کاربردها می تواند به جستجوی متون در بانک های اطلاعاتی متنی، سیستم های دسته بندی اخبار، شناسایی و فیلتر کردن هرزنامه ها، سیستم های تولید دانش و دیگر موارد کاربردی اشاره نمود. همچنین با افزایش روز افزون انبوه اسناد متنی مطالعه، بررسی و ارتقاء سطح کیفی الگوریتم های پردازش متن همچون کلاسه بندی اسناد متنی امری ضروری است. از این رو در این مطالعه سعی بر آن است تا شیوه ای ترکیبی از الگوریتم تشابه کسینوسی^{*} به همراه الگوریتم های کلاسه بندی همچون بیز ساده چند جمله ای (MNB[†]) و ماشین بردار پشتیبانی (SVM[‡]) جهت ارتقاء کیفیت الگوریتم های کلاسه بندی معمول برای استفاده در کلاسه بندی اسناد متنی مورد مطالعه و سنجش قرار گیرد. بهینه سازی و پیاده سازی متدلوژی مورد مطالعه در این تحقیق برای اسناد متنی فارسی به عنوان فعالیت های آتی مد نظر بوده و پیشنهاد می گردد.

کلمات کلیدی: کلاسه بندی متون، تشابه کسینوسی، بیز ساده چند جمله ای، ماشین بردار پشتیبانی

* Cosine Similarity

† Multinomial Naïve Bayes (MNB)

‡ Support Vector Machine (SVM)