Archive of SID

Using One-Class SVM for Scientific Documents Classification (Case study: Iranian Environmental Thesis)

Mohammad Rabiei
PhD Candidate in Information Technology Engineering; Iran University of Science and Technology Email: mo_rabiei@ind.iust.ac.ir

Seyyed Mahdi Hosseini Motlagh*
PhD. in Industrial Engineering; Associate Professor; Iran University of Science and Technology Email: motlagh@iust.ac.ir

Behrouz Minaei Bidgoli
PhD in Computer Science and Engineering; Associate Professor; Iran University of Science and Technology;
Email: b_minaei@iust.ac.ir

Iranian Journal of Information Processing and Management

Received: 16, Oct. 2018 **Accepted: 30, Dec. 2018**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 34 | No. 3 | pp. 1211-1234
Spring 2019

Abstract: The classification of research studies is important in order to identify and analyze the research supply and demand in various fields of science. In particular, the classification of environmental research is essential because of its importance in Iran and its interdisciplinary nature. This research proposes One-Class Classification (OCC) method to classify the research studies in this domain using Support Vector Machine (SVM) and consequently evaluates important parameters affecting the quality of this classification. The results show that the use of descriptive metadata has better performance than the content metadata in order to make a core data set to learn the model. Moreover, the use of the polynomial kernel and the binary weighing of words in the features vector matrix leads to better results than other states. In this paper a new weighing method has been proposed which is superior to the other methods especially in precision criterion. We call this weighing method as NG-TF, which can be used in term-document matrix to determine the indicator terms of scientific domains.

Keywords: Environment, One-Class Classification, Support Vector Machine (SVM), Text Mining, NG-TF Weighing

* Corresponding Author

www.SID.ir

Archive of SID

ارائهٔ روش رده‌بندی تک‌رده‌ای برای شناسایی متون پژوهشی حوزهٔ محیط زیست ایران با استفاده از ماشین بردار پشتیبان[1]

محمد ربیعی
دانشجوی دکتری مهندسی فناوری اطلاعات؛
دانشگاه علم و صنعت ایران mo_rabiei@ind.iust.ac.ir

سید مهدی حسینی‌مطلق
دکتری مهندسی صنایع؛ دانشیار؛
دانشگاه علم و صنعت ایران؛
پدیدآور رابط motlagh@iust.ac.ir

بهروز مینایی بیدگلی
دکتری مهندسی و علوم کامپیوتر؛ دانشیار؛
دانشگاه علم و صنعت ایران b_minaei@iust.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۷/۰۷/۲۴ پذیرش: ۱۳۹۷/۱۰/۰۹ مقاله برای اصلاح به مدت ۳۶ روز نزد پدیدآوران بوده است.

چکیده: رده‌بندی متون پژوهشی به‌منظور شناسایی و تحلیل عرضه و تقاضای پژوهشی در حوزه‌های مختلف علوم اهمیت ویژه‌ای دارد. در این میان رده‌بندی پژوهش‌های حوزهٔ محیط‌زیست به‌دلیل اهمیت فراوان آن در کشور و نیز میان‌رشته‌ای بودن آن ضروری است. این پژوهش روش رده‌بندی تک‌رده‌ای متون پژوهشی این حوزه را با استفاده از ماشین بردار پشتیبان ارائه می‌دهد و به ارزیابی پارامترهای مهم تأثیرگذار در کیفیت این رده‌بندی می‌پردازد. نتایج نشان می‌دهد که استفاده از مجموعهٔ دادهٔ هستهٔ توصیفی در یادگیری مدل، کارایی بهتری نسبت به هستهٔ محتوایی دارد. همچنین، استفاده از هستهٔ چندجمله‌ای و وزن‌دهی دودویی واژه‌ها در ماتریس بردار ویژگی‌ها نتایج بهتری نسبت به حالت‌های معمول دیگر ارائه می‌کند. در این مطالعه، روش جدید وزن‌دهی با نام NG-TF معرفی و ارائه شده است که نتایج ارزیابی آن نسبت به روش‌های دیگر، به‌ویژه در معیار دقت، برتری قابل‌توجهی دارد. از این رو، می‌توان از این روش وزن‌دهی برای تعیین واژگان نمایندهٔ یک حوزهٔ پژوهشی استفاده کرد.

کلیدواژه‌ها: محیط زیست، رده‌بندی تک‌رده‌ای، ماشین بردار پشتیبان، متن‌کاوی، وزن‌دهی NG-TF

فصلنامه
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۴ | شمارهٔ ۳ | ص‌ص ۱۲۱۱-۱۲۳۴
بهار ۱۳۹۸

۱. این پژوهش مستخرج از رسالهٔ دکتری آقای محمد ربیعی در رشتهٔ مهندسی فناوری اطلاعات در دانشگاه علم و صنعت ایران است.

www.SID.ir

Archive of SID

۱. مقدمه و پیشینهٔ پژوهش

رده‌بندی متن، تخصیص یک سند متنی به یک ردهٔ از پیش تعریف‌شده را می‌گویند (Sebastiani 2002). در این میان رده‌بندی متون پژوهشی به‌دلیل آن‌که کاربرد فراوانی در نمایه‌سازی متون داشته و نیز امکان تحلیل و رصد دارایی پژوهشی[1] یک جامعهٔ پژوهشی را برای پژوهشگران و سیاست‌گذاران فراهم می‌کند، دارای اهمیت ویژه‌ای است. تاکنون روش‌های خودکار یا نیمه‌خودکار مختلفی مبتنی بر یادگیری ماشین در این زمینه ارائه شده است. در این روش‌ها لازم است نمونه‌هایی از رده‌های از پیش تعیین‌شده به‌منظور یادگیری مدل رده‌بند در اختیار مدل قرار گیرد تا مدل بر مبنای الگویی که بیشترین تطابق را با مجموعه‌های یادگیری دارد به رده‌بندی نمونه‌های نهایی بپردازد (بصیری، نعمتی و قاسم‌آقایی ۱۳۸۶). با توجه به گستردگی و تنوع متون پژوهشی و نیز ظهور حوزه‌های جدید پژوهشی، شناسایی و یا تولید مجموعهٔ داده‌هایی که با جامعیت کامل بتواند در مرحلهٔ یادگیری ماشین مورد استفاده قرار گیرد، بسیار مشکل و در برخی موارد غیرممکن است. از این رو، لازم است از روش‌های دیگری برای یادگیری ماشین استفاده نمود.

به‌دلیل جغرافیای خاص ایران، محیط زیست به‌عنوان یکی از چالش‌های اساسی کشور همیشه مطرح بوده و بر همین اساس، از سال‌ها پیش، رشته‌ها و گرایش‌های مرتبط با آن در نظام آموزش عالی ایران ایجاد شده است. به همین خاطر، سالانه پژوهش‌های متعددی در این حوزه در کشور انجام می‌شود و به‌خاطر لزوم ارائهٔ پیوست‌های محیط زیستی در بسیاری از پروژه‌های کلان، همچنان نیازهای پژوهشی متعددی نیز در این حوزه مطرح می‌شود (Rabiei, Hosseini-Motlagh & Haeri 2017). ذات میان‌رشته‌ای بودن این حوزه از علم موجب شده است که رده‌بندی پژوهش‌های آن، که لازمهٔ تحلیل اطلاعات و شناخت از دارایی‌ای پژوهشی و نیازهای پژوهشی این حوزه است، پیچیده‌تر باشد. اهمیت بررسی و تحلیل پژوهش‌های حوزهٔ محیط زیست و نیز میان‌رشته‌ای بودن این حوزه موجب شده است که محیط زیست به‌عنوان نمونهٔ مورد مطالعه در این پژوهش مد نظر قرار گیرد. هدف این پژوهش ارائهٔ راهکاری برای رده‌بندی متون پژوهشی است و هرچند که حوزهٔ محیط زیست به دلایل بیان‌شده به‌عنوان مورد مطالعه انتخاب شده، ولی روش ارائه‌شده در این پژوهش مستقل از این حوزه بوده و در موضوعات پژوهشی دیگر نیز قابل استفاده

1. research asset

www.SID.ir

Archive of SID

است. در راستای نیل به این هدف، عوامل تأثیرگذار بر کارایی راهکار ارائه‌شده مورد ارزیابی قرار خواهد گرفت تا در نهایت، روش، ابزارها، و پارامترهای مد نظر به بهترین شیوه انتخاب شوند.

برخی پژوهش‌ها برای شناسایی و انتخاب مجموعهٔ داده‌های یک حوزهٔ پژوهشی، توسط خود مؤلف یا خبرگان آن حوزه، تعدادی کلیدواژه را تعیین نموده و با جست‌وجوی آن کلیدواژه‌ها در پایگاه مد نظر، مجموعهٔ داده‌ها را تشکیل می‌دهند. «قنادی‌نژاد، حیدری، و چینی‌پرداز» با تعیین کلیدواژه‌های انگلیسی و فارسی مرتبط با موضوع علم اطلاعات و دانش‌شناسی و جست‌وجوی آن‌ها در پایگاه‌های علمی منتخب خارجی و داخلی، مجموعهٔ داده‌های این حوزه را شناسایی کرده‌اند (۱۳۹۷). همچنین، با جست‌وجوی کلیدواژه‌های ارائه‌شده توسط ستاد نانو، مقالات ISI این حوزه شناسایی شده و مورد تحلیل قرار گرفته است (تیمورپور، سپهری و پزشک ۱۳۸۸). به روش مشابه با جست‌وجوی عبارت «انرژی تجدیدپذیر» در پایگاه Web of Science، روند پژوهش‌های این حوزه مورد بررسی قرار گرفته است (Castillo, Salas & Ochoa 2018). «فتاحی و نعیمی‌صدیق» در پژوهش خود، که مربوط به تحلیل لاگ جست‌وجوهای کاربران در پایگاه اطلاعات علمی است، متن مورد جست‌وجو را بر اساس کلیدواژه‌های موجود در آن رده‌بندی نموده و بر مبنای نظر خبرگان ۱۹ برچسب مختلف برای آن‌ها در نظر گرفته‌اند و جست‌وجوی کاربران را از ابعاد مختلف در این رده‌ها مورد ارزیابی قرار داده‌اند (۱۳۹۵).

دسته‌ای دیگر از پژوهش‌ها با به‌کارگیری اصطلاح‌نامه[1]های موضوعی آن حوزه، مستندات پژوهشی را گردآوری می‌کنند. «ربیعی، حسینی‌مطلق و حائری» با استفاده از تحلیل لاگ جست‌وجوی کاربران در پایگاه پایان‌نامه‌ها و رساله‌های تحصیلات تکمیلی روند جست‌وجو و تعداد نتایج مربوط به زیرشاخه‌های مختلف حوزهٔ محیط زیست را مورد بررسی قرار داده‌اند. در این بررسی با استفاده از تکنیک‌های متن‌کاوی و پردازش زبان طبیعی، چهار منطقه شامل منطقه‌های دارای روند جست‌وجوی بالا و میزان پاسخ بالا، روند جست‌وجوی بالا و میزان پاسخ پایین، روند جست‌وجوی پایین و میزان پاسخ بالا و روند جست‌وجوی پایین و میزان پاسخ پایین مشخص شده است. در این پژوهش برای تولید و شناسایی مجموعهٔ داده‌های مرتبط با محیط زیست از اصطلاح‌نامهٔ این حوزه

1. thesaurus

www.SID.ir

Archive of SID

استفاده شده و واژگان مرتبط با محیط زیست از این طریق شناسایی و استخراج شده است (Rabiei, Hosseini-Motlagh & Haeri 2017).

دستهٔ سوم با استفاده از فراداده‌های توصیفی مربوط به مستندات پژوهشی مانند نام نشریه، نام دانشگاه یا دانشکده، عنوان کنفرانس و ... پژوهش‌های مربوط به یک حوزهٔ پژوهشی را شناسایی می‌کنند. بررسی مقالات هفت فصلنامهٔ علمی-پژوهشی مرتبط با اقتصاد با هدف تحلیل محتوای این حوزه (آشتیانی، رشیدی و لاریجانی ۱۳۹۱)، تحلیل مقالات منتشرشده در مجلات مرتبط با مدیریت ساخت (Darko & Chan 2016) و یا استفاده از برچسب‌های توصیفی خبرگان برای شناسایی مستنداتی مرتبط با بیماری سرطان و استفاده از آن در مرحلهٔ یادگیری مدل برای شناسایی بیماران مبتلا به آن از این دسته‌اند (Joffe et al. 2015).

در هر سه روش فوق انتخاب مجموعهٔ داده‌های یک حوزه به دانش مؤلف یا خبرگان آن حوزه محدود است. از طرف دیگر، جامعیت داده‌های انتخابی در هر سه حالت مورد ابهام واقع می‌شود. این محدودیت به‌ویژه در مورد موضوعات میان‌رشته‌ای که آن‌ها را می‌توان از ابعاد مختلف و با نگرش‌های گوناگون مورد بررسی قرار داد، بیشتر احساس می‌شود. از این رو، لازم است روشی برای شناسایی و رده‌بندی متون پژوهشی ارائه شود که کمترین نیاز را به دانش افراد خبره داشته و مستقل از تغییرات حوزه‌های دانشی پیرامون بتواند مرز نسبتاً مشخصی برای شناسایی پژوهش‌های یک حوزه ترسیم نماید. این پژوهش به‌دنبال آن است که روشی بر مبنای یادگیری ماشین و مستقل از نظر خبرگان برای شناسایی مجموعهٔ داده‌های یک حوزهٔ پژوهشی ارائه نماید و پارامترهای مورد استفاده در یادگیری ماشین را در شناسایی مجموعهٔ داده‌ها ارزیابی نموده و آن‌ها را بهبود بخشد.

در بخش دوم این مقاله، روش پژوهش بیان شده و در بخش سوم، پیش‌پردازش داده‌های اولیه به‌منظور آماده‌سازی آن‌ها برای رده‌بندی شرح داده شده است. بخش چهارم به رده‌بندی متون پژوهشی با استفاده از ماشین بردار پشتیبان می‌پردازد. کیفیت این رده‌بندی بر اساس معیارهای مختلف در بخش پنجم مورد ارزیابی قرار می‌گیرد و در نهایت در بخش ششم، نتیجه‌گیری پژوهش بیان شده است.

۲. روش پژوهش

این پژوهش با استفاده از متن‌کاوی در پایان‌نامه‌ها و رساله‌های حوزهٔ محیط زیست

www.SID.ir

Archive of SID

و با به‌کارگیری رویکرد رده‌بندی با استفاده از ماشین بردار پشتیبان تک‌رده‌ای به شناسایی متون پژوهشی مرتبط با این حوزه می‌پردازد. برای این منظور، پس از استخراج مجموعهٔ اولیهٔ داده‌های حوزهٔ محیط زیست، مجموعهٔ داده‌های هستهٔ این حوزه به‌منظور استفاده در فرایند یادگیری ماشین ایجاد و سپس، از تکنیک‌های پردازش زبان طبیعی در مراحل پیش‌پردازش و تولید ماتریس بردار ویژگی‌ها استفاده شده و در نهایت، مجموعهٔ داده‌های آزمون ایجاد و مدل بر اساس پارامترهای مختلف مورد ارزیابی و بهبود قرار گرفته است. مراحل مختلف کار در زبان برنامه‌نویسی R انجام شده و در این بین از بسته‌های نرم‌افزاری مانند بستهٔ پردازش زبان طبیعی Kurt (2017)، بستهٔ متن‌کاوی Feinerer, Hornik, & Feinerer (2018) و بستهٔ آمار و یادگیری ماشین Meyer et al. (2018) استفاده شده است.

مجموعهٔ داده‌های اولیه

مجموعهٔ داده‌های مورد استفاده در این پژوهش از طریق پایگاه اطلاعاتی «پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)» که بزرگ‌ترین پایگاه اطلاعاتی پایان‌نامه‌ها و رساله‌های تحصیلات تکمیلی ایران است، استخراج شده است. این پایگاه دارای نزدیک به یک میلیون رکورد علمی است که تقریباً نیمی از آن مربوط به پایان‌نامه‌ها و رساله‌های تحصیلات تکمیلی است و عمدهٔ منابع موجود در این پایگاه به زبان فارسی است (پایگاه اطلاعات علمی/ گنج). رکوردهای استخراج‌شده از طریق جست‌وجوی عبارت‌های شروع‌شونده با «محیط زیست» و «زیست‌محیط» در میان ویژگی‌های «عنوان»، «کلیدواژه»، «چکیده»، «موضوع» و «سازمان» در پایان‌نامه‌ها و رساله‌های کارشناسی ارشد و دکتری به‌دست آمده است. این مجموعهٔ داده‌ها شامل تعداد ۱۶۶۲۶ رکورد است. ویژگی‌های موجود در این مجموعهٔ داده‌ها در جدول ۱، آمده است.

جدول ۱. ویژگی‌های موجود در مجموعهٔ داده‌ها

ویژگی	نام اختصاری	توضیح
شناسه	ID	یک شناسهٔ یکتا برای این رکورد اطلاعاتی است.
عنوان	Title	شامل عنوان پژوهش است.
موضوع ۱	Subject 1	۶ گروه اصلی آموزشی است. مانند «فنی مهندسی»، «علوم پایه»، ...
موضوع ۲	Subject 2	دومین سطح گروه آموزشی است. مانند «مهندسی نفت»، «شیمی»، ...

www.SID.ir

Archive of SID

ویژگی	نام اختصاری	توضیح
چکیده	Abstract	شامل چکیدهٔ پژوهش است.
تاریخ دفاع	Date	سال برگزاری جلسهٔ دفاع از پایان‌نامه یا رسالهٔ دکتری است. مانند «۱۳۹۷»
کد رهگیری	Tracking	یک شناسهٔ کنترلی برای ارتباط با دیگر پایگاه‌های اطلاعاتی است.
پدیدآوران	Authors	نقش و نام پدیدآوران است. مانند «دانشجو: علی ربیعی؛ استاد راهنما: مهدی حسینی»
سازمان	Org	نام سازمانی است که مدرک در آنجا تولید شده است. مانند «دانشگاه تهران»
مقطع	Degree	مقطع تحصیلی مرتبط با مدرک است. مانند «کارشناسی ارشد»، «دکتری تخصصی»، ...
کلیدواژه‌ها	Keywords	شامل کلیدواژه‌های تخصیص داده‌شده به مدرک است.
رشتهٔ مدرک	Field	رشتهٔ تخصیص‌یافته از طرف «ایرانداک» به این مدرک است. مانند «فیزیک»
رشتهٔ دانشجو	FieldSt	رشتهٔ تحصیلی گرایش تحصیلی مربوط به دانشجو است. مانند «فیزیک»
گرایش دانشجو	SubFieldSt	گرایش تحصیلی مربوط به دانشجو است. مانند «فیزیک بنیادی»

۳. پیش‌پردازش

رفع چالش‌های موجود در زبان فارسی

برای رفع چالش‌های مربوط به پردازش زبان فارسی و استانداردسازی متن، جعبه‌ابزاری[1] نوشته شد که متنی را به‌عنوان ورودی گرفته و متن استاندارد فارسی معادل آن را تولید می‌کند. تمامی متون فارسی موجود در مجموعهٔ داده‌ها با استفاده از این ابزار به حالت استاندارد تبدیل شده است. این جعبه‌ابزار موارد زیر را اصلاح می‌کند:

- ◇ فاصله‌های ابتدا و انتهای متن را حذف می‌کند؛
- ◇ تمامی نیم‌فاصله‌های متن را به فاصله تبدیل می‌کند؛
- ◇ فاصله‌های به طول بیش از یک را حذف می‌کند؛
- ◇ کاراکتر «ی» عربی را به «ی» فارسی تبدیل می‌کند؛
- ◇ کارکترهای ویرگول و نقطه-ویرگول انگلیسی را به معادل فارسی آن تبدیل می‌کند؛
- ◇ کاراکترهای نقطه‌گذاری (به جز کاراکترهایی که برای جداسازی اقلام استفاده می‌شوند) را حذف می‌کند؛
- ◇ اعداد فارسی و انگلیسی را حذف می‌کند.

1. toolbox

www.SID.ir

Archive of SID

شناسایی گروه اصلی آموزشی

گروه اصلی آموزشی در آموزش عالی شامل ۶ گروه «فنی و مهندسی»، «علوم پایه»، «علوم انسانی»، «کشاورزی»، «هنر» و «علوم پزشکی» است. این اطلاعات در ویژگی «موضوع ۱» در رکوردها ذخیره شده است، اما در بسیاری از موارد به دلایل مختلف از جمله اشتباهات کاربری در ورود اطلاعات، تنوع نوشتاری برخی از کاراکترها و نیز نقص اطلاعات، محتوای این ویژگی نادرست است. از این رو، در این مرحله علاوه بر یکسان‌سازی کاراکترهای فارسی تلاش شد تا نقص اطلاعاتی این ویژگی با استفاده از ویژگی‌های «رشتهٔ مدرک»، «رشتهٔ دانشجو» و «گرایش دانشجو» برطرف شود و در نهایت، ۶ گروه اصلی آموزشی به همراه گروه «نامشخص» برای آن دسته از ویژگی‌هایی که به روش‌های مختلف امکان شناسایی گروه برای آن‌ها فراهم نشد، شناسایی شدند.

جدول ۲، فراوانی رکوردهای گروه‌های اصلی آموزشی در «پایگاه اطلاعاتی ایرانداک» و نیز در مجموعهٔ داده‌های اولیه و مجموعهٔ هستهٔ توصیفی و مجموعهٔ هستهٔ محتوایی را نشان می‌دهد. در بخش ۴ در مورد چگونگی ساخت مجموعهٔ داده‌های محتوایی و توصیفی که از مجموعهٔ دادهٔ اولیه استخراج شده، به‌طور مفصل بحث شده است.

جدول ۲. فراوانی رکوردهای گروه‌های اصلی آموزشی در مجموعه‌های داده

گروه آموزشی	ایرانداک		مجموعهٔ دادهٔ اولیه		مجموعهٔ دادهٔ محتوایی		مجموعهٔ دادهٔ توصیفی	
	تعداد	درصد	تعداد	درصد	تعداد	درصد	تعداد	درصد
فنی و مهندسی	۸۸۹۰۸	۱۸	۵۹۹۵	۳۶	۵۰۳	۲۰	۱۰۵۸	۴۷
علوم انسانی	۱۶۱۷۱۴	۳۳	۲۸۹۴	۱۷	۸۷۸	۳۵	۳۴۰	۱۵
علوم پایه	۷۶۱۵۹	۱۶	۲۵۷۵	۱۵	۴۰۴	۱۶	۳۴۷	۱۵
کشاورزی	۳۹۲۸۰	۸	۱۳۹۰	۸	۱۱۰	۴	۲۳۴	۱۰
هنر	۱۵۷۸۸	۳	۷۱۱	۴	۱۰۸	۴	۷	۱
علوم پزشکی	۶۱۸۹۷	۱۳	۱۶۴	۱	۵۶	۲	۷	۱
نامشخص	۴۶۴۹۸	۹	۲۸۹۷	۱۷	۴۲۳	۱۷	۲۵۱	۱۱
مجموع	۴۹۰۲۴۴	۱۰۰	۱۶۶۲۶	۱۰۰	۲۴۸۲	۱۰۰	۲۲۴۴	۱۰۰

جدول ۲، نشان می‌دهد که حوزهٔ محیط زیست یک حوزهٔ میان‌رشته‌ای است که گروه‌های آموزشی فنی و مهندسی، علوم انسانی و علوم پایه بیشترین سهم را از پژوهش‌های

www.SID.ir

Archive of SID

این حوزه به خود اختصاص داده‌اند و گروه‌های هنر و علوم پزشکی کمترین سهم را از پژوهش‌های این حوزه در اختیار دارند.

ساخت سبدواژه[1]ها برای هر مستند

برای هر مستند لازم است سبدی از واژه‌های مرتبط تهیه شود که این سبد نمایندهٔ[2] مناسبی از محتوای آن سند باشد. از این رو، مجموع N-gramهای کلیدواژه و عنوان استخراج، و به‌عنوان سبدواژه‌ها در نظر گرفته شده است. N-gram مجموعه‌ای شامل N عنصر پشت سر هم در یک نوشتار یا صحبت را گویند که می‌تواند حرف، آوا، واژه و... باشد (Erkan & Radev 2004).

معادلهٔ ۱
$$n\_gram_i(St,n) = word_i word_{i+1} \dots word_{i+n-1}$$

با توجه به معادلهٔ ۱، در این پژوهش منظور از، N-gram یک عبارت شامل N واژهٔ پشت سر هم است. عناوین Uni-gram، Bi-gram، Three-gram و Quad-gram به‌ترتیب برای N=1, 2, 3, 4 مصطلح است.

استخراج N-gramهای کلیدواژه

با توجه به این‌که کلیدواژه‌ها کنترل‌شده هستند و پیش‌تر توسط افراد خبره به هر رکورد اختصاص داده شده‌اند، فرض شده است که کلیدواژه‌ها به‌درستی انتخاب شده و نمایانگر خوبی برای یک مستند هستند. از این رو، پالایش خاصی روی آن‌ها صورت نپذیرفت و هر کلیدواژه به‌عنوان یک N-gram برای رکورد انتخاب شده است.

استخراج N-gramهای عنوان

برای استخراج N-gramهای عنوان، به این شکل عمل شده است که تمامی N-gramها برای N=1, 2, 3, 4 استخراج شده و سپس، با استفاده از دو دسته از ایست‌واژه[3]ها این N-gramها پالایش شده است. گروه اول از ایست‌واژه‌ها شامل ۱۳۰۷ واژهٔ کم‌اهمیت و پرتکرار زبان فارسی شامل عدد، قید، ضمیر، کلمهٔ ربط، فعل و واژه‌های عمومی که در بیشتر متون پژوهشی وجود دارند (مانند «بررسی»، «تأثیر»، «ارزیابی» و ...) است و تمامی Uni-gramهایی که برابر با این واژه‌ها بوده‌اند، حذف شده‌اند. اما برای پالایش N-gramهایی با N>1 از

1. bag of words 2. representer 3. stopword

www.SID.ir

Archive of SID

گروه دوم ایست‌واژه‌ها شامل ۳۹۲ عدد، قید، ضمیر و کلمهٔ ربط زبان فارسی استفاده شده است؛ به این شکل که اگر واژهٔ ابتدایی یا انتهایی یک N-gram شامل این موارد باشد، آن N-gram از لیست حذف شود، به‌عبارت دیگر:

معادلهٔ ۲

$$Acptble\_NG(St,n) = \begin{cases} \bigcup_{i=1}^{k} n\_gram_i(St,1) | n\_gram_i(St,1) \notin FullStpW \\ \bigcup_{i=1}^{l} n\_gram_i(St,n) | \\ \{word_i(n\_gram_i(St,n)), word_n(n\_gram_i(St,n))\} \cap ShortStpW = \emptyset \end{cases}$$

در معادلهٔ ۲، سبدواژه‌های عنوان از میان تعداد K عدد Uni-gram و l عدد N-gram دیگر انتخاب شده است. همان‌طور که در این معادله آمده، در مورد عبارت‌هایی که بیش از یک جزء داشته باشند، به شرطی آن عبارت در سبد واژه‌ها قرار می‌گیرد که واژه‌های ابتدایی یا انتهایی آن جزء ایست‌واژه‌های مورد نظر نباشند، اما وجود ایست‌واژه در واژه‌های میانی عبارت مانع از قرارگیری آن در سبد واژه‌ها نمی‌شود. به این ترتیب، از عنوانی مانند «تحقیق و توسعه در آب و هوای ایران»، عبارت‌هایی مانند «تحقیق و توسعه» یا «آب و هوا» در سبدواژه‌ها باقی می‌مانند، اما «و توسعه در» یا «و هوای ایران» از این سبد حذف خواهند شد. در نهایت، سبدواژه‌های هر مستند از مجموع n-gramهای عنوان و کلیدواژهٔ آن مستند به‌دست آمد:

معادلهٔ ۳

$$BoW(Doc) = \left(\bigcup_{n=1}^{4} Acptble\_NG(Title(Doc),n)\right) \bigcup keywords(Doc)$$

تولید ماتریس بردار ویژگی‌های متون

ماتریس بردار ویژگی‌های متون با استفاده از تابع وزنی که اهمیت هر یک از واژه‌های موجود در سبد واژگان را برای هر مستند تعیین می‌کند، تشکیل شده است. این ماتریس که به ماتریس واژه–مستند[1] مصطلح است، عمدتاً یک ماتریس خلوت[2] خواهد بود (Feinerer, Hornik & Feinerer 2018). برای سهولت در انجام محاسبات و نیز بهبود

1. term-document-matrix 2. sparse

www.SID.ir

Archive of SID

عملکرد الگوریتم‌های رده‌بندی لازم است از واژه‌های نادری که فراوانی آن‌ها در مستندات بسیار پایین است، صرف نظر شود. روش‌های مختلف نظارتی و بدون ناظر برای انتخاب ویژگی‌ها ارائه شده است. از مهم‌ترین روش‌های نظارتی می‌توان به بهرهٔ اطلاعاتی[1] و اطلاعات متقابل[2] اشاره کرد و از مهم‌ترین روش‌های بدون ناظر انتخاب ویژگی، که عمدتاً بر محاسبهٔ امتیاز ویژگی‌ها بر اساس یک روش اکتشافی تمرکز دارند، می‌توان روش فرکانس اسناد[3] را نام برد (بصیری، نعمتی و قاسم‌آقایی ۱۳۸۶). از آنجا که هرچه تعداد واژه‌های موجود در ماتریس بردار ویژگی افزایش یابد، تعداد مستنداتی که به‌واسطهٔ آن‌ها شناسایی خواهند شد نیز افزایش خواهد یافت، پیدا کردن یک نقطهٔ بهینه برای انتخاب تعداد واژه‌ها اهمیت زیادی دارد. برای این منظور در این پژوهش، پس از آن‌که واژه‌ها از نظر تعداد دفعات تکرار آن‌ها در مستندات مرتب شدند، در یک روش اکتشافی، پارامتری به‌عنوان ارزش واژه[4] برای آن‌ها تعریف شده است. برای آن‌که علاوه بر تعداد دفعات تکرار واژه، اهمیت تعداد اجزای تشکیل‌دهندهٔ واژه نیز در این پارامتر مد نظر قرار گیرد، ارزش واژه از حاصل‌ضرب این دو عامل به‌دست آمده است.

معادلهٔ ۴
$$A(Term) = \{S | \forall i \in S, Term \in BoW(Doc_i)\}$$

$$Index(Term) = \max_{S \in A(Term)} |S|$$

$$Lenght(Term) = \max n | Term \in n - gram_1(Term, n)$$

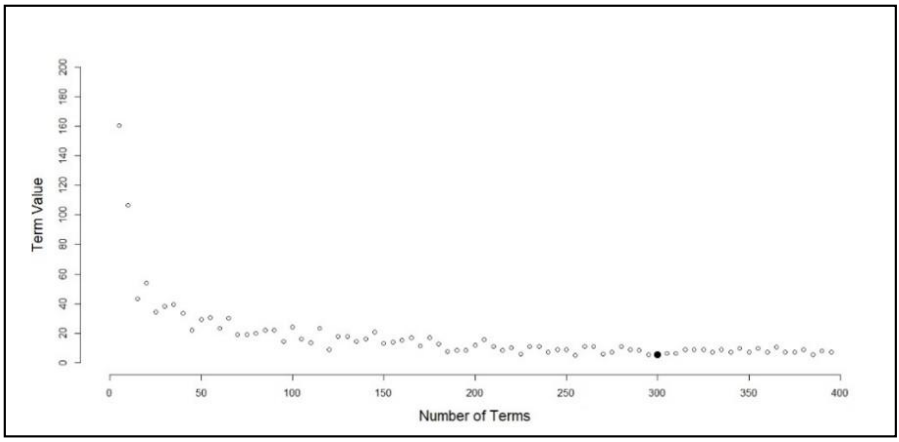
$$TermValue(Term) = Lenght(Term) * Index(Term)$$

شکل ۱، مقادیر ارزش واژه را برای تعداد واژه‌های مختلف در ماتریس بردار ویژگی‌ها نشان می‌دهد.

1. information gain
2. mutual information
3. Document frequency
4. term value

www.SID.ir

Archive of SID



شکل ۱. ارزش واژه برای تعداد واژه‌های پرتکرار ماتریس ویژگی‌ها

همان‌طور که در شکل ۱، مشخص شده، ارزش واژه در ماتریس بردار ویژگی برای واژه‌های پس از ۳۰۰ واژهٔ پُرارزش بسیار اندک است و نمودار آن تقریباً نزدیک به صفر است. از این رو، ۳۰۰ واژهٔ پُرارزش در این ماتریس به‌عنوان واژه‌های بردار ویژگی‌های متون انتخاب شده است. جدول ۳، واژه‌هایی را که دارای بیشترین مقدار ارزش واژه در ماتریس ویژگی‌ها هستند، در حوزهٔ محیط زیست نشان می‌دهد.

جدول ۳. واژه‌های دارای بیشترین مقدار ارزش واژه در حوزهٔ محیط زیست

ردیف	واژه	ارزش واژه	ردیف	واژه	ارزش واژه
۱	محیط زیست	۲۰۸	۱۱	کیفیت آب	۵۰
۲	آموزش محیط زیست	۱۳۸	۱۲	آلودگی خاک	۴۸
۳	سیستم اطلاعات جغرافیایی	۹۶	۱۳	پوشش گیاهی	۴۶
۴	جذب سطحی	۸۸	۱۴	اثرات زیست‌محیطی	۴۵
۵	تصفیهٔ فاضلاب	۶۸	۱۵	آلودگی	۴۳
۶	آلودگی هوا	۶۲	۱۶	آب زیرزمینی	۴۲
۷	ارزیابی اثرات زیست‌محیطی	۶۰	۱۷	خاک	۳۹
۸	حفاظت محیط زیست	۵۷	۱۸	فلز سنگین	۳۶
۹	سنجش از دور	۵۷	۱۹	آگاهی زیست‌محیطی	۳۶
۱۰	توسعهٔ پایدار	۵۲	۲۰	آلودگی آب	۳۲

www.SID.ir

Archive of SID

٤. رده‌بندی تک‌رده‌ای با استفاده از ماشین بردار پشتیبان

روش‌های سنتی رده‌بندی متون، در زمان یادگیری مدل، نیازمند در اختیار داشتن توزیع مناسبی از نمونه‌های کلاس‌های مثبت (هدف) و نمونهٔ داده‌های منفی (پرت) هستند. تولید چنین داده‌هایی نیازمند استفاده از دانش افراد خبره برای برچسب‌زنی نمونه‌هاست که فرایندی بسیار پرهزینه و زمان‌بر است. اما صرف نظر از زمان و هزینهٔ تولید این مجموعهٔ داده، مشکل اصلی این است که تولید مجموعهٔ داده‌هایی که به‌صورت جامع نمایانگر تمام حالت‌هایی باشد که جزو داده‌های هدف نیستند، امری بسیار دشوار و در بسیاری از کاربردها غیرممکن است (Khan & Madden 2014). به‌عنوان مثال، در این پژوهش لازم است نمونه‌هایی از انواع پژوهش‌هایی را که با حوزهٔ محیط‌زیست مرتبط نیست، در اختیار مدل قرار داد تا مدل بتواند در فرایند یادگیری، متون مرتبط با این حوزه را از دیگر حوزه‌ها تشخیص دهد. تولید چنین نمونه‌هایی تقریباً ناممکن است. برای این منظور، می‌توان از الگوریتم‌های رده‌بندی تک‌رده‌ای استفاده کرد که مهم‌ترین هدف این الگوریتم‌ها، آموزش مدل با استفاده از نمونه‌های یک کلاس (عمدتاً کلاس هدف) است (Khan & Madden 2014; Tax 2001).

هرچند برای رده‌بندی تک‌رده‌ای متون، روش‌هایی مانند رده‌بندی «بیزین»[1] (Liu et al. 2002)، رده‌بندی شبکهٔ عصبی[2] (Manevitz & Yousef 2000) یا رده‌بندی بر پایهٔ الگوریتم ژنتیک (Peng, Zuo & He 2006) نیز ارائه شده است، اما استفاده از ماشین بردار پشتیبان تک‌رده‌ای[3] برای رده‌بندی متون نتایج بسیار بهتری داشته است. همچنین، ماشین بردار پشتیبان در حالت‌هایی که بردار ویژگی داده‌ها دارای ابعاد بالاست، نسبت به روش‌های دیگر کاراتر است. در داده‌های متنی هر یک از واژه‌های موجود در تمامی سبد واژگان رکوردها به‌عنوان یکی از ابعاد ماتریس بردار ویژگی در نظر گرفته می‌شود. از این رو، در این حالت بردار ویژگی دارای ابعاد بالا خواهد بود و بنابراین، استفاده از ماشین بردار پشتیبان کارایی بیشتری داشته و کاربرد آن بسیار رایج است (Khan & Madden 2014).

علاوه بر کاربرد گستردهٔ ماشین بردار پشتیبان تک‌رده‌ای در رده‌بندی متون، این روش در حوزه‌های مختلف دیگر مانند تشخیص چهره، خلاصه‌سازی فیلم (Mygdalis et al. 2015)، تشخیص بیماری (Retico et al. 2016) و ... کاربردی گسترده دارد.

1. Naive Bayesian Classification
2. Neural Network Classification
3. One Class Support Vector Machine (OSVM)

www.SID.ir

Archive of SID

تولید مجموعهٔ داده‌های هسته

برای تولید مجموعهٔ دادهٔ هسته که به‌منظور یادگیری ماشین بردار پشتیبان مورد استفاده قرار خواهد گرفت، ویژگی‌های فراداده‌ای مورد بررسی قرار گرفته است. بر اساس تقسیم‌بندی «گلیلند» ویژگی‌های فرادادهٔ سه حوزهٔ محتوا، بافت و ساختار یک شیء اطلاعاتی را شامل می‌شود که ویژگی‌های محتوایی، ویژگی‌های ذاتی یک شیء را دربردارد و نشان می‌دهد که یک شیء شامل چیست. ویژگی‌های توصیفی، ویژگی‌های بیرونی و بافتی یک شیء را که کیستی، چرایی، کجایی و چگونگی آن را توصیف می‌کند و ویژگی‌های ساختاری به توصیف ساختار و شکل آن می‌پردازد (Gilliland 2008). از این رو، در این پژوهش دو مجموعهٔ دادهٔ محتوایی و توصیفی استخراج شده است. مجموعهٔ اول شامل رکوردهایی است که عبارت «محیط زیست» یا «زیست‌محیط» در ویژگی‌های محتوایی (عنوان و کلیدواژهٔ) آن‌ها موجود است و دستهٔ دوم وجود این عبارت‌ها را در ویژگی‌های توصیفی (موضوع ۱، موضوع ۲، رشتهٔ مدرک، رشتهٔ دانشجو و گرایش دانشجو) تضمین می‌کند. مجموعهٔ داده‌های هستهٔ محتوایی شامل ۲۴۸۲ رکورد و مجموعهٔ داده‌های هستهٔ توصیفی شامل ۲۲۴۴ رکورد شد. این دو مجموعهٔ داده دارای ۶۴۳ رکورد مشترک هستند.

تعیین پارامترهای ماشین بردار پشتیبان

برای استفاده از ماشین بردار پشتیبان ابتدا لازم است که علاوه بر تعیین هستهٔ ماشین، مقدار پارامترهای «گاما»[1] و «نو»[2] تعیین شود. از آنجا که در ادامهٔ این پژوهش هسته‌های مختلف مورد ارزیابی قرار می‌گیرد، تعیین پارامترهای «گاما» و «نو» در این بخش مورد توجه است. «گاما» ضریب هسته برای حالت‌های پایه شعاعی[3]، چندجمله‌ای[4] و سیگموید[5] است. هرچه مقدار «گاما» بیشتر باشد، الگوریتم تلاش می‌کند برازش را دقیقاً بر اساس مجموعهٔ داده‌های آموزشی[6] انجام دهد و این امر موجب تعمیم یافتن خطا و وقوع مشکل بیش‌برازش[7] می‌شود. همچنین، پارامتر «نو» نشان‌دهندهٔ حد بالای داده‌های آموزش است که مدل اجازه دارد آن‌ها را به‌عنوان دادهٔ پرت[8] در نظر بگیرد (Mack, Roscher & Waske 2014). تعیین مقدار مناسب این دو پارامتر در کارایی مدل تأثیر زیادی دارد. از این رو، روش‌های مختلفی برای تعیین آن‌ها ارائه شده است. یکی از این روش‌ها اجرای مدل با مقادیر

1. Gama
2. Nu
3. radial basis function (RBF)
4. polynomial
5. sigmoid
6. training data set
7. over-fitting
8. outlier data

www.SID.ir

Archive of SID

مختلف این دو پارامتر و سپس، بررسی مدل و انتخاب مناسب‌ترین مقادیر است (Lameski et al. 2015). این روش که در اصطلاحاً جست‌وجوی شبکه‌ای[1] نام دارد، برای انتخاب مقادیر مناسب «گاما» و «نو» به‌کار گرفته شد و مقدار ۰/۹۷ برای «گاما» و مقدار ۰/۱ برای «نو» بهترین عملکرد را نشان داد.

۵. ارزیابی کیفیت رده‌بند

تعیین معیارهایی برای کیفیت رده‌بندی

به‌منظور ارزیابی کیفیت مدل رده‌بند از ماتریس درهم‌ریختگی[2] استفاده شده است. سطرهای این ماتریس دربردارندهٔ وضعیت نمونه‌ها در دنیای واقعی است و ستون‌های این ماتریس پیش‌بینی مدل از وضعیت نمونه‌ها را نشان می‌دهد. جدول ۴، اجزای این ماتریس را نشان می‌دهد.

جدول ٤. ماتریس درهم‌ریختگی

	مثبت پیش‌بینی‌شده	منفی پیش‌بینی‌شده
مثبت واقعی	True Positive (TP)	False Negative (FN)
منفی واقعی	False Positive (FP)	True Negative (TN)

با استفاده از عناصر ماتریس درهم‌ریختگی شاخص‌های ارزیابی مختلف مانند دقت[3] و بازخوانی[4] به‌دست خواهد آمد. همچنین، شاخص F از ترکیب دو شاخص قبل به‌دست خواهد آمد. شاخص دقت نشان می‌دهد که چه میزان از نمونه‌های انتخابی درست هستند و شاخص بازخوانی بر این مفهوم دلالت دارد که چه میزان از نمونه‌های صحیح موجود انتخاب شده‌اند. در برخی از منابع شاخص صحت[5] نیز به دقت ترجمه شده است که مفهومی کاملاً متفاوت داشته و بر میزان نمونه‌هایی اشاره دارد که سیستم در تشخیص آن‌ها موفق بوده است. این شاخص‌ها به‌صورت دقیق‌تر و با استفاده از عناصر ماتریس درهم‌ریختگی به شکل زیر تعریف می‌شوند:

1. grid search
2. confusion matrix
3. precision
4. recall
5. accuracy

www.SID.ir

Archive of SID

معادلهٔ ۵

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F\_Measure = \frac{2TP}{2TP + FP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

از آنجا که برای آموزش مدل رده‌بند تک‌رده‌ای فقط به نمونه‌های کلاس مثبت دسترسی وجود دارد، یادگیری مدل صرفاً با استفاده از کلاس مثبت صورت می‌پذیرد. از این رو، برای ارزیابی کارایی مدل در بخش یادگیری از بین شاخص‌های ارزیابی عنوان‌شده در معادلهٔ ۵ فقط معیار بازخوانی قابل محاسبه است (Mack, Roscher & Waske 2014).

تولید مجموعهٔ داده‌های آزمون[1]

به‌منظور بررسی کارایی مدل رده‌بندی، تولید مجموعهٔ داده‌های آزمون ضروری است. برای این منظور مجموعه‌ای شامل ۴۰۰۰ رکورد انتخاب شد. این مجموعه شامل ۳۰۰۰ رکورد از داده‌های پرت بود. این رکوردها با توجه به جدول ۲، و در نظر گرفتن موضوعاتی که کمترین ارتباط را با مجموعهٔ داده‌های اولیه و مجموعهٔ داده‌های هسته داشته است، از موضوعات آموزشی هنر، پزشکی، و علوم پایه (هر یک ۱۰۰۰ رکورد) انتخاب شده است. همچنین، ۱۰۰۰ رکورد مرتبط با موضوع محیط زیست نیز در این مجموعه وجود دارد که ۵۰۰ رکورد آن به‌صورت تصادفی از بین داده‌های هستهٔ محتوایی و ۵۰۰ رکورد آن از داده‌های هستهٔ توصیفی انتخاب شده است.

کیفیت رده‌بند بر اساس داده‌های هستهٔ مورد استفاده در یادگیری ماشین بردار پشتیبان

همان‌طور که پیش‌تر بیان شد، برای یادگیری مدل تک‌رده‌ای، صرفاً از نمونه‌های کلاس مثبت (رکوردهایی که به‌طور قطع مربوط به حوزهٔ محیط زیست هستند) استفاده می‌شود. برای این منظور، دو مجموعهٔ دادهٔ هسته در یادگیری مدل استفاده شده است.

1. test data

www.SID.ir

Archive of SID

جدول ۵، نتایج ارزیابی کیفیت رده‌بند بر اساس هر یک از این دو مجموعهٔ یادگیری را نشان می‌دهد.

جدول ۵. ارزیابی مدل رده‌بند بر اساس نوع دادهٔ مورد استفاده در مرحلهٔ یادگیری[1]

مجموعهٔ دادهٔ یادگیری	آزمون مدل			
	صحت	شاخص F	دقت	بازخوانی
هستهٔ محتوایی (۲۴۸۲ رکورد)	۰/۸۴	۰/۶۲	۰/۷۸	۰/۵۱
هستهٔ توصیفی (۲۲۴۴ رکورد)	۰/۸۶	۰/۶۷	۰/۸۶	۰/۵۵

همان‌طور که جدول ۵، نشان می‌دهد، تمامی شاخص‌های ارزیابی حاکی از آن است که هستهٔ توصیفی برای یادگیری مدل مناسب‌تر است. از بین شاخص‌های مورد ارزیابی، شاخص دقت در هستهٔ توصیفی بهبود قابل ملاحظه‌ای نسبت به هستهٔ محتوایی دارد. به‌عبارت دیگر، مدل با استفاده از رکوردهای مجموعهٔ توصیفی، سبدواژه‌های مناسب‌تری را، که نمایندهٔ بهتری از حوزهٔ محیط زیست است، شناسایی نموده است. از این رو، رکوردهایی که در مرحلهٔ آزمون به‌عنوان رکوردهای این حوزه توسط مدل شناسایی شده‌اند، تطابق بیشتری با واقعیت دارند و ۸۶ درصد رکوردهای شناسایی‌شده به‌طور قطع مربوط به حوزهٔ محیط زیست هستند. این در حالی است که همین معیار در زمان استفاده از هستهٔ محتوایی برای یادگیری مدل، به ۷۸ درصد خواهد رسید. از این رو، در ادامهٔ فرایند ارزیابی کارایی مدل، یادگیری مدل بر مبنای هستهٔ توصیفی انجام خواهد شد.

کیفیت رده‌بند بر اساس هستهٔ مورد استفاده در ماشین بردار پشتیبان

همان‌طور که پیش‌تر بیان شد، ماشین بردار پشتیبان با هسته‌های مختلف قابل فراخوانی است. در رده‌بندی تک‌رده‌ای استفاده از ۴ هستهٔ خطی[2]، شعاعی[3]، چندجمله‌ای[4] و سیگموید متداول است (Khan & Madden 2014). جدول ۶، نتایج ارزیابی کیفیت رده‌بند بر مبنای هر یک از هسته‌های فوق را نشان می‌دهد.

۱. در این آزمون، بردار پشتیبان با مقادیر Gama=0.97 و nu=0.1 و کرنل polynomial فراخوانی شده است.
2. linear 3. radial basis 4. polynomial

www.SID.ir

Archive of SID

جدول ٦. ارزیابی مدل رده‌بند بر اساس هستهٔ مورد استفاده در ماشین بردار پشتیبان

هسته	آزمون مدل			
	صحت	شاخص F	دقت	بازخوانی
خطی	۰/۸۴	۰/۵۶	۰/۸۹	۰/۴۱
شعاعی	۰/۷۶	۰/۱۷	۰/۶۳	۰/۱۰
چندجمله‌ای	۰/۸۶	۰/۶۷	۰/۸۶	۰/۵۵
سیگموید	۰/۲۴	۰/۳۹	۰/۲۴	۰/۹۷

جدول ۶، نشان می‌دهد که در شاخص‌های مورد ارزیابی، استفاده از هستهٔ سیگموید و شعاعی ضعیف‌ترین کارایی را برای مدل رقم زده است. هستهٔ سیگموید تمایل بیشتری به پذیرش بیشتر متون به‌عنوان متون مرتبط با محیط زیست دارد. به همین دلیل، معیار دقت در این هسته پایین است. در مقابل، هستهٔ شعاعی تمایل زیادی دارد که عمدهٔ متون را به‌عنوان دادهٔ پرت در نظر گیرد. به همین دلیل، معیار بازخوانی در این حالت پایین است. اما شاخص‌های ارزیابی‌شده نشان می‌دهد که هسته‌های خطی و چندجمله‌ای کارایی مناسبی را در رده‌بندی تک‌رده‌ای مستندات پژوهشی دارند. با توجه به این‌که هستهٔ چندجمله‌ای در شاخص F عملکرد بالاتری نسبت به هستهٔ شعاعی دارد و در دیگر شاخص‌ها تقریباً با شاخص شعاعی برابر است، از این رو، در ادامهٔ فرایند ارزیابی کارایی مدل، از هستهٔ چندجمله‌ای در ماشین بردار پشتیبان استفاده خواهد شد.

کیفیت رده‌بند بر اساس نوع وزن‌دهی واژه‌ها در ماتریس بردار ویژگی

وزن واژه‌ها در ماتریس بردار ویژگی تعیین‌کنندهٔ اهمیت آن واژه برای یک متن در مقایسه با واژه‌ها و متون دیگر است. سه روش متداول در وزن‌دهی واژه‌ها در ماتریس بردار ویژگی وجود دارد که متداول‌ترین آن‌ها روش Binary، روش TF و روش TF-IDF است. در روش Binary وجود یا نبودِ واژه در سبدواژه‌های متن مد نظر قرار می‌گیرد. در روش TF تعداد دفعات تکرار واژه در سبدواژه‌های متن مورد نظر است، و در روش TF-IDF علاوه بر تعداد دفعات تکرار واژه در یک متن، تعداد متونی که این واژه در سبد واژگان آن‌ها مشاهده شده نیز مد نظر قرار می‌گیرد. معادلهٔ ۶، چگونگی محاسبهٔ وزن در روش TF-IDF را بیان می‌کند:

www.SID.ir

Archive of SID

معادلهٔ ۶
$$TF\_IDF(Term) = TF(Term) * \left[\log \frac{n}{Docs(Term)} + 1\right]$$

در معادلهٔ ۶، متغیر n نمایانگر تعداد تمام واژه‌های موجود در پیکرهٔ مورد بررسی و تابع Docs (Term) نشانگر تعداد مستنداتی است که عبارت مورد نظر در سبد واژگان آن‌ها آمده است. علاوه بر روش‌های وزن‌دهی بالا، روش وزن‌دهی دیگری نیز در این پژوهش معرفی شده است. در این روش علاوه بر تعداد دفعات تکرار یک عبارت در یک متن (TF) با استفاده از مفهوم N_gram و اندازهٔ یک عبارت که در معادلهٔ ۴ بیان شده، به تعداد اجزای تشکیل‌دهندهٔ آن عبارت نیز توجه شده است. NG-TF1 حالتی را نشان می‌دهد که دفعات تکرار یک عبارت به اندازهٔ تعداد اجزای آن عبارت اهمیت خواهد داشت.

معادلهٔ ۷
$$NG\_TF1(Term) = TF(Term) * Lenght(Term)$$

$$NG\_TF2(Term) = TF(Term) * \sqrt{Lenght(Term)}$$

روش NG-TF2 برای تعدیل وزن اندازهٔ واژه از جذر آن استفاده کرده است. جدول ۷، نتایج ارزیابی کیفیت رده‌بند بر مبنای روش‌های مختلف وزن‌دهی واژه‌ها را نشان می‌دهد.

جدول ۷. ارزیابی مدل رده‌بند بر اساس نوع وزن‌دهی ماتریس بردار ویژگی

وزن‌دهی	آزمون مدل			
	صحت	شاخص F	دقت	بازخوانی
Binary	۰/۸۶	۰/۷۰	۰/۸۰	۰/۶۲
TF	۰/۸۶	۰/۶۷	۰/۸۶	۰/۵۵
TF-IDF	۰/۷۸	۰/۳۹	۰/۶۷	۰/۲۷
NG-TF1	۰/۸۳	۰/۵۲	۰/۹۳	۰/۳۶
NG-TF2	۰/۸۸	۰/۶۹	۰/۹۲	۰/۵۶

جدول ۷، نشان می‌دهد که روش NG-TF بهترین نتیجه را داشته و پس از آن روش Binary و TF نتایج قابل قبولی داشته‌اند. در مقابل، روش TF-IDF به‌دلیل اهمیت ویژه‌ای که به واژگان نادر می‌دهد، باعث شده است که نرخ بازخوانی در این روش، نتیجهٔ نامناسبی داشته باشد. روش‌های NG-TF به‌دلیل این‌که موجب شناسایی واژه‌های خاص به‌عنوان بردار

www.SID.ir

Archive of SID

ویژگی شده، موجب شده تا معیار دقت نسبت به روش‌های دیگر بالاتر باشد. به‌عبارت دیگر، روش‌های NG-TF ارائه‌شده علاوه بر این‌که به فراوانی تکرار عبارت‌ها اهمیت داده، طول این عبارت‌ها نیز مد نظر قرار گرفته و از آنجا که در بسیاری از موارد عبارت‌های کلیدی یک حوزه دارای طول بیش از یک واژه است. این روش وزن‌دهی توانسته است عبارت‌های خاصی را که دقیقاً نمایانگر یک حوزهٔ پژوهشی هستند، با دقت ۹۳ درصد استخراج کند که نسبت به روش‌های دیگر قابل ملاحظه است. از این رو، این روش برای شناسایی واژگان خاص یک حوزهٔ پژوهشی می‌تواند مورد استفاده قرار گیرد.

بررسی موردی نتایج آزمون در حالت NG-TF2 نشان می‌دهد که نتایج آزمون در فضای واقعی می‌تواند حتی بهتر از نتایج بیان‌شده در جدول ۷، باشد. جدول ۸، برخی از مواردی را که در آزمون به‌عنوان FP قلمداد شده، نشان می‌دهد.

جدول ۸. نمونه‌هایی از نتایج آزمون با روش NG-TF2

عنوان	گروه آموزشی	نتیجه
طراحی پل گالری بر روی رودخانهٔ زاینده‌رود اصفهان	هنر	FP
بررسی و ارزیابی سیاست ایجاد کمربند سبز در کنترل و هدایت توسعهٔ شهری ...	هنر	FP
سیری در نقش‌مایه‌های گیاهی و جانوری دورهٔ صفوی، مطالعهٔ موردی تالار ...	هنر	FP
تحلیل پراکندگی و مکان‌یابی بهینهٔ فضای سبز شهری (پارک‌ها) ...	هنر	FP
شبیه‌سازی منابع آب زیرزمینی دشت عقیلی با استفاده از مدل ریاضی ...	علوم پایه	FP
زمین شیمی زیست‌محیطی و منشأ عناصر سنگین در دریاچهٔ مهارلو ...	علوم پایه	FP
شناسائی باکتری‌های آلوده‌کنندهٔ سبزیجات مصرفی	علوم پزشکی	FP

همان‌طور که جدول ۸، نشان می‌دهد، اگرچه این رکوردها و برخی رکوردهای دیگر مشابه، به‌دلیل این‌که از گروه‌های آموزشی کمتر مرتبط با محیط زیست انتخاب شده‌اند به‌عنوان FP در نظر گرفته شده است، اما عناوین آن‌ها نشان می‌دهد که مدل به‌درستی این موارد را به‌عنوان رکوردهای مرتبط شناسایی کرده است. از این رو، مقادیر به‌دست‌آمده در جدول ۷، حالت بدبینانه‌ای را نشان می‌دهد و عملکرد مدل در شناسایی نمونه‌های واقعی بهتر از نتایج به‌دست‌آمده خواهد بود.

www.SID.ir

Archive of SID

٦. نتیجه‌گیری

دسترسی به نمونهٔ داده‌هایی که نمایانگر حوزه‌های مختلف پژوهشی باشند و روش‌های یادگیری ماشین بتوانند از طریق کشف الگو از آن‌ها، به تشخیص و رده‌بندی متون پژوهشی بپردازد، بسیار مشکل و پرهزینه بوده و در برخی موارد غیرممکن است. رده‌بندی تک‌رده‌ای به‌عنوان یک راه حل ارائه شده است. برای این منظور ساخت سبدواژه‌ها برای هر مستند از اهمیت ویژه‌ای برخوردار است که در این پژوهش این سبد با استفاده از N-Gramهای کلیدواژه و عنوان هر مستند تولید شده و برای تعیین ابعاد ماتریس بردار ویژگی از مفهوم ارزش واژه استفاده شده است. در نهایت، پس از تعیین پارامترهای ماشین بردار پشتیبان، مستندات پژوهشی حوزهٔ محیط زیست رده‌بندی شد.

ارزیابی کیفیت رده‌بندی تک‌رده‌ای منابع پژوهشی حوزهٔ محیط زیست نشان می‌دهد که استفاده از مجموعهٔ دادهٔ هستهٔ توصیفی برای یادگیری بردار ماشین پشتیبان نتایج بهتری را به نسبت استفاده از مجموعهٔ دادهٔ هستهٔ محتوایی به‌دست می‌آورد. همچنین، در ارزیابی نوع هستهٔ مورد استفاده مشخص شد که هستهٔ سیگموید تمایل زیادی به پذیرش داده‌های آزمون به‌عنوان نمونه‌های مرتبط با حوزهٔ محیط زیست دارد و در مقابل، هستهٔ شعاعی با سخت‌گیری زیاد، عمدهٔ داده‌های آزمون را به‌عنوان دادهٔ پرت در نظر می‌گیرد. به همین دلیل، معیار دقت در هستهٔ سیگموید و معیار بازخوانی در هستهٔ شعاعی بسیار پایین است. در این میان، هستهٔ چندجمله‌ای با مقدار دقت ۰/۸۶ و مقدار صحت ۰/۶۷ برای معیار F کارایی مناسبی را نشان می‌دهد. در نهایت، روش‌های مختلف وزن‌دهی به ماتریس بردار ویژگی در ارزیابی کیفیت رده‌بند مورد سنجش قرار گرفت و روش NG-TF که در این پژوهش ارائه شده، نسبت به روش‌های متداول پیشین نتایج بهتری کسب کرد. در شرایطی که معیار دقت از اهمیت بالایی برخوردار باشد، روش وزن‌دهی NG-TF دارای عملکرد بهتری در شناسایی واژه‌های خاص و کلیدی یک حوزهٔ پژوهشی دارد.

ارزیابی روش ارائه‌شده در این پژوهش برای حوزه‌های دیگر علوم، به‌ویژه دیگر حوزه‌های میان‌رشته‌ای و بررسی قدرت تفکیک حوزه‌های نزدیک به هم و همچنین، ارزیابی روند همگرایی یا واگرایی حوزه‌های پژوهشی مرتبط با هم و نیز چگونگی تشکیل، رشد، ترکیب موضوعات تشکیل‌دهنده یک حوزهٔ پژوهشی با استفاده از روش ارائه‌شده در این پژوهش، می‌توانند در مطالعات آتی مورد بررسی قرار گیرد. همچنین، تمرکز بر روش وزن‌دهی ارائه‌شده در این پژوهش و کاربرد آن در دیگر حوزه‌های

www.SID.ir

Archive of SID

متن‌کاوی مانند مدل‌سازی موضوعی، تصحیح متون، بازیابی اطلاعات و ... که بر مبنای تولید ماتریس شباهت کلمات بنا نهاده شده، زمینهٔ تحقیقات آتی را فراهم می‌نماید.

قدردانی

نویسندگان این مقاله لازم می‌دانند از مسئولان «آزمایشگاه متن‌کاوی و یادگیری ماشین «ایرانداک» که امکانات لازم برای انجام این پژوهش را فراهم نمودند، مراتب قدردانی خود را به عمل آورند.

فهرست منابع

حسن لاریجانی، حجت‌اله، و اعظم رشیدی آشتیانی. ۱۳۹۱. تحلیل محتوای نشریه‌های علمی-پژوهشی در حوزهٔ موضوعی اقتصاد در سال‌های ۱۳۸۵-۱۳۸۹. *فصلنامه علمی پژوهشی برنامه‌ریزی و بودجه* ۱۶ (۳): ۱۳۳-۱۵۷.

بصیری، محمداحسان، شهلا نعمتی، و ناصر قاسم‌آقایی. ۱۳۸۶. *مقایسهٔ دسته‌بندی متون فارسی با استفاده از الگوریتم‌های Knn و Fknn و انتخاب ویژگی‌ها بر اساس بهرهٔ اطلاعات و فرکانس سند*. کنفرانس ملی سالانهٔ انجمن کامپیوتر ایران، دانشگاه صنعتی شریف.

پایگاه اطلاعات علمی ایران/ گنج. https://ganj-beta.irandoc.ac.ir/. پژوهشگاه علوم و فناوری اطلاعات ایران. (دسترسی در ۱۳۹۷/۰۷/۲۰).

تیمورپور، بابک، محمدمهدی سپهری، و لیلا پزشک. ۱۳۸۸. روشی نوین برای دسته‌بندی هوشمند متون علمی (مطالعهٔ موردی مقالات فناوری نانو متخصصان ایران). *سیاست علم و فناوری* ۲ (۲): ۱-۱۴.

فتاحی، سمیه، و علی نعیمی صدیق. ۱۳۹۵. تحلیل رفتار اطلاع‌یابی پژوهشگران در موتور جست‌وجوی سامانهٔ ملّی اطلاعات پایان‌نامه‌ها/ رساله‌های دانش‌آموختگان داخل کشور (گنج). *فصلنامه علمی پژوهشی مدیریت اطلاعات* ۵ (۲): ۳۱-۵۸.

قنادی‌نژاد، فرزانه، غلامرضا حیدری، و رحیم چینی‌پرداز. ۱۳۹۷. تحلیل محتوای متون مربوط به اولویت‌های پژوهشی در علم اطلاعات و دانش‌شناسی. *پژوهشنامه کتابداری و اطلاع‌رسانی* ۸ (۱): ۵۵-۷۴.

References

Castillo, C. E. C., W. P. G. Salas, & G. V. Ochoa. 2018. Research Trend in Renewable Energy Resource: A Detail Bibliometric Study. *Contemporary Engineering Sciences* 11 (75): 3739-3746.

Darko, A., & A. P. C. Chan. 2016. Critical analysis of green building research trend in construction journals. *Habitat International* 57: 53-63. doi: https://doi.org/10.1016/j.habitatint.2016.07.001

Erkan, G., & D. R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research* 22: 457-479.

Feinerer, I., K. Hornik, & M. I. Feinerer. 2018. *Package 'tm'*. Corpus. Retrieved from https://cran.r-project.

www.SID.ir

Archive of SID

org/package=tm (accessed April 04, 2019).

Gilliland, A. J. 2008. Setting the stage. *Introduction to metadata* 2: 1-19.

Joffe, E., E. J. Pettigrew, J. R. Herskovic, C. F. Bearden, & E. V. Bernstam. 2015. Expert guided natural language processing using one-class classification. *Journal of the American Medical Informatics Association* 22 (5): 962-966.

Khan, S. S., & M. G. Madden. 2014. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review* 29 (3): 345-374.

Kurt, H. 2017. Natural Language Processing Infrastructure (Version 0.1-11): *CRAN*. Retrieved from https://cran.r-project.org/web/packages/NLP/index.html (accessed April 04, 2019).

Lameski, P., E. Zdravevski, R. Mingov, & A. Kulakov. 2015. *SVM parameter tuning with grid search and its impact on reduction of model over-fitting Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing* (pp. 464-474) Cham: Springer.

Liu, B., W. S. Lee, P. S. Yu, & X. Li. 2002. Partially supervised classification of text documents. Paper presented at the *International Conference on Machine Learning (ICML*). Sydney, Australia.

Mack, B., R. Roscher, & B. Waske. 2014. Can i trust my one-class classification? *Remote Sensing* 6 (9): 8779-8802.

Manevitz, L. M., & M. Yousef. 2000. Document classification on neural networks using only positive examples (poster session). *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. Athen, Greece.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F ,.Chang, C.-C., . . . Meyer, M. D. (2018). *Package 'e1071.'*

Mygdalis, V., A. Iosifidis, A. Tefas, & I. Pitas. 2015. Exploiting subclass information in one-class support vector machine for video summarization. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Brisbane, Australia.

Peng, T., W. Zuo, & F. He. 2006. Text classification from positive and unlabeled documents based on ga. *7th International Meeting High Performance Computing for Computational Science*. Rio de Janeiro, Brazil.

Rabiei, M., S.-M. Hosseini-Motlagh & A. Haeri. 2017. Using text mining techniques for identifying research gaps and priorities: a case study of the environmental science in Iran. *Scientometrics* 110 (2): 815-842.

Retico, A., I. Gori, A. Giuliano, F. Muratori, & S. Calderoni. 2016. One-class support vector machines identify the language and default mode regions as common patterns of structural alterations in young children with autism spectrum disorders. *Frontiers in neuroscience* 10: 306.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR*) 34 (1): 1-47.

Tax, D. M. J. 2001. One-class classification: concept-learning in the absence of counter-examples. Ph. D. thesis. Delft University of Technology, Stevinweg, The Netherlands.

www.SID.ir

Archive of SID

محمد ربیعی

متولد سال ۱۳۶۲، دانشجوی دکتری مهندسی فناوری اطلاعات از دانشگاه علم و صنعت ایران است. ایشان هم‌اکنون عضو هیئت علمی گروه کسب‌وکار الکترونیک پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است.

متن‌کاوی، پردازش زبان طبیعی، داده‌کاوی اطلاعات علم و فناوری و تحلیل رفتار کاربران از جمله علایق پژوهشی وی است.

سید مهدی حسینی‌مطلق

متولد سال ۱۳۵۸، دارای مدرک تحصیلی دکتری مهندسی صنایع از دانشگاه تربیت مدرس است. ایشان هم‌اکنون دانشیار دانشکدهٔ مهندسی صنایع دانشگاه علم و صنعت ایران است.

کاربردهای تحقیق در عملیات در سیستم‌های سلامت، طراحی شبکه‌های لجستیک و زنجیرهٔ تأمین، مدل‌های زمان‌بندی و مسیریابی در زنجیرهٔ تأمین، مدل‌های هماهنگی و زنجیرهٔ تأمین و برنامه‌ریزی احتمالی و بهینه‌سازی استوار از جمله علایق پژوهشی وی است.

بهروز مینایی بیدگلی

متولد سال ۱۳۴۱، دارای مدرک تحصیلی دکتری علوم و مهندسی کامپیوتر با تخصص هوش مصنوعی و داده‌کاوی از دانشگاه ایالتی میشیگان آمریکاست. ایشان هم‌اکنون دانشیار دانشکدهٔ مهندسی کامپیوتر دانشگاه علم و صنعت ایران و سرپرست گروه پژوهشی فناوری‌های بازی‌های رایانه‌ای و نیز آزمایشگاه داده‌کاوی است.

محاسبات نرم، یادگیری ماشین، بازی‌های رایانه‌ای، داده‌کاوی، متن‌کاوی، و پردازش زبان طبیعی، زمینه‌های پژوهشی مورد علاقه ایشان است.

www.SID.ir