

به نام خدا



وزارت علوم، تحقیقات و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران  
(ایرانداک)

## خلاصه گزارش طرح پژوهشی سازمانی

### عنوان طرح پژوهشی

تهیه مجموعه داده استاندارد فارسی برای مسئله استخراج خودکار کلیدواژه از اسناد علمی

|                                       |                                                                        |                        |            |
|---------------------------------------|------------------------------------------------------------------------|------------------------|------------|
| تاریخ شروع طرح پژوهشی                 | ۱۳۹۸/۰۶/۱۰                                                             | تاریخ پایان طرح پژوهشی | ۱۴۰۰/۰۱/۱۸ |
| نام و نام خانوادگی مجری طرح پژوهشی    | آزاده محبی                                                             |                        |            |
| نام و نام خانوادگی همکاران طرح پژوهشی | مرضیه زرین‌بال، عمار جلالی‌منش، زهرا دهرسرای، علی سلیمی‌صدر، حمید حسنی |                        |            |

### درباره طرح پژوهشی

(بیشینه ۳۰۰ واژه درباره مقدمه، مسئله، و هدف‌های پژوهش و آنچه که انجام، بررسی، یا تعیین شده است.)

با توجه به حجم داده‌ها و مستندات ایجاد شده در پایگاه‌های تخصصی علمی مانند گنج، نیاز است که روش‌های هوشمند و خودکار برای افزایش دقت و سرعت تخصیص کلیدواژه به این مستندات به کار گرفته شود. توسعه و بکارگیری این روش‌ها نیازمند وجود معیارهای استاندارد برای ارزیابی و بهبود آنهاست. یکی از این معیارها، وجود یک مجموعه داده استاندارد است که حاوی مجموعه‌ای از اسناد و کلیدواژه‌های تخصیص داده شده درست و کامل به آنها است. اگر چنین مجموعه داده‌ای موجود نباشد، نمی‌توان به راحتی الگوریتم‌ها و روش‌های استخراج خودکار کلیدواژه از مستندات علمی را با یک معیار مشخص با هم مقایسه کرد و بکار گرفت. در برخی از پژوهش‌های انجام شده در حوزه استخراج خودکار کلیدواژه از مستندات فارسی، از مجموعه داده‌هایی استفاده شده لیکن این مجموعه داده‌ها یا داده‌های علمی نیستند و یا در صورت استفاده از داده‌ها علمی، تنها کلیدواژه‌های نویسنده به عنوان کلیدواژه‌های تخصیص داده شده به یک سند علمی در آنها در نظر گرفته شده است. این در حالی است که در بسیاری از پژوهش‌های انجام شده در زبان انگلیسی، مجموعه داده‌های استاندارد و متعددی برای این منظور وجود دارد که در اغلب پژوهش‌ها از آنها استفاده می‌شود. بنابراین در این طرح پژوهشی، مجموعه داده‌ای از اسناد علمی (پایان‌نامه‌ها و رساله‌ها) گنج به همراه کلیدواژه‌های مرتبط با آنها ایجاد می‌شود. مدارک نمایه‌شده در گنج، می‌تواند منبعی غنی برای ساخت مجموعه داده برای مساله استخراج کلیدواژه از مستندات علمی فارسی باشد. لیکن کلیدواژه‌های تخصیص داده شده به هر مدرک در فرایند نمایه‌سازی اغلب محدود هستند و حتی در برخی از موارد به دلیل محدودیت‌های موجود در فرایند نمایه‌سازی، کامل نیستند. علاوه بر آن، مشخص نیست که هر کلیدواژه تخصیص داده شده به هر مدرک تا چه میزان به آن مدرک مرتبط است.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>در فرایند ایجاد مجموعه داده در این طرح، ابتدا از مجموعه‌ای از اسناد پایگاه گنج که کلیدواژه‌های تخصصی نمایه‌ساز و نویسنده را دارند استفاده شده است. سپس براساس فرایندی ماشینی مجموعه کلیدواژه وسیع‌تری، به عنوان کلیدواژه‌های کاندید برای هر سند در نظر گرفته شده و در انتها نیز متخصصین نمایه‌ساز، با بررسی مجدد هر سند و کلیدواژه‌های کاندید، کلیدواژه‌های تخصصی مناسب را از بین آنها انتخاب کرده‌اند. در نهایت ۲۸۳۸ داده در چهار حوزه موضوعی فنی و مهندسی، علوم انسانی، علوم و علوم کاربردی، و هنر در مجموعه داده قرار گرفته‌اند که هر یک به طور متوسط دارای ۱۰ کلیدواژه هستند. در انتها نیز پیشنهادهایی برای چگونگی بهره‌برداری از مجموعه داده استاندارد توسط جامعه علمی و انتشار آن ارائه شده است.</p>                                                                                                                                                                                                                                                                                                                                   |
| <p><b>روش پژوهش</b></p> <p>(بیشینه ۳۰۰ واژه درباره گام‌های پژوهش و چگونگی گردآوری و تحلیل داده‌ها (ابزار، جامعه، نمونه، روش نمونه‌گیری، روش‌های تجزیه و تحلیل داده‌ها).)</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| <p>در این پژوهش، مجموعه داده‌ای برای مساله استخراج خودکار کلیدواژه از مستندات علمی ایجاد شد. این مجموعه داده مشتمل بر ۲۸۳۸ سند از حوزه‌های موضوعی علوم انسانی، فنی و مهندسی، علوم و علوم کاربردی، و هنر است. هر سند شامل عنوان، چکیده و کلیدواژه‌های پایان‌نامه‌ها و رساله‌هایی است که در پایگاه گنج ثبت و نمایه شده‌اند. برای تخصیص کلیدواژه به اسناد، چهار روش استخراج کلیدواژه برای هر سند بکار گرفته شده است و براساس روش پیشنهادی در این پژوهش، کلیدواژه‌های پیشنهادی همه روش‌ها با هم یکپارچه شده‌اند و در نهایت برای هر سند یک فهرست ۲۴ تا ۳۰ تایی کلیدواژه بدست آمده است. سپس در قالب یک سامانه ارزیابی کلیدواژه‌ها، نتایج بدست آمده توسط متخصصین نمایه‌ساز ارزیابی شده‌اند و کلیدواژه‌های مناسب از بین کلیدواژه‌های پیشنهادی انتخاب شده‌اند. همچنین برای هر کلیدواژه انتخابی، میزان ارتباط آن با سند براساس سه عدد ۱، ۲، و ۳ مشخص شده‌اند. علاوه بر آن، در صورت نیاز، کلیدواژه‌های جدیدی نیز برای هر سند پیشنهاد شده‌اند. در نهایت عملکرد الگوریتم‌های مختلف روی مجموعه داده، براساس شاخص‌های دقت، بازخوانی و MAP@K بررسی شده‌اند.</p> |
| <p><b>یافته‌های طرح پژوهشی</b></p> <p>(بیشینه ۵۰۰ واژه درباره یافته‌ها و دستاوردهای کلیدی پژوهش در پاسخ به پرسش‌ها یا بررسی فرضیه‌های پژوهش.)</p> <p>در این پژوهش، مجموعه داده‌ای متشکل از ۲۸۳۸ سند علمی ایجاد شده که هر برای هر سند، عنوان، چکیده و کلیدواژه‌های آن آمده است. علاوه بر آن، میزان ارتباط هر کلیدواژه با سند نیز مشخص شده است. ۶ الگوریتم مطرح استخراج کلیدواژه روی مجموعه داده ایجاد شده، پیاده‌سازی شدند و دقت و بازخوانی آنها مورد مطالعه قرار گرفتند. در نهایت این مجموعه داده می‌تواند در بسیار از مسائل حوزه متن کاوی نظیر استخراج کلیدواژه، دسته‌بندی، خوشه‌بندی و خلاصه‌سازی خودکار مورد استفاده قرار گیرد.</p>                                                                                                                                                                                                                                                                                                                                                                                                          |
| <p><b>نتیجه‌گیری و پیشنهادها</b></p> <p>(بیشینه ۳۰۰ واژه بحث و نتیجه‌گیری و بیان پیشنهادهای برگرفته از پژوهش.)</p> <p>روش‌های مختلف در حوزه استخراج کلیدواژه‌ها بر پژوهش‌های آتی می‌تواند مجموعه داده وسیع‌تری را با رویکردی که در این طرح پیشنهاد شده ایجاد کرد. همچنین می‌توان علاوه بر خلاصه و عنوان، اسنادی را در مجموعه داده قرار داد که متن کامل آنها در دسترس باشد. علاوه بر آن، سایر الگوریتم‌های استخراج کلیدواژه می‌تواند روی مجموعه داده ایجاد شده پیاده‌سازی شود و نتایج دقت و بازخوانی آنها ارزیابی گردد.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |