

فهرست مطالب

[كليات تحقيق 8](file:///D:\project\قرباني%20داده%20كاوي\نهایی.docx#_Toc491254818)

[1-1 مقدمه 8](#_Toc491254819)

[1-2 چرا کلمات کلیدی؟ 9](#_Toc491254820)

[1-3 بيان مسئله 12](#_Toc491254821)

[1-4 اهمیت انجام پژوهش و وجه تمایز آن با سایر پژوهشها 13](#_Toc491254822)

[1-5 اهداف پژوهش 14](#_Toc491254823)

[1-6 فرضیه ها یا سوالات پژوهش 15](#_Toc491254824)

[1-7 امکانات و محدودیت های پژوهش 15](#_Toc491254825)

[1-8 ساختار پايان نامه 16](#_Toc491254826)

[فصل دوم: مروري بر مباني و پيشينه تحقيق 17](#_Toc491254827)

[2-1 مقدمه 18](#_Toc491254829)

[2-2 معرفي كلمات كليدي 19](#_Toc491254830)

[2-2-1 اهميت كلمات كليدي 19](#_Toc491254831)

[2-3 اهمیت کلمات کلیدی در بهینه سازی وب سایت (سئو) 20](#_Toc491254832)

[2-3-1 کلمات کلیدی سئو 21](#_Toc491254833)

[2-3-2 موتورهای جستجو چگونه کار می کنند؟ 24](#_Toc491254834)

[2-4 بازاريابي محتوا 25](#_Toc491254835)

[2-5 روشهاي مختلف استخرج كلمات كليدي 28](#_Toc491254836)

[2-6 ريشهيابي 30](#_Toc491254837)

[2-6-1 الگوي ساختاري (روشهای مبتنی بر قواعد زبان) 30](#_Toc491254838)

[2-6-2 الگوي غیر ساختاری 31](#_Toc491254839)

[2-7 الگوریتم ژنتیک 32](#_Toc491254840)

[2-7-1 ساختار الگوريتم‏هاي ژنتيكي 33](#_Toc491254841)

[2-7-2 عملگر انتخاب 35](#_Toc491254842)

[2-7-3 عملگر آميزش 36](#_Toc491254843)

[2-7-4 عملگر جهش 38](#_Toc491254844)

[2-7-5 روند كلي الگوريتم‏هاي ژنتيكي 39](#_Toc491254845)

[2-8 پيشينه­ي تحقيق 40](#_Toc491254846)

[2-8-1 روشهاي مبتني بر آمار تكرار كلمات 41](#_Toc491254847)

[2-8-2 روشهاي مبتني بر دسته­بندي و يادگيري ماشين 43](#_Toc491254848)

[2-8-3 روشهاي مبتني بر گراف 45](#_Toc491254849)

[2-8-4 كارهاي انجام شده بر روي متون فارسي 46](#_Toc491254850)

[فصل سوم: معرفی روش پیشنهادی 48](#_Toc491254852)

[3-1 مقدمه 48](#_Toc491254853)

[3-2 روش پیشنهادی 49](#_Toc491254854)

[3-3 پالایش کننده 51](#_Toc491254855)

[3-4 تجزیه کننده­ي جمله 53](#_Toc491254856)

[3-5 بستهبند 53](#_Toc491254857)

[3-6 صافی کلمه­ها 53](#_Toc491254858)

[3-6-1 کلمات ایست 54](#_Toc491254859)

[3-7 ایجاد بردار 63](#_Toc491254860)

[3-7-1 نرمال سازي بردار 64](#_Toc491254861)

[3-7-2 انتخاب كلمات با حروف بزرگ 64](#_Toc491254862)

[3-7-3 انتخاب كلمات مصور 65](#_Toc491254863)

[3-8 ايجاد ارتباط معنايي كلمات به كمك آنتولوژي 65](#_Toc491254864)

[3-8-1 آنتولوژي opencyc 66](#_Toc491254865)

[3-8-2 آنتولوژي WordNet 66](#_Toc491254866)

[3-9 مدلسازی الگوریتم ژنتیک برای روش پیشنهادی 67](#_Toc491254867)

[3-9-1 تابع ارزیابی الگوریتم ژنتیک 69](#_Toc491254868)

[فصل چهارم: پیاده­سازی و ارزیابی روش پیشنهادی 70](#_Toc491254869)

[4-1 مقدمه 70](#_Toc491254871)

[4-2 مجموعهي آزمايشي 71](#_Toc491254872)

[4-3 پكيج **jena** 71](#_Toc491254873)

[4-3-1 نحوه اضافه نمودن jena در eclipse 72](#_Toc491254874)

[4-4 پكيج **JWI** 74](#_Toc491254875)

[4-5 نرمافزار **Protégé** 76](#_Toc491254876)

[4-6 مراحل پیادهسازی 79](#_Toc491254877)

[4-6-1 مرحله اول 79](#_Toc491254878)

[4-6-2 مرحله دوم 80](#_Toc491254879)

[4-6-3 مرحله سوم 82](#_Toc491254880)

[4-7 معيار ارزيابي 82](#_Toc491254881)

[4-8 نتایج ارزیابی 84](#_Toc491254882)

[فصل پنجم: نتیجه­گیری و پیشنهادات 89](#_Toc491254883)

[5-1 مقدمه 89](#_Toc491254885)

[5-2 نتیجه­گیری 90](#_Toc491254886)

[5-3 پیشنهادات آینده 91](#_Toc491254887)

[6 مراجع 92](#_Toc491254888)

فهرست شکل­ها

[شکل 2-1 یک نمونه تلفیق (آمیزش) 37](#_Toc491255117)

[شکل 2-2) تلفیق دو نقطه­ای 38](#_Toc491255118)

[شکل 2-3) تلفیق جامع 38](#_Toc491255119)

[شکل 2-4) نمونه­ای از عمل جهش بر روی یک بیت 39](#_Toc491255120)

[شکل 2-5) فلوچارت ساده از الگوریتم ژنتیک 40](#_Toc491255121)

[شكل 3-1: مراحل مختلف روش پيشنهادي 50](#_Toc491255122)

[شكل 3-2: كد تابع HTML2Text 52](#_Toc491255123)

[شكل 4-1) نرمافزار protégé 76](#_Toc491255124)

[شكل 4-2) قسمتي از آنتولوژي كه توسط نرمافزار Protege اضافه شده است 78](#_Toc491255125)

[شکل 4-3) ارزیابی روش پیشنهادی 88](#_Toc491255126)

**فهرست جداول**

[جدول 3-1: لیست تعدادی از کلمات ایست زبان انگلیسی 55](#_Toc491255312)

[جدول 4-1: نقش ریشه­یابی در کاهش تعداد کلمات 80](#_Toc491255313)

[جدول 4-2: تعریف اختصارات در ارزیابی 84](#_Toc491255314)

[جدول 4-3) ارزیابی مرحله اول روش پیشنهادی 86](#_Toc491255315)

[جدول 4-4) ارزیابی مرحله دوم روش پیشنهادی 87](#_Toc491255316)

[جدول 4-5) ارزیابی مرحله دوم روش پیشنهادی 87](#_Toc491255317)

[جدول 4-6) ارزیابی کلی روش پیشنهادی 88](#_Toc491255318)

چكيده

رشد فزاینده اسناد متنی بر روی وب ضرورت استفاده از ابزارهای پردازش خودکار متون را بیش از پیش آشکار می­سازد. طبیعت غیر ساخت یافته متون، استفاده از روش­های سنتی استخراج اطلاعات از پایگاه داده های رابطه­ای را غیر ممکن می­سازد. استخراج کلمات کلیدی متون می­تواند کمک زیادی به موتورهای جست­وجو برای استخراج دقیق متون کند. در این تحقیق یک روش جدید استخراج کلمات کلیدی از متون معرفی می­شود. این روش شامل سه مرحله است. در مرحله اول پیش پردازش­های اولیه روی متون انجام می­گیرد و توسط روش­های آماری لغات اولیه استخراج می­شوند در مرحله دوم با استفاده از آنتولوژی کلمات مفهومی به لغات اضافه می­شوند و در مرحله سوم توسط الگوریتم ژنتیک کلمات نهایی استخراج می­شوند. پیاده­سازی مرحله اول و دوم به زبان جاوا و مرحله سوم در محیط متلب انجام گرفته است. نتایج بدست آمده بر اساس معیارهای استاندارد نشان می­دهد که روش پیشنهادی بهتر از روش­های مشابه عمل می­کند

**کلمات کلیدی: کلمات کلیدی، موتور جست­وجو، آنتولوژی، الگوریتم ژنتیک**

1

# كليات تحقيق

## مقدمه

با توجه به ازدیاد اسناد متنی در وب، نیاز به پردازش خودکار آنها روز به روز افزایش می-یاد. اما پردازش خودکار آن­ها نیاز به پیش­پردازش­های اولیه و خلاصه­سازی متون دارد. به عنوان مثال یک موتور جست­وجو نمی­تواند تمامی صفحات یک وب سایت را کلمه به کلمه بازید کند و تمامی اطلاعات را ایندکس کند. بنابراین لازم است از هر متن تعدادی کلمه کلیدی که بیان کننده­ی محتوای آن سند هستند استخراج شود تا موتورهای جست­وجو فقط با آن کلمات سروکار داشته باشند. دقیقا مثل زمانی که شما دنبال مطلبی در یک کتاب می­گردید در نگاه اول، فهرست مطالب را نگاه می­کنید. این فهرست یک خلاصه ایندکس شده از مطالب اصلی کتاب است. کلمات کلیدی نیز برای صفحات وب همچین نقشی دارند. در این پایان نامه یک روش جدید استخراج کلمات کلیدی از متون معرفی می­شود.

## چرا کلمات کلیدی؟

کلمات کلیدی مرتبط آن­هایی هستند که به گوگل اطلاع می­دهند که شما یک موضوع را به طور کامل پوشش می­دهید. برای مثال اگر من درباره آیفون می­نویسم، احتمالا بصورت غریزی از عباراتی مثل "تکنولوژی"، "اپ" و "آی او اس" در مقاله­ام استفاده می­کنم. گوگل با اسکن میلیاردها صفحه در اینترنت که اکثریت آن­ها روی آیفون تمرکز کرده­اند می­داند که چه کلمات مرتبطی احتمال دارد در یک مقاله راجع به آیفون باشند. پس استفاده از آن عبارت الزاما به گوگل نشان می­دهد که من گستره اطلاعات کافی­ راجع به موضوع اصلی ارائه می­دهم [1].

جهت بازاریابی محتوا چگونه کلمات کلیدی را می بایست جستجو کنید؟ چگونه رتبه­بندی سایت­تان را پیگیری و به تجزیه و تحلیل رقبایتان بپردازید؟ چگونه ROI شما اندازه­گیری می­شود؟ اگر در مورد موفقیت در کسب وکار آنلاین­تان در بازاریابی محتوا جدی هستید، تمامی این سوالات مهم را از خود بپرسید. البته، شما می­توانید از انجام این وظایف سرباز زنید، اما مطمئن باشید که بودجه تان را به هدر می دهید و همچنان نیازمند نظارت بیشتر هستید [2].

ابزارهای بهینه­سازی موتور جستجو شما را قادر به انجام این اعمال موثر جهت بازاریابی محتوا می نماید. بنابراین شما می­توانید با زمان و منابع بیشتر در بخش­های دیگری از کسب و کار خود تمرکز نمایید. هیچ چیز مثل محتوای با کیفیت برای سئو خوب نیست. اما تا زمانی که این محتوا براساس پژوهش بر روی کلمات کلیدی نباشد آنچان در جایگاه وب سایت در موتورهای جستجو تاثیر ندارد [1].

عده­ای از متخصصان سئو که بیشتر بر اساس تجربه عمل می کنند؛ معتقدند چگالی کلمات کلیدی امر مهمی در رتبه­بندی سایت به شمار می­آید. اما دستۀ دوم معتقدند که چگالی کلمات کلیدی دیگر در سئو تاثیری ندارد و با پیشرفت تکنولوژی معنايي موتور جستجوی گوگل، بالا بردن چگالی کلمات، تنها باعث ناخوانا شدن محتوا و حتی در مواقعی باعث جریمه شدن یک سایت می­شود.

در دنیای همواره در تغییر بازاریابی دیجیتالی و وجود برندهای پرقدرتی همچون گوگل، شرط موفقیت و پیشرو بودن در فعالیت­های بازاریابی محتوای وب، آموختن و به روزنگه داشتن آموخته­های خود در حوزه بازاریابی دیجیتالی براساس تغییرات و به روز رسانی باید باشد که توسط گوگل به عنوان پر قدرتمندترین موتور جستجو مطرح می­شود. از دیدگاه گوگل محتوایی مرتبط است که نیازهای کاربر يعني جستجو کننده در صفحات وب را پاسخگو باشد. و مدت زمان ایستایی کاربر در صفحه وب سایت یا وبلاگ شما را افزایش دهد. همچنین چگونگی نمایش محتوا در یک صفحه از وب سایت یا وبلاگ، نیز می­تواند در تجربه کاربری بازدیدکننده از صفحه شما تاثیرگذار باشد [7].

از آنجایی که بیش از پیش مخاطبین از طریق تلفن های هوشمند و تبلت به صفحات وب سایت ها و وبلاگ ها راه می­یابند، واکنشگرا بودن وب سایت­ها و وبلاگ­ها یک الزام است. محتوای با کیفیت از طریق وب سایت واکنشگرا می تواند به راحتی براساس سایز دستگاه در حال استفاده توانایی تغییر سایز و بهینه شدن برای خواندن را داشته باشد. عدم واکنشگرا بودن سایت منجر می­شود که بازدیدکننده صفحه شما به راحتی نتواند مطالب را بخواند و در نتیجه صفحه را به سرعت ترک نماید. و بالاخره اینکه کلمات کلیدی مهم ترین بخش سئو می باشد؛ بنابراین باید دربارهء آنها قبل از شروع تهیه محتوا کاملاً فکر کنید. این امر در ابتدا می­تواند سخت و پر فشار به نظر برسد. حجم بالایی از لغات و عباراتی که شما فکر می­کنید افراد به دنبال آن هستند اما ابزارهای قابل اعتمادی مانند Google Ad words هستند که می­توانید از آنها برای انتخاب بهترین کلمه کلیدی کمک بگیرید. به همین منظور تحقیق حاضر با بهره گیری از الگوریتم ژنتیک در دو بخش استخراج بهینه کلمات کلیدی و بهبود نتایج برمبنای معیارهای استاندارد سنجش توانایی روش در بازاریابی محتوا ارائه خواهیم داد. در بخش اول یک روش استخراج کلمات کلیدی ارائه شده، که از بازخورد موتورهای جستجوی محبوب در قبال کلمات استخراج شده برای بهینه سازی روش استفاده می کند. همچنین با استفاده از نتایج بدست آمده از روش معرفی شده در این بخش، می توان به بررسی معیارهای اهمیت یک کلمه از دید موتورهای جستجو پرداخت. در بخش دوم برای بهبود معیار دقت یک پس پردازش مستقل از روش استخراج کلمات کلیدی ارائه شده که با کاهش قابل قبول معیار بازخوانی معیار دقت و پارامتر F \_Measure را بطور محسوسی افزایش می دهد، ارائه خواهیم داد.

## بيان مسئله

با وجود آنکه هدف اصلی آنالیز متون فهم محتوای متنی و حتی کشف دانش است اما این مسئله هنوز در دنیای الگوریتم­ها و کامپیوترهای امروزی مسئله پیچیده­ای به نظر می­رسد که ما را بر آن می­دارد که در سطوح پایین­تر به دنبال زیر وظایفی باشیم که البته هر کدام از این زیر وظایف به خودی خود ارزش زیادی دارند. یکی از این زیر وظایف­ها استخراج بهینه کلمات کلیدی است که نقش بسزایی در عملی شدن سایر فرآیندها دارد در تعریف کلمات کلیدی باید گفت، به مجموعه­ای از کلمات یا عبارات کوتاه که بیان کننده منظور اصلی یک متن می باشند کلمات کلیدی آن متن می­گویند [10].

جستجو و انتخاب کلمه کلیدی مناسب اولین گام برای هر فرد یا سازمانی است که با هدف بازاریابی محتوا می خواهد وب سایت خود را مطرح کند و مخاطبینی را جذب نماید ، شاید به نظر برخی افراد انتخاب کلمه کلیدی مناسب امری ساده و کم اهمیت باشد. ولی ما در اینجا می­خواهیم بگوئیم که این کار بسیار مهم است زیرا این کلمات هستند که ترافیک را به سوی وب سایت هدایت می­کنند. بنابراین، تشخیص نقش و اهمیت کلمه و رابطه­ی بین کلمات، شاخص­های مهمی در کاوش متن می­باشند که کلید حل مسائل مشخصی در پردازش متن است. از این رو در تمامی سیستم­های پردازش اطلاعات، کلمه و روابط بین کلمات، پردازش می­شود. تجزیه و تحلیل کلمه کلیدی مهمترین بخش از سئو می­باشد. شبیه کلمه کلیدی جستجو شده برای بازاریابی آنلاین. یکی از اهداف مخاطبان سئو این است که جستجوگران بجای موتور جستجو جایگاه کلمه کلیدی را تعیین می­کنند. این کار توسط برخی از ابزارهای آنالیز کلمه کلیدی مانند Google ad words و پیشنهادات کلمه کلیدی یوتیوب، تجزیه و تحلیل لاگ پرس و جو و ... ارائه شده است [21].

برای انتخاب بهینه کلمات کلیدی، ابزارهای زیادی موجود می باشد که با توجه به جستجوی عبارت مورد نظر می­توانید کلمات کلیدی مرتبط به زمینه جستجوی خود را بیابید. مشکلی که در این ابزار وجود دارد این است که نتیجه آنها برای تنها یک موتور جستجو می باشد. بنابراین مهمترین سوال پیش روی محقق این است که برای انتخاب بهینه کلمات کلیدی در بازاریابی محتوا چه روشی را می توان ارائه داد؟

## اهمیت انجام پژوهش و وجه تمایز آن با سایر پژوهش­ها

با توجه به اهمیت کلمه کلیدی در سئو و تاثیر آن در رتبه­بندی سایت، موضوع انتخاب کلمه کلیدی مناسب، ارزش بررسی دارد. در این زمینه کارهایی جهت انتخاب کلمه کلیدی صورت گرفته است اما در این طرح در مورد انتخاب کلمات کلیدی مناسب با عنوان بازاریابی محتوای سایت و تاثیر آن بر روی رتبه بندی موتور جستجو با استفاده از رویکرد الگوریتم ژنتیک بعنوان نوآوری تحقیق حاضر مطرح می باشد.

از نظر کاربرد، این تحقیق از نوع تحقیقات کاربردی می­باشد، زیرا نتایج حاصل از آن دارای کاربرد خاصی می­باشد و از آن می­توان برای رفع پاره­ای از مشکلات وب سایت­ها از نظر بازخورد در موتورهای جستجو استفاده نمود.

از نظررویکرد تحقیق، جزء تحقیقات کیفی و کمی می­باشد.

از نظر هدف، جزء تحقیقات اکتشافی است، زیرا به دنبال استخراج و انتخاب بهینه کلمات کلیدی در بازاریابی محتوای وب سایت­ها است.

## اهداف پژوهش

کانال­های بازاریابی به سرعت تکامل یافته­اند و بطور چشم­گیری در موتور­های جستجو بهینه شدند. در مراحل ابتدایی بهینه­سازی سایه­ای از کلمات کلیدی بود و به افزایش سرعت به عقب پایان کد و هرزه نگاره وارد لینک می­شد اسپم دهی در میزان لینک دریافتی و چگالی کلمات کلیدی مورد استفاده تا زمانی که رتبه­بندی برای کلمات کلیدی که شما می­خواهید اغاز شود.

خوشبختانه گوگل از کسانی که شیوه تقلب­های این چنینی و تقلب در الگوریتم جستجوی خود را شناسایی کرد. این مقوله اهمیت بسیار بالای ارائه یک تکنیک خوب برای انتخاب کلمات کلیدی مناسب در محتوای وب را نشان می­دهد، با تغییرات جزئی در الگوریتم ژنتیک و ارائه یک راهکار مناسب می توان اهداف زیر را به منصه ظهور رساند:

1. بررسی تاثیر استخراج بهینه کلمات کلیدی در بازاریابی محتوای وب سایت­ها و هم چنین بررسی نقش بازخورد موتورهای
2. جستجوی محبوب در قبال کلمات استخراج شده در انتخاب روش بهنه که بسیار ضروری می­باشند.
3. بهینه­سازی روند انتخاب کلمه کلیدی با استفاده از الگوریتم ژنتیک و انتخاب یک رویکرد مناسب با ارتقای الگوریتم
4. تحلیل معیارهای اهمیت یک کلمه از دید موتورهای جستجو براساس روش بهینه ارائه شده.

## فرضیه ها یا سوالات پژوهش

1. استخراج بهینه کلمات کلیدی در بازاریابی محتوای وب سایت موثر است.
2. استفاده از بازخورد موتورهای جستجوی محبوب در قبال کلمات استخراج شده می­توان روش بهنه را ارائه داد.
3. با استفاده از نتایج بدست آمده از روش معرفی شده می توان به بررسی معیارهای اهمیت یک کلمه از دید موتورهای جستجو پرداخت.
4. با استفاده از الگوریتم ژنتیک می­توان کلمات کلیدی مناسب را انتخاب کرد.

## امکانات و محدودیت های پژوهش

روش گردآوری داده­ها در انتخاب بهینه کلمات کلیدی ، مطالعات کتابخانه­ای شامل مقالات مرتبط، پایگاه های اینترنتی، کتاب­های مرتبط می باشد. که می توان از ابزار­هایی مانند Google Keyword Planner فیش­برداری از مقالات علمی مرتبط مراجعه به پایگاه داده­های مربوط به وب سایت­های معتبر و همچنین مصاحبه و مشورت با تعدادی از خبرگان، کارشناسان، مدیران و متخصصان وب بهره جست. و اینکه روش تجزیه و تحلیل اطلاعات برای پیاده­سازی مدل ارائه شده از نرم افزار متلب بهره خواهیم برد. برای تست استفاده از سیستم مدیریت محتوای وردپرس و در پایان برای تحلیل کلی و ارائه یک رویکرد مناسب استفاده از آزمونهای رگرسیون خطی با استفاده از نرم افزار spss سرلوحه کار قرار می­گیرد.

از جمله محدودیت­هایی که در این امر سر راه پژوهش ما قرار می­گیرد، بسیار جدید بودن مبحث بازاریابی محتوای وب و کمبود مقالات و رفرنس در این زمینه می­باشد و اینکه هنوز به طور جدی مقاله یا پایان نامه­ای به این موضوع خاص اشاره نکرده است و در این موضوع با کمبود منابع مواجهه هستیم.

## ساختار پايان نامه

اين پايان نامه شامل 5 فصل است. در فصل اول كليات تحقيق شامل بيان مسئله، اهداف و فرضيات تحقيق بررسي مي­شود. در فصل دوم ابتدا مباني تحقيق بيان مي­شود و سپس توضيحاتي در مورد كارهاي انجام شده و معايب و مزاياي آن­ها داده مي­شود. در فصل سوم جزئيات روش پيشنهادي و نحوه پياده سازي آن مي­شود. در فصل چهارم ارزيابي و مقايسه روش پيشنهادي با روش­هاي ديگر انجام مي­شود و در نهايت در فصل پنجم، نتيجه­گيري روش پيشنهادي مطرح شده و پيشنهاداتي براي بهبود كار مطرح مي­شود.

# فصل دوم: مروري بر مباني و پيشينه تحقيق

2

# مروري بر مباني و پيشينه تحقيق

## مقدمه

از اختراع خط تاكنون نوشتار جايگاه ويژه­اي در ثبت و انتقال دانش و تجربيات بشر به نسل­هاي بعد داشته است. متن مجموعه­اي از روابط غير ساختار يافته از عينيت­هاي جهان واقعي يا ذهنيت­هاي نويسنده و يا جمع اين دو است. در واقع بشر زماني كه خواست آنچه را كه مي­انديشد، ثبت كند، قدم به دنياي خط و نگارش نهاد و از خود گنجينه­ي نوشتاري از علوم، دانش و معرفت را به يادگار گذاشت. اما در اين ميان پيدايش وب، اين كار را بسيار سهل و آسان­تر كرد و اين قابليت را فراهم كرد كه افراد بسيار راحت­تر نوشتارهاي خود را با هم به اشتراك بگذارند. بخش قابل توجهی از اطلاعات قابل دسترس در پایگاه داده­های متنی که شامل مجموعه بزرگی از اسناد منابع مختلف مثلاً مقالات خبری، کتاب­ها، ایمیل­ها و صفحات وب ذخیره شده­اند. پایگاه داده­های متنی به علت افزایش مقدار اطلاعات موجود به فرم الکترونیکی سریع رشد می­کنند.

امروزه بیشتر اطلاعات در صنعت، کسب و کار و سازمانهای دیگر به صورت الکترونیکی و به فرم پایگاه داده متنی ذخیره شده­اند. اما مخزن اصلي اين همه داده متني اينترنت و صفحات وب است. بدلیل حجم فزاینده اطلاعات وب سايت­ها نیاز به ابزارهایی که دانش مفید و مختصری از اطلاعات متنی موجود را بصورت خودکار استخراج کند بیش از هر زمان دیگری احساس می­شود از جمله کارهایی که در این زمینه می­توان انجام داد عبارت است از : طبقه بندی خودکار متون و یا به عبارتی برچسب زنی موضوعی متون، خلاصه سازی متون و غيره است اما مقدمه این کارها این است که بتوان کلمات کلیدی متون را استخراج کرد.

## معرفي كلمات كليدي

كلمات كليدي در يك متن به كلماتي گفته مي­شود كه بيانگر يك مفهوم كلي از متن يا منظور نويسنده از آن متن باشد. مجموعه­ای از لغات مهم در یک مستند هستند که توصیفی از محتوای سند را فراهم می­کند و برای اهداف مختلفی قابل استفاده هستند، به عنوان مثال در مقالات علمی برای اینکه خواننده بتواند درک مناسبی از مقاله داشته باشد، کلمات کلیدی عنوان می­شوند [11].

هر كلمه كليدي را مي­توان با عبارت كليدي تشكيل شده از حداقل دو كلمه نيز در نظر گرفت. كلمات كليدي در دو گروه دسته­بندي مي­شوند: كلمات كليدي تابعي[[1]](#footnote-2) و كلمات كليدي آموزنده[[2]](#footnote-3). كلمات كليدي تابعي يا غير مفيد براي كلمات دستوري يا مرتبط با زبان استفاده مي­شوند و ارتباط كمي با محتويات سند دارند. اين نوع از كلمات كليدي بايد حذف شوند و در نظر گرفته نشوند. كلمات كليدي آموزنده ارتباط خيلي قوي و معنايي با محتويات و معني متن سند دارند. مرز بين كلمات كليدي تابعي و آموزنده خيلي واضح و سخت نيست و ما مي­توانيم يك مرز فازي را براي آن­ها در نظر بگيريم [12].

1. Functional [↑](#footnote-ref-2)
2. Informative [↑](#footnote-ref-3)