
بسم الله الرحمن الرحيم

جمهوری اسلامی ایران

دانشگاه جامع علمی کاربردی

مرکز آموزش عالی علمی-کاربردی جهاد دانشگاهی مشهد

جهاد دانشگاهی
۱۳۵۹
واحد مشهد

# معرفی خلاصه‌سازی ماشینی متن و معرفی روش‌های خلاصه‌سازی متون در وب

## پایان نامه کارشناسی رشته برنامه نویسی تحت وب

**طیبه هنرور دوقلعه**

**سیده زکیه اتحادی**

**سعید راه داران**

**شهره راد رحیمی**

استاد راهنما:

**دکتر محمد عبدالهی**

**۱۳۹۵**

جمهوری اسلامی ایران

دانشگاه جامع علمی کاربردی

مرکز آموزش عالی علمی-کاربردی جهاد دانشگاهی مشهد

جهاد دانشگاهی
۱۳۵۹
واحد مشهد

# پایان نامه کارشناسی رشته برنامه نویسی تحت وب خانم ها طیبه هنرور دوقلعه، سیده زکیه اتحادی، شهره راد رحیمی و آقای سعید راه داران

# تحت عنوان معرفی خلاصه سازی ماشینی متن و معرفی روش‌های خلاصه سازی متون در وب

در تاریخ ..... توسط کمیته تخصصی زیر مورد بررسی و تصویب نهایی قرار گرفت.

۱- استاد راهنمای پایان نامه دکتر محمد عبدالهی

۲- استاد مشاور پایان نامه مهندس علی نیک فرجام

۳- استاد داور

۴- استاد داور

معاون شعبه:

# تشکر و قدردانی

به مصداق «من لم یشکر المخلوق لم یشکر الخالق » بسی شایسته است از استاد فرهیخته و فرزانه جناب آقای دکتر محمد عبدالهی که با کرامتی چون خورشید ، سرزمین دل را روشنی بخشیدند و گلشن سرای علم و دانش را با راهنمایی های کار ساز و سازنده بارور ساختند ، تقدیر و تشکر نماییم.

همچنین از پدر و مادران عزیز ، دلسوز و مهربانمان که آرامش روحی و آسایش فکری را فراهم نمودند تا با حمایت های همه جانبه در محیطی مطلوب ، مراتب تحصیلی و نیز پایان نامه درسی را به نحو احسن به اتمام برسانیم ، سپاسگزاری می نماییم.

# تقدیم به

ماحصل آموخته هایم را تقدیم می کنم به آنان که مهر آسمانی‌شان آرام بخش آلام زمینی‌ام است
به استوارترین تکیه گاهم ، دستان پرمهر پدرم
به سبزترین نگاه زندگیم ، چشمان سبز مادرم
که هرچه آموختم در مکتب عشق شما آموختم و هرچه بکوشم قطره‌ای از دریای بی‌کران مهربانیتان را سپاس نتوانم بگویم.
امروز هستی‌ام به امید شماست و فردا کلید باغ بهشتم رضای شما
ره‌آوردی گران سنگ‌تر از این ارزان نداشتم تا به خاک پایتان نثار کنم، باشد که حاصل تلاشم نسیم گونه غبار خستگیتان را بزداید.

بوسه بر دستان پرمهرتان

# فهرست مطالب

# فهرست جدول‌ها

# فهرست شکل‌ها

# چکیده پایان نامه

هر تکنولوژی جدیدی که عرضه می‌شود سنگ بنایش شناخت دقیق علوم حاکم برآن می‌باشد . پژوهشگر با شناخت کامل به تسط در حوزه تحقیقاتی خود می‌رسد و پس از تخصص ایده افکار فرد را در برمی‌گیرد و درنهایت منجر به تولید و نوآوری می‌شود. تحقیقی که پیش رو دارید اصول تکنولوژی خلاصه‌سازی خودکار را به طور جامعی بیان می‌کند . تمام شاخه ها و زیر شاخه ها به دقت برسی شده و تقریبا تمامی محاسبات لازم برای شروع طراحی یک سیستم در اختیار مخاطب قرار می‌گیرد .

شما با مطالعه هر فصل با بنیان خلاصه‌سازی خودکار آشنا می‌شوید و می‌آموزید این فناوری پیشرفته در مراحل اولیه چه طور خود را با علوم مختلف دیگر پیوند می‌زند و شما را وادار می‌کند برای آشنایی با اصول آن علوم مختلف دیگر را هم بیاموزید. ما تحقیق را بگونه‌ای پیش می‌بریم که مخاطب ابتدا اصول کلی را بیاموزد و سپس مشکلاتی که بر سر راه طراحی این فناوری می‌باشد را دریابد، چون این فناوری پیوندی ناگسستنی با زبان شناسی دارد در نتیجه مشکلات آن در زبانی همانند انگلیسی با زبانی همانند فارسی متفاوت می‌باشد و به ناچار باید با مشکلات هر کدام آشنا شود ،کاری که در این تحقیق  به صورت مفصل دنبال کردیم و آن را در اختیار مخاطب قرار دادیم.

یکی از چالش های بزرگ در این حوزه علمی چگونگی امتیازدهی به جملات برای تعیین قسمت‌های مهم یک متن می‌باشد که در نهایت متن منتخب به صورت خلاصه به کاربر تحویل داده می‌شود. البته هنوز هم این چالش وجود دارد و همه پژوهشگران در سرتاسر جهان در حال مطالعه و نقد و برسی این چالش هستند که هرچه می‌گذرد سیستم‌هایی با عملکرد بهتر عرضه می‌شود . در این تحقیق مهم‌ترین و کاربردی‌ترین محاسبات مورد نقد و برسی قرار گرفت و به طور مفصل با توضیح هر فرمول در تحقیق ثبت شد تا مخاطب به راحتی بتواند اصول یاد گرفته در فصل های اولیه را با این فصل تلفیق کند و بنیان مدل سازی سیستم خلاصه‌ساز خودکار را به صورت علمی و عملی بیاموزد. این تحقیق تنها به سیستم های خلاصه‌ساز خودکار فارسی نمی‌پردازد بلکه سیستم‌هایی که به زبان فارسی وجود دارد را بررسی می‌کند که نمونه بارز آن ققنوس و پارسینا می‌باشد سپس هر دو مورد نقد قرار می‌گیرند و مزایا و معایب هر کدام برای مخاطب به سادگی همراه با آمار بیان می‌شود و بعد از اینکه سیستم های موجود فارسی توضیح داده شد قدرتمندترین سیستمهای خارجی نیز مورد برسی قرار می‌گیرند و قسمتهای مختلف هر کدام با زبانی ساده در اختیار مخاطب قرار می‌گیرد. امید میرود با این تحقیق بتوانیم ضعف‌های سیستم‌های فارسی را برطرف کرده هم چنین بتوانیم سیستم های قدرتمندتری را طراحی کنیم که این بزرگترین هدف و مورد انتظارترین نتیجه‌ای است که می‌توانیم برای این تحقیق در نظر داشته باشیم.

کلمات کلیدی:

خلاصه‌سازی متن ، خلاصه‌سازی استخراجی ، خلاصه‌ساز، روش آماری

# فصل اول

## معرفی موضوع تحقیق

## ۱ – ۱ مقدمه

علم و دانش هر روز به پیشرفت خود در راستای بی‌نهایت حرکت می‌کند و مفاهیم جدیدی به روی انسان قرن بیستم می‌گشاید به طوری که تخصص داشتن در همه حوزه های علمی از عهده هیچ کس برنمی‌آید و جمع‌آوری علوم در یک کتاب به هیچ عنوان میسر نیست، بلکه هر علمی میلیون‌ها کتاب مخصوص خود را می‌طلبد، که درنتیجه بیانگر دنیایی بسیار گسترده است. اکنون در عصر حاضر یافتن پاسخ سوالات گوناگون از علوم متفاوت چگونه است؟

با ظهور الکترونیکی کردن علوم این کار تا حد نسبتا زیادی آسان شد. ظهور اینترنت نیز کمک شایانی به توسعه علوم کرد و اکنون با جست‌وجو در اینترنت می‌توان به کلیه دانش‌های سرتاسر این کره خاکی دست یافت.

اما آیا همین مقدار کافی است؟پاسخ بسیار روشن است. یافتن مقالات متنوع صرفا ما را در هنگام جستجوی یک مطلب خاص یاری نمی‌کند زیرا مطالعه مطالب متفاوت برای دستیابی به پاسخ مطلوب کاری بسیار زمان‌بر است و در مواردی هم مفید واقع نمی‌شود، زیرا ممکن است پاسخ خود را در آن مقاله بدست نیاوریم و تنها زمان زیادی صرف مطالعه آن کرده باشیم.

آیا راهکاری برای این مشکل می‌توان یافت؟

مشکل فوق را با طرح جدیدی که به تازگی مطرح شده می‌توان برطرف کرد. هرچند که هنوز این ایده به طور کامل به منتهای کمال خود نرسیده ولی پیشرفت‌های چشمگیری را به جهان ارائه داده است . این طرح خلاصه‌سازی اتوماتیک می‌باشد که مطلب مورد نظر را به صورت خودکار خلاصه‌سازی شده در دسترس مخاطبین قرار می‌دهد. مهمترین مزیت استفاده از خلاصه‌سازی ، کاهش زمان خواندن متن است. یک خلاصه خوب ، باید موضوعات گوناگون یک سند را بدون داشتن افزونگی بیان کند. ابزارهای خلاصه‌سازی می‌توانند برای تشخیص عناوین و موضوعات کلیدی یک متن مورد استفاده قرار گیرد. از جمله کاربردهای خلاصه‌سازی متن می‌توان به خلاصه‌سازی پرونده‌های پزشکی بیماران، سرویس‌های صوتی برای ناشنوایان، بازیابی اطلاعات، خلاصه‌سازی فایل‌های صوتی و تصویری و کاربردهای دیگر اشاره کرد.

## ۱–۲ اهمیت و ضرورت تحقیق

امروزه خلاصه‌سازی اتوماتیک مورد بررسی و تجزیه و تحلیل‌های زیادی قرار می‌گیرد . نمونه‌های مطرحی در جهان وجود دارد که به عنوان نسخه‌های اولیه طرح مذکور خودنمایی می‌کند با این که این نمونه محصول مایکروسافت می‌باشد و Microsoft word ' s Anfo Summarize است و قدرت نسبتا زیادی دارد اما هنوز نتوانسته است رضایت کامل را به دنبال داشته و انتظارات را تحقق بخشد . در این تحقیق هدف ما آشنایی فارسی زبانان با این طرح جدید و بسیار کارآمد می‌باشد تا بلکه بتوانیم در آینده خلاصه‌ساز اتوماتیک فارسی را راه‌اندازی کنیم زیرا سیستم‌های موجود در نسخه زبان انگلیسی نیز مشکلات بسیار زیادی دارند و می‌دانیم پیاده‌سازی این سیستم به زبان فارسی مشکلاتی به مراتب بیشتر از نسخه زبان انگلیسی خواهد داشت. این بدین معناست که باید با دقت بیشتری با این تکنولوژی آشنا شده و نکات علمی آن را فراگیریم تا بتوانیم از این تکنولوژی به نحو احسن استفاده کنیم .لازم به ذکر است که کمپانی‌های اروپایی تمرکز خود را بر روی ایجاد خلاصه‌ساز اتوماتیک انگلیسی و فرانسوی قرار داده‌اند و این ما هستیم که باید با تحقیق و بررسی این سیستم را به زبان فارسی آماده کنیم .

آماده‌سازی این سیستم می‌تواند تاثیر بسیار زیادی در رشد علمی کشور به همراه داشته باشد. با داشتن بستری مناسب برای محققین و پژوهشگران راه رسیدن به پاسخ سوالات راهورتر شده و هر کسی می‌تواند با صرف زمانی بسیار اندک به مطلب موردنظر خود دست یابد و لازم نیست ساعتها وقت برای جستجوی مطلب موردنظر خود بگذارد. یکی از دلایل پیشرفت کشورهای غربی توجه بسیار زیاد آنها به مدیریت زمان و راه‌های کاهش صرف زمان‌های کم کاربرد است . اتفاقا همین دلیل هم باعث توجه مهندسین نرم افزار و هوش مصنوعی به سیستم خلاصه‌ساز اتوماتیک شده است.

## ۱–۳ سوالات تحقیق

سوالاتی که در این تحقیق به آنها پاسخ داده می‌شود از این قرار است:

۱ – متن در یک سیستم الکترونیکی چگونه مورد بررسی قرار می‌گیرد ؟

۲ – چگونه می‌توان قسمت‌هایی از متن را جداسازی کرد ؟

۳ – چگونه می‌توان قسمت‌های مهم متن را تشخیص داد ؟

۴- چگونه می‌توان قسمت‌های مهم متن را از متن کامل جدا کرد؟

۵- چه ویژگی‌هایی را می‌توان استخراج کرد؟

## ۱–۴ فناوری تعریف شده در تحقیق

هدف ما در این تحقیق ارائه مطالبی سودمند برای مهندسین می‌باشد. به همین دلیل سعی کردیم بهترین و معروف‌ترین خلاصه‌سازهای جهان را مورد نقد و بررسی قرار دهیم . مزایا و معایب هر یک را به صورت علمی بررسی می‌کنیم. نقاط ضعف و قوت آنها را به مخاطبینمان معرفی می‌کنیم و بگونه‌ای مطالب تحقیق را توسعه می‌دهیم که پس از پایان مطالعه این تحقیق هر نوع خلاصه‌سازی که به شما معرفی شد به راحتی فرآیند اجرای آن را بشناسید و آن را از لحاظ علمی به چالش بکشید . هدف نهایی ما نیز این است که با آشنایی شما با سیستم‌های فوق گام‌های خود را جهت طراحی سیستمی مبتنی بر ساپورت زبان فارسی بردارید .

این سیستم ها شامل :

- Microsoft word ' s Anfo Summarize
- Biritish telecom prosum
- FRUMP
- Auto Summerized
- GistaSumm
- FarsiSum

لازم به ذکر است در تحقیق به خلاصه‌ساز FarsiSum به عنوان خلاصه‌ساز فارسی نیز توجه داشتیم و آن را نیز مورد بررسی قرار دادیم.

## ۱–۵ خلاصه فصل

با گسترش روزافزون حجم اطلاعات موجود در وب و افزایش چشم گیر مقالات منتشر شده در زمینه‌های مختلف علمی ، دسترسی درست و مطالعه اطلاعات مورد نیاز، همواره یکی از مشکلات محققان و پژوهشگران قرن ۲۱ می‌باشد. اینکه چه طور از یک طرف با این حجم انبوه از داده‌ها و از طرفی دیگر با زمان محدودی که در اختیار داریم ، بتوانیم مطالب مورد نیاز خود را مطالعه کنیم و یا اینکه چه‌طور می‌توان در روز چندین کتاب را مطالعه نمود و یا اینکه آیا می‌توان سیستمی طراحی نمود که بتواند با داده‌های موجود به تمامی سوالات ما پاسخ دهد، اینها سوالاتی است که پاسخ آنها را می‌توان در یک سیستم خلاصه‌ساز متن جستجو کرد.