دانشکده مهندسی کامپیوتر

بررسی روش‌های خلاصه‌سازی خودکار متون و پیاده‌سازی یک نمونه عملی برای زبان فارسی

پایان‌نامه برای دریافت درجه کارشناسی

در رشته مهندسی کامپیوتر - گرایش نرم‌افزار

نام دانشجو:

محمد عبدوس

استاد راهنما:

دکتر خنجری

آبان‌ماه ۱۳۹۲

دانشکده مهندسی کامپیوتر

بررسی روش‌های خلاصه‌سازی خودکار متون و پیاده‌سازی یک نمونه عملی برای زبان فارسی

پایان‌نامه برای دریافت درجه کارشناسی

در رشته مهندسی کامپیوتر گرایش نرم‌افزار

نام دانشجو:

محمد عبدوس

استاد راهنما:

دکتر خنجری

آبان‌ماه ۱۳۹۲


بسم الله الرحمن الرحيم

تأییدیه‌ی هیأت داوران جلسه‌ی دفاع از پایان‌نامه

نام دانشکده: دانشکده مهندسی کامپیوتر

نام دانشجو: محمد عبدوس

عنوان پایان‌نامه یا رساله: **بررسی روش‌های خلاصه‌سازی خودکار متون و پیاده‌سازی یک نمونه عملی برای زبان فارسی**

تاریخ دفاع: ۱۳۹۲/۰۸/۲۵

رشته: مهندسی کامپیوتر

گرایش: نرم‌افزار

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضا
۱	استاد راهنما	دکتر عین‌ا... خنجری	استادیار	دانشگاه علم و صنعت ایران	
۲	استاد داور داخلی	دکتر حسن نادری	استادیار	دانشگاه علم و صنعت ایران	

تقدیم به پدر بزرگوار و مادر مهربانم

آن دو فرشته ای که از خواسته هایشان گذشتند، سختی ها را به جان خریدند و خود را سپر بلای مشکلات و ناملایمات کردند تا من به جایگاهی که اکنون در آن ایستاده ام برسم و تقدیم به تمامی علاقمندان و دانشجویان رشته مهندسی کامپیوتر

تشکر و سپاس

من لم یشکر المخلوق لم یشکر الخالق

از زحمات استاد ارجمند جناب آقای دکتر خنجری که من را در تمامی مراحل تدوین این پایان‌نامه کمک نمودند و همچنین از پدر و مادرم که همراه همیشگی من بوده‌اند سپاسگزارم.

در پایان، از کلیه اساتیدی که تا این لحظه از محضر ارزشمندشان کسب فیض نموده‌ام، صمیمانه تشکر کرده و از درگاه ایزد منان سلامتی، بهروزی و شادکامی این عزیزان را خواستارم.

چکیده

در این پایان‌نامه ابتدا با رویکردهای مطرح در حوزه خلاصه‌سازی خودکار متون آشنا می‌شویم و سپس چندین روش مطرح برای خلاصه‌سازی خودکار متون را نام‌برده و آن‌ها را توضیح می‌دهیم. از دو منظر می‌توان به مقوله خلاصه‌سازی خودکار متون پرداخت. یک منظر آن این است که برای خلاصه‌سازی از بحث‌های آماری استفاده کرد که روش پیاده‌سازی شده در این پایان‌نامه نیز از همین روش استفاده کرده است یا اینکه از روش‌های مطرح در حوزه پردازش زبان‌های طبیعی استفاده کرد که این زمینه نیازمند تحقیق بیشتر در این عرصه می‌باشد. در نهایت یک نمونه عملی برای زبان فارسی پیاده‌سازی شده است.

واژه‌های کلیدی: خلاصه‌سازی، رویکرد، متن، پردازش زبان طبیعی، آماری

فهرست

فصل ۱:

مقدمه

۱-۱- شرح مسئله

مبحث خلاصه‌سازی [1]متن یکی از مسائل مهم حوزه علوم کامپیوتر می‌باشد. دلیل اهمیت این مبحث نیز به علت کاربرد زیاد این مقوله می‌باشد. به دلیل اهمیت این موضوع کارهای زیاد و گسترده‌ای بر نحوه خلاصه‌سازی انجام گرفته است که در این پایان‌نامه به چندین روش از آن‌ها اشاره می‌کنیم.

۱-۲- مقدمه

امروزه اکثر افراد از طریق متنها و نوشته ها دانش و آگاهی خود را افزایش می‌دهند. در این بین بیشتر اطلاعات به صورت الکترونیکی در دسترس افراد قرار میگیرد و برای یافتن سریع و مناسب اطلاعات مورد نیاز خواند کامل متون بزرگ نامناسب است. همچنین زمان نیز از اهمیت ویژه ای برخوردار است و انسانها باید از زمان خود بهترین استفاده را داشته باشند. برای صرفه جویی در وقت و همچنین کسب آگاهی از اطلاعات موجود بهترین راه این است که خلاصه‌ای از اطلاعات مورد نیاز به افراد داده شود و افراد به جای صرف وقت در خواندن کل این مطالب فقط خلاصه‌ای از آن را مورد استفاده قرار دهند. مبحث خلاصه سازی به دلیل کاربرد و اهمیت زیاد آن به یکی از مسائل مهم در حوزه متن کاوی و پردازش متن در سرتاسر جهان تبدیل شده است. این تصور میشود که برای خلاصه‌سازی متون نیاز به فهم کامل متن میباشد و ماشین باید همانند انسان متن را بفهمد و خلاصه آن را در خروجی بیاورد، اما چنین نگرشی برای ماشین عملا امکان پذیر نیست. یعنی نباید چنین توقعی را از سیستم خلاصه ساز داشت که همانند انسان تولید خلاصه کند. اما کارهای زیادی از سال ۱۹۵۰ تا کنون در بحث خلاصه‌سازی اتوماتیک انجام شده و در حال پیشرفت است اما کارهای بسیار زیادی نیاز است تا سیستمی طراحی و تولید شود که همانند انسان خلاصه تولید کند و به هوشمندی انسانی دست پیدا کند. در حال حاضر اکثر خلاصه‌سازها فقط جملاتی از متن اصلی را به عنوان جملات خلاصه انتخاب میکنند و دقیقا همان جملات را بدون هیچگونه تغییری در خروجی میاورند. در واقع در این نوع سیستمها متن ورودی به جملاتی شکسته میشود و این جملات به مجموعه ای از جملات خلاص و جملات غیر خلاصه تبدیل میشود. در این پایان نامه ابتدا با رویکردهای مطرح در حوزه خلاصه‌سازی خودکار متون آشنا میشویم و سپس چندین روش مطرح در خلاصه‌سازی خودکار متون را نام برده و آنها را توضیح میدهیم. از دو منظر میتوان به مقوله خلاصه‌سازی خودکار متون پرداخت. یک منظر آن این است که برای خلاصه‌سازی از بحثهای آماری استفاده کرد که روش پیاده سازی شده در این پایان نامه نیز از همین روش استفاده کرده است یا اینکه از روشهای مطرح در حوزه پردازش زبانهای طبیعی[2]استفاده کرد که این

[1] Summarization
[2] Natural Language Processing

زمینه نیازمند تحقیق بیشتر در این عرصه میباشد. در نهایت یک نمونه عملی برای زبان فارسی پیاده سازی شده است.

فصل ۲:

تعاریف و مفاهیم مبنایی

۲-۱- تعریف خلاصه‌سازی

خلاصه‌سازی خودکار سند، یعنی تولید یک نسخه مختصر تر از سند اصلی توسط یک برنامه کامپیوتری به نحوی که ویژگی‌ها و نکات اصلی سند اولیه حفظ شود. بنا بر تعریف ارائه‌شده در استاندارد ISO 215 سال ۱۹۸۶، خلاصه «یک بازگویی مختصر از سند» می‌باشد. خلاصه‌سازی مطالب فرایندی است که از یک سند نسخه‌ای فشرده پدید می‌آورد و برای کاربر اطلاعات مفیدی فراهم می‌آورد؛ به عبارت دیگر خلاصه‌سازی ایده اصلی یک سند را در فضایی کمتر بیان می‌کند. اگر تمامی جمله‌های یک سند از اهمیت یکسانی برخوردار باشند، خلاصه‌سازی مؤثر نخواهد بود و هر گونه کاهش در اندازه آن سند منجر به کاهش نسبی در محتواهای آگاهی‌دهنده‌ی آن سند می‌گردد. خلاصه‌سازی تکنیکی است برای تشخیص اجزای آگاهی‌دهنده، که ایده‌ی اصلی یک سند را ارائه می‌دهد. خلاصه سازی متن فرآیند شناسایی برجسته ترین اجزای منبع و گرد آوری آن در حجمی کمتر میباشد، که این اجزا بیشترین تطابق و تعلق به مفهوم و موضوع مطرح شده در سند را داشته باشد.

۲-۲- تاریخچه خلاصه‌سازی

کارهای انجام‌شده از سال ۱۹۵۰ تا ۱۹۹۰ را می‌توان بر اساس نوع نگرش به دو دسته طبقه‌بندی کرد: نگرش‌های آماری و نگرش‌های هوش مصنوعی. سیستم هایی که از نگرش‌های آماری استفاده می‌کنند سعی در پیدا کردن نشانه‌هایی دارند که می‌تواند جمله‌های مهم و باارزش را نشان دهد. این نشانه‌ها می‌تواند تعداد تکرار کلمات، موقعیت جمله، عبارت‌های نشانه و یا ساختار نحوی یک جمله باشد. بر اساس این نشانه‌ها سیستم اهمیت هر جمله را با اختصاص دادن یک امتیاز نشان می‌دهد. جمله‌هایی با بالاترین امتیاز به عنوان جمله‌های باارزش تر در نظر گرفته‌شده و استخراج می‌شوند. در مقابل نگرش‌های مبتنی بر هوش مصنوعی سعی در فهم معانی جملات با استفاده از تکنیک‌های پردازش زبان دارد. با تکنیک‌هایی که برای پردازش زبان وجود دارد فهم عمیق یک متن در صورتی امکان‌پذیر است که ساختار متن ثابت باشد و یا از الگوهای مشخصی باشد که اغلب استفاده می‌شود.

سال ۱۹۹۳ را می‌توان شروع دوره جدیدی از تحقیقات در زمینه خلاصه‌سازی متن دانست. از سال ۱۹۹۳ به بعد تحقیقات حول چهار نگرش آماری، مبتنی بر دانش، مبتنی بر فهم سطحی و نگرش ترکیبی انجام گرفته است.

در سال ۱۹۹۴ Salton و همکارانش از یک مدل فضای برداری در بازیابی اطلاعات برای اندازه‌گیری میزان شباهت جمله‌ها و پیدا کردن جمله‌های مهم استفاده کردند. در سال بعد مسئله خلاصه‌سازی به صورت یک

مسئله دسته‌بندی آماری با استفاده از الگوریتم Bayesian توسط سه محقق به نامهای Pedersen،Kupiec و Chen مطرح شد. در سال ۱۹۹۵ Radev و McKeown از نتایج سیستم های استخراج اطلاعات برای ساخت خلاصه‌ای روان استفاده کردند. Lin و Hovy تکنیک آماری موقعیت جمله را در سال ۱۹۹۷ به کاربردند. Knight و Marcu در سال ۲۰۰۰ از تکنیک های آماری برای فشرده‌سازی جملات استفاده کردند. در همان سال Berger و Mittal نیز از تکنیک های مشابهی برای خلاصه‌سازی صفحات وب بهره بردند. کارهایی با نگرش فهم سطحی توسط Hirst و Morris در سال ۱۹۹۱، Benbrahim و Ahmad در سال ۱۹۹۵،Barzilay و Elhadad در سال ۱۹۹۷ و Baldwin و Morton در سال ۱۹۹۸ انجام گرفت. Svore در سال ۲۰۰۷ استفاده از تکنیک های یادگیری ماشین را برای خلاصه‌سازی پیشنهاد داد.

از اوایل سال ۲۰۰۰ بحث خلاصه‌سازی مبتنی بر کاربر و یا خلاصه‌سازی شخصی سازی شده مطرح‌شده و تا به امروز نیز مقالات بسیاری در این زمینه منتشر شده است. ایده اصلی خلاصه‌سازی شخصی سازی شده و یا مبتنی بر کاربر این است که کاربران مختلف با توجه به دانش و پس زمینه اطلاعاتی که دارند، دیدگاه‌های متفاوتی روی اسناد یکسان دارند. اکثر مقالاتی که جدیداً در خلاصه‌سازی متن ارائه می‌شوند، سعی می‌کنند به نوعی بحث شخصی سازی را در نظر بگیرند. [1]

۲-۳- مزایا و کاربردهای خلاصه‌سازی

پیشرفت اینترنت در سال ۱۹۹۰ و پیشرفت روزافزون پایگاه داده‌های الکترونیکی نیاز به توسعه و تکمیل ابزار جستجوی اطلاعات را بیشتر نمود. در انتهای دهه ۱۹۹۰ پروژه‌های تحقیقاتی روی ایجاد سیستمهای خلاصه‌سازی اتوماتیک در کشور آمریکا و کشورهای اروپایی تعریف شد از کابردهای خلاصه‌سازی می‌توان به خلاصه‌سازی اتوماتیک اخبار و ارسال آن‌ها از طریق پست الکترونیکی یا پیامک اشاره نمود. از دیگر کاربردهای آن خلاصه‌سازی تحقیقاتی، تجاری و خلاصه‌سازی صفحات وب برای آنکه در صفحه موبایل قابل نمایش باشد، است.

در یک سیستم بازیابی اطلاعات، نمایش خلاصه‌ای که به صورت اتوماتیک ایجاد شده است مناسب و مفید می‌باشد، در این صورت کاربر می‌تواند سریعاً تصمیم بگیرد که چه سندی مطلوب و ارزش باز کردن دارد. گوگل تا حدودی این کار را انجام می‌دهد و اطلاعات مختصری به همراه نتایج جستجو نشان می‌دهد. به خاطر علاقه روزافزون دولت، بخش بازرگانی، بخش دانشگاهی به این صنعت، تحقیقات و محصولات خلاصه‌سازی زیادی در سراسر دنیا موجود می‌باشد. سیستمهای سنتی بین سالهای ۱۹۵۰ تا ۱۹۸۰ برای جستجوی اطلاعات علمی استفاده می‌شد. سیستمهای خلاصه‌ساز امروزی در فیلدهای جدیدی همانند

صنعت مخابرات، ویراستارها، سیستمهای فیلترینگ و یادگیری زبان خارجی مورد استفاده قرار می‌گیرد. کار اصلی سیستم خلاصه‌ساز کمک به کاربر برای پیدا کردن اطلاعات مورد نیازش است. [3]

انسان ها با توجه به هوش و شعور ذاتی خود قادر به درک و فهم مفاهیم موجود در متن و ارتباط بین آنها می باشند و این در حالی است که انجام این عملیات توسط ماشین کار بسیار دشوار و پیچیده ای می باشد. از طرفی دیگر، انسان ها با توجه به سطح دانش و پس زمینه ی اطلاعاتی که دارند دید متفاوتی از خلاصه ی یک متن یکسان دارند. به عنوان مثال کسی که سال ها در زمینه شبکه های کامپیوتری به تحقیق و مطالعه پرداخته است با کسی که به تازگی قصد تحقیق و مطالعه در زمینه ی شبکه های کامپیوتری را دارد، متفاوت بوده و خلاصه ای که این دو فرد از یک متن در زمینه ی شبکه های کامپیوتری تولید می کنند قطعا یکسان نخواهد بود. زمینه های کاربردی خلاصه‌سازی خودکار متن گسترده است. با رشد قابل ملاحظه میزان اطلاعات در اینترنت، انتخاب اطلاعات مرتبط، کار مشکلی است. اطلاعات بطور همزمان روی بسیاری از کانالهای رسانه ای با نسخه های مختلف منتشر می شود. برای مثال یک صفحه روزنامه، صفحه خبر در وب، پیغامهای SMS، پخش اخبار رادیو و روزنامه سخنگو برای کسانیکه مشکل بینایی دارند. تنظیم اطلاعات برای کانالها و فرمتهای مختلف یک کار ویرایشی مهم است که خصوصا در خلاصه کردن متن اصلی نقش دارد.

خلاصه‌سازی خودکار متن می تواند این فرایند را کاملا خودکار و یا حداقل با تولید خلاصه پیش نویس، دراین روند یاری کند. همچنین می توان برای ارائه اسناد در زبانهای دیگر، ابتدا آنها را خلاصه و سپس ترجمه نمود که در بسیاری از موارد این امر برای برقراری ارتباطات یک سند در زبان خاص کافی است و از این رو نیازی به ترجمه هر سند به صورت دستی نیست و کار مترجمین انسان به این ترتیب ساده می شود. خلاصه‌سازی خودکار متن را همچنین می توان برای خلاصه کردن متن پیش از خوانده شدن توسط ایجاد کننده خودکار سخن، برای کاهش زمان مورد نیاز برای انتقال حقایق کلیدی، استفاده کرد.

از سیستم خلاصه ساز به ویژه می توان به منظور آماده سازی اطلاعات برای استفاده در دستگا ه ها ی کوچک موبایل، مانند PDA که نیازمند کاهش قابل توجه محتوی است مورد استفاده قرار داد.

در موتورهای جستجوگر معروف نظیر گوگل، هنگامی که عمل جستجو صورت می گیرد چند خط توضیح در مورد صفحه ای که به آن لینک داده شده است نیز آورده می شود که در واقع خلاصه ای از صفحه ی لینک داده شده می باشد. هرچقدر این خلاصه دقیق تر باشد به کاربر کمک می کند تا انتخاب دقیق تری داشته باشد. با توجه به این که فعالیت اصلی در خلاصه‌سازی خودکار متن، تشخیص مفاهیم می باشد، بنابراین اگر بتوانیم مفاهیم را به درستی تشخیص دهیم می توانیم از آن برای کاربردهای دیگر نیز استفاده نماییم. به عنوان مثال می توانیم سیستمی طراحی نماییم از روی یک مقاله، فایل ppt مناسب را ایجاد نماید؛ و یا از

روی یک فایل power point گزارشی مبسوط تهیه نماید. می توان سیستمی طراحی نمود که قادر به تشخیص مقالات مشابه باشد و یا اینکه سیستمی طراحی نماییم که میزان کیفیت متن نوشته شده را بررسی نماید. [3]

به هر حال علیرغم اینکه سال هاست در این زمینه مطالعه و پژوهش صورت می گیرد، اما هنوز سیستمی که قادر به خلاصه‌سازی تمام انواع متون با کیفیت بالا باشد، ارائه نشده است.

فصل ۳:

مروری بر کارهای مرتبط

۳-۱- انواع خلاصه‌سازی

سیستم های خلاصه ساز معمولا از دیدگاه های مختلفی تقسیم بندی می شوند. از دیدگاه آقای hovy سیستم های خلاصه‌سازی خودکار را می توان بر حسب منبع، هدف و خروجی به سه دسته عمده تقسیم بندی نمود.

۳-۱-۱- منبع خلاصه‌سازی

- **زبان**

روش خلاصه‌سازی بسته به اینکه بخواهیم متون تک زبانه و یا چند زبانه باشد، متفاوت می باشد. چراکه زبان های مختلف با یکدیگر تفاوت داشته و بسیاری از ویژگی ها که در یک زبان صادق است در زبان دیگر ممکن است صادق نباشد.

- **تعداد سند**

روش خلاصه‌سازی می تواند تک سند یا چند سندی باشد. در حالت چند سندی، پیچیدگی های کار بیشتر می شود. در این حالت وروی چندین سند می باشد که هر کدام از آنها ممکن است در مورد یک موضوعی بوده که همین امر یکی از مشکلات خلاصه‌سازی چند سندی می باشد. از طرف دیگر ممکن است تعدادی از اسناد حاوی مطالب ضد و نقیض باشند. وجود این دو مشکل باعث شده است که خلاصه‌سازی چند سندی، پیچیده تر از روش تک سند باشد.

- **نوع متن**

بسته به این که نوع متن، اخبار، علمی، گزارش و ... باشد روش بر خورد هم متفاوت می باشد. به عنوان مثال در خلاصه سازهای اخبار یکی از ساده ترین روش هایی که استفاده می شود این است که تیتر متون به همراه جمله های اول پاراگراف ها به عنوان خلاصه استفاده می شود و این در حالی است که این روش برای متون علمی جوابگو نمی باشد.

- **طول خلاصه**

طول خلاصه می توان کوتاه و یا طولانی باشد. البته در اکثر روش های خلاصه‌سازی جدید، میزان خلاص سازی خود یک پارامتر قابل تنظیم می باشد؛ یعنی اینکه کاربر می تواند تعیین کند که سند اصلی چند درصد خلاصه شود

- **جنس ورودی**

امروزه روش های خلاصه‌سازی متعددی برای انواع غیر متنی مثل ویدئو، صوت، عکس، نقشه و ... معرفی شده است. در این پایان‌نامه فقط به نوع متنی می پردازیم.

۳-۱-۲- هدف

- **قصد**

متن خلاصه به چه منظوری تولید می شود. هدف خلاصه‌سازی، هشدار، پیش نمایش، آگاهی‌دهنده، زندگی نامه و ... در تعیین روش خلاصه‌سازی تاثیرگذار می باشد.

- **کاربر**

اینکه خلاصه‌سازی برای قشر خاصی از کاربران تهیه می شود و یا اینکه برای عموم باشد می تواند تاثیر گذار باشد.

۳-۱-۳- خروجی خلاصه‌سازی

مشخص کردن فرمت خروجی خلاصه و اطلاعاتی که در آن قرار دارد. تقسیم بندی های زیر را برای خروجی داریم

- **اشتقاق:** روش خلاصه‌سازی می تواند چکیده ای و یا استخراجی باشد.

روش استخراجی :در روش استخراجی که عموما اکثر روش های خلاصه‌سازی هم از این نوع می باشند قسمت هایی از متن به عنوان چکیده انتخاب شده و سپس با یک چیدمان مناسب در کنار همدیگر قرار گرفته و به عنوان خلاصه تلقی می شود. اکثر این روش ها جملات را جدا کرده و به آنها امتیازدهی کرده و سپس جمله های با بالاترین امتیاز را به عنوان خلاصه انتخاب می کنند؛ بنابراین در این روش ساختار جمله ها تغییری نمی کند. سادگی این روش مسائلی را نیز به همراه دارد. یکی از آشکارترین مسائل حفظ حالت اولیه متن است. شکستن متون و نادیده گرفتن اتصال بین جمله ها میتواند خطرناک باشد. در این روش شانس کمی وجود دارد که اطلاعات مهم از بین نروند.

روش چکیده ای :در این روش که بسیار نزدیک به مدل ذهنی انسان می باشد ساختار جمله های چکیده ممکن است به کلی عوض شود. در این روش ممکن است جملات در متن اصلی موجود نباشند. در حقیقت متن خلاصه استنتاجی از متن اولیه است. از آنجا که برای ایجاد چکیده باید متن اصلی را فهمید و درک کرد لذا این نوع خلاصه‌سازی میتواند از تراکم و فشردگی بیشتری نسبت به خلاصه‌سازی استخراجی برخوردار باشد. در این روش ابتدا متن اصلی مورد تحلیل و بررسی قرار میگیرد و بعد از ارزیابی مفاهیم با

هم ادغام میشوند. این اطلاعات با هم ترکیب شده و تولید متن خلاصه را میدهند.

تا کنون بیشتر تحقیقاتی که در زمینه خلاصه‌سازی متون به صورت اتوماتیک انجام شده است، تولید خلاصه هایی از نوع استخراجی بوده است. [3]

فرمت خروجی: متن، تصویر، نقشه، ویدئو و...

فصل ۴:

روش‌ها و مراحل خلاصه‌سازی

۴-۱- مراحل خلاصه‌سازی

خلاصه‌سازی متن از سه گام اصلی تشکیل شده است .مرحله پیش پردازش متن، مرحله پردازش متن وتولید خلاصه .مرحله پیش پردازش شامل حذف اطلاعات اضافی و غیر مهم و ریشه یابی میباشد؛ به عبارت دیگر، پیش پردازش متن در این بخش انجام میشود .بعد از پیش پردازش، بخش مهم متن باقی میماند .بخش اصلی خلاصه‌سازی همان تفسیر متن و در حقیقت مرحله پردازش متن میباشد .این مرحله شامل پیدا کردن و امتیازدهی به کلمه های مهم میباشد .در مرحله آخر، خلاصه متن بر اساس جمله ها و امتیازدهی مربوطه ایجاد خواهد شد .مطابق با این مراحل تکنیک پیشنهادی نشان داده خواهد شد[۳] .

۴-۱-۱- پیش پردازش متن:

اهمیت تحلیل متن برای هر کاربردی در پردازش زبان طبیعی کاملا روشن است اما اهمیت درک متن برای هر کاربردی متفاوت است؛ مثلا درک سوال در یک سیستم پرسش و پاسخ برای سیستم خیلی مهم تر از تحلیل و درک متن در یک سیستم تشخیص صحبت میباشد .به هر حال مهمترین قسمت در هرسیستم کاربردی پردازش زبان طبیعی تحلیل متن ورودی خواهد بود .سطوح مختلف تحلیل متن که بیانگر هدف تحلیل میباشند را می توان به صورت ذیل تقسیم کرد:

تصحیح متن :تبدیل جمله به فرم استاندارد .

لازم به ذکر است حسب پردازشی که روی متن انجام خواهد پذیرفت ممکن است یک و یا چندین مورد از موارد بالا در مرحله پیش پردازش متن انجام نشود.

بخشی از تحلیل متن استخراج اطلاعات است .این استخراج اطلاعات میتواند با استفاده از برچسب زنی کلمات متن باشد .برچسب زنی کمک قابل توجهی به تجزیه صحیح جملات و تصمیم گیری در مورد کلماتی که در شرایط مختلف معانی متفاوتی دارند، خواهد نمود .تعیین کلمات کلیدی و استخراج واژه های موجود در متن از جمله دیگر کارهایی است که در بخش استخراج اطلاعات در تحلیل متن به کار میآید.

در پیش پردازش متن کار یکنواختسازی متن انجام میشود .هدف از یکنواخت سازی تبدیل متن به یک فرم یکنواخت میباشد تا فرایند خلاصه‌سازی بطور صحیح و با دقت لازم انجام شود .از اینرو یکنواخت سازی متن شامل مراحل ذیل میباشد:

تجزیه مورفولوژی :در تجزیه مورفولوژی هر کلمه از نظر حضور در واژگان بررسی میشود .اگرکلمه ای در واژگان نباشد از قوانین مورفولوژی برای رسیدن به اصل واژه استفاده میشود؛ مثلا یکی از قوانین این قسمت که برای پیدا کردن لغت اصلی استفاده میشود به صورت زیر است: اسم+ها > اسم

به عنوان مثال دیگر میتوان تجزیه افعال مرکب که حاصل صرف فعل میباشد را مورد توجه قرار داد.
همانند : داشتم میرفتم که از مصدر رفتن است.
کار دیگری که در این قسمت انجام میشود و از اهمیت خاصی برخوردار است جایگزینی کلمات معادل همانند کلمات unix و یونیکس است .در متون فارسی این اشکال به وفور مشاهده میشود و بیشتر به خاطر استفاده از واژه های بیگانه و واژه های هم معنی که در جامعه جا افتاده نشات میگیرد .مساله دیگر آن است که خیلی از واژه هایی که از یک واژه ریشه ساخته شده اند، معانی متفاوتی دارند .به عنوان مثال میتوان به کلمه داده در «داده های کامپیوتری» و «داده شده است» اشاره نمود .هر چند که در هردو حالت، «داده »از ریشه «دادن» استخراج شده است، با اینحال کلمه داده های کامپیوتری مبین "اطلاعات داده شده "میباشد و لذا یکسان فرض کردن واژه داده در عبارات داده های کامپیوتری با "داده "در" داده شده است"، صحیح نیست .برای حل این مشکل باید از مدلهای آماری دو کلمه ای و قوانین مورفولوژی احتمالی استفاده کرد تا اطلاعات وابسته به متن در اختیار ماشین قرار گیرد مورفولوژی با توجه به این مدلهای پیچیده تر بطور صحیح صورت بگیرد .
مشکل دیگری از این نوع، کلماتی با اشکال متفاوت و معنای متفاوت میباشند که از یک ریشه استنتاج میشوند؛ یعنی در اصل کلمات متفاوتی هستند ولی پس از مورفولوژی به کلمات معادلی تبدیل میشوند.
رفع ابهام: موثرترین انواع رفع ابهاماتی که در خلاصه‌سازی مطرح میباشد رفع ابهام مربوط به ارجاعات و رفع ابهام ضمایر میباشد .با بررسی متون فارسی مشخص میشود که ضمایر شخصی به ندرت استفاده میشوند بخصوص در کتب و بیشتر از ضمایر اشاره به همراه نوع اشاره استفاده میشود، همانند این سیستم، این تکنیک، همان نتایج .اکثر این ارجاعات رجوع به ماقبل میباشد؛ به عبارت دیگر مناسبترین عبارت مورد رجوع از جملات قبل به نام عبارت کاندیدا انتخاب شود و به جای عبارت مبهم قرار گیرد .تشکیل یک عبارت از جمله ماقبل که شامل کلمه مورد رجوع باشد به اطلاعات دیگری همانند همنشینی کلمات نیاز دارد؛ مثلا در جملات "سیستمهای اطلاعات مدیریت" ." ..... این سیستمها"، تشخیص اینکه سیستمهای اطلاعات مدیریت، مورد ارجاع میباشد و نه سیستمهای اطلاعات با استفاده از چند – وزنی قابل انجام است؛ بنابراین به طور کلی ابهاماتی که در زمینه تحلیل متن وجود دارد شامل ابهام در تعیین عنوان و موضوع اصلی یک متن میباشد، ابهام در معنای جملاتی که چندین معنی متفاوت دارند و همچنین ابهام در تجزیه جمله نیز اتفاق میافتد .این مساله در حالتی اتفاق میافتد که جمله با چندین قانون قابل اشتقاق باشد و معمولا تنها یکی از این قوانین صحیح است.
تشخیص و حذف کلمه های بی اثر :در این مرحله، کلمات بی اثر که از مرحله قبل برچسب خورده اند و یا توسط محاسبات جدید عضو کلاس کلمات بی اثر خواهند شد از متن اصلی حذف میشوند.

ریشه یابی کلمات و یافتن کلمات کلیدی :به دو صورت میتوان ریشه یابی انجام داد .در نوع اول از الگوریتمهای ریشه یاب استفاده میشود و در نوع دوم از پایگاه داده هایی که ریشه لغات را داشته باشند استفاده خواهد شد . در زبان انگلیسی از الگوریتم پورتر استفاده میشود[۶] .

۴-۱-۲- پردازش متن

در این بخش به موجودیتهای متن که در مرحله پیش پردازش متن ایجاد شده‌اند امتیاز داده میشود.

یافتن موضوع متن :تفسیر هر قسمت یا بخش از متن برای یافتن مفهوم، موضوع یا عنوانی که متن در توضیح آن آمده است.

ارتباط بین قسمتهای متن: یافتن ارتباطهای منطقی معنایی و ظاهری بین جملات مختلف بر اساس تشابه یا تفاوت بین جملات یا معنای آنها.

تجزیه جمله:استخراج اطلاعاتی مانند فعل یا اسم یا حرف بودن کلمه

درک معنی متن :تبدیل هر جمله به یک فرم انتزاعی که بیانگر معنای جمله باشد .هدف از تحلیل در این مورد پی بردن به معنای جمله و یا جملات است.

۴-۱-۳- تولید خلاصه

در این بخش کارهای ذیل انجام میشود:

۱- مشخص نمودن اندازه خلاصه :در سیستمهای خلاصه‌ساز اندازه خلاصه بر اساس تعداد جملات، تعداد کلمات و یا درصدی از سند ابتدایی مشخص میشوند .به عنوان مثال خلاصه‌ساز به طور پیش فرض خلاصه ٪۲۵ایجاد مینماید و کاربر میتواند اندازه خلاصه را تغییر دهد .این مقادیر نسبی میباشند .

۲-استخراج جملات بر اساس امتیازهای داده شده

۳- ایجاد خلاصه: برای آنکه پیوستگی متن حفظ شود، جمله‌های استخراجی بر اساس ترتیبشان در متن در خلاصه قرار داده می شوند.

۴-۲- رویکردهای خلاصه‌سازی

۴-۲-۱- روش‌های کلاسیک

در این روش مشخصه هایی همچون فرکانس کلمات، موقعیت جملات که از روی خلاصه های تولید شده توسط انسان اقتباس شده اند، برای کلمات یا جملات محاسبه شده و بر اساس آنها به جمله امتیاز داده میشود. معیار امتیاز دهی در این روش خلاصه های انسانی میباشد. این رویکردنیاز به درک سطحی متن دارد ومعمولاً درگیر تحلیل نحوی جملات می باشد و برای استخراج جملات مهم از ویژگی های موقعیت جملات

در یک پاراگراف، فرکانس تکرار کلمات، شناسایی کلمات کلیدی، شناسایی کلمات منطبق بر عنوان و ... استفاده می‌کنند. به این علت به این روش رویکرد کلاسیک میگویند چون پایه ای برای رویکردهای مدرن محسوب می شود.

۴-۲-۲- روش های TF_ISF

این روش معادل معیار فرکانس کلمه –معکوس فرکانس سند در مفهوم بازیابی اطلاعات (IR) است. در خلاصه‌سازی متن یک سند d داریم و میخواهیم مجموعه ای از جملات مرتبط شامل شده در خلاصه مستخرج شامل برخی جمله های dرا انتخاب کنیم؛ بنابراین شبیه TF-IDF در بازیابی اطلاعات در اینجا معیاری وجود دارد که فرکانس کلمه –معکوس فرکانس جمله (TF-ISF) نامیده میشود. فرکانس کلمه تعداد تکرار کلمه در متن میباشد. فرکانس جمله تعداد جملات سند است که حاوی کلمه مشخص هستند. این مشخصه برای تمام کلمات هر جمله محاسبه میشود. وزن هر جمله از مجموع وزن کلمات آن جمله تقسیم بر تعداد کلمات آن بدست میاید و در نهایت جملات با امتیاز بیشتر به عنوان بخشی از خلاصه انتخاب میشوند. [4]

فرمول این روش به صورت زیر میباشد:

$$\mathrm{tf}_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \qquad (۱)$$

$$\mathrm{isf}_i = \log \frac{N}{n_i} \qquad (۲)$$

به طوریکه $tf_{i,j}$ بیانگر تعداد تکرار کلمه و isf_i بیانگر عکس تعداد تکرار جمله از کلمه Iام میباشد. در رابطه ۲، N تعداد کل جملات و n_i تعداد جملاتی است که در آن کلمه iام وجود دارد.

حال وزن هر کلمه به صورت زیر محاسبه می شود:

$$\mathrm{w}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{isf}_i$$

در نهایت وزن جملات را از تقسیم مجموع وزن کلمات هر جمله بر تعداد کلمات آن جمله بدست میاوریم.

برای مثال وزن جملات در متن زیر به این صورت تولید میشود :

۱- هفته صرفه‌جویی در مصرف آب آغاز شد .

۲- هدف از تعیین هفته‌ای تحت این عنوان دعوت مردم به استفاده بهینه و جلوگیری از هدردادن منابع آب اعلام شده است.

۳- طی این هفته آموزشهای لازم جهت مصرف بهینه از طریق گوناگون به مردم ارائه می شود.

تعداد تکرار کلمات و تعداد جملات حاوی آن کلمه :

هفته : ۳ و ۲	صرفه‌جویی : ۱ و ۱	در : ۱ و ۱	مصرف : ۲ و ۲	آب : ۲ و ۲	آغاز : ۱ و۱
شد : ۱ و ۱	هدف : ۱ و ۱	از : ۳ و ۲	تعیین : ۱ و ۱	تحت : ۱ و ۱	این : ۲ و۲
عنوان : ۱ و۱	دعوت : ۱ و ۱	مردم : ۲ و۲	به : ۲ و ۲	استفاده : ۱ و۱	و : ۱ و ۱
جلوگیری : ۱ و ۱	هدردادن : ۱ و ۱	منابع : ۱ و ۱	اعلام : ۱ و ۱	شده است: ۱ و۱	طی : ۱ و ۱
آموزشهای : ۱ و۱	لازم : ۱ و ۱	جهت : ۱ و ۱	بهینه : ۱ و ۱	گوناگون : ۱ و ۱	ارائه : ۱ و ۱
می‌شود : ۱ و ۱	طریق : ۱و۱				

حال ورن کلمات را طبق فرمول بالا محاسبه کردیم :

هفته : ۰.۱۷۶	صرفه‌جویی : ۰.۴۷۷	در : ۰.۴۷۷	مصرف : ۰.۱۷۶	آب : ۰.۱۷۶	آغاز : ۰.۴۷۷
شد : ۰.۴۷۷	هدف : ۰.۴۷۷	از : ۰.۱۷۶	تعیین : ۰.۴۷۷	تحت : ۰.۴۷۷	این : ۰.۱۷۶
عنوان : ۰.۴۷۷	دعوت : ۰.۴۷۷	مردم : ۰.۱۷۶	به : ۰.۱۷۶	استفاده : ۰.۴۷۷	و : ۰.۴۷۷
جلوگیری : ۰.۴۷۷	هدردادن : ۰.۴۷۷	منابع : ۰.۴۷۷	اعلام : ۰.۴۷۷	شده است: ۰.۴۷۷	طی : ۰.۴۷۷
آموزشهای : ۰.۴۷۷	لازم : ۰.۴۷۷	جهت : ۰.۴۷۷	بهینه : ۰.۴۷۷	گوناگون : ۰.۴۷۷	ارائه : ۰.۴۷۷
می‌شود : ۰.۴۷۷	طریق : ۰.۴۷۷				

سپس وزن جملات را از مجموع وزن کلمات تشکیل دهنده آن جمله تقسیم بر تعداد کلمات آن جمله به دست میاوریم . وزن جملات از قرار زیر است :

جمله اول : ۰.۳۴۸

جمله دوم : ۰.۳۵۶۶

جمله سوم : ۰.۳۷۶

همانگونه که ملاحظه میشود وزن جملات دوم و سوم بیشتر است و برای خروجی مناسبتر است.

۴-۲-۳- تکنیک های یادگیری ماشینی[1]

در این روش با مجموعه ای از سندهای آموزشی و خلاصه های استخراجی آنها فرآیند خلاصه‌سازی به عنوان یک مسئله طبقه بندی مدل میشود. در فاز یادگیری داده آموزش که همان جمله‌ها میباشند به الگوریتم طبقه بندی داده میشود. جملات به جملات خلاصه و جملات غیر خلاصه بر اساس مشخصه هایی که دارا

[1] Machine Learning Technique

هستند طبقه بندی میشوند. احتمالهای طبقه بندی به صورت آماری از مجموعه آموزش یاد گرفتـه میشـود. مهمترین نکته‌ای که در یادگیری ماشینی اهمیت دارد انتخاب و استخراج ویژگیهاست که با اسـتفاده از ایـن ویژگیها بتوان دسته بندی و انتخاب جملات را انجام داد. معمولا از الگوریتم ژنتیک برای اسـتخراج ویژگیهـا استفاده میشود. در هر موضوعی ویژگیهای مورد استفاده متفاوت میباشد. برای نمونه ویژگیهایی مثـل طـول جمله، مکان جمله(چندمین جمله متن)، شباهت به عنوان مـتن، شـباهت بـه کلمـات کلیـدی، شـباهت بـه جملات دیگر، شباهت به جمله مرکزی متن(جمله ای که اهمیت آن از همه بیشتر اسـت) بـرای اسـتفاده در سیستمهای خلاصه‌سازی که از بحث یادگیری ماشین بهره گرفته اند، استفاده میشود.

از دیگر ویژگیهای مطرح میتوان به داشتن اسامی مشهور، تکرار چندین باره یک عبارت خاص، تکرار کلمـات غیر ضروری نام برد. معمولا برای استفاده از این روش از الگوریتمهای معروفی مثل Naive Bayes و C4.5 نام برد. برای راحتی کار در الگوریتم Naive Bayes فرض مستقل بودن ویژگیها از هم را هم در نظر میگیرند. این تکنیک نسبت به روشهای دیگر خلاصه‌سازی انعطاف پذیری بیشتری دارد.

۴-۲-۴- روش های مبتنی بر گراف[1]

در این روش بعد از گامهای پیش پردازشی برای مشخص کردن عناوین (اسامی مهـم و..) در مـتن، سـند بـه شکل گرافی غیر جهت دار که جملات، گره های تشکیل دهنـده آن هسـتند ارائـه میشـود و تئـوری گـراف میتواند به راحتی برای تجسم شباهت بین سندی و درون سندی بکار رود. در واقع در این روش به هر جمله از سند یک گره اختصاص داده میشود و اگر یک جمله با جمله ای دیگر ارتباط داشته باشد بین آن دو گـره یک یال رسم میشود. معیار ارتباط بین جملات داشتن کلمات کلیدی بیشتر، نزدیک بودن به عنوان متن و ... میباشد. در نهایت گره هایی را که از درجه بیشتری برخوردار هستند به عنوان خروجی انتخاب میشوند. برای مثال در این روش پس از آنکه جملات از یک متن استخراج شد، آنها را در یک گراف ترسیم میکنیم که این گراف یک گراف غیر جهت‌دار و وزن‌دار میباشد. ارتباط بین نودها بر اساس میزان شباهت بین جملات است. میزان شباهت بین جملات طبق ویژگیهایی مثل کلمه کلیدی، شباهت به عنوان متن، تعداد کلمات پر تکرار متن و ... بیان میشود. [4]

برای نمونه جهت خلاصه سازی چند سندی از روش گراف استفاده شده است.۳ سند ورودی به همـراه مـتن آنها به سیستم داده شده است و سیستم به شکل گرافی آن را مدل میکند.داده هـای ورودی بـه شـکل زیـر میباشند :

[1] Graph Based Approaches

SNo	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, "will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq "did not end" and that Britain is still "ready, prepared, and able to strike Iraq."
10	d5s2	In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq "will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

همانگونه که در این شکل مشخص است تعداد سندهای ما برابر است با ۵ عدد و تعـداد جملـه هـا ۱۱ عـدد است. سپس با استفاده از فرمول شباهت کسینوسی که در زیر بیان شده است شباهت بـین جمـلات را بـه دست آوردیم :

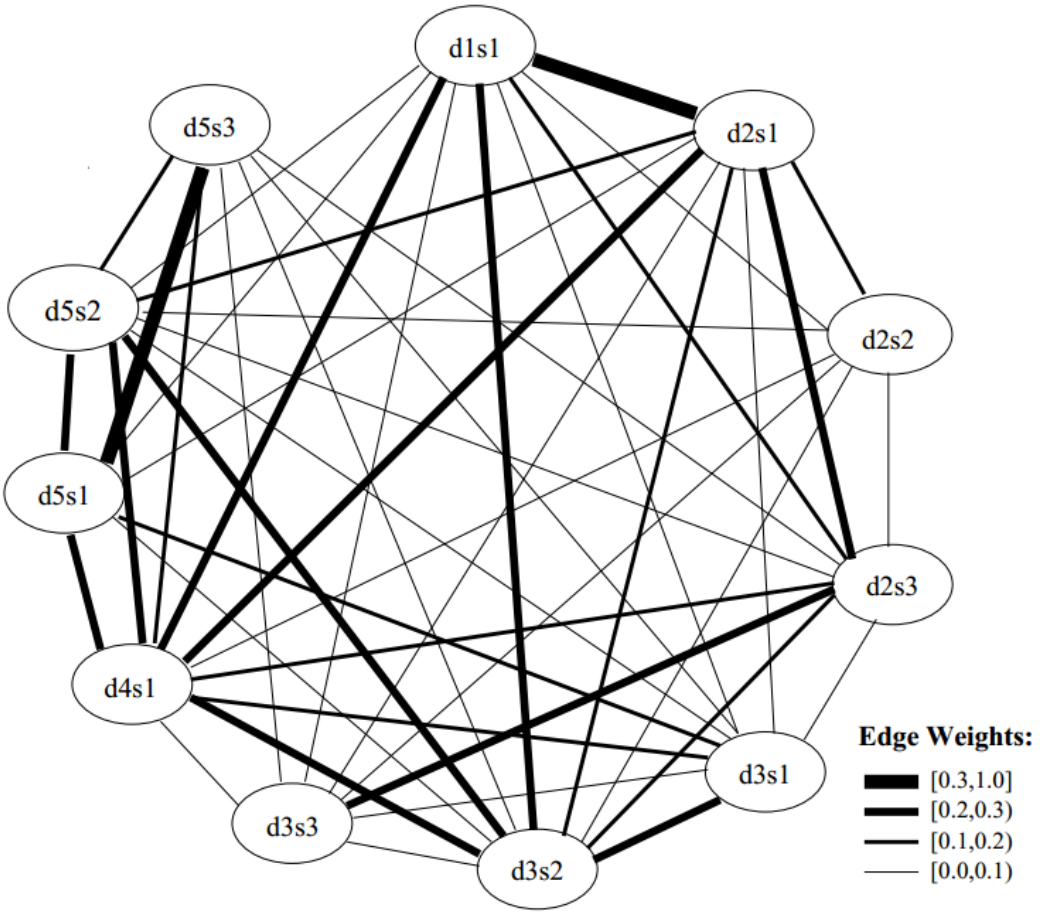
$$\text{idf-modified-cosine}(x,y) = \frac{\sum_{w \in x,y} \text{tf}_{w,x}\text{tf}_{w,y}(\text{idf}_w)^2}{\sqrt{\sum_{x_i \in x}(\text{tf}_{x_i,x}\text{idf}_{x_i})^2} \times \sqrt{\sum_{y_i \in y}(\text{tf}_{y_i,y}\text{idf}_{y_i})^2}}$$

در این شکل $\text{tf}_{w,s}$ تعداد رخداد کلمه W در جمله S می‌باشد و X ,Y نیز دو جمله از سند میباشند. شباهت

بین هر دو جمله در این مثال طبق ماتریس زیر میباشد :

	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00

همانطور که در ماتریس بالا مشخص است بیشترین میزان شباهت ۱ میباشد که برای جمله های عینا مانند هم این اتفاق میفتد. این ماتریس یک ماتریس متقارن است به این معنی که شباهت جمله I و J برابر است با شباهت بین جمله های J و I . حال با استفاده از ماتریس بالا به رسم گراف میپردازیم. تعداد نودهای این گراف برابر با تعداد جمله های پیکره است. وزن بین یالها نیز با محاسبه شباهت کسینوسی قابل محاسبه است . شکل آن نیز به این صورت است :



همانطور که مشخص است یالهای پر رنگ تر نشاندهنده شباهت بیشتر میباشند.حال برای بدست آوردن خلاصه نودهایی را که وزن اتصالات آن بیشتر است به عنوان خروجی در نظر میگیریم. در این مثال جمع وزن یالهای متصل به هر گره به صورت زیر میباشد :

جمله ۱	جمله ۲	جمله ۳	جمله ۴	جمله ۵	جمله ۶	جمله ۷	جمله ۸	جمله ۹	جمله ۱۰	جمله ۱۱
۲.۳۲	۲.۳	۱.۲۸	۲.۲	۲.۶۱	۲.۳۹	۱.۵۵	۲.۸۱	۲.۲۲	۲.۱۳	۱.۹

۴-۲-۵- روشهای زنجیره لغوی

زنجیره های لغوی به عنوان خوشه هایی از کلمات از نظر معنایی مرتبط با هم تعریف شده اند؛ مثلا {خانه،

سرا، اتاق} یک زنجیره است در حالی که "خانه و سرا " هم معنی هستند، "اتاق" بخشی از یک "خانه " است. در حالت کلی بیشتر خلاصه سازهای مبتنی بر زنجیره لغوی روش یکسانی را به صورت زیر دنبال میکنند:

۱-تولید زنجیره های لغوی

۲- امتیاز دادن به زنجیره ها

۳- یافتن قویترین این زنجیره ها که برای ارزش دهی و استخراج جملات کلیدی متن به کار میرود.

زنجیره های لغوی به عنوان خوشه هایی ازکلمات که از نظر معنایی مرتبط با هم هستند مانند هم معنی، مخالف، اشتقاق و شمول تعریف شده اند .در روش زنجیره های لغوی معمولاً برای ایجاد زنجیره ها WordNet را مورد استفاده قرار میدهند. دو الگوریتم مختلف برای ایجاد زنجیره ها وجود دارد:رویه غیرمبهم حریصانه وغیر حریصانه. در رویه غیر مبهم حریصانه که زنجیره‌ی یک کلمه فقط به وسیله زنجیره‌های کلمات قبل ازآن در متن مشخص میشود به این صورت که بر اساس روابط تعیین شده اگر زنجیره ی مرتبط با کلمه در زنجیره های از قبل موجود یافت شود آن را در همان زنجیره درج می کند و در غیر این صورت برای آن زنجیره ی جدید می سازد .در مقابل الگوریتم غیرحریصانه تا هنگامی که همه ی کلمات درمتن پردازش شود منتظرمی ماند سپس با توجه به تمام لغات متن زنجیره ی هر کدام را یافته یا ایجاد می کند[4] .

۴-۲-۶- روش پیاده سازی شده

این روش مخلوطی از دو روش یادگیری ماشین وTF-IDF است. با توجه به اینکه در این روش نیاز به داده های عظیم برای یادگیری وجود داشت از مجموعه داده های همشهری استفاده کردیم. از ویژگی های مهم این مجموعه این بود که داده های خبری در همه زمینه های اجتماعی، سیاسی، فرهنگی، هنری، ورزشی و ... وجود داشت و به همین دلیل برای داده های یادگیری مناسب میباشد.

این داده ها مجموعه ای از خبرهای روزنامه همشهری از سال ۷۵ تا ۸۰ میباشد. ابتدا خبرها از این مجموعه جدا شد. خروجی این قسمت مشتمل بر ۵۰۰۰ فایل بود که در هر فایل یک خبر وجود داشت. سپس با استفاده از معیار TF_IDFبه هر کلمه موجود یک وزن داده شد. مراحل وزن دهی به هر کلمه به این صورت محاسبه شده است.

۱. تعداد تکرار هر کلمه در کل فایلها محاسبه میشود که به آن Term Frequencyمیگویند.

۲. تعداد فایلهایی که شامل کلمات هستند نیز محاسبه شده است که به آن Document Frequencyمیگویند.

۳. با استفاده از فرمول

Log(1+Term Frequency) * log(N/Document Frequency) که Nبیانگر تعداد کل اسناد میباشد وزن هر کلمه به دست میاید. علت این که از معکوس فرکانس سند استفاده میشود این است که ممکن است تعداد تکرار یک کلمه زیاد باشد اما کلمه ارزش چندانی نداشته باشد؛ مانند کلمات "به"،"از" و... . سپس به صورت زوجهای کلمه و وزن در یک فایل ذخیره کردیم.

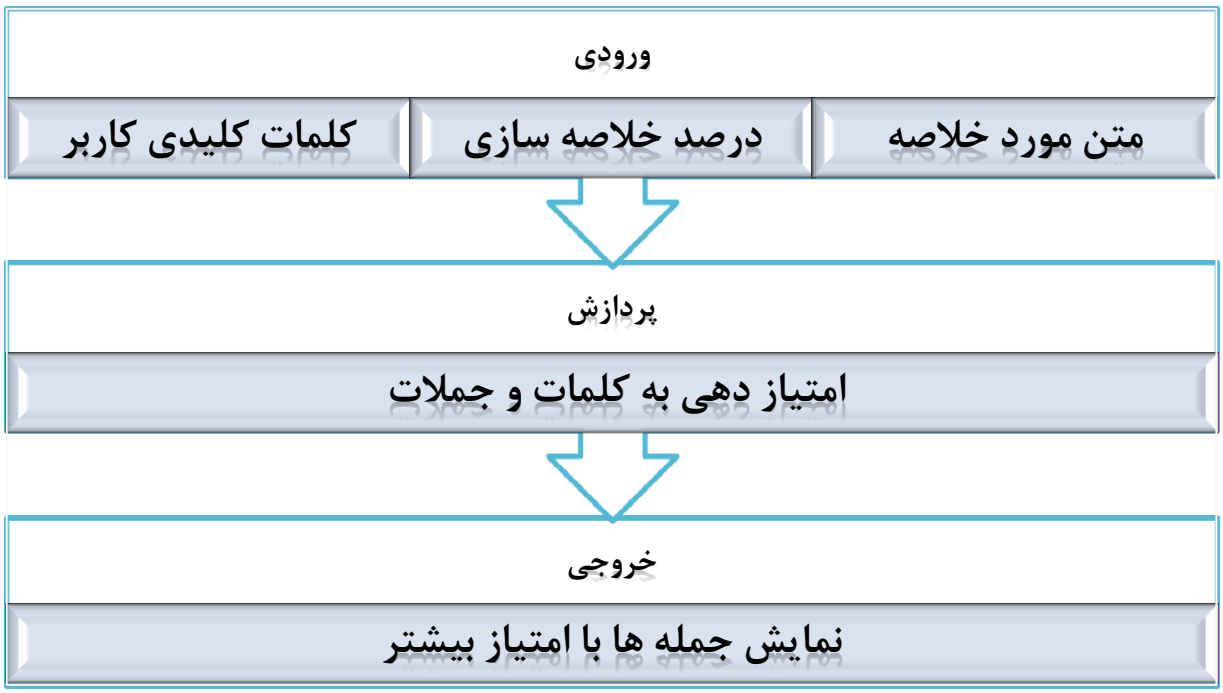
۴. حال برای هر جمله از متن ورودی وزن آن را به دست میاوریم که در واقع وزن جمله میباشد. وزن هر جمله عبارت است از میانگین وزن کلمات آن جمله یعنی برای هر جمله مجموع وزنهای کلمات تشکیل دهنده آن جمله بدست میاید و بر تعداد کلمات تقسیم میشود. البته لازم به ذکر است که لیستی از کلماتی که اهمیتی ندارند و StopWordنامیده میشوند نیز تهیه شده است که اگر کلمه ای در جمله StopWordبود وزن آن کلمه صفر در نظر گرفته میشود. همچنین ممکن است بعضی از کلمات در لیست کلمات وزن دار وجود نداشته باشند برای چنین کلماتی وزن ۱.۵ در نظر گرفته شده است.

۵. به کلمات کلیدی کاربر که در ورودی وارد میکند نیز امتیاز ویژه ای تعلق میگیرد.

۶. همچنین به جملات خاص نیز امتیاز داده شده است. در حال حاضر جملات خاص ما جملات حاو ی عبارات اشاره هستند، به جملات حاوی عبارات اشاره‌ی فارسی شامل" بنابراین" ،" نتیجه "،"موضوع "و" معرفی "امتیاز مثبت و به جملات حاوی کلمات" مثال" ، مثلا "و" نمونه " علامت منفی نسبت داده می شود.

سپس جملات را بر اساس وزن آنها مرتب میکنیم و جملات با ارزش را به عنوان متن خلاصه بر میگردانیم. لازم به ذکر است که معمولا جمله اول در هر متن از ارزش بالاتری برخوردار است به همین علت جمله اول از وزن بالاتری نسبت به بقیه جملات برخوردار است. همچنین تعداد جملات براساس انتخاب کاربر میباشد اما به صورت نرمال ۲۵ درصد جملات خلاصه میشود.

مراحل خلاصه سازی طبق نمودار زیر میباشد :



مثال : برای نمونه مراحل خلاصه سازی متن زیر به این صورت می‌باشد:

متن منبع :

شبکه های کامپیوتری (بزرگراههای اطلاعات رسانی) شهری، کشوری و جهانی تحول بزرگی در معاملات سهام شرکتها بوجود آورده اند و هرکس با هر مبلغ پول، بدون داشتن نیاز به دلالان، معامله گر سهام برای خود شده است. هر روز تنها در آمریکا ۶۴۰ میلیون سهم معامله می شود. میزان روزانه معاملات سهام در خاور دور و اروپا حدود یک میلیارد است. در معاملات سهام از طریق دلال، وی راهنمائی و توصیه لازم را می کند، ولی در معاملات کامپیوتر مستقیم، یک معامله گر انفرادی باید به خودش تکیه کند به ابتکار خودش اقدام کند و ریسک زیاد است. مشاور معامله گران انفرادی، صفحات اقتصادی روزنامه ها هستند که سهام روبه ترقی را توصیه می کنند. هنوزتاثیر این نوع معاملات الکترونیک بر مراکز بزرگ بورس مثلا وال استریت روشن نیست. سیستم کامپیوتر مبلغ خرید را بلافاصله از حساب معامله گر کسر و مبلغ فروش را به آن اضافه می کند. بنابراین معامله گر انفرادی (تنها) سعی می کند که در اوقات فراغت از کامپیوتر دور نشود و از توصیه های روزنامه ها غافل نماند. این معاملات مرز نمی شناسد و جنبه جهانی دارد.

حال جمله ها از متن استخراج میشوند:

جمله ها :

۱. شبکه های کامپیوتری (بزرگراههای اطلاعات رسانی) شهری، کشوری و جهانی تحول بزرگی در معاملات سهام شرکتها بوجود آورده اند و هرکس با هر مبلغ پول، بدون داشتن نیاز به دلالان، معامله گر سهام برای خود شده است.
۲. هر روز تنها در آمریکا ۶۴۰ میلیون سهم معامله می شود
۳. میزان روزانه معاملات سهام در خاور دور و اروپا حدود یک میلیارد است
۴. در معاملات سهام از طریق دلال، وی راهنمائی و توصیه لازم را می کند، ولی در معاملات کامپیوتر مستقیم، یک معامله گر انفرادی باید به خودش تکیه کند به ابتکار خودش اقدام کند و ریسک زیاد است
۵. مشاور معامله گران انفرادی، صفحات اقتصادی روزنامه ها هستند که سهام روبه ترقی را توصیه می کنند

۶. شبکه های کامپیوتری (بزرگراههای اطلاعات رسانی) شهری، کشوری و جهانی تحول بزرگی در معاملات سهام شرکتها بوجود آورده اند و هرکس با هر مبلغ پول، بدون داشتن نیاز به دلالان، معامله گر سهام برای خود شده است
۷. هنوزتاثیر این نوع معاملات الکترونیک بر مراکز بزرگ بورس مثلا وال استریت روشن نیست
۸. سیستم کامپیوتر مبلغ خرید را بلافاصله از حساب معامله گر کسر و مبلغ فروش را به آن اضافه می کند
۹. بنابراین معامله گر انفرادی (تنها) سعی می کند که در اوقات فراغت از کامپیوتر دور نشود و از توصیه های روزنامه ها غافل نماند
۱۰. این معاملات مرز نمی شناسد و جنبه جهانی دارد

حال با استفاده از وزنهای کلمات که از قبل محاسبه شده است و همچنین با استفاده از کلمات ایست فارسی[1] و کلمات کلیدی کاربر اقدام به وزن دهی به جملات میکنیم.وزنهای این جملات به صورت زیر میباشد:

۱. [شبکه های کامپیوتری (بزرگراههای اطلاعات رسانی) شهری، کشوری و جهانی تحول بزرگی در معاملات سهام شرکتها بوجود , ۰.۲۲۴۲۹۹۰۶۵۴۲۰۵۶۱]
۲. [هر روز تنها در آمریکا ۶۴۰ میلیون سهم معامله می شود, ۰.۲۰۵۸۸۲۳۵۲۹۴۱۱۷۶]
۳. [این معاملات مرز نمی شناسد و جنبه جهانی دارد, ۰.۲۰۴۵۴۵۴۵۴۵۴۵۵]
۴. [هنوزتاثیر این نوع معاملات الکترونیک بر مراکز بزرگ بورس مثلا وال استریت روشن نیست, ۰.۱۹۷۵۳۰۸۶۴۱۹۷۵۳۱]
۵. [بنابراین معامله گر انفرادی (تنها) سعی می کند که در اوقات فراغت از کامپیوتر دور نشود و از توصیه های روزنامه ها غافل نماند, ۰.۱۹۵۱۲۱۹۵۱۲۱۹۵۱۲]
۶. [مشاور معامله گران انفرادی، صفحات اقتصادی روزنامه ها هستند که سهام روبه ترقی را توصیه می کنند, ۰.۱۹۳۵۴۸۳۸۷۰۹۶۷۷۴]
۷. [سیستم کامپیوتر مبلغ خرید را بلافاصله از حساب معامله گر کسر و مبلغ فروش را به آن اضافه می کند, ۰.۱۹۳۵۴۸۳۸۷۰۹۶۷۷۴]
۸. [میزان روزانه معاملات سهام در خاور دور و اروپا حدود یک میلیارد است, ۰.۱۵۹۰۹۰۹۰۹۰۹۰۹۰۹]
۹. [آورده اند و هرکس با هر مبلغ پول، بدون داشتن نیاز به دلالان، معامله گر سهام برای خود شده است, ۰.۱۴۸۳۵۱۶۴۸۳۵۱۶۴۸]
۱۰. [در معاملات سهام از طریق دلال، وی راهنمائی و توصیه لازم را می کند، ولی در معاملات کامپیوتر مستقیم، یک معامله گر انفرادی باید به خودش تکیه کند به ابتکار خودش اقدام کند و ریسک زیاد است, ۰.۱۴۸۳۵۱۶۴۸۳۵۱۶۴۸]

همانطور که مشاهده میشود جملاتی که طول آنها بزرگتر است (از تعداد کلمات بیشتری تشکیل شده است) وزن کمتری را نسبت به سایر جملات دارند.

۴-۳- روش های ارزیابی خلاصه ساز ها

ارزیابی خلاصه ها و سیستم‌های خلاصه سازی متن رویه پیچیده ای است. ضرورت وجود معیاری برای مقایسه خلاصه ها چه از نظر کلمات بکاررفته در آنها و چه از لحاظ خاص بودن ، احساس میشود. در خلاصه سازی خودکار متن، ممکن است چندین خلاصه خوب برای یک متن خاص وجود داشته باشند که به این ترتیب عمل ارزیابی این خلاصه ها در مقایسه با یک خلاصه مرجع ثابت و تغییرناپذیر، کار ارزیابی را با مشکل روبرو میکند. همچنین با توجه به اینکه نرخهای فشردگی مختلف برای انواع مختلفی از متون مناسب

[1] StopWords

است، روشهای ارزیابی که امکان ارزیابی برای نرخهای مختلف می دهند را باید مورد توجه قـرار داد. معمـولا برای ارزیابی سیستمها از دو ارزیابی درونی و بیرونی استفاده میکنند.در ادامه توضـیحاتی از ایـن دو ارزیـابی بیان میشود.

۴-۳-۱- ارزیابی درونی:

ارزیابی درونی سیستم خلاصه را بدون توجه به هدف نهایی آن مورد سنجش قرار می دهد. در عوض ، توجه برروی فاز تولید در چرخه عمر یک خلاصه است. اکثر روشهای ارزیابی خلاصه درونی هستند و اغلب با یک استاندارد طلایی مقایسه می شوند. ارزیابی درونی توجه اصلی اش بروی پیوستگی و اطلاع رسانی خلاصه ها است و درنتیجه تنها کیفیت های خروجی را مورد سنجش قرار می دهد.

- **پیوستگی خلاصه :**

متن خلاصه ای که از طریق روشهای مبتنی بر استخراج تولید می شوند، گاهی از بعضی بی ارتباطی های معنایی در دنباله ی جملات متوالی رنج می‌برند. یک راه برای سنجش پیوستگی خلاصه ، رتبه بندی یا درجه بندی جملات برحسب میزان پیوستگی شان است و سپس درجه جملات خلاصه با امتیازات خلاصه های مرجع ، با امتیازات جملات منبع ، یا با امتیازات سایر سیستم های خلاصه ساز ، مقایسه شود .

- **دقت و بازخوانی جمله :**

بازخوانی تعداد جملات خلاصه مرجع که در خلاصه تولید شده حضور دارند را مشخص می کند. بـه همین ترتیب می توان دقت را به صورت تعداد جملات خلاصه تولید شده که در خلاصه مرجع وجود دارند، تعریف کرد. بازخوانی و دقت معیارهای استاندارد در بازیابی اطلاعات هستند و اغلب از ترکیب آنها، تحت عنوان F_measureیاد می شود. recallبرابر است با تقسـیم تعـداد جملاتـی کـه توسـط سیستم درست تشخیص داده شده است بر تعداد جملاتی که توسط سیستم معیار درست تشـخیص داده شده اند. Precision برابر است با تقسیم تعداد جملاتی که توسط سیستم درست تشخیص داده شده اند بر تعداد کل جملاتی که توسط سیستم برای خلاصه ایجاد شده اند.

مشکلات اصلی که این معیارها برای خلاصه سازی متن دارند آن است که قـادر بـه تشـخیص بـین خلاصه های ممکن ولی یکسان از نظر کیفیت نیستند و همچنین خلاصه هایی که محتـوای بسـیار متفاوتی دارند ممکن است امتیازات مشابهی دریافت کنند .

$$\text{Precision Rate} = \frac{\text{number of correctly selected sentences}}{\text{total number of selected sentences}}$$

$$\text{Recall Rate} = \frac{\text{number of correctly selected sentences}}{\text{total number of correct sentences}}$$

در روابط فوق دقت بیانگر درصد درستی جملات انتخاب شده نسبت به کل جملات انتخاب شـده و فراخـوانی درصد درستی جملات انتخاب شده نسبت به خلاصه مطلوب (خلاصه انسانی) نشان مـی دهـد .بـرای ارزیـابی خلاصه ساز پیاده سازی شده از این معیار استفاده شده است. با توجه بـه اینکـه منبـع معیـاری بـرای زبـان فارسی موجود نبود، از سیستم خلاصه ساز ایجاز که توسط آزمایشگاه فناوری وب دانشـگاه فردوسـی مشـهد ایجاد شده بود، استفاده کردیم. در این ارزیابی متن مورد خلاصه را هم به سامانه ایجاز و هـم بـه نـرم افـزار پیاده سازی شده دادیم و با تغییر در درصد فشرده سازی معیارهـای دقـت و فراخـوانی را محاسـبه کـردیم. جداول دقت و فراخوانی به شرح زیر است:

جدول (۴-۱) ارزیابی معیار دقت سامانه

نرخ فشرده سازی	Precision
20%	62%
30%	65%
40%	58%
50%	54%
60%	59%

جدول (۴-۲) ارزیابی معیار فراخوانی سامانه

نرخ فشرده سازی	Recall
20%	46%
30%	49%
40%	53%
50%	62%
60%	64%

۴-۳-۲- ارزیابی بیرونی :

برخلاف ارزیابی درونی ، در ارزیابی بیرونی توجه برروی کاربر نهایی معطوف می شود .درنتیجه در این روش میزان مؤثر بودن و قابلیت پذیرش خلاصه های تولید شده با بعضی روشها ، مثل ارزیابی رابطه ای یا قابلیت فهم در خواندن ، سنجیده می شود. همچنین اگر خلاصه به نوعی شامل مجموعه دستوراتی باشد یک روش ممکن برای ارزیابی آن ، بررسی قابلیت رسیدن به نتیجه با پیروی از دستورات خواهد بود. سایر روشهای ممکن برای سنجش ، جمع آوری اطلاعات در یک مجموعه بزرگ از اسناد است ، میزان تلاش و زمان موردنیاز برای پس‌ویرایش خلاصه تولید شده توسط ماشین برای بعضی مقاصد خاص ، یا تاثیر سیستم خلاصه ساز برروی سیستمی که جزئی از آن است ، برای مثال بازخورد مرتبط در یک موتور جستجو و یا یک سیستم پرسش‌پاسخ ، می باشد .

چندین سناریوی بازی به عنوان روشهای سطحی برای ارزیابی خلاصه ، پیشنهاد داده شده که ترتیب های مختلفی دارند. در میان آنها بازی Shannon ، بازی سوال ، بازی دسته بندی و کلمات کلیدی انجمنی می‌توان نام برد.

بازی Shannon که نوعی از معیارهای سنجش Shannonدر تئوری اطلاعات است ، تلاشی برای تعیین کیفیت محتوی اطلاعات بوسیله حدس لغت بعدی (حرف یا کلمه) می باشد ، و به این ترتیب متن اصلی را مجددا ایجاد می کند. این ایده از معیارهای Shannon از تئوری اطلاعات اقتباس شده است ، که در آنجا از سه گروه مخبر خواسته می شود قطعات مهم از متن اصلی را (با مشاهده متن کامل ، یک خلاصه تولید شده و یا حتی هیچ متنی) به صورت حرف به حرف یا کلمه به کلمه مجددا تولید کنند. سپس معیار حفظ اطلاعات با تعداد ضربه های کلیدی که برای ایجاد مجدد قطعه اصلی طول می کشد ، .سنجیده می شود .

هدف از بازی سوال ، آزمایش میزان فهم خواننده از خلاصه و توانایی آن برای نقل وقایع کلیدی متن اصلی است . این عمل ارزیابی در دو مرحله انجام می شود . ابتدا آزمایشگر مقاله های اصلی را می خواند و بخشهای مرکزی آن را علامت گذاری می کند. سپس از عبارات مهم بخشهای مرکزی متن ، سوالاتی طرح می کند. و در مرحله بعد ، ارزیاب سوالات را سه مرتبه پاسخ می دهد ؛ یکبار بدون مشاهده هیچ متنی ، پس از مشاهده یک خلاصه ساخته شده توسط سیستم و درانتها پس از مشاهده متن اصلی .خلاصه ای که به خوبی وقایع کلیدی مقاله را نقل کرده باشد ، باید قادر به پاسخگویی به بیشتر سوالات باشد .

۴-۳-۳- محیط ارزیابی خلاصه ها

محیط ارزیابیSEE ، محیطی است که در آن ارزیابها می توانند کیفیت یک خلاصه را در مقایسه با یک خلاصه مرجع مورد سنجش قرار دهند. متونی که درگیر ارزیابی هستند، با شکسته شدن به لیستی از قطعات (عبارات، جملات و...) مورد پیش پردازش قرار می گیرند. برای مثال هنگامی که یک سیستم مستخرج با سایز قطعه جمله را ارزیابی می کنیم، ابتدا متون با شکسته شدن به جملات آماده سازی می شوند .

در طول فاز ارزیابی، هر دو خلاصه در دو صفحه مجزا نشان داده می شوند و واسطه ای برای ارزیابی درنظرگرفته شده تا برروی محتوا و کیفیت خلاصه ها قضاوت کنند. برای سنجش محتوی، ارزیاب از میان خلاصه مورد ارزیابی، قطعه به قطعه حرکت می کند و برروی یک یا چند واحد مرتبط در خلاصه مدل کلیک می کند .

۴-۴- معرفی چند سیستم خلاصه ساز معروف

۴-۴-۱- AutoSummerized

سیستم AutoSummerizedکه در Microsoft word قرار داده شده است به صورت ذیل خلاصه ایجاد مینماید:

۱-ضریب وزنی به کلمه ها بر اساس تعداد تکرار کلمات داده میشود .در شمارش تعداد تکرار کلمات، ریشه یابی و کلمه های کمکی در نظر گرفته میشود ولی کلمه های توقف حذف نمیشوند.

۲- تجزیه و تقسیم متن به جملات

۳- انتصاب ضرایب وزنی به جملات .ضریب وزنی جمله به وسیله جمع ضرایب وزنی کلمات تشکیل دهنده آن بدست میآید.

۴-امتیاز دهی به جملات بر اساس وزن هر جمله

۵- استخراج Nجمله بر اساس اندازه خلاصه .اندازه خلاصه به طور پیش فرض ۲۵ %میباشد.

GistSumm –۲–۴–۴

GistSumm خلاصه ساز استخراجی میباشد که از سه بخش تکه سازی متن، امتیاز دهی بـه جملـه و ایجـاد خلاصه تشکیل شده است .امتیازدهی به جملات مطابق با روش کلمه کلیدی است، به این صورت که فرکانس تکرار کلمات هر جمله جمع میشود و بر اساس طول جمله نرمال میشود که به آن روش میانگین کلمه کلیدی گفته میشود .جمله ای که بالاترین امتیاز را داشته باشد جمله ای است که به بهترین حالت مـتن را توصـیف خواهد نمود .انتخاب دیگر جمله ها بر اساس میزان ارتباط با این جمله و یا از طریق میزان کل ارتباطی که هر جمله با کل محتوای متن دارد بدست میآید .میزان ارتباط با جمله اصلی از طریـق میـزان مشـاهده همزمـان کلمات در هر جمله و جمله اصلی مشخص خواهد شد که بر ایـن اسـاس میتـوان از پیوسـتگی مـتن خلاصـه مطمئن شد .میزان ارتباطی که هر جمله با کل محتوای متن دارد از طریق انتخاب جملاتی است که امتیاز آنها از یک حدی بالاتر باشد .

در روش دیگری موارد ذیل به عنوان خصوصیات پردازش متن در نظر گرفته شده است

۱–شباهت به عنوان :از شباهت cosine استفاده شده است

۲–شباهت به کلمه های کلیدی :جمله هایی که ارتبـاط و بسـتگی بیشـتری بـه کلمـه هـای کلیـدی دارنـد، مرتبطتر هستند و باید در خلاصه انتخاب شوند.

۳- ارتباط جمله به جمله: برای هر جمله s شباهت بین S و تمام جملات محاسبه و با هم جمع شده و نرمال میشود.

۴–مشاهده کلمه های خاص همانند اسم مکان، انسان و زمان میزان اهمیت جمله را بالا میبرد.

۵–مشاهده حروف اضافه :وجود اطلاعات غیر مهم را در متن نشان میدهد.

۶–مشاهده اطلاعات غیر لازم همانند" زیرا"،" علاوه بر آن که"در ابتدای جملات.

یک درخت دودویی از روی خصوصیات بالا تشکیل میشود که هر سطحی مربوط به یک خصوصیت است. بـه این صورت، هر جمله ای در مسیر درست خود در درخت قرار خواهد گرفت. پـس از انجـام یـک سـری پـیش پردازش همانند حذف کلمه های توقف، ریشه یابی و یکسان سازی، برای هر جملهای اسـتفاده و خصوصـیات بالا را مشخص کردهاند و سپس از طبقه بندی کنندهnaive bayse استفاده کرده اند و در مجموعه آمـوزش و تست آن را اعمال کرده اند و در مرحله بعدی به جای آنکه مقادیر این ۶ خصوصیت را برای هر جمله صـفر یـا یک در نظر بگیرد به صورت فازی در نظر گرفته شده است .برای ارزیابی ۱۰ متن از تافل انتخاب شـده و بـه ۵ نفر داده شده است تا متن را خلاصه نمایند و نتایج با روش برداری و فازی مقایسه شده و نتایج بهتری داشـته

است.

۴-۴-۳- سیستم خلاصه ساز FarsiSum

یک خلاصه ساز فارسی مطرح به نام FarsiSum وجود دارد که بر پایه خصوصیات آماری متن پایه گذاری شده است و خصوصیات زبان شناختی متن را به جز لیست کلمه های توقف برای پیش پردازش در نظر نمیگیرد.

FarsiSumنسخه تغییر یافته ی سیستم خلاصه‌سازی متون سوئدی به نامSweSum برای پوشش زبان فارسی میباشد. خلاصه ی خروجی SweSum از نوع مستخرج است و برای زبان های سوئدی، نروژی، دانمارکی، اسپانیایی، انگلیسی، فرانسوی و آلمانی پیاده سازی یا نمونه سازی شده است .این خلاصه ساز متنی در فرمت textیا HTML را به عنوان ورودی از روزنامه یا گزارش دریافت کرده و خلاصه ی آن را تولید میکند به این منظور ابتدا محدوده ی جملات و کلمات را تعیین کرده و کلمات مخفف شده برچسب می خورند . سپس فرکانس ریشه ی کلمات دارای مقوله های نحوی کلاس باز (مانند اسامی، افعال و صفات) با استفاده از یک واژگان ایستا محاسبه می شود. سپس از آن نوبت به محاسبه‌ی امتیاز جملات متن است.

سیستم برای این امتیازدهی موارد زیر را لحاظ میکند:

-اولین خط در متن باید در خلاصه وارد شود، زیرا امتیاز بسیار بالایی دارد(مقدار پیشفرض آن ۱۰۰۰ است)

امتیاز موقعیت به نوع متن بستگی دارد SweSum دو نوع متن روزنامه و گزارش را پشتیبانی میکند .

موقعیت خط در گزارش روزنامه اهمیت کمتری دارد .در روزنامه مهم ترین بخش متن، اولین خط به اضافه‌ی چند خط دیگر به ترتیب نزولی است.

-مقادیر عددی مثل تاریخ ها در یک سند مهم است .یک مقدار ثابت به امتیاز خط شامل اعداد اضافه می‌شود(مقدار پیش فرض آن 1 است)

متن پررنگ در سند HTML مهم است و امتیاز بالاتری دارد.

در متن برخی از روزنامه‌های سوئدی آن آغاز یک پاراگراف را نشان میدهد و اولین جمله هر پاراگراف معمولا پًاراگراف را معرفی می کند(مقدار پیشفرض آن 100 است)

-جملات شامل کلمات کلیدی (کلمات با بیشترین تکرار در متن) امتیاز بیشتری از جملاتی که کلمات کلیدی کمتری را شامل میشوند، دارند .این کلمات توسط کاربر از طریق سوال به سیستم داده میشوند.

-تمام پارامترهای بالا در تابع ترکیبی با وزنهای قابل اصلاح (مقدار پیشفرض) برای به دست آوردن امتیاز کل هر جمله گذاشته میشود .در این تابع امتیاز هر کلمه از حاصلضرب تعداد تکرار آن در وزن کلمه‌ی کلیدی حاصل میشود و امتیاز جمله از حاصل نرمالسازی مجموع امتیاز کلمات آن بدست می‌آید.

-جملات طبق امتیاز محاسبه شده مرتب شده و خطوطی که بیشترین امتیاز را دارند تا سقف اندازه برش تعیین شده توسط کاربر (درصد کلمه یا حرف در خروجی) در فایل خروجی نوشته می شوند .این فایل شامل تمام خطوط غیر متنیHTML ، تمام خطوط متن امتیاز داده شده و اطلاعات آماری در مورد خلاصه، تعداد کلمات، تعداد خطوط، کلمات با بیشترین فرکانس و ... می‌باشد.

۴-۴-۴- مشکلات زبان فارسی

مسائل و خصوصیات نگارشی زبان فارسی، امر پیش پردازش اتوماتیک خلاصه‌سازی متون را دچار چالش می‌نماید که از آن جمله میتوان به موارد ذیل اشاره نمود:

-ساختارهای تولید واژگان در بسیاری از موارد از ترکیب چند واژهی مستقل و با معنی، در کنار یکدیگر به وجود آمده اند در حالی که اگر میان این اجزاء فاصله درج شود، تبدیل به دو لغت و اگر نیمفاصله درج شود تبدیل به یک لغت خواهند شد .این مسئله در انگلیسی بسیار بسیار کمتر رخ میدهد.

-انواع مختلف نگارش برای یک کلمه همانند اتاق و اطاق و یا به کاربردن همزه به صورت های مختلف همانند مسئول و مسوول

-استفاده از کلمات اروپایی به صورتهای مختلف همانند سورس وsource

-استفاده از" ا "و" آ "به جای یکدیگر همانند" فرایند "و "فرآیند"

-بازگذاشتن دست نویسندگان در فاصله گذاری میان کلمات.

-عدم وجود دستورالعمل قطعی برای استفاده از نیم فاصله.

-عدم وجود قواعدی ثابت برای فاصله‌گذاری ترکیبات؛ استفاده از دستورالعمل مبتنی بر لغت مانند تک هجایی بودن، بسیط گونه بودن

- تنوع استفاده از" می "چسبان و غیر چسبان همانند کلمات" میتواند "و" میتواند."

-تنوع نحوه به کاربردن" ها "چسبان و غیر چسبان همانند" آنها"،" آنها "و" آن ها."

-تنوع نگارش" ی "اضافه در کلمات مختوم به" ه "همانند" خانه سبز "و" خانهی سبز."

عدم در نظر گرفتن مسائل نگارشی بیانشده باعث میشود تا تفکیک لغات نا مفید و غیر اصلی از لغات مفید به آسانی انجام پذیر نباشد، ریشه لغات به درستی تشخیص داده نشود و یا لغات همسان، غیرهمسان تشخیص داده شود که در مرحله امتیازدهی به لغات تاثیر خواهد گذاشت.همچنین عدم وجود پیکره ای که شامل متن به همراه خلاصه آنها باشد کار ارزیابی خلاصه سازها را با مشکل روبرو کرده است.در زبان انگلیسی مجموعه خلاصه های تولید شده توسط انسان به نام DUC وجود دارد که میتواند به عنوان معیار ارزیابی استفاده

شود. این مجموعه شامل ۱۰۰۰ سند به همراه خلاصه های آنها میباشد که در میان آنها همه نوع متن اعم از علمی، اجتماعی، سیاسی، فرهنگی و ... دیده میشود.

از دیگر مشکلات زبان فارسی کاربرد یک لغت در متون مختلف است که معنای مختلفی را نسبت به زمینه متن تولید میکند. برخلاف زبان انگلیسی که در آن هم حروف و هم لغات کاملا متمایز از یکدیگرند، در زبان فارسی پیوستگی میان برخی علائم با لغات وجود دارد و علاوه بر آن تنوع نگارش در کلمات نیز موجود میباشد. ریشه یابی نیز در زبان فارسی چالشهای خاص خود را دارد .به عنوان مثال در یک لغت به هم پیوسته هم بن فعل، شناسه، علامت زمان فعل و حتی شناسه‌های مفعولی میتوان داشت که کار پردازش لغات را پیچیده تر مینماید به طوری که نمیتوان از دانش، تجربه و نرم‌افزارهای موجود در این زمینه استفاده نمود و تولید نرم‌افزاری که قادر به حل تمامی این پیچیدگیها باشد، فرایندی زمانبر و مستلزم تلاش فراوان میباشد . تفاوتهای ذاتی زبانهای گسستهای مانند انگلیسی با زبانهایی مانند فارسی که با یکدیگر تفاوتهای بنیادین در قواعد دستوری دارند، منجر به آن شده است که ادعای اعمال تغییرات در ساختار یک نرم‌افزار انگلیسی و به دست آوردن نتایج خوب برای زبان فارسی لزوما امکان پذیر نمیباشد و مستلزم آزمایشهای فراوان برای اثبات صحت آن خواهد بود. همچنین از لحاظ گرامری در زبان فارسی نمیتوان قواعدی را مانند زبان انگلیسی تعریف کرد.برای نمونه به پاره ای از مشکلات آن اشاره میکنیم :

۱. وجود لغات ترکیبی چند جزئی همانند" آبسردکن"
۲. در زبان فارسی، ضمائر مفعولی و نشانه‌های جمع به لغات متصل میشوند .این مسئله نیز منجر به پیچیدگی تفکیک لغات برای رایانه میشود زیرا حروف ضمائر مفعولی و نشانه‌های جمع با تعدادی از وندها و نیز واژگان فارسی مشابهت دارند در نتیجه رایانه نمیتواند به سادگی در مورد بخش اصلی لغت قضاوت کند. به عنوان نمونه،" ان "در" درختان "نشانه‌ی جمع است در حالی که" ان "در" خوران " نشانه‌ی صفت فاعلی میباشد.
۳. در زبان فارسی قواعد تبدیل بن مضارع به بن ماضی و بالعکس به صورت قانونمند وجود ندارد،در نتیجه رایانه برای آنکه بتواند لغات حاصل از بن مضارع یا بن ماضی را تشخیص دهد(مثلا برای آنکه بداند" ان "در لغت" خوران "نشانه‌ی جمع است یا نشانه‌ی صفت فاعلی)، هیچ راه مشخصی وجود ندارد.
۴. ابهام ساختاری به نحوی که یک کلمه می تواند معانی مختلف داشته باشد .به عنوان مثال، شیر سه معنی متفاوت دارد .شیر حیوان، شیر آب، شیر خوراکی.
۵. عدم وجود قاعده خاص برای تشخیص اسامی و مکانهای خاص همانند آنچه در زبان انگلیسی موجود میباشد.

۶. ابهام در معنی کلمه به علت نبود اطلاعات آوایی همانند"مَرد "و" مُرد"
۷. عدم وجود دستورالعمل قطعی برای استفاده از نیم‌فاصله.
۸. عدم وجود قواعدی ثابت برای فاصله‌گذاری ترکیبات
۹. ناآگاهی رایانه از نقش لغت در جمله همانند" آیین نامه نوشتن "و" آیین‌نامه رانندگی ."از آنجا که در این ترکیبات، تمام کلمات به تنهایی معنی دارند، رایانه قادر به تعیین مرز لغات نمیباشد.
۱۰. وجود کلمه های ترکیبی و امکان در نظر گرفته شدن دو کلمه مجزا .به عنوان مثال سیب زمینی که کلمه مرکبی است که از ۲ کلمه سیب و زمینی تشکیل شده‌است.
۱۱. وجود اشتباهات نگارشی در مستندات فارسی همانند نگارش با کاراکترهای متفاوت همانند آئین‌نامه و آیین‌نامه، به کار بردن همزه به صورتهای مختلف، و نیاز به یکسان سازی این موارد برای پیش‌پردازش متن به صورت کارا.
۱۲. ابهام در دستور خط زبان فارسی : پیوسته نویسی کلمات مرکب ترکیب شده با پسوند
۱۳. فقدان هر نوع دادگان بزرگ مانند یک گرامر کامل، یک مجموعه از جملات تجزیه شده نمونه ویا آمارهای ارزشمند از کاربرد لغات در جملات مختلف و جایگاههای مختلف در جمله.
۱۴. همانند سایر زبانهای هند و اروپایی وجود پیشوند و پسوند و استفاده غلط از واژه‌هایی که طبق عرف تغییر معنا داده‌اند.

از جمله مسائل دیگر در زبان فارسی نبود حروف کوچک و بزرگ همانند زبان انگلیسی برای تشخیص اسامی خاص (همانند نام روز، هفته، ماه و مناطق حغرافیایی) میباشد. تشخیص اختصارات موجود در زبان فارسی از دیگر چالشهای موجود میباشد.

۴-۵- نتیجه گیری

مبحث خلاصه‌سازی یکی از مباحث مهم حوزه متن کاوی و پردازش زبانهای طبیعی میباشد.خلاصه‌سازی به عنوان یکی از ابزارهای عمدهٔ تجزیه و تحلیل و سازماندهی مدارک مطرح است .در سال های اخیر تمرکز و تأکید بسیاری بر ایجاد سیستم ها و رویکردهای خلاصه‌سازی خودکار صورت گرفته است. در هر کدام از روش ها و سیستم ها تلاش بر این بوده است که خلاصه‌های حاصل شده به خلاصه‌های دستی و انسانی نزدیک شود. دو مبنای اصلی خلاصه‌سازی در این حوزه موجود میباشد که عبارتند از خلاصه‌سازی چکیده و خلاصه‌سازی استخراجی. در خلاصه‌سازی چکیده خروجی سیستم باید مانند خلاصه انسانی باشد و در آن جملات تغییر شکل میابند اما در خلاصه‌سازی استخراجی قالب اصلی جملات حفظ میشود و جملات دستخوش تغییرات نمیشوند. از آنجایی که خلاصه‌سازی چکیده نیاز به فهم متن در سطح بالا دارد، امروزه

تمرکز خلاصه‌سازی در سراسر دنیا روی خلاصه‌سازی استخراجی میباشد .به هر میزان که از اطلاعات زبانشناختی زبان فارسی در امر خلاصه‌سازی استفاده شود به همان میزان خلاصه دقیقتری به وجود خواهد آمد ولی این مساله در قبال پرداخت هزینه و تلاش بیشتری خواهد بود. در این پایان نامه ابتدا به تعریف و تاریخچه خلاصه‌سازی پرداختیم و پس از آن به بررسی رویکردها و روشهایی در حوزه خلاصه‌سازی متون پرداختیم. روشها و رویکردهای مختلفی برای خلاصه‌سازی متون موجود میباشد که شامل دو قسمت آماری و پردازش زبانهای طبیعی میباشد. در ادامه توضیحاتی را در مورد روش پیاده سازی شده که با استفاده از روش آماری و با معیار وزن دهی TF-IDF انجام شده است، بیان کردیم. برای ارزیابی خلاصه سازهای زبان فارسی مجموعه استانداردی از متون و خلاصه‌های ایده‌آل آنها همانند آنچه در زبان انگلیسی موجود میباشد، موجود نیست که همین امر کار ارزیابی را مشکل مینماید. در پایان نیز به معرفی چند نمونه از سیستمهای خلاصه‌سازی متن پرداختیم.

منابع

[1] Eduard Hovy. "*Text Summarization*".

[2] Vishal Gupta, Gurpreet Singh Lehal. "*A Survey Of Text Extractive Techniques*". Department Of Computer Science,Punjabi University Patiala

[۳] "فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی"، ویرایش ۱۳۸۸/۰۳/۱،۲۴/۰

[۴] زهره کریمی، مهرنوش شمس فرد،"سیستم خلاصه ساز خودکار متون فارسی"، انجمن کامپیوتر ایران،۱۳۸۵

[5] Nasim Argha.Reza Mirshahi, "*Summarization PersianText As An Important Application*",Department Of Computer Science And Research Branch Of IAU.

[۶] فاطمه پورغلامعلی، محسن کاهانی، آصف پورمعصومی."خلاصه‌سازی چکیده ای مبتنی بر مشابهت جملات". آزمایشگاه فناوری وب دانشگاه فردوسی مشهد.۱۳۹۱/۶/۱۵

[7] Shuhua Liu. "*Theory and Methods on Text Summarization*".Institute for advanced management systems research.2004.

[8] Sankar K .Sobha L . "An Approach To Text Summarization*".Mit Campus,Anna University,Chennai-44.*

[9] Hongyan Jing. "Sentence Reduction For Automatic Text Summarization",Department Of Computer Science Columbia University,New York,NY 10027,USA.

[10] Juan Ramos. "Using TF-IDF to Determine Word Relevance in Document Queries",Department Of Computer Science, Rutgers University,23515 BPO Way.

[11] Rafeeq Al-Hashemi. "Text Summarization Extraction System(TSES) Using Extracted Keywords",Faculty of Information Technology,Al-Hussain Bin Talal University,Jordan.

پیوستها

پاره ای از کلمات ایست فارسی[1]:

به خوبی	به درشتی	به دلخواه	به راستی	به رغم	به روشنی	به زودی	به سادگی	به سرعت	به شان	به شدت	به طور کلی
طوری که	به علاوه	به قدری	به کرات	به گرمی	به مراتب	به ناچار	به هرحال	به هیچ وجه	به وضوح	به ویژه	بهت
	بهش	بود	بویژه	بی	بی آنکه	بی اطلاعند	بی تردید	بی تفاوتند	بی نیازمندانه	بی هدف	بیرون
	بیشتر	بیگمان	بین	پ	پارسال	پارسایانه	پاره ای	پایین ترند	پدرانه	پدیده هاست	پرسان
	پروردگارا	پریروز	پس	پس از	پس فردا	پشت	پشتوانه اند	پشیمونی	پهن شده	پی	پی در پی
	پیداست	پیرامون	پیش	پیوسته	ت	تا	تازه	تاکنون	تحت	تحریم هاست	تر
	تصریحاً	تعدادی	تعمدا	تقریبا	تقریباً	تک تک	تلویحاً	تمام	تمام قد	تماما	تمامشان
	تند تند	تنها	تو	توؤماً	توسط	توی	ث	ثالثاً	ثانیاً	ج	جداً
	جداگانه	جدیدا	جرمزاست	جز	جلو	جلوی	جمع اند	جمعی	جنابعالی	جنس اند	جهت
	چ	چاپلوسانه	چت	چته	چرا	چرا که	چشم بسته	چطور	چقدر	چکار	چگونه
	چنانچه	چند	چند روزه	چندان	چنده	چندین	چنین	چه	چه بسا	چه طور	چو
	چی	چیزهاست	چیزیست	چیست	چیه	ح	حاشیه ای	حاضرم	حاکیست	حال	حتما
	حتی	حداقل	حداکثر	حدود	حسابگرانه	حضرتعالی	حقیرانه	حکماً	حول	خ	خالصانه
	خداحافظ	خداست	خسته ای	خصوصاً	خواسته	خواهد	خوب	خود	خود به خود	خودبه خودی	خودت
	خودتان	خودتو	خودش	خودشان	خودم	خودمان	خودمو	خوش	خوشبختانه	خویش	خویشتنم
	خیره	خیلی	د	دا	دااام	داما	داخل	داراست	دارد	دامم	در
	در بارهٔ	در ثانی	در کل	در کنار	در مجموع	در نهایت	در واقع	دراین میان	درباره	درحالی که	درحالیکه
	درست و حسابی	درسته	درصورتی که		درعین حال	درواقع	دریغ	دریغا	دسته دسته	دشمنیم	دقیقا
	دم	دهد	دو روزه	دوباره	دیر	دیرت	دیرم	دیروز	دیشب	دیگر	دیگران
	دیگه	دیوانه ای	دیوی	ذ	ذاتاً	ر	را	راجع به	راحت	راست	راستی
	رفتارهاست	رنجند	رهگشاست	رو	رواست	روبروست	روز به روز	روزانه	روزه ای	روزه ایم	روزه ست
	روش	روی	رویش	ز	زشتکارانند	زنند	زهی	زودتر	زیاد	زیاده	زیر
	زیرچشمی	ژ	س	ساده اند	ساکنند	سالانه	سالته	سالم تر	سالهاست	سپس	سخت
	سر	سراپا	سراسر	سرانجام	سری	سریع	سریعاً	سه باره	سهواً	سیاه چاله هاست	سیخ
	ش	شاهدند	شاهدیم	شاید	شبهاست	شخصا	شخصاً	شد	شدن	شده	شدیدا
	شما	شماری	شماست	شمایند	شود	شوراست	شوقم	شیرین	شیرینه	شیک	ص
	صددرصد	صرفا	صرفاً	صریحاً	صندوق هاست		ض	ضمناً	ط	طبعاً	طبیعتا
	طلبکارانه	طی	ظ	ظاهرا	ظاهراً	ع	عاجزانه	عاقبت	عبارتند	عجب	عجولانه
	عقب	علاوه بر	علاوه بر آن	علاوه برآن	علناً	علی الظاهر	علی رغم	علیه	عمدا	عمداً	عمدتا
	عمده	عملا	عملاً	عملی اند	عموم	عموما	عموماً	عنقریب	عیناً	غ	غالبا
	غیرقانونی	ف	فاقد	فبها	فر	فردا	فعلاً	فقط	فلان	فلذا	ق
	قاطبه	قاطعانه	قاعدتاً	قانوناً	قبلا	قبلاً	قبلند	قدر	قدری	قضایاست	قطعا
	ک	کارند	کاش	کاشکی	کاملا	کاملاً	کجا	کجاست	کدام	کرده	کلا
	کلیشه هاست		کلیه	کم کم	کما□اینکه	کمتر	کمتره	کمی	کنار	کنارش	کنایه ای
	کنم	کنند	که	کی	گ	گاه	گاهی	گرچه	گرفتارند	گونه	گویی
	لااقل	لاجرم	لب	لذا	لزوماً	لطفا	لیکن	م	ما	مادامی	ماست
مامان گویان	مانند	متؤسفانه	متاسفانه	متفاوتند	مثل	مثلا	مجبورند	مجدداً	مجموعاً	محتاجند	محکم
	محکم تر	مخالفند	مخصوصاً	مدام	مدتهاست	مذهبی اند	مرا	مرتب	مردانه	مردم اند	مستحضرید
	مستقیما	مستند	مشت	مشترکاً	مشغولند	مطمانا	مطمانم	مع الاسف	مع ذلک	معتقدم	معتقدند
	معتقدیم	معدود	معذوریم	معلومه	معمولا	معمولاً	معمولی	مغرضانه	مفیدند	مقدار	مقصرند
	مقصری	مکرر	مکرراً	مگر	مگر آن که	مگر این که	ممیزیهاست	من	منتهی	منطقی	منی

[1] Persian Stop Words

	مواجهند	موجودند	مورد	می	میان	میزان	ن	ناامید	ناخواسته	ناراضی اند	ناگزیر	ناگهان
	نبش	نخست	نخودی	ندارد	نزد	نزدیک	نظیر	نفرند	نمی	نه	نه تنها	نهایتـــا
	نهایتاً	نوع	نوعاً	نیازمندند	نیز	نیمی	ه	ها	های	هایی	هر	هـــر از
گاهی	هر چند	هر چند که هر چه		هرچند	هرچه	هرکس	هرگاه	هرگز	هستند	هق هق کنان		هـــــم
	هم اکنون	هم اینک	همان	همان طور که		همان گونه که		همانند	همانها	همچنان	همچنـــــان کـــــه	
	همچون	همچین	همدیگر	همزمان	همگان	همگی	همه	همهٔ	همه اش	همه روزه	همه ساله	همـــــه
شان	همواره	همیشه	همین	همین که	هنگامی که هنوز		هوی	هی	هیچ	هیچ گاه	هیچکـــــــــــدام	
	هیچکس											

(عکسهایی از نرم‌افزار پیاده‌سازی شده)


نرم افزار خلاصه ساز متون
نرم افزار خلاصه ساز متون
پردازش
درصد خلاصه سازی
کلمات کلیدی :
دولت آمریکا تهران
متن ورودی
خاتمی تصریح کرد: همیشه گفته‌ام، معنای حمایت از دولت‌ها این نیست که لب قروبندیم، زیرا جامعه زنده آن است که انتقاد داشته باشد.
وی خاطر نشان کرد: خوش گمانی نسبت به دولت سبب می‌شود، مسئولان یا پشت گرمی در عرصه‌های مختلف کار کنند و بنابراین در عرصه داخلی و خودی ما باید اساس را بر خوش گمانی بگذاریم تا شیرینی زندگی لمس شود.
نماینده مردم استان کرمان در مجلس خبرگان رهبری اذعان داشت: در قضایی که جور و ستم غلبه دارد و فرهنگ آنها زورگویی است در این قضا خوش گمانی محکوم است و باید اساس را بر بدبینی گذاشت.
به غربی‌ها بدبین هستیم
وی افزود: به همان اندازه که ما به دولت خودمان اعتماد داریم به همان اندازه به دولت‌های غربی و قدرت‌های شش‌گانه بدبین هستیم.
متن خلاصه شده


نرم افزار خلاصه ساز متون
نرم افزار خلاصه ساز متون
پردازش
درصد خلاصه سازی
کلمات کلیدی :
دولت آمریکا تهران
متن ورودی
خاتمی تصریح کرد: همیشه گفته‌ام، معنای حمایت از دولت‌ها این نیست که لب قروبندیم، زیرا جامعه زنده آن است که انتقاد داشته باشد.
وی خاطر نشان کرد: خوش گمانی نسبت به دولت سبب می‌شود، مسئولان یا پشت گرمی در عرصه‌های مختلف کار کنند و بنابراین در عرصه داخلی و خودی ما باید اساس را بر خوش گمانی بگذاریم تا شیرینی زندگی لمس شود.
نماینده مردم استان کرمان در مجلس خبرگان رهبری اذعان داشت: در قضایی که جور و ستم غلبه دارد و فرهنگ آنها زورگویی است در این قضا خوش گمانی محکوم است و باید اساس را بر بدبینی گذاشت.
به غربی‌ها بدبین هستیم
وی افزود: به همان اندازه که ما به دولت خودمان اعتماد داریم به همان اندازه به دولت‌های غربی و قدرت‌های شش‌گانه بدبین هستیم.
متن خلاصه شده
به گزارش خبرگزاری فارس از کرمان، آیت‌الله سیداحمد خاتمی پیش از ظهر امروز در همایش ملی سبک زندگی در تالار وحدت دانشگاه شهید باهنر کرمان اظهار داشت: سه عنصر در سبک زندگی اسلامی حرف نخست را می‌زند که اگر نباشد، هیچ‌گاه زندگی دینی شکل نمی‌گیرد
مجری صالح کیست
وی با طرح این سئوال که این مجری صالح کیست؟ خاطر نشان کرد: پیامبر و امام مجری صالح هستند و بر اساس متون موجود اجرای قانون به ولایت فقیه می‌رسد
عضو خبرگان رهبری با بیان اینکه با پیروزی انقلاب اسلامی دو عنصر نخست در در سبک زندگی اسلامی شکل گرفت، اذعان داشت: درود بر شهیدان که تحقق‌بخش این دو عنصر بودند
وی متذکر شد: عنصر سوم که در سبک زندگی اسلامی کلیدی است، عبارت از رفتار اسلامی آحاد جامعه با یکدیگر است که این عنصر زمین مانده است
امام جمعه موقت تهران اذعان داشت: نمی‌گوییم، رفتارها اسلامی نیست، رفتارها اسلامی هست، اما نمی‌توانیم بگوییم آن توفیقی که در عنصر نخست و دوم داشتیم در این بخش هم داریم
حمایت مردم از حکومت یک سرمایه است
امام جمعه موقت تهران با بیان اینکه خانواده از دیدگاه قرآن دو عنصر اساسی دارد، یادآور شد: در عرصه اجتماعی و تعامل با یکدیگر باید اساس را بر خوش‌گمانی گذاشت
وی اذعان داشت: آقایی می‌آید، هشت سال حکومت می‌کند و می‌رود، اما این آرزو به دل آدم می‌ماند که یک بار بگوید، بیخشید، آنجا اشتباه کردم
خاتمی گفت: آرزو به دل شدیم یک مسئول دیوی هم در کنار مقام معظم رهبری بیاید و استغفرالله بگوید