Khazarmi University

Mathematical Researches

Print ISSN: 2588-2546
Online ISSN: 2588-2554

Homepage: https://jemr.khu.ac.ir/

Automatic Text Summarization Based on The Power of Reconstructing Sentences from Each Other in a Sparse Reconstruction Framework

M.Rezghi[1], F. **Mohammadian**[2] ✉, F. **Behzad**[3]

1. Corresponding Author, Department of Computer Science, Faculty of Mathematical Sciences, Tarbiat Modares University,Tehran, Iran. ✉E-mail: rezghi@modares.ac.ir
2. Department of Computer Science, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran. E-mail: mfaieg@yahoo.com
3. Department of Computer Science, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran. E-mail: farzaneh.behzad@yahoo.com

Article Info

Article type:
Research Article

Article history:
Received: 21 July 2020
Accepted: 22 April 2022
Published online:
6 February 2024

Keywords:
Almost contact structure,
B-metrics,
Natural metric,
Sphere bundle,
Structure tensor.

ABSTRACT

Introduction

In recent years, the rapid growth of textual data has made extracting valuable summaries with a minimum volume from this massive volume of data inevitable. Due to a large amount of data, text summarization by humans is very time-consuming and practically impossible. Therefore, artificial intelligence in text summarization has been one of the essential branches of text mining and natural language processing. Among the existing methods, the methods that select meaningful sentences based on their role in sparsely reconstructing other sentences have obtained better results. These methods have two main terms, one term related to reconstructing each sentence by others modeled by the l2 norm. The other one is a regularization term that controls the sparseness of the reconstruction coefficients modeled by the group sparse norm. This sparseness allows a limited number of sentences to participate in the construction of each sentence. The reconstruction function based on the L2 norm causes all keywords to play an equal role in sentence reconstruction, which may cause the outlier words to change the result of the summary. Therefore, to improve the summary's quality obtained in this article, we rewrite the penalty function with L1 norm. This substituting allows us to have a sparse reconstruction error. Due to the sparseness property of L1, the reconstruction error corresponding to most words is good and is close to zero in this method. Still, for some words (outlier), it allows this error to be significant, which reduces the method's sensitivity to the outlier words. The implementation results show that the proposed method provides faster and higher quality summaries based on ROUGE and F-measure criteria than the previous methods.

Material and Methods

In this paper, we introduced a new loss function for text summarization with sparse viewpoint.

Results and discussion

The method in [5], used the following model for text summarization

$$\min \sum_{i=1}^{n} \|s_i - Sw_i\|_2^2 + \lambda \|W\|_{2,1} \;\; s.t \;\; diag(W) = 0 \, , W \geq 0, \;\; S \epsilon R^{n \times m}$$

Here the first term denotes the summation of reconstruction errors of sentences by others. Also, the second term controls the sparseness of the coefficients in

the reconstruction. After solving this minimization problem, they sort the sentences based on the norm of the rows of the matrix W.

The reconstruction function based on the L2 norm causes all keywords to play an equal role in sentence reconstruction, which may cause the outlier words to change the result of the summary. Therefore, to improve the summary's quality obtained in this article, we rewrite the penalty function with the L1 norm as follows:

$$min_{wi}\|s_i - Sw_i\|_1 + \lambda\|w_i\|_1 \quad , i = 1, \ldots, n$$

Due to the sparseness property of L1, the reconstruction error corresponding to most words is close to zero. At the same time, for some words (outlier), it allows this error to be significant, which reduces the method's sensitivity to the outlier words. To evaluate the performance of the proposed method, all the texts of the DUC 2002 dataset, which is 115 documents, were compared by both test methods and the results obtained by the F-measure criterion. The obtained results show the quality of the proposed method in obtaining appropriate summaries compared to a technique based on the group sparse method.

Conclusion

The Main results of the papers could be summarized as follows:

- Using the L1 norm in the reconstruction term can filter anomaly data in the text summarization.
- Experimental results confirm that the proposed method generally gives results better than the Group sparse norm method.
- Examining the hand summaries, we concluded that the participant sentences in the hand summaries are not the only ones with the most keywords. Sometimes sentences with fewer keywords are also meaningful and participate in the handwriting summary. The property of the Group sparse method in [5] is that it tries to have all the keywords within the selected sentences. If the sentence has a specific keyword and the number of keywords is small, it is a strong candidate for selection in the selected sentences. But the proposed method doesn't have such drawbacks.

How to cite: Rezghi. M, Mohammadian. F, & Behzad. F. (2023). Automatic Text Summarization Based on The Power of Reconstructing Sentences from Each Other in a Sparse Reconstruction Framework. *Mathematical Researches*, 9 (3), 111 – 135.

Publisher: Kharazmi University

پژوهش‌های ریاضی
شاپا چاپی: ۲۵۸۸-۲۵۴۶
شاپا الکترونیکی: ۲۵۸۸-۲۵۵۴
Homepage: https://mmr.khu.ac.ir/

خلاصه‌سازی خودکار متن مبتنی بر قدرت بازسازی جملات از روی همدیگر در یک بازسازی تنک

منصور رزقی[1]✉، فائقه محمدیان [2]، فرزانه بهزادی ابراهیمی [3]

۱. نویسنده مسئول، گروه علوم کامپیوتر، دانشکده علوم ریاضی، دانشگاه تربیت مدرس، تهران، ایران. رایانامه: rezghi@modares.ac.ir
۲. نویسنده مسئول، گروه علوم کامپیوتر، دانشکده علوم ریاضی، دانشگاه تربیت مدرس، تهران، ایران. رایانامه: mfaieg@yahoo.com
۳. گروه ریاضی، گروه علوم کامپیوتر، دانشکده علوم ریاضی، دانشگاه تربیت مدرس، تهران، ایران. رایانامه: farzaneh.behzad@yahoo.com

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۳۹۹/۴/۳۱
تاریخ پذیرش: ۱۴۰۱/۲/۲
تاریخ انتشار: ۱۴۰۲/۱۱/۱۷

واژه‌های کلیدی:
داده‌کاوی،
متن‌کاوی،
خلاصه‌سازی خودکار متون،
متون چند گانه،
نمایش تنک گروهی

چکیده

رشد سریع و پیوسته شبکه جهانی وب باعث شده است فرآیند استخراج اطلاعات مفید با حجم کمینه، از میان مجموعه‌ی اسناد بزرگ چالش جدی این روزها باشد. خلاصه‌سازی اسناد برای انسان امری بسیار زمان‌بر و دشوار است، ولذا نیاز به یک سیستم خلاصه‌سازی قدرتمند را برای کاهش حجم متون و همچنین سرعت بالاتر دسترسی به اطلاعات مفید را آشکار می‌کند. اخیرا سیستم خلاصه‌سازی مبتنی بر رویکرد نمایش تنک ارائه شده است که سعی بر آن دارد تا هر جمله را با ترکیب خطی از جملات دیگر به صورت تنک بازسازی کند. در این رویکرد زیر مجموعه‌ای از جملات متن اصلی که حاوی اطلاعات مهم متن می‌باشد را انتخاب کرده و به عنوان خلاصه به خروجی می‌فرستد. همچنین نیاز است کم‌ترین تعداد از جملات متن که حداکثر بازسازی سایر جملات متن را داشته باشد انتخاب شود، که استفاده از رویکرد نمایش تنک این هدف محقق می‌کند. این مدل از یک تابع جریمه مبتنی بر نرم $L2$ برای کنترل بازسازی جملات و یک عامل منظم ساز تنک مبتنی بر نرم یک تشکیل شده است. تابع بازسازی بر اساس نرم $L2$ سبب می‌شود که تمام کلمات کلیدی نقش مساوی در بازسازی جملات داشته باشند که این امر ممکن است باعث شود کلمات پرت نتیجه خلاصه‌سازی را عوض کنند. بنابراین برای بهبود کیفیت خلاصه به دست آمده در این مقاله تابع جریمه را با نرم $L1$ بازنویسی می‌کنیم. این امر باعث می‌شود تا میزان خطای متفاوتی برای هر کدام از کلمات در بازسازی جملات اختصاص یابد که موجب کمتر شدن حساسیت روش به کلمات پرت می‌شود. نتایج پیاده سازی نشان می‌دهند که روش پیشنهادی نسبت به روش‌های قبلی خلاصه‌ای سریع و با کیفیت بالا بر مبنای معیارهای ROUGE و F-measure ارائه می‌دهد

استناد: نام خانوادگی، منصور رزقی ، فائقه محمدیان ، فرزانه بهزادی ابراهیمی (۱۴۰۲). خلاصه‌سازی خودکار متن مبتنی بر قدرت بازسازی جملات ازروی همدیگر در یک بازسازی تنک. *پژوهش‌های ریاضی*، ۹ (۳)، ۱۱۱ – ۱۳۵.

ناشر: دانشگاه خوارزمی

مقدمه

با پیشرفت علم و تکنولوژی و افزایش چشم‌گیر اطلاعات، با مشکل انبوه داده‌ها مواجه هستیم [1] . این فراوانی اطلاعات باعث می‌شود تحقیقات و جست‌وجو در مورد موضوعی خاص با مشکل مواجه شود که به آن سرریز اطلاعات گفته می‌شود. این امر دسترسی ما به داده‌ها را پیچیده می‌کند؛ بنابراین لازم است روشی پیدا کنیم که دسترسی به اطلاعات مورد نظر را ساده کند. بهترین روش خلاصه کردن و سپس طبقه‌بندی اطلاعات است. به طور کلی داده‌ها به دو دسته‌ی ساختاریافته و غیرساختاریافته تقسیم می‌شوند. داده‌های ساختار یافته به کدهایی گفته می‌شود که دارای فرمت مشخصی هستند و به گونه‌ای نوشته می‌شوند که قابل درک برای موتورهای جستجو باشند. داده‌هایی مانند ویدیو یا تصویر یا متن که بایستی پردازش‌های اضافه‌تری بر روی آن‌ها انجام شود تا قابل فهم برای کامپیوتر باشند، داده‌های بدون ساختار می‌گویند. حجم زیادی از داده‌های موجود در دنیای امروز مانند داده‌های صفحات وب و داده‌های شبکه‌ی اجتماعی داده‌های بدون ساختار هستند و سالانه چندین برابر می شود. در نتیجه، ما با یک مشکل اجتناب‌ناپذیر و چالش برانگیزی از اطلاعات اضافی مواجه هستیم. بنابراین بهره برداری مناسب از چنین منابع اطلاعاتی و تصمیم گیری هوشمندانه بر اساس آنها یکی از نیازهای مهم است. پایگاه داده‌های متنی شامل مجموعه‌ی بزرگی از اسناد و منابع مختلف (مانند مقالات، صفحات خبری، کتاب‌ها، ایمیل‌ها و صفحات وب و...) است. افزایش چشمگیر این نوع اطلاعات وجود ابزارهایی برای ارزیابی خودکار منابع متنی را بیش از هر زمان دیگری آشکار می سازد. در این میان خلاصه‌سازی خودکار متون یکی از راهکارهایی است که با حذف اطلاعات تکراری و کم اهمیت از اتلاف وقت کاربران می‌کاهد. [2] داده‌کاوی[1]، به مفهوم استخراج اطلاعات نهان یا الگوها و روابط مشخص با حجم زیادی از داده‌ها که در یک یا چند بانک اطلاعاتی بزرگ ذخیره شده‌اند، گفته می‌شود. متن‌کاوی[2]، به داده‌کاوی‌ای که بر روی متن انجام شود اشاره دارد. [3] از مسائل مهم متن‌کاوی می‌توان به استخراج اطلاعات[3]، ردیابی موضوع[4]، خلاصه‌سازی[5]، رده‌بندی[6]، خوشه-بندی[7] اشاره کرد. مسئله خلاصه‌سازی می‌تواند به عنوان یک زیر مسئله، در مسائل دیگر مطرح شود. به عنوان مثال در خوشه‌بندی متون می‌توانیم خلاصه‌ای از متن مورد نظر را که به صورت فشرده و فاقد متون تکراری و حشو است، به عنوان ورودی برای خوشه‌بندی در نظر بگیریم.[4] خلاصه‌سازی متون توسط انسان با وجود داشتن مزایایی از قبیل صحت و جامعیت، مستلزم صرف وقت و هزینه‌ی بالایی است. همچنین خلاصه‌سازی اسناد بزرگ به صورت دستی برای انسان کاری دشوار است. لذا وجود یک سیستم خلاصه‌ساز خودکار باعث صرفه‌جویی در زمان و هزینه‌ی مصرفی در تولید متن خلاصه خواهد شد، هرچند که ممکن است از لحاظ کیفیت با خلاصه تولید شده توسط انسان برابری نکند. با توجه به اهمیت خلاصه‌سازی در این مقاله ما به بررسی این مسئله‌ی می‌پردازیم.

[1] Data Mining
[2] Text Mining
[3] Information extraction
[4] Topic tracking
[5] Summarization
[6] Categorization
[7] Clustering

یک خلاصه‌ی خوب باید دارای حداکثر اطلاعات مهم متن، قابلیت فهم بالا و حجم کمینه باشد، هم‌چنین موضوعات گوناگون یک فایل متنی را با کمترین میزان تکرار پوشش دهد. خلاصه‌ای که بتواند حداکثر این ویژگی‌ها را حفظ کند، از کیفیت بالاتری برخوردار خواهد بود.[5] به‌طور کلی می‌توان گفت، کیفیت یک خلاصه می‌تواند نزدیکی آن به خلاصه‌ی تولید شده توسط انسان تعبیر شود. برای سنجش کیفیت و دقت الگوریتم برای خلاصه‌ی تولید شده از معیارهای متعددی استفاده می شود که یکی از شناخته شده‌ترین معیارهای ارزیابی سیستم‌های خلاصه‌ساز خودکار، معیارهای ارزیابی ROUGE[1] می‌باشدکه در حوزه‌های پردازش زبان طبیعی و بازیابی اطلاعات نیز مورد استفاده قرار می‌گیرد.[6] معیارهای ارزیابیROUGE کیفیت خلاصه خودکار را با مجموعه‌ی خلاصه‌های دستی تولید شده توسط انسان مقایسه می‌کند. این مقایسه از طریق هم‌پوشانی واحدهای متنی مانند nتایی‌ها، رشته‌ی کلمات، جفت کلمات و ... بین خلاصه‌ی خودکار سیستمی و خلاصه‌ی انسانی صورت می‌گیرد که با توجه به حساسیت هرکدام به جنبه‌های متفاوت از هم متمایز می‌شوند. [7]

سیستم‌های خلاصه‌ساز از منظر منابع به خلاصه‌سازهای تک‌سندی یا خلاصه‌سازهای چندسندی، و از منظر نوع خلاصه به خلاصه‌ی استخراجی[2] و خلاصه چکیده‌ای[3] تقسیم‌بندی می‌شوند. در خلاصه‌سازی استخراجی جملات مهم متن ورودی عینا در خلاصه می‌آیند. از جمله‌ی این روش‌ها می‌توان روش‌های مبتنی بر استخراج جملات، روش‌های مبتنی بر تحلیل آماری، روش‌های یادگیری ماشین را اشاره کرد. این روش در مقایسه با نقطه‌ی مقابل آن، یعنی خلاصه‌سازی چکیده‌ای، از پیچیدگی کمتری برخورداراست. یکی از مهمترین چالش‌ها برای یک سیستم خلاصه‌ساز استخراجی، مرحله‌ی پیش‌پردازش است که در آن متن مورد نظر برای استخراج جملات خلاصه، عمدتا براساس عملگرهای پردازش زبان طبیعی نظیر ریشه‌یابی، حذف کلمات توقف، برچسب‌زنی نقش کلمات و تعیین کلمات کلیدی پردازش می‌شود. پس از آن، به هر جمله در متن امتیازی تعلق می‌گیرد و درنهایت، جملات با امتیاز بالا انتخاب می‌شوند. از آنجا که کلمات کلیدی نقش به سزایی در تعیین امتیاز جملات ایفا می‌کنند، مرحله‌ی پیش‌پردازش اهمیت ویژه‌ای در خلاصه‌ساز استخراجی دارد. خلاصه‌سازی چکیده‌ای با تغییر در ساختار جملات سعی درتولید جملات جدید برای خلاصه دارد. [8]

تاکنون پژوهش‌های متعدد و روش‌های گوناگونی برای خلاصه‌سازی متن به روش استخراجی و چکیده‌ای ارائه شده که در ادامه به چند مورد از آنها اشاره می‌شود. در مقاله [9] خلاصه‌سازی استخراجی با روش دومرحله‌ای مبتنی بر ایجاد گراف و بهینه‌سازی روی تابع هدف انجام شده است. مجموعه داده ورودی در این DUC2002 - DUC2001 می‌باشد و نتایج به دست آمده با روش‌های خلاصه‌سازی دیگری مانند لکس رنک [10] و ماشین بردار پشتیبان [11] و مدل بهینه‌سازی تحولی فازی [12] مقایسه شده و در نهایت نشان داده شده کارایی بهتری نسبت به این روش‌ها داشت. یکی دیگر از روش‌های خلاصه‌سازی خوشه‌بندی است که برای هر موضوع یک خوشه در نظر گرفته می‌شود. شباهت میان جمله‌ها براساس مجموعه‌ای از پارامترها بررسی می‌شوند و سپس عبارت‌های مشابه در یک خوشه قرار می‌گیرند. در هر خوشه، به جمله‌هایی که شباهت بیشتری به عنوان خوشه دارند، امتیازهای بیشتری اختصاص می‌یابد و درنهایت، می‌توانند برای خلاصه انتخاب شوند. [13]

[1] Recall-Oriented Understudy for Gisting Evaluation
[2] Extractive
[3] Abstractive

در این مقاله با انجام الگوریتم خوشه‌بندی و استخراج جملات پرتکرار بر روی مجموعه داده‌ای که شامل صد مقاله تصادفی از میان مقالات پزشکی است، به نتایج موفقی رسیده‌اند.

همچنین از الگوریتم‌های شبکه عصبی مانند شبکه عصبی بازگشتی RNN[1] برای تولید خلاصه‌ی اسناد استفاده شده است که می‌توان به مقالات[15] [14] اشاره داشت. در زمینه‌ی خلاصه‌سازی متن به روش چکیده‌ای نیز روش‌هایی متعددی وجود دارد که به یکی از آنها در مقاله [16] اشاره شده است. این روش با کاهش- افزایش ابعاد منابع اسناد، متن خلاصه را تولید می‌کند. در مقاله [17] فرآیند تولید خلاصه بر مبنای موازی نمودن یک الگوریتم بهینه‌سازی مشهور تحت عنوان الگوریتم کلونی زنبور عسل انجام می‌شود. این الگوریتم بر اساس هوش جمعی و رفتار هوشمندانه جمعیت طراحی شده است.

روش‌های گوناگونی برای خلاصه‌سازی متن وجود دارد و رویکرد مبتنی بر جبرماتریسی که اساس اکثر این روش‌هاست، مجموعه‌ای از جملات را که نماینده‌ی مناسبی از متن اصلی بوده انتخاب می‌کند و به خروجی می‌فرستد. تعداد این جملات کمینه بوده و حداکثر بازسازی متن اصلی را دارند. رویکرد مبتنی بر جبرماتریسی، برای انجام خلاصه‌سازی از تجزیه ماتریسی استفاده می‌کنند که می‌توان به روش‌های[18] NMF و[19] LSA اشاره کرد.

همانطور که اشاره کردیم یک خلاصه خوب باید حجم کمینه‌ای داشته باشد و در عین حال از لحاظ مفهومی شامل مهمترین جملات متن که می‌توانند سایر جملات پوشش دهند، باشد. در مقاله [5] تعدادی از جملات متن اصلی را به عنوان خلاصه استخراج کرده، به‌طوری که جملات انتخاب شده حداکثر قابلیت بازسازی متن اصلی را دارا باشند. این جملات از افزونگی کم‌تری برخوردار بوده و تعدادشان در کمترین حالت ممکن است. رویکرد این مقاله تجزیه ماتریس با استفاده از روش نمایش تنک و روش پیشنهادی نمایش تنک گروهی می‌باشد. این رویکرد به دنبال یافتن ترکیب خطی از مجموعه مشاهدات قبلی (دیکشنری) است که از کم‌ترین تعداد ستون‌های این ماتریس استفاده کند. در سال‌های اخیر نمایش تنک به دلیل این که اغلب مجموعه داده‌ها شامل داده‌های تکراری و شبیه به هم هستند و می‌توان داده‌ها را به صورت ترکیب خطی از هم نوشت، اهمیت و موفقیت زیادی در انواع کاربردهای پردازش تصویر، نمونه‌برداری، فشرده‌سازی، بازیابی و طبقه‌بندی داده‌ها کسب کرده است.

در این مقاله هر جمله به صورت ترکیب خطی از سایر جملات نشان داده می‌شود. یعنی همه‌ی جملات متن با وزن‌های متفاوت، در ساخت هر جمله‌ی دیگر نقش دارند که در نهایت جملاتی به عنوان خلاصه انتخاب می‌شوند که بتوانند سایر جملات را خوب بازسازی کنند. پس جملاتی که دارای ضرایب خطی بیشتری در معادله خطی یک جمله‌اند، اهمیت زیادی دارند. برای انجام این کار کلمات کلیدی متن که با روش TF-IDF [20] مشخص شده، ماتریس جملات بر اساس کلمات کلیدی (ماتریس S) را تشکیل می‌دهد. نتیجه‌ی تجزیه این ماتریس، یک ماتریس تنک (W) که شامل وزن جملات است.

در مقاله [5] که از نرم L_2 در $||S-SW||$ برای بازسازی جملات استفاده شده است، باعث می‌شود جواب آن یعنی W در اکثر مولفه‌هایش مقداری ناچیز داشته باشد، یعنی هر جمله نسبت به تمام کلمات (حتی کلمات پرت و کم اهمیت) بازسازی شود. برای حل این مشکل و بهبود کیفیت خلاصه‌ی تولید شده، در این مقاله از نرم L_1 که تنک‌تر است و در بیشتر مواقع

[1] Recurrent Neural Network

صفر می‌باشد، استفاده شده است. این امر سبب می‌شود اولا تعداد جملات انتخابی کمینه شود و دوما فقط کلمات مهم در بازسازی همه‌ی جملات نقش داشته باشند.

۱. تعریف نمایش تنک در خلاصه‌سازی متن

در بحث‌های متن‌کاوی هر جمله را به صورت یک بردار نمایش می‌دهند. فرآیند نمایش یک داده متنی (جمله) با یک بردار توسط روش‌های مختلفی از حوزه پردازش زبان‌های طبیعی انجام می‌شود که یکی از آن‌ها روش TF-IDF است. با این روش هر جمله به صورت یک بردار نمایش داده می‌شود که مولفه‌های هر بردار میزان اثر کلمه کلیدی را در هر جمله مربوطه نشان می‌دهد. تشخیص کلمات کلیدی و میزان اهمیتشان در سند با روش TF-IDF انجام می‌گیرد. در مسئله خلاصه‌سازی متن ورودی‌ها مجموعه از بردارهای s_i است که جملات را نشان می‌دهند. مجموعه این جملات (بردارها) را با ماتریس تنک "ترم-جمله" $S = [s_1, \dots, s_n] \in R^{m \times n}$ نمایش می‌دهند که m نشان دهنده بعد هر جمله یعنی تعداد کلمات کلیدی سند و n تعداد جملات است. در حالت کلی هدف خلاصه‌سازی پیدا کردن یک زیر مجموعه از بردارهای ماتریس S است که نشان‌دهنده کلیت بردارهای ماتریس S باشند. مولفه‌های برداری یک جمله را کلمات کلیدی مستخرج از اسناد تشکیل داده‌اند، به این طریق که هر عنصر این بردار با یک مقداردهی خواهد شد اگر که کلمه‌ی متناظر در جمله‌ی منظور یافت شود، در غیر این‌صورت مقدار صفر می‌گیرد.[5]

براساس موفقیت روش‌های مبتنی بر نمایش تنک در حوزه‌های مختلف تجزیه و تحلیل داده‌ها، اخیرا روش تنک در مسئله خلاصه‌سازی نیز مورد استفاده قرار گرفته است. [21]

رویکرد نمایش تنک [5] به معنی یافتن ضرایب یک ترکیب خطی از داده‌های مشاهده شده قبلی است (که به آن دیکشنری می‌گوییم) به صورتی که تعداد صفرهای این ضرایب خطی بیشینه باشد. به عبارتی دیگر از کمترین تعداد ستون‌های این دیکشنری استفاده کنیم.

مدل‌سازی مسئله خلاصه‌سازی با استفاده از نمایش تنک، جملات اسناد را به عنوان بردار (برای رایانه قابل فهم است) در نظر گرفته که معمولاً به طور طبیعی یا تنک می‌باشد یا قابل تبدیل شدن به پایه‌های تنک را دارند. می‌خواهیم بنا به روش استخراجی تعداد کمی از این بردارها را به عنوان خلاصه‌ای از یک متن ارائه دهیم.

ایده روش بازسازی تنک در خلاصه‌سازی متن این است که هر جمله به صورت تنک توسط بردارهای دیگر بازسازی می‌کند. حال اگر جمله‌ای در بازسازی تنک همه جملات نقش بیشتری ایفا کند حتماً می‌تواند به عنوان یکی از خلاصه‌های متن انتخاب شود.

به صورت ریاضی روش بازسازی تنک در خلاصه‌سازی به صورت زیر است. اگر s_i جمله i ام و n تعداد جملات متن در نظر گرفته شود:

$$\forall s_i \in S \implies s_i \approx \sum_{j=1}^{n} w_{j,i} s_j \qquad (۱)$$

که $w_{j,i}$ نقش جمله‌ی j در بازسازی جمله i را نشان می‌دهد. عبارت به این معناست که جمله i ام را می‌توانیم با یک ترکیب خطی از سایر جملات با وزن‌های مختلف بازسازی کنیم. حالت ایده آل برای ما این است که هر یک از جملات متن را بتوان توسط جملات انتخاب شده به عنوان خلاصه بازسازی کرد، به طوری که خطای حاصل از این تقریب باید حداقل شود یعنی:

$$min \sum_{i=1}^{n} \left\| s_i - \sum_{j=1}^{k} w_{j,i} s_j \right\|_2^2 = \sum_{i=1}^{n} \|s_i - Sw_i\|_2^2 \qquad s.t \;\; diag(W) = 0 \,.\, W \geq 0 \tag{۲}$$

که در آن $w_i = \begin{bmatrix} w_{1i} \\ \vdots \\ w_{ni} \end{bmatrix}$ است. همچنین اگر W را به این صورت $W = [w_1, \dots, w_n]$ تعریف کنیم، آنگاه معادله را می‌توان به صورت زیر نوشت:

$$min \|S - SW\|_F^2 \tag{۳}$$

چون هر s_i نباید در ساختن خود نقش داشته باشد، در مقاله [5] از شرط $diag(W) = 0$ استفاده کرده‌اند. همچنین ادعا شده است که برای حذف جملات اضافی و افزونگی از شرط W>=0 استفاده کرده‌اند. از آنجا که هر سطر ماتریس W میزان اثر یک جمله در ساختن بقیه را نشان می‌دهد چون هدف خلاصه سازی است پس در معادله ۳ باید سعی شود تا تعداد جملات کمتری در ساختن بقیه مشارکت داشته باشد که این با اضافه کردن شرطی که سعی کند بیشتر سطرهای W صفر شوند امکان‌پذیر است.

در مقاله [5] برای اعمال این شرط از نرم $\|W\|_{2,1}$ که به صورت زیر تعریف می‌شود:

$$\|W\|_{2,1} = \sum_{i=1}^{k} \|W(i,:)\|_2 \tag{۴}$$

در این معادله $W(i,:)$ نشان دهنده‌ی سطر iام ماتریس W است و $\|W\|_{2,1}$ به عنوان عامل جریمه در معادله ۳ استفاده شده است. در ماتریس $W \epsilon R^{n \times n}$ سطر iام نشان دهنده نقش جمله iام در ساختن همه جملات است. حال چون می‌خواهیم تعداد جملات کمتری در ساختن همه جملات نقش داشته باشند بنابراین دلخواه ما این است که W حاصل از معادله ۳ حداکثر سطرهایش برابر با صفر باشد.

اضافه کردن شرط ۴ در بازسازی جملات بر روی W که باعث می‌شود که ماتریس W درسطرها تنک بوده و تعداد جملات انتخابی برای خلاصه که حداکثر بازسازی را دارد کمینه باشد. بنابراین مدل ارائه شده در مقاله [5] به صورت زیر نوشته شده است:

$$\min \textstyle\sum_{i=1}^{n} \|s_i - Sw_i\|_2^2 + \lambda \|W\|_{2,1} \;\; s.t \;\; diag(W) = 0 \,, W \geq 0 \,, \; S \epsilon R^{n \times m} \tag{۵}$$

نویسندگان با نتایج پیاده‌سازی، مزیت روش پیشنهادی خود در مقایسه با روش‌های دیگر را نشان داده‌اند. نویسندگان برای حل این مدل از روش نستروف استفاده کرده اند که شامل یک تکرار درونی و یک تکرار بیرونی است و نشان داده اند که پیچیدگی این مدل برابر با $O(kln^3)$ می باشد. در اینجا n تعداد جملات و k,l تعداد تکرارهای درونی و بیرونی هستند.

اما این مدل یک ضعف عمده دارد. استفاده از نرم L2 در بازسازی جمله در مدل معادله ۵ باعث می‌شود که بنابه خاصیت نرم L2 اکثر درایه‌های $r_i = s_i - Sw_i$ جای ممکن یکنواخت و غیر صفر شوند. دلیل این اتفاق این است که در بهینه سازی نرم L2 تلاش می‌کند تا نسبت به تک تک درایه‌ها کمینه شدن را انجام دهد. بنابراین در این مسئله سعی می‌شود بازسازی جمله si توسط بقیه برای هر کلمه نیز اتفاق بیفتد. اگر کلمه‌ای به صورت اتفاقی(نویز) در این جمله باشد این تلاش باعث می شود تا نتیجه به شدت به این کلمه بی‌ربط حساس شود. در ادامه روشی ارائه می‌دهیم که این حساسیت را نسبت به مدل بالا کمتر داشته باشد

۲. روش پیشنهادی

در این بخش در پی انجام فرآیندهایی برای بهبود روش ذکر شده هستیم، به طوری که هم مزیت‌های آن را حفظ کند و از معایب آن بکاهد و باعث بهبود کیفیت خلاصه و دقت الگوریتم شود. با بررسی عمیق‌تر مدل، روش پیشنهادی استفاده از نرم L_1در ترم بازسازی و منظم سازی می‌باشد. همانطور که پیش‌تر توضیح دادیم، استفاده از نرم L_2 در مسئله‌ی بهینه‌سازی باعث می‌شود که بنابه خاصیت نرم L_2 اکثر درایه‌های $r_i = s_i - Sw_i$ تا جای ممکن یکنواخت باشد در واقع سعی می‌کند هر کدام از کلمات را به خوبی بازسازی کند. بنابرین اگر در جمله‌ی مانند s_i یک کلمه‌ی نامربوط اضافه شده باشد این حالت حتما در کیفیت بازسازی تاثیر خواهد داشت. استفاده از نرم L_1 در مینیمم‌سازی تابع جریمه $\|s_i - Sw_i\|_1$ سبب می‌شود که اکثر r_i ها در بیشتر مواقع صفر خواهد بود و در برخی مواقع بزرگ می‌باشد. بنابراین استفاده از نرم L_1 باعث می‌شود که حساسیت فرآیند خلاصه‌سازی نسبت به کلمات پرت و نامربوط از بین برود. علاوه‌براین چون می‌خواهیم کمترین تعداد از جملات در بازسازی نقش داشته باشند، اگر برای مینیمم‌سازی $\|w_i\|$ از نرم L_1 استفاده کنیم سبب می‌شود که جواب حاصل تنک بوده و در بیشتر مواقع صفر باشد. بنابراین مدل پیشنهادی به‌صورت زیر است:

$$min_w \|s_i - Sw_i\|_1 + \lambda\|w_i\|_1 \quad (۶)$$

در این مدل چون هر جمله نباید در ساختن خودش نقش داشته باشد باید شرط $w_{ii} = 0$ را داشته باشیم. با تعریف نمادهای زیر

$$\bar{S}_i = [s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_n] \in R^{m\times(n-1)} \quad (۷)$$

$$\bar{w}_i = [w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_n] \quad (۸)$$

معادله ۶ با مسئله زیر معادل خواهد شد:

$$min_{\overline{w}_i}||s_i - \bar{S}_i\overline{w}_i||_1 + \lambda||\overline{w}_i||_1 \quad (۹)$$

بنابراین برای هر جمله یک مسئله به‌فرم بالا باید حل شود. در ادامه روش حلی برای این مسائل پیشنهاد می‌کنیم.
چون هر دو ترم بازسازی و منظم‌سازی هر دو نرم L_1 هستند، بنابراین برای حل آن از روش‌های معمول نمی‌توانیم استفاده کنیم. برای حل این مسئله باید آنها را تغییر داد. یکی از روش‌های ساده در حل مسائل با نرم یک، روش تکراری است که در هر مرحله نرم یک را با نرم دو تقریب زده و مسئله را به یک مسئله کمترین مربعات تبدیل می‌کند. نام این روش کمترین مربعات وزندار تکراری IRLS[1] است. این روش سابقه طولانی در حل مسائل خطی با نرم یک دارد. همچنین از این روش در حل مسائلی که عبارت منظم سازی نرم یک دارند استفاده شده است[22,23]. در ادامه ما از ایده تقریب نرم یک با نرم دو در یک فرآیند تکراری برای حل مسئله ۱۰ استفاده می‌کنیم.

برای هر $x \in R^n$ می‌توان نشان داد که:

$$\|x\|_1 = \|L(x)x\|_2^2 \quad (۱۰)$$

است که در آن که:

$$L(x) = diag\left(\frac{1}{\sqrt{|x_i| + \delta(x_i)\varepsilon}}\right)_{i=1,\dots,n} \quad (۱۱)$$

$$\delta(x_i) = \begin{cases} 1, & x_i = 0; \\ 0, & x_i \neq 0 \end{cases} \quad (۱۲)$$

در آن ε یک عدد کوچک نزدیک به صفر فرض می‌شود.

در رویکرد IRLS فرض بر این است که اگر در مرحله k یک فرآیند تکراری $\overline{w}^k$ معلوم باشد آنگاه بروزرسانی جدید $\overline{w}$ نزدیک $\overline{w}^k$ فرض می‌شود. یعنی اگر $\overline{w}_i = w_i^k + \Delta\overline{w}_i^k$ آنگاه $\Delta\overline{w}_i^k \simeq 0$.

بنابراین $L(\overline{w}) \simeq L(\overline{w}_k)$ خواهد بود و داریم:

$$\|\overline{w}_i\|_1 = \left\|L(\overline{w}_i)\overline{w}_i\right\|_2^2 \simeq \left\|L(\overline{w}_i^k)\overline{w}_i\right\|_2^2 \quad (۱۳)$$

حسن این روش تبدیل نرم L_1 به نرم L_2 در هر مرحله از این روش است. این ایده در بسیاری از کاربردهای مرتبط با نرم L_1 استفاده شده است. از این رویکرد برای تقریب نرم منظم ساز و نیز عبارت بازسازی در روش یادگیری دیکشنری در مقاله [24] استفاده شده است که ما نیز در اینجا از آن رویکرد استفاده می‌کنیم.

[1] Iteratively reweighted least squares

بنابراین با داشتن $\overline{w}^k$ برای تخمین جدید از جواب معادله از روش مبتنی بر IRLS روابط زیر را خواهیم داشت:

$$\|\overline{w}_i\|_1 \simeq \left\| L(\overline{w}_i{}^k)\overline{w}_i \right\|_2^2 \quad (۱۴)$$

اگر $\rho_i^k = s_i - \bar{S}_i \overline{w}_i$ آنگاه:

$$\left\|s_i - \bar{S}_i \overline{w}_i\right\|_1 \simeq \left\| L_1(\rho_i^k)(s_i - \bar{S}_i \overline{w}_i) \right\|_2^2 \quad (۱۵)$$

بنابراین معادله ۱۰، در مرحله k+1ام به صورت زیر تقریب زده می‌شود:

$$w^{k+1} = argmin \left\| L_1(\rho_i^k)(s_i - \bar{S}_i \overline{w}_i) \right\|_2^2 + \lambda \left\| L(\overline{w}_i{}^k)\overline{w}_i \right\|_2^2 \quad (۱۶)$$

حال اگر قرار دهیم $L_1^i = L_1(\rho_i^k)$ و $L_2^i = L(\overline{w}_i{}^k)$

در اینصورت مسئله ۱۶ به صورت مسئله‌ی مینیم‌سازی به صورت زیر تبدیل می‌شود:

$$min_{\overline{w}} \left\| L_1^i(s_i - \bar{S}_i \overline{w}_i) \right\|_2^2 + \lambda \left\| L_2^i \overline{w}_i \right\|_2^2 = min_{\overline{w}_i} \left\| \begin{bmatrix} L_1^i(\bar{S}_i \overline{w}_i - s_i) \\ \sqrt{\lambda} L_2 \overline{w}_i \end{bmatrix} \right\|_2^2, \quad (۱۷)$$

$$L_1^i \in R^{m\times m}, L_2^i \in R^{(n-1)\times(n-1)}$$

که در نهایت مسئله‌ی کمترین مربعات زیر را خواهیم داشت:

$$min_{\overline{w}} \left\| \begin{bmatrix} L_1^i \bar{S}_i \\ \sqrt{\lambda} L_2 \end{bmatrix} \overline{w}_i - \begin{bmatrix} L_2^i s_i \\ 0 \end{bmatrix} \right\|_2^2, \quad L_1^i \bar{S}_i \in R^{m\times(n-1)} \quad (۱۸)$$

اگر A_i و b_i را به‌صورت زیر تعریف کنیم:

$$A_i = \begin{bmatrix} L_1^i \bar{S}_i \\ \sqrt{\lambda} L_2 \end{bmatrix} \qquad b_i = \begin{bmatrix} L_2^i s_i \\ 0 \end{bmatrix} \quad (۱۹)$$

خواهیم داشت:

$$\overline{w}_i^{k+1} = argmin_{\overline{w}_i} \|A_i \overline{w}_i - b_i\|_2^2 \quad (۲۰)$$

در اینجا با داشتن تقریب تکرار kام از جواب توانستیم با تقریب نرم یک، در مرحله k+1 الگوریتم پیشنهادی تقریب دیگری از جواب را بدست آوریم. لذا این فرآیند تا برآورده کردن شرط توقف ادامه پیدا خواهد کرد.

الگوریتم کلی روش پیشنهادی به صورت زیر تعریف می‌شود:

الگوریتم کلی روش پیشنهادی

مفروضات: ماتریس S، پارامتر $\boldsymbol{\lambda}$ ، مقدار اولیه $\overline{\boldsymbol{w}}_{\boldsymbol{i}}^{\boldsymbol{0}}$ برای هر i که ما برابر بردار یک در نظر گرفتیم.

۱. مراحل زیر را تا زمانی که $\|\mathbf{S}-\mathbf{SW}\|_{\mathbf{F}}^{\mathbf{2}} \leq \boldsymbol{\varepsilon}\|\mathbf{S}\|_{\mathbf{F}}^{\mathbf{2}}$ تکرار کنید.

۱-۱. به ازای هر i در بازه‌ی ۱ تا n (تعداد جملات) مراحل زیر را انجام دهید.

۱-۱-۱. برای هر $\overline{\boldsymbol{w}}_{\boldsymbol{i}}^{\boldsymbol{k}}$ ، $\boldsymbol{L}_{\mathbf{1}}^{\boldsymbol{i}}$ و $\boldsymbol{L}_{\mathbf{2}}^{\boldsymbol{i}}$ را بروزرسانی کنید.

۱-۱-۲. متغیرهای $\mathbf{b_i}$ و $\mathbf{A_i}$ را با توجه به فرمول ۱۹ بروزرسانی کنید.

۱-۱-۳. $\overline{\boldsymbol{w}}_{\mathbf{i}}^{\mathbf{k+1}} = \mathbf{argmin}\|\mathbf{A}_{\mathbf{i}}\overline{\boldsymbol{w}}_{\boldsymbol{i}} - \mathbf{b}_{\mathbf{i}}\|_{\mathbf{2}}^{\mathbf{2}}$

خروجی: سطرهای W، بر اساس نرم L_2 آنها مرتب‌سازی شده و به تعداد دلخواه جملات انتخاب می‌شوند.

در اینجا شرط توقف بصورت $\|\mathbf{S}-\mathbf{SW}\|_{\mathbf{F}}^{\mathbf{2}} \leq \boldsymbol{\varepsilon}\|\mathbf{S}\|_{\mathbf{F}}^{\mathbf{2}}$ شده است. در ماتریس W برای هر i، عنصر i ام ستون i صفر بوده و بقیه عناصر این ستون همان عناصر $\overline{w}_i$ خواهند بود.

در اینجا دستگاه ۱-۱-۳ را می توان با روشهای مستقیم یا روشهای تکراری حل کرد. در روش مستقیم هزینه حل دستگاه از مرتبه $O(n^3)$ و در روش تکراری مانند LSQR با k تا تکرار هزینه برابر $O(kn^2)$ خواهد بود. از آنجا که این دستگاه در هر مرحله تقریبی است لذا حل دقیق آن نیز مد نظر نخواهد بود و می توان از تعداد تکرارهای بسیار کمی استفاده نمود. لذا در روش مستقیم و تکراری، کل هزینه برای الگوریتم بالا برابر با $O(n^4)$و $O(kn^3)$ خواهد بود. در داده های زیر چون تعدادجملات کم بوده از روش مستقیم QR برای حل دستگاه استفاده گردیده است که در بخش ۵ توضیحات بیشتر داده شده است.

۳. بررسی الگوریتم پیشنهادی

برای بررسی صحت عملکرد روش پیشنهادی، تمامی متون مجموعه‌ی داده‌ای DUC 2002 را که ۱۱۵ سند است، توسط هر دو روش آزمایش و نتایج بدست آمده توسط معیار F-measure مقایسه شده است.

به عنوان مثال برای خلاصه کردن یکی از سندها (سند شماره ۱۰۶)، اولین مرحله ساخت کلمات کلیدی است، که رویکرد مورد استفاده در این مقاله برای استخراج کلمات کلیدی، روش TF-IDF است. تعداد این کلمه‌های کلیدی ۵ درصد (جدول ۱) از تمام کلمات متن تمامی سندها انتخاب شده است.

کلمه	TF-IDF	TF	DF
Tornado	29.02963	11	8
Storm	8.958797	5	17
Jersey	8.833317	3	6
injured	7.783641	4	16
Cafeteria	2.275172	2	3
Quayle	7.275172	2	3
Illinois	6.664409	2	4
Delaware	6.664409	2	4
Storms	6.591674	3	12
York	6.437752	4	20
thunderstorms	5.888878	2	6

جدول ۱- کلمه‌های کلیدی با بیشترین مقادیر TF-IDF

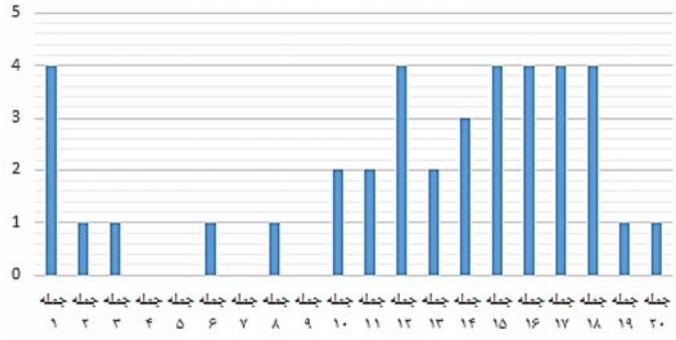
توجه شود که مقدار DF از بین تمامی اسناد مجموعه‌ی داده‌ای DUC 2002 محاسبه شده است که تعداد آنها ۱۱۵ سند است.

با استفاده از کلمات کلیدی به‌دست‌آمده از مرحله‌ی قبل، ماتریس S برای این سند ۲۰ جمله‌ای به صورت ماتریس «کلمه-جمله» که محور عمودی شامل کلمات کلیدی که با روش TF-IDF به دست آمده و محور افقی شامل همه جملات سند ورودی به صورت شکل ۱ ساخته خواهد شد.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}	s_{13}	s_{14}	s_{16}	s_{17}	s_{18}	s_{19}	s_{20}
tornado	1	1	1	0	0	1	0	1	0	1	1	0	1	0	1	1	1	0	0
storm	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1
jersey	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0
injured	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
cafeteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
quayle	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
illinois	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
delaware	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
storms	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
york	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0
thunderstorms	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

شکل ۱- ماتریس ترم- جمله (S) برای سند ۱۰۶

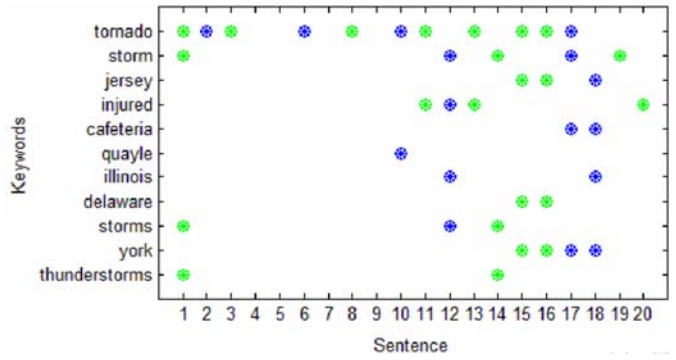
قبل از اجرای الگوریتم برای این سند، می‌توانیم میزان پراکندگی کلمه‌های کلیدی نسبت به جملات را در نمودار شکل۲ مشاهده کنیم. جمله‌های S_1 ، S_{12} ، S_{14} ، S_{15} ، S_{16} ، S_{17} ، S_{18} دارای تعداد کلمات کلیدی یکسانی هستند، بنابراین باید با توجه به اهمیت کلمات کلیدی‌شان برای خلاصه انتخاب شوند. در شرایط یکسان برای جملات، بعد از مرتب‌سازی جملات بر اساس نرم سطری ماتریس W، جملات در رتبه‌های بالاتر به خروجی فرستاده می‌شوند تا تعداد k ی از پیش تعیین شده‌ی جملات انتخاب شوند.



شکل ۲- نمودار جملات سند و تعداد کلمه‌های کلیدی که در هر جمله

و همچنین این جملات انتخاب شده را بر روی ماتریس S نیز با شکل ۳ نشان می‌دهیم. دو دسته نقاط در شکل۳ وجود دارد که با رنگ‌های آبی و سبز نشان داده شده است. نقاط آبی رنگ کلمات کلیدی موجود در جملات انتخاب شده برای سیستم خلاصه‌ساز و نقاط سبز رنگ کلمات کلیدی سایر جملات هستند.

با توجه به این شکل مشخص است که در روش پیشنهادی ملاک انتخاب جملات، تعداد کلمات کلیدی نیست و ممکن است جملاتی که از کلمات کلیدی کم‌تری برخوردار هستند اولویت بالاتری داشته‌باشند. چرا که در روش پیشنهادی، جملاتی را که نسبت به برخی کلمات خوب بازسازی نمی‌شوند ولی نسبت به‌برخی کلمات دیگرشان خوب بازسازی می‌شوند را از دست نمی‌دهد (از چنین کلماتی چشم‌پوشی می‌کند).



شکل ۳- تصویر ماتریس S و جملات انتخاب شده پس از خلاصه‌سازی به روش پیشنهادی(نقاط آبی جملات انتخاب شده در خلاصه سازی و رگ سبز بقیه جملات را نشان می دهند)

با استفاده از نرم L_1 در تابع جریمه اکثر مواقع بردار حاصل مقدار نزدیک به‌صفر و در برخی مواقع غیر صفر می‌شود، چرا که می‌خواهیم از تاثیر برخی کلمات که به‌درستی در کلمات کلیدی انتخاب نشده‌اند چشم‌پوشی کنیم. در حالی که در روش قبل در اکثر مواقع مقداری کوچک به‌خود می‌گیرد، به‌این معنی که کلمات در بازسازی تاثیر یکسان دارند.

به عنوان مثال در جدول ۲ مقدار مانده بازسازی $r_i = s_i - Sw_i$ برای یک جمله از داده‌های بالا را در نظر گرفته و مقادیر بازسازی آن توسط دو روش را بدست آمده‌است. مقادیر نشان دهنده در هر سطر میزان اهمیت بازسازی هر کلمه در آن جمله را نشان می‌دهد. با توجه به جدول نشان داده می‌شود که روش پیشنهادی کلمات Tornado و Storm در بازسازی این جمله نقش زیادی دارد.

با توجه به جدول زیر واضح است که روش پیشنهادی در اکثر کلمات توانسته جمله مزبور را به خوبی بازسازی کند و مانده این روش تنک است. در حالی‌که روش تنک گروهی به این صورت نبوده و بازسازی را به خوبی انجام نداده است.

شماره سطر	روش پیشنهادی	روش تنک گروهی
۱	-۰/۹۵	-۲
۲	۰/۳	-۳
۳	۰	۰
۴	۰	-۱
۵	۰	-۱
۶	۰	۰
۷	۰	-۱
۸	۰	۰
۹	۰/۰۴	-۱
۱۰	۰	-۱
۱۱	۰/۰۵	۰

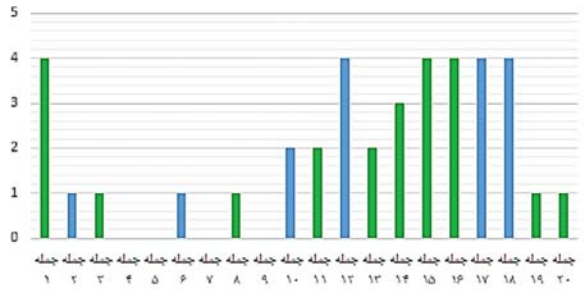
جدول ۲- مقایسه‌ی مقدار $s_i - Sw_i$

همچنین کیفیت خلاصه سازی روش ارایه شده در مقایسه با تنک گروهی را با معیارهای F-Measure و ROUGE بر روی سند ۱۰۶ مجموعه داده DUC2002 در جدول ۳ نشان داده شده است. این نتایج نشان دهنده پیشرفت چشم‌گیر خلاصه سازی توسط روش پیشنهادی هستند.

روش	شماره سند	تعداد جملات اصلی	تعداد جملات خلاصه	معیار F-Measure درصد	معیار ROUGE درصد
پیشنهادی	۱۰۶	20	10	۸۵/۰۲	۹۱/۳۲
تنک گروهی	۱۰۶	20	10	۶۸/۳۹	۷۹/۰۸

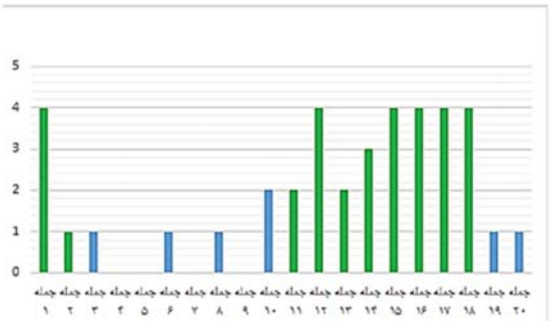
جدول ۳ - مثالی از ورودی و نتیجه الگوریتم پیشنهادی

اکنون با مقدار پارامتر $\lambda=1$، الگوریتم یافتن خلاصه را بر روی این داده اجرا می‌کنیم. نتیجه‌ی اجرای الگوریتم روش پیشنهادی، در نمودار میله‌ای شکل ۴ مشخص می‌شود. در شکل زیر جملاتی که با نمودارهای میله‌ای به رنگ سبز مشخص شده‌اند، جملات انتخاب شده به روش پیشنهادی می‌باشد.

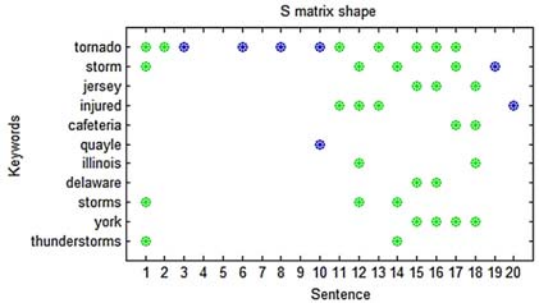


شکل ۴- نمودار میله‌ای تعداد کلمات کلیدی در جملات و جملات انتخاب شده در خلاصه‌سازی با روش پیشنهادی(سبز: جملات انتخابی- آبی: بقیه جملات)

برای روش تنک گروهی، خلاصه‌ی بدست آمده دارای جملات به شکل۵ می‌باشد. همانطور که مشخص است، در روش تنک گروهی جملاتی که از کلمات کلیدی بیشتری برخوردارند، جملات پراهمیت‌تر شناخته می‌شوند. محور افقی جملات و محور عمودی تعداد کلمات کلیدی را مشخص می‌کند همچنین در شکل ۵ جملاتی که با نمودارهای میله‌ای به رنگ سبز مشخص شده‌اند، جملات انتخاب شده به روش تنک گروهی می‌باشد. ماتریس کلمات کلیدی-جمله نیز برای آن نیز به‌صورت(شکل ۶) خواهد بود.



شکل ۵- نمودار میله‌ای تعداد کلمات کلیدی در جملات و جملات انتخاب شده در خلاصه‌سازی با تنک گروهی (سبز: جملات انتخابی- آبی: بقیه جملات)

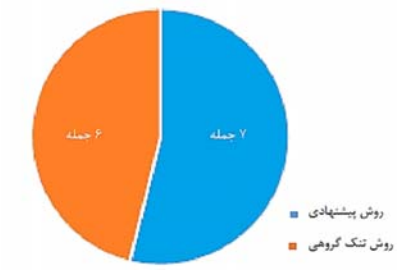


شکل ۶- تصویر ماتریس S و جملات انتخاب شده پس از خلاصه‌سازی با روش تنک گروهی (آبی: جملات انتخابی- سبز: بقیه جملات)

با توجه به شکل ۶ متوجه می‌شویم که در این روش جملات در شرایط یکسان، به‌ترتیب انتخاب می‌شوند تا میزان k تعداد از جملات برای خلاصه تکمیل شود. همچنین برای جملاتی که در تعداد کلمات کلیدی مشترک هستند، برای اینک بهترین بازسازی را داشته باشد، جملاتی را که در کلمات کلیدی با سایر جملات اشتراکات بیشتری دارند را انتخاب می‌کند.

به ازای هر دو روش، سعی شده بهترین مقدار λ بدست بیاید و نتایج با یکدیگر مقایسه شود.

در نهایت، برای مقایسه با خلاصه‌ی دستی و تعداد جملاتی که درست در خلاصه تشخیص داده شده‌اند، می‌توانیم از نمودار شکل ۷ زیر استفاده کنیم.



شکل ۷- مقایسه جملات درست تشخیص داده‌شده در خلاصه سیستمی

۱،۳. مقایسه پیچیدگی زمانی الگوریتم پیشنهادی

برای مقایسه زمانی الگوریتم پیشنهادی با الگوریتم تنک گروهی، می‌بایست توجه شود که هرکدام از این الگوریتم‌ها یک بخشی اصلی دارند که عمده زمان الگوریتم در آن گرفته می‌شود، این بخش اصلی در روش قبلی، قسمت یافتن و بهینه‌سازی

افکنش و روش تنک گروهی است و در روش پیشنهادی، حل دستگاه معادلات خطی می‌باشد. بنابراین در این بخش از دو دیدگاه، مقایسه زمانی را بین این دو روش انجام می‌دهیم:

- زمان سپری شده در کل اجرای الگوریتم‌ها با یکدیگر مقایسه می‌شود.
- زمان سپری شده در قسمت اصلی الگوریتم با یکدیگر مقایسه می‌شود.

برای مقایسه زمانی بین این دو روش، بین تمامی ۱۱۵ سند مجموعه‌ی داده‌ای DUC2002 قسمت اصلی و کل الگوریتم در هر دو روش محاسبه شد و میانگین آن‌ها به شکل جدول ۴ به‌دست آمد.

روش	زمان اجرای کلی الگوریتم (ثانیه)	زمان سپری شده در قسمت اصلی الگوریتم
تنک گروهی	۱,۰۷۹ ثانیه	۰,۰۰۸۲ ثانیه
پیشنهادی	۰,۷۱ ثانیه	۰,۰۰۳۸ ثانیه

جدول ۴- مقایسه‌ی میانگین زمان سپری شده در محاسبه خلاصه‌ی متن در هر دو روش

همانطور که در این جدول نشان داده شده، روش پیشنهادی زمان اجرای کمتری نسبت به روش تنک گروهی دارد. اگر چه در استفاده از روش QR برای حل دستگاه، پیچیدگی روش پیشنهادی بیشتر از روش قبلی است اما چون ابعاد داده ها کوچک بوده و از طرفی در پیاده سازی در نرم افزار متلب دستور آماده QR استفاده شده است در حالیکه در روش قبلی مجبور بودیم از چندین دستور تکرار در متلب استفاده کنیم لذا در پیاده سازی در نرم افزار متلب سرعت روش پیشنهادی بیشتر بوده است.

۲,۳. نتایج بعد از تغییر در ورودی

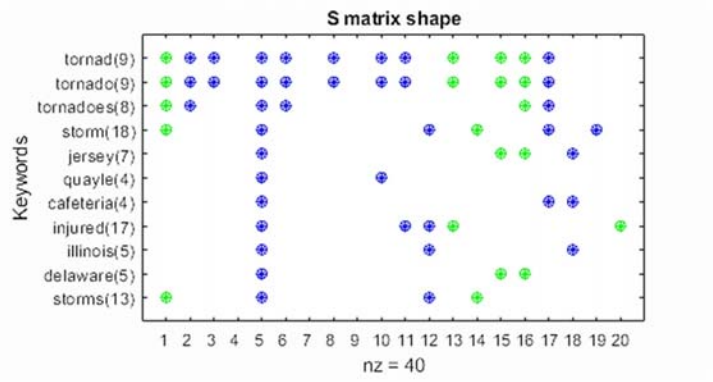
همانطور که اشاره کردیم روش پیشنهادی نسبت به کلمات پرت حساس نیست و در صورت وجود کلمه‌ی پرت و نامناسب در جمله‌ای، اخلالی در متن خلاصه خروجی ایجاد نخواهد شد. با بررسی خلاصه‌های مرجع، در این قسمت می‌خواهیم نشان دهیم که اگر یک جمله در سند شامل تمام کلمات کلیدی باشد در اینصورت این جمله از لحاظ معنایی فاقد ارزش و معنا است و لذا نباید در متن خلاصه باشد. برای بررسی این مورد ماتریس S را به صورت زیر تغییر می‌دهیم(شکل ۸):

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}	s_{13}	s_{14}	s_{16}	s_{17}	s_{18}	s_{19}	s_{20}
tornado	1	1	1	0	1	1	0	1	0	1	1	0	1	0	1	1	1	0	0
storm	1	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	1	0	1
jersey	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	1	0
injured	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0
cafeteria	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0
quayle	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
illinois	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0
delaware	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
storms	1	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0
york	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0
thunderstorms	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0

شکل ۸- ماتریس ترم-جمله S با قرار دادن کلمات پرت در جمله S_5

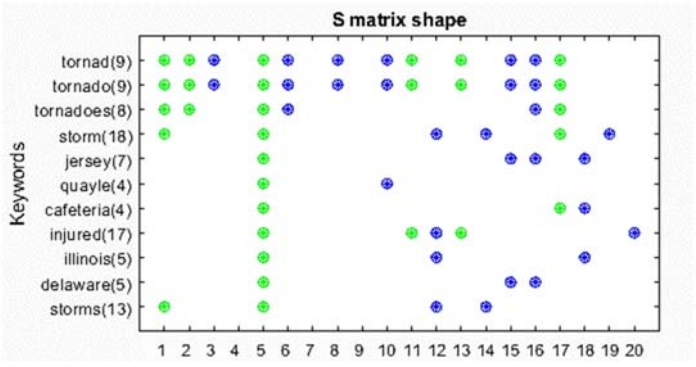
همانطورکه مشخص است جمله پنجم متن سند ۱۰۶ شامل همه‌ی کلمات کلیدی متن می‌باشد. در روش پیشنهادی دیدیم که طبق خاصیتی که نرم L_1 دارد، جملاتی را به عنوان خلاصه انتخاب می‌کند که بتوانند بهترین بازسازی از سایر جملات را انجام داشته باشند. در واقع جملاتی مهم‌تر از سایر جملات هستند که کلمات کلیدی آن‌ها ارتباط معنایی و مفهومی منطقی-تری با یک دیگر داشته باشند، بنابراین فقط داشتن تعداد زیادی از کلمه‌ی کلیدی میزان اهمیت جملات را نشان نمی‌دهد.

نتیجه‌ی فرآیند انتخاب جملات مهم متن برای تولید متن خلاصه برای هر دو روش پیشنهادی تصویر شکل ۹ و روش تنک گروهی در تصویر شکل ۱۰ مشخص شده است.



شکل ۹- تصویر ماتریس S و جملات S_{19}،S_{18}،S_{16}،S_{14}،S_{12}،S_1 انتخاب شده پس از خلاصه‌سازی با روش پیشنهادی بعد از تغییر ورودی (سبز: جملات انتخابی- آبی: بقیه جملات)

در شکل ۹ نقاط سبز رنگ کلمات کلیدی موجود در جملات انتخاب شده برای سیستم خلاصه‌ساز به روش پیشنهادی و نقاط آبی رنگ کلمات کلیدی سایر جملات هستند. در نتیجه در روش پیشنهادی جمله ۵ به عنوان جمله خلاصه انتخاب نشد.



شکل ۱۰- تصویر ماتریس S و جملات S_{17} ،S_{13}،S_{11}،S_5،S_2، S_1 انتخاب شده پس از خلاصه‌سازی با روش تنک گروهی پس از تغییر در ورودی(سبز: جملات انتخابی- آبی: بقیه جملات)

در شکل ۱۰ نیز نقاط سبز رنگ کلمات کلیدی موجود در جملات انتخاب شده برای سیستم خلاصه‌ساز به روش تنک گروهی و نقاط آبی رنگ کلمات کلیدی سایر جملات هستند. همانطور که مشخص است در این روش جمله ۵ به عنوان جمله‌ی خلاصه انتخاب شده است که نشان می‌دهد این روش فقط به تعداد کلمات کلیدی هر جمله در سند اهمیت می‌دهد چه بسا این جملات فاقد معنی باشد.

با توجه به ایجاد تغییر در ورودی کیفیت خلاصه سازی روش ارایه شده در مقایسه با تنک گروهی را با معیارهای F-Measure و ROUGE بر روی سند ۱۰۶ مجموعه داده DUC2002 در جدول ۵ نشان داده شده است. همچنین این نتایج نیز نشان دهنده بهبود کیفیت خلاصه سازی توسط روش پیشنهادی هستند.

روش	شماره سند	تعداد جملات اصلی	تعداد جملات خلاصه	معیار F-Measure درصد	معیار ROUGE درصد
پیشنهادی	۱۰۶	۲۰	۱۰	۶۴/۱۷	۶۷/۵۲
تنک گروهی	۱۰۶	۲۰	۱۰	۴۷/۰۲	۵۰/۵۱

جدول ۵- نتیجه الگوریتم پیشنهادی پس از تغییر ورودی

در حالت کلی با بررسی نتایج می‌بینیم که روش پیشنهادی، بهبودی بر روش تنک گروهی می‌باشد و با تغییر در فرآیند حل مسئله، بهبودهایی از نظر کارایی و سرعت الگوریتم داده‌ایم.

در این قسمت به بررسی روش پیشنهادی بر روی یک سند پرداخته‌ایم. نتایج حاصل از روش خلاصه‌سازی پیشنهادی و روش تنک گروهی با ایجاد تغییر در سند را با خلاصه‌های تولیدی دستی(مرجع)، مقایسه نموده و نتایج آن‌ها را نمایش دادیم.

۳,۳ . نتایج کلی

برای اینکه از عملکرد بهتر روش خلاصه‌سازی پیشنهادی اطمینان حاصل شود، در این قسمت می‌خواهیم نتایج دقت سیستم پیشنهادی با روش تنک گروهی که بر روی مجموعه داده‌های خبری BBC و مجموعه داده DUC2002 اجرا نموده ایم، مقایسه و نتایج آن را نشان دهیم.

- مجموعه داده‌های خبری BBC

این مجموعه داده خلاصه استخراجی اخبار BBC از سال ۲۰۰۴ تا ۲۰۰۵ می‌باشد که شامل ۲۲۲۵ سند در ۵ حوزه خبری تجاری ، سرگرمی، سیاسی، ورزشی و فناوری هست.

- مجموعه داده DUC2002

مجموعه داده‌ای DUC تاکنون ۷ مجموعه از داده‌ها را با عنوان‌های DUC2001 تا DUC2007 ارائه کرده است که هر کدام از آن‌ها با هدف استفاده در امور خاص و برای کمک در ارزیابی روش‌های خلاصه‌سازی خودکار متن و بررسی آن‌ها انتشار یافته‌اند.

مجموعه داده‌ی مورد استفاده در این مقاله DUC2002 می‌باشد که شامل ۱۱۵ سند است.

ابتدا در جدول ۶ نتایج خلاصه‌های تولیدی با منابع مشخص شده را با نتایج خلاصه‌های سیستم پیشنهادی تولید کرده است، مقایسه نموده‌ایم.

مقدار میانگین F-Measure برای مجموعه داده‌های خبری BBC		
زمینه‌های خبری	**روش پیشنهادی درصد**	**تنک گروهی در صد**
تجاری	۴۷/۲۵	۴۶/۵۶
سرگرمی	۵۹/۸۲	۵۷/۴۷
سیاسی	۵۸/۱۶	۵۷/۸۱
ورزشی	۵۸/۴۹	۵۷/۴۴
فناوری	۵۷/۴۱	۵۴/۲۴

جدول ۶- جدول مقایسه نتایج روش پیشنهادی با روش تنک گروهی در مجموعه داده‌هایBBC

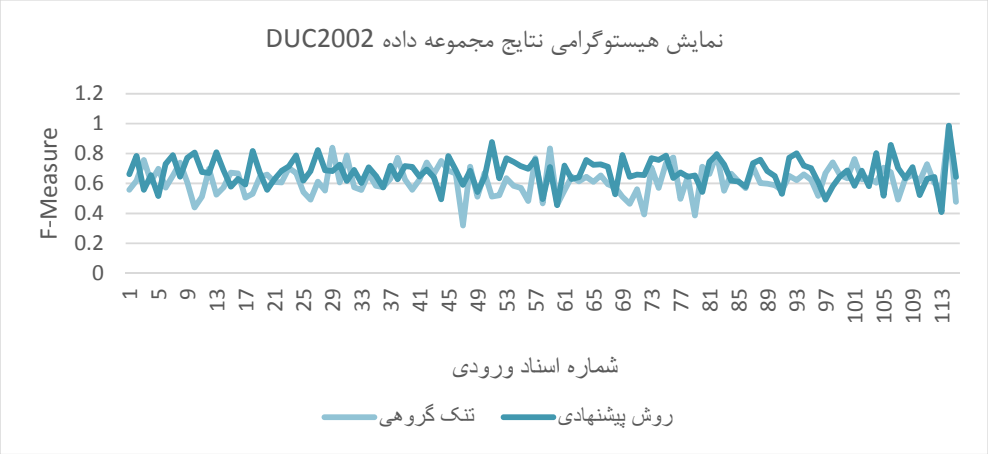
همان‌طور که مشخص است، سیستم پیشنهادی در تولید خلاصه با میانگین دقت بالاتری در همه حوزه‌های خبری عملکرد بهتری دارد.

در جدول ۷ نتایج خلاصه‌های تولیدی با ورودی مشخص شده را با نتایج خلاصه‌های سیستم پیشنهادی تولید کرده است، نشان می‌هیم.

نتایج مجموعه داده ۱۱۵ سندی DUC2002		
F-Measure	**روش پیشنهادی درصد**	**تنک گروهی درصد**
بیشترین مقدار	۸۷/۶۲	۸۵/۵۳
کمترین مقدار	۴۵/۴۷	۳۱/۰۱
میانگین	۶۷/۴۱	۶۲/۱۲
میانه	۶۸/۱۶	۶۱/۹۴

جدول ۷- جدول مقایسه نتایج روش پیشنهادی با روش تنک گروهی در مجموعه داده‌های DUC2002

برای نمایش بهتر نتایج نمودارهای هیستوگرامی (شکل ۱۱) برای نتایج به دست آمده با سیستم پیشنهادی و تنک گروهی که برای مقایسه با هم انتخاب نموده‌ایم، رسم کرده‌ایم.



شکل ۱۱- نمایش هیستوگرامی نتایج مجموعه داده DUC2002

در شکل ۱۱ محور عمودی میزان دقت در معیار F-Measure و محور افقی شماره سند ورودی در مجموعه ۱۱۵ سندی DUC2002 را نشان می‌دهد. نمودار هیستوگرامی روش پیشنهادی با رنگ آبی پرنگ و روش تنک گروهی با رنگ آبی کم رنگ نشان داده شده است. همانطور که مشاهده می‌کنید نمودار روش پیشنهادی در اکثر سندها بالاتر از نمودار روش تنک گروهی قرار گرفته است که نشان می‌دهد در اکثر سندهای ورودی دقت خلاصه تولید شده به روش پیشنهادی بیشتر از روش تنک گروهی است.

همانطور که نشان داده شد، روش پیشنهادی در این مجموعه داده‌ها نیز در تولید خلاصه با میانگین دقت بالا، عملکرد بهتری دارد.

درحالت کلی، هر دو این روش‌ها برای یافتن خلاصه جملات سعی بر حل معادله زیر دارند:

$$\min\|S - SW\|_1^2 + \lambda\|W\|_1 \qquad (۲۵)$$

در نتیجه هردو این روش‌ها از مزایای رویکرد تنک بهره‌مند هستند. در روش تنک گروهی به دلیل دیدگاه ماتریسی به انتخاب ستون‌های ماتریس S ستون‌های به‌دست آمده معمولا دارای بیشترین کلمه‌های کلیدی هستند تا بتواند بیشترین بازسازی کل ماتریس S را داشته باشند و در صورتی که الگوریتم مجبور به انتخاب از بین دو ستون با یک میزان کلمه کلیدی شود، معمولا ستونی را انتخاب می‌کند که بیشترین میزان بازسازی را داشته باشد. بنابراین جمله‌ای انتخاب می‌شود که در کلمات کلیدی با سایر جملات اشتراک بیشتری داشته باشد.

از طرفی با بررسی خلاصه‌های دستی به این نتیجه رسیدیم که جملات شرکت‌کننده در خلاصه‌های دستی تنها جملات با بیشترین کلمه‌های کلیدی نیستند. گاهی جملاتی با کلمات کلیدی کم‌تر نیز اهمیت بالایی داشته و در خلاصه‌دستی شرکت

می‌کنند. ویژگی اکثر این جملات به این صورت است که سعی بر آن دارند که تمام کلمات کلیدی را در درون جملات انتخابی داشته باشند. در واقع اگر جمله‌ای کلمه‌ی کلیدی خاصی را داشته باشد و تعداد کلمه‌های کلیدی آن کم باشد، کاندید قوی ای برای انتخاب در جملات انتخابی می‌باشد.

در روش پیشنهادی، به دلیل بررسی ستون‌های S به صورت جداگانه و بررسی میزان بازسازی ماتریس توسط آن، معمولا جملاتی انتخاب می‌شوند که اولا تعداد بالایی از جملات را بازسازی کنند، هم‌چنین برخی جملات که از اهمیت بالایی برخوردارند ولی نسبت به برخی از کلمات نمی‌توانند خوب بازسازی شوند، در خلاصه ظاهر شوند. چرا که روش پیشنهادی از این کلمات در بازسازی چشم‌پوشی می‌کند.

References

1. A. Aker et al. Multi-document summarization techniques for generating image descriptions: A comparative analysis. Multi-source, Multilingual Information Extraction and Summarization. Springer, Berlin, Heidelberg, (2013**)**. 299-320.

2. T. Hosseinikhah, A. Ahmadi, and A. Mohebi, A New Persian Text Summarization Approach Based on Natural Language Processing and Graph Similarity, Iranian Journal of Information Processing and Management, **92** (2018), 885-914.

3. S. M. Weiss, N. Indurkhya, and T. Zhang. Fundamentals of predictive text mining. Springer, 2015.

4. S. J. Ker., and J. N. Chen. A Text Categorization Based on a Summarization Extraction, ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval. 2000**.**

5. R. He, J. Tang, P. Gong, Q. Hu, and,B, Wang, Multi-document summarization via group sparse learning, Information Sciences, **349** (2016), 12-24.

6. L. Eldén, Matrix methods in data mining and pattern recognition. Society for Industrial and Applied Mathematics, 2007.

7. C. Y. Lin, , and J. Franz, Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), 2004.

8. Reeve, Lawrence H., Hyoil Han, and Ari D. Brooks. The use of domain-specific concepts in biomedical text summarization, Information Processing & Management, **43**.6 (2007), 1765-1776.

9. R. M. Alguliyev, M. A. Ramiz, COSUM: Text summarization based on clustering and optimization, Expert Systems, **36**.1 (2019), e12340.

10. X. Wen, L. Shao, Y. Xue, and, W. Fang, A rapid learning algorithm for vehicle classification, Information sciences, **295** (2015), 395-406.

11. Gu, Bin, V. S. Sheng, K.Y. Tay, W. Romano, and, S. Li, Incremental support vector learning for ordinal regression, IEEE Transactions on Neural networks and learning systems, **26**.7 (2014), 1403-1416.

12. W. Song, J. Z. Liang, and S. C. Park, Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering, Information Sciences, **273** (2014), 156-170.

13. O. Rouane, H. Belhadef, and M. Bouakkaz, Combine clustering and frequent itemsets mining to enhance biomedical text summarization, Expert Systems with Applications, **135** (2019), 362-373.

14. M. Dey, and D. Dipankar, A Deep Dive into Supervised Extractive and Abstractive Summarization from Text, Data Visualization and Knowledge Engineering. Springer, Cham, (2020), 109-132.

15. W. Kryściński, N. S. Keskar, and B. McCann, Neural text summarization: A critical evaluation. arXiv preprint arXiv:1908.08960 (2019).

16. U. Khandelwal, K. Clark, D. Jurafsky, and L. kaiser, Sample efficient text summarization using a single pre-trained transformer, arXiv preprint arXiv:1905.08836 (2019).

17. J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Perez, Parallelizing a multi-objective optimization approach for extractive multi-document text summarization, Journal of Parallel and Distributed Computing, **134** (2019), 166-179.

18. H. T. Fan, J. W. Hung, X. Lu, S. S. Wang, and Y. Tsao, Speech enhancement using segmental nonnegative matrix factorization 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014**.**

19. M. G. Ozsoy, F. N. Alpaslan, Text summarization using Latent Semantic Analysis, Journal of Information Science, **37** (2011), 405-417.

20. N. Bhatiaand A. Jaiswal, Automatic text summarization and it's methods-a review, 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence). IEEE, 2016.

21. P. Unnikrishnan, V. K. Govindan, and S. D. M . Kumar. Enhanced sparse representation classifier for text classification, Expert Systems with Applications, **129** (2019), 260-272.

22. Björck, Åke. Numerical methods in matrix computations. Vol. 59. Cham: Springer, 2015.

23. P. Parvasideh, and M. Rezghi. A novel dictionary learning method based on total least squares approach with application in high dimensional biological data, Advances in Data Analysis and Classification, **15** (2021), 575-597.

24 S. Mukherjee, B. Rupam, and S.S. Chandra Sekhar, ℓ1-K-SVD: A robust dictionary learning algorithm with simultaneous update, Signal Processing, **123** (2016), 42-52.