الگوریتمی مبتنی بر گراف برای خوشه‌بندی سوره‌های قرآن کریم

دکتر بهروز مینایی[1]، مریم سادات متّقی[2*]
[1]دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران، [2]پژوهشکده‌ی اعجاز قرآن دانشگاه شهید بهشتی

چکیده

قرآن کتاب نازل‌شده از طرف خداست و تا به امروز اندیشمندان و پژوهش‌گران مختلفی در جهت شناخت قرآن و فهم آن تلاش نموده‌اند. در دسترس بودن سیستم‌های رایانه‌ای فرصت مغتنمی است که با افزایش سرعت پژوهش‌گران در پیمودن مسیر، آن‌ها را در رسیدن به قله‌های بلندتری یاری کند. خوشه‌بندی یکی از روش‌هایی است که برای فهم ساختار داده به کار می‌رود. در این مقاله به خوشه‌بندی سوره‌های قرآن کریم بر اساس هم‌وقوعی کلمات در آن پرداخته و برای دست‌یابی به این هدف از یک رویکرد موجود مبتنی بر گراف استفاده نموده‌ایم. در پژوهش جاری ابتدا هر سوره را به صورت یک گراف غیرجهت‌دار و وزن‌دار بازنمایی کرده، سپس بردار هر سوره را بر اساس گراف سوره تشکیل داده‌ایم و پس از آن سوره‌ها را خوشه‌بندی نموده‌ایم. برای ارزیابی کیفیت خوشه‌بندی از معیار نیم‌رخ استفاده کرده‌ایم. بر اساس این معیار در بهترین خوشه‌بندی در بین اجراهای مختلف مقدار نیم‌رخ ۰/۹۱ به دست‌آمده است. این پژوهش زیرساخت ساختاری مناسبی برای توصیف لایه معنایی سوره‌ها و آیات قرآن پیش روی پژوهش‌گران حوزه زبان‌شناسی محاسباتی در دامنه علوم قرآنی فراهم می‌سازد.

کلیدواژه‌ها: خوشه‌بندی متن، بازنمایی شبکه‌ای متن، گراف متن، زیرگراف پرتکرار، قرآن‌کاوی رایانشی

A Graph-based Algorithm for Clustering Qur’anic Surahs

Behrouz Minaei-Bidgoli[1], Maryam Sadat Mottaghi[2*]
[1]School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran,
[2]Qur’an Miracle Research Institute, Shahid Beheshti University, Tehran, Iran

Abstract

The Holy Qur'an is revealed from God Almighty. Up to now many scholars and researchers have tried to understand the Holy Qur'an and comprehend it. The availability of computer systems is a great opportunity to help researchers reach higher peaks by speeding them up in their way. Clustering is one of the methods has been used to understand the structure of the data. In clustering, we want to divide samples of data into groups so that the members of each cluster are similar together and are different from the members of the other clusters. Clustering of Qur'anic surahs has been the subject of some computer studies on the Qur'an. In these studies, different approaches have been considered to vectorizing the surahs. In a study, *Thabet* formed vectors of each surah by considering some stems of Qur'anic words as features and the normalized probability of their occurrences in the surah as feature values and clustered just 24 surahs due to the sparseness of the obtained data matrix. With a similar approach in vectorizing the surahs, *Moisl* calculated the minimum surah length threshold per feature in order to solve the problem of shorter surahs by using some concepts of statistical sampling theory, and could cluster more surahs. Instead of using words as features, *Sharaf* considered 13 features including existence of referring to the story of *Adam* and *Ebliys*, number of the phrase «يا أَيُّهَا الَّذينَ آمَنُوا» (O you who believe), and determined the method of measuring each feature. Then, he formed data matrix and clustered the Qur'anic surahs. In another study, *Sufi* et al. considered the topics identified for each verse in the Tafsir Rahnama as features and constructed a binary data matrix based on the presence or absence of that topic in the Tafsir of that surah and applied clustering. In this article, we have clustered the surahs of the Holy Qur'an based on the co-occurrence of words in it. To achieve this goal, we have used an existing graph-based approach. In the present study, we first represent each surah as a weighted undirected graph. Then we form the vector of each surah by considering closed frequent sub-graphs as features and relative occurrence of them in each surah as feature values, and eventually cluster the surahs. We used the Silhouette score to evaluate the quality of clustering. Based on this criterion, in the best clustering among different runs, the Silhouette score of 0.91 was obtained. This

---

1* نویسنده‌ی عهده‌دار مکاتبات

research provide a proper structural infrastructure for specifying the semantic layer of Holy Qur'an surahs for computational linguistics researchers in the domain of Qur'anic studies.

Keywords: Document Clustering, Text Graph, Frequent subgraph, Computational Qur'an mining

۱- مقدمه:

۱-۱- تعریف مسأله

قرآن، کتاب نازل‌شده از طرف خداست و همه‌ی انسان‌ها مخاطب آن هستند. منافع حاصل از شناخت قرآن و فهم عمیق معارف بلند آن می‌تواند دامنه‌ی وسیعی داشته باشد. ابزارهای رایانه‌ای در این مسیر امکان حرکت سریع‌تر و دقیق‌تر پژوهش‌گر قرآنی را فراهم می‌کند. برای مثال ممکن است او را در یافتن شواهد درستی یا نادرستی نظریه‌های موجود یا ساختن فرضیات جدید کمک کند یا او را در ارائه و انتقال مفاهیم موجود یاری نماید. به‌خصوص با گسترش اسلام در سطح جهان مخاطبان قرآن با زبان‌ها، فرهنگ‌ها و نظام فکری و تجربیات متفاوت از دنیا با آن آشنا شده و نیاز به توسعه‌ی کمی و کیفی مسیر شناخت قرآن در آینده بیش‌تر احساس خواهد شد. آشنایی با قابلیت‌های موجود و تعریف مسائل برآمده از نیازهای خاص این حوزه توسط جامعه‌ی قرآنی می‌تواند منجر به جهت‌دهی، افزایش کیفیت و سرعت رشد این‌گونه پژوهش‌ها شود.

امروزه انواع مختلف داده‌ها با حجم زیادی در دسترس است. مطالعه‌ی این داده‌ها و کشف الگوهای درون آن‌ها می‌تواند قدرت شناخت و پیش‌بینی متخصصان را در مسائل مختلف بالا ببرد. یکی از انواع داده‌ای متن و متن‌کاوی یکی از شاخه‌های داده‌کاوی است که در کاربردهای مختلفی مانند خلاصه‌سازی متن، استخراج اطلاعات، تشخیص نظر و غیره و در حوزه‌های متفاوتی چون شبکه‌های اجتماعی، زیست‌شناسی، مدیریت و غیره استفاده شده است. یکی از این کاربردها خوشه‌بندی متن است.

در خوشه‌بندی داده‌ها به گروه‌هایی تقسیم می‌شوند که مفید یا معنادار و یا مفید و معنادار هستند؛ با این هدف که نمونه‌ها در یک خوشه به هم شبیه و با سایر خوشه‌ها متمایز باشند [۳۰]. هرچه اعضای درون خوشه‌ها به هم شبیه‌تر و اعضای بین خوشه‌ها از هم متفاوت‌تر باشند، نتیجه‌ی خوشه‌بندی بهتر است. از خوشه‌بندی می‌توان برای فهم داده استفاده کرد؛ در این کاربرد خوشه‌ها ساختار طبیعی داده را به خود می‌گیرند [۳۰]. خوشه‌بندی برای متون می‌تواند در سطوح گوناگون مانند حرف، کلمه، جمله، پاراگراف و متن انجام بگیرد.

قرآن به صورت متنی در اختیار بشر قرار گرفته‌است. متن قرآن که کلام خداوند رحمان رحیم، حکیم و علیم و در نتیجه مهم‌ترین کتاب است، موضوع پژوهش‌های مختلفی بوده و انگیزه‌ی شناخت بیش‌تر و عمیق‌تر قرآن و امید کشف افق‌های جدید در اعجاز قرآن مثلاً کشف نظمی خاص در آن وجود دارد. متن قرآن از ۱۱۴ سوره با اندازه‌های مختلف تشکیل شده است. ظاهر کلام خدا در قرآن شبیه یک متن عادی که مفاهیم مختلف با رعایت ترتیب معمول در کنار هم آمده باشند، نیست. بنابراین با توجه به سبک بیان مطالب و اهمیت خود قرآن نیاز به درک ارتباطات بین مفاهیم مطرح‌شده بیش‌تر از متون دیگر احساس می‌شود. این ارتباط هم از نظر ارتباط درون آیات، هم ارتباط بین آیات و هم ارتباط بین سوره‌ها دارای اهمیت است.

در این پژوهش می‌خواهیم سوره‌ها را خوشه‌بندی کنیم تا بتوانیم روابط درون خوشه‌ها و بینشان را تحلیل نماییم. اگر سوره‌ها هم در درون گروه‌های به دست‌آمده با هم تناسب معنایی یا مکانی یا وجه تناسب دیگری جز مبنای اولیه‌ی خوشه‌بندی داشته باشند و هم ارتباط گروه‌ها سازگار با یک هدف کلی باشد، نتیجه جالب توجه است؛ حتی اگر به گونه‌ای باشد که خوشه‌بندی‌های صحیح مختلفی وجود داشته باشد. در ادامه برخی مفاهیم مورد نیاز برای تعریف مسأله آورده شده است.

۱-۱-۱- فضای ویژگی و کاهش بعد

یک روش توصیف داده‌ها این است که ویژگی‌های مهم برای مسئله‌ی مورد نظر را تعیین کرده و هر نمونه‌ی داده را با مقادیری که برای هر کدام از ویژگی‌های تعیین‌شده دارند، بازنمایی کنیم. در این روش هر نمونه به صورت برداری در یک فضای چند بعدی است که به تعداد ویژگی‌های تعیین‌شده بعد دارد. در متن‌کاوی به جهت بازنمایی برداری متن، ویژگی‌ها بسته به کاربرد مورد نظر از روش‌های مختلفی مانند کیسه‌ی کلمات[1]، چندتایی[2] و موضوعات تعیین می‌شوند.

گاهی تعداد ویژگی‌های مورد استفاده در توصیف برداری داده‌ها بسیار زیاد است. الگوریتم‌هایی مانند **PCA** وجود دارند که با استفاده از آن‌ها می‌توان تعداد ویژگی‌ها را کاهش داد. با این روش فضای ویژگی‌ها به فضای دیگری با تعداد ویژگی کمتر نگاشت می‌شود. در این پژوهش هر سوره به صورت‌برداری از زیرگراف‌های پرتکرار بسته در مجموعه‌ی گراف متنی سوره‌ها در نظر گرفته شده است. در این‌جا ماتریس داده حاوی مقادیری در بازه‌ی صفر تا یک است که هر سطر آن اطلاعات یک نمونه از داده‌ها و هر ستون آن اطلاعات یک ویژگی را برای هر نمونه نشان می‌دهد. هر خانه‌ی این ماتریس، تعداد نسبی ویژگی متناظر با ستون مربوطه در نمونه‌ی متناظر با سطر مربوطه را نمایش می‌دهد.

۱-۱-۲- خوشه‌بندی[3]

در دسته‌ای از مسائل تشخیص الگو می‌خواهیم در مجموعه بردارهای داده شده‌ی $\mathbf{x}$، گروه‌های متشکل از نمونه‌های مشابه درون داده را پیدا کنیم. این‌گونه مسائل که به روش بی‌ناظر[4](بدون توجه به یک گروه‌بندی از پیش تعیین‌شده) انجام می‌گیرد، خوشه‌بندی نامیده می‌شود [۱۰]. به عبارت دیگر در خوشه‌بندی، نمونه‌های داده بر اساس اطلاعات موجود در خود داده به گروه‌هایی تقسیم می‌شود؛ با این هدف که نمونه‌های درون گروه به هم شبیه (یا مرتبط) و متفاوت(یا غیر مرتبط) با نمونه‌های دیگر گروه‌ها باشد. هرچه این شباهت درون‌گروهی و تفاوت بین گروهی بیشتر باشد، خوشه‌بندی بهتر و متمایزکننده‌تر است [۳۰]. بنابراین هر نوع گروه‌بندی خوشه‌بندی محسوب نمی‌شود.

خوشه‌بندی سلسله مراتبی روشی است که در آن خوشه‌بندی تنها بر مبنای معیار شباهت(یا فاصله)، بدون نیاز به اطلاعات دیگری از داده انجام می‌شود[5] [۸]. این معیار

1 Bag-Of-Words
2 N-gram
3 Clustering
4 Unsupervised
۵ در مقابل مثلاً در برخی روش‌ها به دنبال کمینه‌سازی یک تابع خطا هستیم.

می‌تواند بسته به مسئله متفاوت باشد. برای مثال می‌توان از معیارهای فاصله‌ی اقلیدسی[1]، جاکارد[2] و غیره استفاده نمود. انتخاب معیار شباهت در نتیجه‌ی خوشه‌بندی تأثیر قابل توجهی دارد [۲۷].

در خوشه‌بندی سلسله مراتبی تجمعی ابتدا هر نمونه در یک خوشه‌ی جداگانه قرار دارد. در هر مرحله دو خوشه انتخاب شده و با هم ادغام می‌شوند. بنابراین علاوه بر معیار شباهت بین نمونه‌ها باید یک معیار شباهت هم برای ادغام دو خوشه انتخاب شود. برای مثال ممکن است برای تعیین شباهت بین دو خوشه، میانگین(روش پیوند میانگین[3])، بیشینه(روش پیوند کامل[4]) یا کمینه‌ی شباهت(روش تک پیوند[5]) بین دو به دوی نمونه‌ها محاسبه شده و دو خوشه‌ای برای ادغام انتخاب شوند که شباهت بینشان بیشینه باشد. و یا در روش دیگری(روش وارد[6]) خوشه‌ها بر اساس مقدار بهینه‌ی یک تابع هدف ادغام شوند. ادغام خوشه‌ها تا رسیدن به یک خوشه‌ی واحد ادامه می‌یابد. حاصل خوشه‌بندی سلسله مراتبی معمولاً به صورت یک نمودار درختی[7] ارائه می‌شود. می‌توان برای انتخاب یکی از خوشه‌بندی‌های به‌دست‌آمده، درخت حاصل را در یک سطح مناسب برش زده و خوشه‌بندی با تعداد خوشه‌های آن سطح به عنوان نتیجه درنظر گرفته شود.

در این پژوهش مسئله، خوشه‌بندی سوره‌های قرآن است. در خوشه‌بندی سلسله مراتبی، تغییر در معیار شباهت بین نمونه‌ها و شباهت بین خوشه‌ها می‌تواند منجر به نتایج متفاوتی شود.

۱-۱-۳- گراف

گراف بدون جهت[8] **G** به صورت زوج مرتب **(V,E)** تعریف می‌شود که در آن **V** یک مجموعه‌ی متناهی و **E** یک مجموعه از زیرمجموعه‌های دو عنصری از **V** است [۴]. مجموعه‌ی **E** را مجموعه‌ی رئوس[9] یا گره‌ها[10] و مجموعه‌ی **V** را مجموعه‌ی یال‌ها[11] یا پیوندها[12] می‌نامند. یک گراف را می‌توان به صورت مجموعه‌ای از نقاط و خطوط متصل‌کننده‌ی آن‌ها نمایش داد که در آن هر نقطه نماینده‌ی یک عضو **V** و هر خط نماینده‌ی یک یال از **E** است. مسائل مختلفی را می‌توان به صورت گراف بازنمایی نمود. برخی مسائل با گراف بازنمایی می‌شوند، سپس با استفاده از خواص یا الگوریتم‌های مربوط به گراف‌ها مورد مطالعه قرار می‌گیرند. از گراف در بازنمایی شبکه‌های اجتماعی [۳۳]، زیستی [۲۲]، حمل و نقل [۱۵]، ساختار روابط خانوادگی [۲] و ... استفاده شده است.

- زیرگراف: برای گراف داده‌شده‌ی $G_1=(V_1,E_1)$، گراف $G_2=(V_2,E_2)$ را زیرگراف آن گوییم هرگاه V_1 زیرمجموعه‌ی V_2 و E_1 زیرمجموعه‌ی E_2 باشد، به گونه‌ای که یال‌های موجود در مجموعه‌ی E_2 تنها با رئوس موجود در V_2 حادث باشند [۴]. همچنین G_1 را گراف بالادستی[13] G_2 می‌نامیم [۲۹].
- مجموعه‌ی زیرگراف‌های پر تکرار: برای مجموعه‌ی داده‌شده‌ی متشکل از چند گراف $\{G_1, G_2, \ldots, G_n\} = D$، پشتیبان زیرگراف **g**، تعداد گراف‌هایی در **D** را نشان می‌دهد که **g** در آن‌ها وجود دارد. به هر زیرگرافی که مقدار پشتیبان بزرگ‌تر یا مساوی با حد تعیین‌شده‌ی کمینه‌ی پشتیبان دارد، زیرگراف پرتکرار گویند [۳۲].
- زیرگراف پرتکرار بسته: به زیرگراف پرتکراری که در مجموعه‌ی زیرگراف‌های پرتکرار، هیچ گراف بالادستی با مقدار پشتیبان برابر مقدار پشتیبان خودش برایش وجود نداشته باشد، زیرگراف پرتکرار بسته گویند [۲۹].

در این پژوهش برای بازنمایی متن هر سوره از یک گراف استفاده شده است. زیرگراف‌های پرتکرار بسته هم به عنوان ویژگی‌ها در فضای برداری در نظر گرفته شده‌اند.

۱-۲- پیشینه‌ی پژوهش

پژوهش‌گران در مقالاتی به حل مسائل مورد نیاز پژوهش‌های قرآنی با استفاده از ابزارهای رایانه‌ای پرداخته‌اند. برخی از آن‌ها در راستای نیاز به ارائه‌ی خدمات در بازنمایاندن دانش موجود به وجود آمده‌اند؛ مانند برخی نرم‌افزارهای قرآنی و سیستم‌های پرسش و پاسخ قرآنی [۶]، [۱۴]، [۱۶]، [۱۸]. همچنین استفاده از ابزارهای رایانه‌ای گاه به صورت بسترسازی برای پژوهش‌های آینده بوده است؛ مانند تلاش‌هایی که در راستای ساخت پیکره‌های قرآنی انجام شده است [۱۱]، [۱۳]، [۱۹]، [۲۵]، [۲۶]. برخی از تحقیقات هم در راستای فهم و شناخت بهتر قرآن می‌باشد [۱]، [۳]، [۱۷]، [۲۰]، [۲۸]،[۳۱]. در این نوع پژوهش‌ها با استخراج خودکار یا نیمه‌خودکار دانش از قرآن، فرصت تحلیل و فرضیه‌سازی در اختیار پژوهشگران قرآنی قرار داده می‌شود. می‌توانیم با استفاده از ابزارهای رایانه‌ای دانش را استخراج کنیم، اگر حاصل در تفسیر و پژوهش افراد متخصص بود، می‌تواند توضیحی بر بیان متخصص باشد و اگر نبود، ممکن است دانش جدیدی کشف شده باشد. برخی پژوهش‌گران مسلمان و غیر مسلمان در سال ۲۰۱۰ مسئله‌ی فهم قرآن را به عنوان چالش جدیدی برای علم کامپیوتر و هوش مصنوعی مطرح کرده‌اند [۹].

در پژوهش‌های صورت گرفته در راستای فهم و شناخت قرآن موارد مختلفی برای مطالعه انتخاب شده‌اند. در بعضی پژوهش‌ها کلمه، آیه، سوره یا ارتباط بینشان مورد مطالعه قرار گرفته‌است. در ادامه پژوهش‌هایی مطرح می‌شوند که سوره‌های قرآن را خوشه‌بندی یا رده‌بندی کرده‌اند.

برای استخراج دانش می‌توان از خوشه‌بندی استفاده کرد. *تبت* [۳۱] در مقاله‌ای به خوشه‌بندی سوره‌های قرآن پرداخته است. در این پژوهش ابتدا کلمات یک‌بار تکرار و کلمات دستوری مثل حروف تعریف و اضافه حذف شده و کلمات به بن کلمه[14] تبدیل می‌شوند. در این حالت کلمه‌های دارای یک بن، یک کلمه محسوب می‌شوند. سپس هر کلمه به عنوان یک ویژگی و فرکانس به‌هنجارشده‌ی[15] وقوع هر کلمه در یک سوره بر اساس طول سوره‌ها به عنوان مقادیر ویژگی‌ها در نظر گرفته شده، هر سوره به شکل برداری بازنمایی می‌شود. در این مرحله ماتریس داده‌ها با ۱۱۴ سطر و ۳۶۷۲ کلمه به‌دست‌می‌آید. از آن‌جا که طول (تعداد کلمات) سوره‌های قرآن بسیار باهم متفاوت است، ماتریس داده‌ها خلوت بوده یا به عبارت دیگر تعداد عناصر صفر آن زیاد می‌باشد. برای رفع این مشکل به‌جای ۱۱۴ سوره تنها

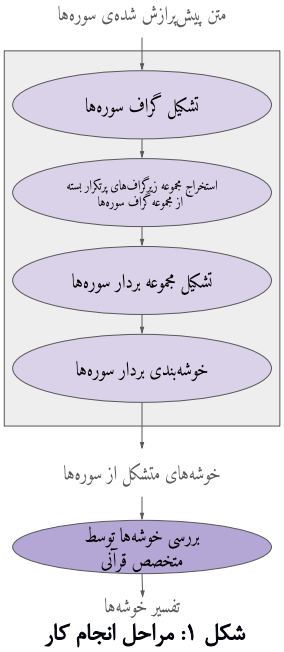
1 Euclidean
2 Jaccard
3 Average method
4 Complete link method
5 Single link method
6 Ward's method
7 Dendrogram
8 Undirected
9 Vertex
10 Node
11 Edge
12 Link
13 Super-graph
14 Stemming
15 Normalized

۲۴ سوره که به نسبت طولانی‌ترند نگه‌داشته شده و باقی سوره‌ها کنار گذاشته می‌شوند. همچنین براساس نمودار واریانس هر کلمه، تنها ۵۰۰ ویژگی نگه داشته‌می‌شود. در مرحله‌ی بعد خوشه‌بندی سلسله‌مراتبی با معیار فاصله‌ی اقلیدسی[1] و وارد بر روی ماتریس داده‌ی حاصل اعمال می‌گردد. نتیجه‌ی گزارش‌شده شامل دو خوشه‌ی اصلی **A** و **B** می‌باشد که **A** خود از دو خوشه‌ی **C** و **D** تشکیل شده است. تمام سوره‌های خوشه‌ی **A** به جز نحل و زمر، مدنی و همه‌ی سوره‌های خوشه‌ی **B** مکی هستند. برای مقایسه‌ی خوشه‌ها، از بردار میانگین نمونه‌های هر خوشه استفاده شده است. بدین منظور در ابتدا ۵۰ ویژگی با بیشترین واریانس درنظر گرفته شده و ویژگی با بیشترین فرکانس یعنی کلمه‌ی «الله» از بین آن‌ها حذف شده است. نویسنده با ارائه‌ی نمودار مقایسه‌ای بردارهای میانگین در دو خوشه‌ی **A** و **B**، با درنظرگرفتن ۴۹ ویژگی باقیمانده نتیجه گرفته که کلمات «قال»، «قل»، «آیة» و «آیات» در خوشه‌ی **B** بیشتر از **A** تکرار شده‌اند. همچنین تکرار کلمات «مؤمن»، «امن» و «إتّق» در خوشه‌ی **A** بیشتر از **B** بوده است. به علاوه از نمودارمشابه برای خوشه‌ی **C** و **D** نتیجه گرفته شده که در سوره‌های خوشه‌ی **C** به مخاطب قرار دادن پیامبر اسلام برای ارائه‌ی شواهد پیامشان و در سوره‌های خوشه‌ی **D** به خطاب مومنان درمورد پاداش عمل صالحشان بیشتر پرداخته‌شده است.

مویزل [۲۰] ، در مقاله‌ای راهی برای حل مشکلی که تبت [۳۱] در مقابل سوره‌های کوچک داشت، ارائه داد. او ابتدا مانند تبت [۳۱]، ماتریس داده‌ای ۱۱۴ سطری و ۳۶۷۲ ستونی تشکیل داد. سپس مقادیر هر سطر (متناظر با هر سوره) را با تقسیم بر مجموع مقادیر آن سطر به‌هنجار نمود. مویزل به کمک مفاهیمی از نظریه‌ی نمونه‌برداری آماری، حداقل طول مورد نیاز برای نمونه‌ها به ازای ویژگی‌ها را محاسبه کرده و به صورت جدولی نمایش داد. برای مثال با این محاسبات، حداکثر تعداد سوره‌ای که می‌توان در این شکل از بازنمایی مسأله خوشه‌بندی کرد، برابر ۸۷ سوره و تنها یک ویژگی است؛ زیرا کمترین طول نمونه‌ی مورد نیاز در جدول مربوطه ۵۶ بوده، ۸۷ سوره طول بیشتر مساوی ۵۶ دارند و به جز ویژگی شماره‌ی یک، بقیه‌ی ویژگی‌ها به طول بیشتری احتیاج دارند. حد آستانه‌ی کمینه‌ی طول سوره‌ها بسته به هدف پژوهش تعیین می‌شود. او برای پژوهش خود حد آستانه‌ی کمینه‌ی طول ۳۰۰ را انتخاب کرد و بر مبنای آن برای بازنمایی بردار ۴۷ سوره‌ی انتخابی از ۹ ویژگی «الله»، «لا»، «رب»، «قال»، «کان»، «یوم»، «ناس»، «یومئذ» و «شر» استفاده نمود. سپس خوشه‌بندی سلسله مراتبی را بر ماتریس داده‌ای با ابعاد ۹* ۴۷ اعمال نمود. اگرچه خوشه‌بندی سلسله‌مراتبی با معیار شباهت اقلیدسی برای نمونه‌ها در سه حالت معیار شباهت بین خوشه‌ای پیوند میانگین، وارد و پیوند تکین انجام شده است، اما فقط نتایج معیار وارد، بدون برش در مقاله گزارش شده است.

صدیق و همکاران [۲۸] در پژوهشی با استفاده از مدلسازی موضوع، تلاش کرده‌اند تا موضوعات سوره‌ها را به صورت خودکار به کمک آمار استخراج نموده و درنتیجه سوره‌های قرآن را بر این اساس تقسیم‌بندی کرده‌اند. در مدلسازی موضوع فرض می‌شود هر متن موضوعاتی دارد و هر موضوع از کلماتی تشکیل شده است. آن‌ها برای برداری کردن سوره‌ها روشی مانند تبت [۳۱] را در پیش گرفته‌اند. نویسندگان ابتدا در مرحله‌ی پیش‌پردازش، متن قرآن را با استفاده از یک تشخیص‌دهنده‌ی مرز کلمه برای زبان عربی قطعه‌بندی[2] کرده و پس از به‌هنجارسازی و حذف ایست‌واژه‌های[3] عمومی و نیز خاص پیکره و پر تکرارترین کلمه یعنی «الله»، کلماتی را که در کمتر از ۱۰٪ سوره‌ها حضور دارند را نیز حذف می‌کنند. در این مرحله با حضور سوره‌های کوتاه، ماتریس خلوت تشخیص داده شده و بنابراین مانند تبت [۳۱] تنها ۲۴ سوره نگه داشته می‌شود. از آن‌جا که در ماتریس داده‌ی باقی‌مانده هنوز هم با بیش از ۱۰۰۰۰ کلمه و ۲۴ سوره، تعداد صفرها زیاد است، کلماتی که در کمتر از ۲۵٪ سوره‌ها بوده‌اند از ماتریس حذف شده و تنها ۴۱۷ کلمه باقی می‌ماند. پس از آن روش **LDA**[4] در سه حالت ۲، ۵ و ۱۰ موضوع برای ماتریس داده‌ها اجرا می‌شود. در نتیجه به ازای هر حالت، برای هر سوره موضوعی تعیین شده‌است. البته طبیعتاً هر متن ممکن است به بیش از یک موضوع ارتباط داشته باشد، اما آن‌چه گزارش می‌شود، مرتبط‌ترین موضوع است. نویسندگان برای نتایج حالت دو موضوعی، تطبیق با رده‌بندی مکی- مدنی را گزارش کرده‌اند. همچنین ۱۵ کلمه‌ی اول هر موضوع نیز در جدولی نمایش داده شده است. در هر دو حالت ۵ و ۱۰ موضوعی یک موضوع وجود دارد که کلمات آن گفتگوی بین حضرت موسی (علیه‌السلام) و فرعون راجع به دعوت و یک موضوع پاسخ مومنان و غیر مومنان به دعوت را نشان می‌دهد.

به جای رویکرد آماری تنها، می‌توان معنا را نیز به بازنمایی سوره‌ها اضافه کرد. شرف [۲۱] در فصلی از پایان‌نامه‌ی خود به رده‌بندی سوره‌های مکی-مدنی پرداخته است. او ۱۳ ویژگی از جمله وجود اشاره به داستان آدم و ابلیس، داشتن حروف مقطعه، تکرار عبارت «یا أَیُّهَا الَّذینَ آمَنُوا» و ... را در نظر گرفت و روشی برای اندازه‌گیری هریک تعیین نمود. سپس بردار هر سوره را تشکیل داد. در مرحله‌ی بعد رده‌بند **J48** را با ۹۳ سوره‌ای که در فهرست رده‌بندی مرجعش در مکی-مدنی بودنشان توافق نظر وجود دارد، آموزش داده تا ۲۱ سوره‌ی اختلافی آن فهرست را در دو دسته‌ی مکی یا مدنی قرار دهد. پس از آن با به حساب



شکل ۱: مراحل انجام کار

Figure 1: The research workflow

1 Euclidean

2 Tokenization

3 Stop word

4 Latent Drichelete Allocation

آوردن مرجع ضمائر از **QuranA** [۲۵]، تعداد شمارش‌ها در مقادیر ویژگی‌ها تغییر یافت و رده‌بندی دوباره اجرا شد. این‌بار پیش‌بینی برای سوره‌ی شماره‌ی ۱۳ تغییر کرده و نتیجه نسبت به قبل از ۶ مورد مغایر با فهرست در ۲۱ سوره به ۵ مورد کاهش یافت. او یک‌بار هم بردار سوره‌ها را با استفاده از روش **EM**[1] خوشه‌بندی نمود. از ۷ خوشه‌ی به‌دست‌آمده، در ۴ خوشه، سوره‌های مکی-مدنی مطابق فهرست مرجع برای ۹۳ سوره، از هم جدا شده‌اند. همچنین نویسنده در فصل دیگری از پایان‌نامه‌ی خود ابزاری به نام **Qursim** [۲۶] ارائه کرده‌است که می‌تواند بر اساس ارجاع متقابل بین آیات سوره‌ها گراف ارتباط معنایی بین سوره‌ها را به دست آورد.

صوفی و همکاران [۳]، در مقاله‌ای سوره‌های قرآن را بر اساس موضوع خوشه‌بندی کرده‌اند. آن‌ها با استفاده از موضوعات مشخص‌شده برای هر آیه در تفسیر راهنما [۵]، جدول موضوعات-آیات را به صورت یک ماتریس دودویی ۶۲۳۶*۱۶۶۲ تشکیل داده و سپس با تجمیع سطرهای مربوط به آیات هر سوره ماتریس سوره‌ها- موضوعات را ساخته‌اند. در مرحله‌ی بعد خوشه‌بندی سلسله مراتبی بر اساس معیار جاکارد و وارد روی ماتریس داده‌ها اجرا شده است. معیار فاصله‌ی جاکارد بین دو سوره نسبت موضوعات مشترک بین آن دو به کل موضوعات مطرح‌شده در دو سوره را نشان می‌دهد. از درخت حاصل سطح دلخواهی برای برش انتخاب شده و خوشه‌بندی حاصل مورد بررسی قرار گرفته است. نتیجه شامل ۷ خوشه می‌باشد که از نظر تقسیم‌بندی مکی- مدنی مرجع مقاله، یک خوشه کاملاً یک‌دست و ۵ خوشه دارای تنها یک سوره‌ی متفاوت از نظر مکی-مدنی بوده است.

آکتاس و *آکباس* [۷] نیز در پژوهشی به رده‌بندی مکی-مدنی سوره‌های قرآن پرداخته‌اند. آن‌ها پس از پیش‌پردازش سوره‌ها، هر یک از آن‌ها را به صورت یک شبکه بازنمایی کرده‌اند. در شبکه‌ی سوره هر کلمه به صورت یک گره نمایش داده شده و چنانچه هر دو گره‌ای در سوره در یک فاصله‌ی تعیین‌شده با هم آمده باشند، بین آن دو یالی برقرار می‌شود. وزن هر گره تعداد تکرار آن کلمه در سوره و وزن هر یال تعداد هم‌وقوعی دو کلمه‌ی دوسر یال را نشان می‌دهد. ماتریس داده‌ها در این پژوهش با درنظرگرفتن کلمات به عنوان ویژگی‌ها ساخته می‌شود و مقادیر ویژگی‌ها برای هر سوره با استفاده از مقدار **DFF**[2] هر کلمه در گراف سوره محاسبه می‌گردد. سپس کاهش بعد بر ماتریس داده‌ها اجرا شده و رده‌بند با تعدادی از سوره‌ها، آموزش داده می‌شود. پس از آن مکی یا مدنی بودن سوره‌های باقیمانده با استفاده از رده‌بند آموزش دیده تشخیص داده می‌شود. نتیجه با اندازه پنجره‌های ۱ تا ۳، (همان فاصله‌ی تعیین‌شده) بیش از ۹۸٪ با فهرست مرجع مطابقت داشته و نسبت به روش چندتایی ۵٪ بهتر بوده است.

یکی از چالش‌های استفاده از خوشه‌بندی این است که انتخاب متفاوت انواع الگوریتم‌ها، روش‌ها و پارامترها برای یک نوع بازنمایی برداری منجر به خوشه‌بندی‌های مختلفی می‌شود. تبت [۳۱] در پژوهش خود اشاره کرده بود که انتخاب ترکیب دیگری به جای معیار فاصله‌ی اقلیدسی و وارد در خوشه‌بندی ممکن است به نتایج دیگری منجر شود. مویزل [۲۰] نیز در مقاله‌ی خود ترکیبات اقلیدسی با سه معیار تک پیوند، پیوند کامل و پیوند میانگین را علاوه بر وارد انجام داد و تفاوت اما اشتراک گسترده بین نتایج را گزارش نمود. به علاوه با استفاده از یک حد آستانه‌ی طول متفاوت با مقدار تعیین‌شده در پژوهش مویزل [۲۰] نتایج دیگری می‌تواند حاصل شود. همچنین بازنمایی برداری مسأله نیز در نتیجه مؤثر است. پیش‌فرض‌های تعریف مسأله هم گاه باعث بروز خطاست. در پژوهش شرف [۲۱] بیان شده که نتیجه ممکن است به دلایلی از جمله وجود آیه‌های مکی در سوره‌های مدنی و طبیعت برخی سوره‌ها دارای خطا باشد. یکتا نبودن نتایج برای متن مهم و مقدس قرآن بازدارنده‌ی آزمودن چنین روش‌هایی نیست؛ بلکه باید در تحلیل نتایج، انتخاب‌های مختلف انجام‌شده در مسأله را درنظر گرفت و به احتمال بروز خطا نیز توجه داشت.

۱-۳- اهمیت طرح

در این پژوهش با به کارگیری یک رویکرد مبتنی بر گراف به خوشه‌بندی سوره‌های قرآن پرداخته‌ایم. در خوشه‌بندی، نتایج به دست‌آمده ساختار داده را توصیف می‌کند، بعد با مطالعه‌ی نتایج می‌توان برای فهم دلیل حصول چنین نتیجه‌ای تلاش کرد. می‌توان گفت در این‌جا خوشه‌بندی آزمون یک فرض جزئی و مشخص نیست بلکه ارائه‌ی توصیفی از داده بر مبنای روش به کارگرفته شده می‌باشد. یافتن وجه ارتباط بین سوره‌های هر خوشه و نیز فهمیدن دلیل تفکیک این چنینی سوره‌ها ممکن است به تذکر یا تأیید دانش قبلی یا توجه به نکته‌های کم‌تر پرداخته‌شده و یا ساختن یک فرضیه یا دیدگاه جدید بیانجامد. این کار شبیه این است که با استفاده از تلسکوپ سیارات و منظومه‌ها را در حد امکانات ابزارمان مشاهده نماییم، سپس به تحلیل و تفسیر و تطبیق مشاهدات با دانش قبلی بپردازیم و حتی تلاش کنیم که از روی نشانه‌های مشاهده شده به آن‌چه وجود دارد، اما مستقیماً مشاهده نشده پی ببریم.

بازنمایی به کارگرفته‌شده در این پژوهش به ماتریس داده‌ی خلوت منجر نمی‌شود و امکان خوشه‌بندی تمام سوره‌ها وجود داشته است. همچنین ویژگی‌های در نظرگرفته‌شده بر گرفته از ساختار خود قرآن و نه فقط دانش و سلیقه‌ی متخصص بوده؛ هرچند نظر متخصص در تعیین فراپارامترها مؤثر بوده است. به علاوه معیار برش درخت سلسله مراتبی به دست‌آمده بر اساس یک معیار کیفیت خوشه‌بندی انجام شده و نتایج گزارش شده‌اند.

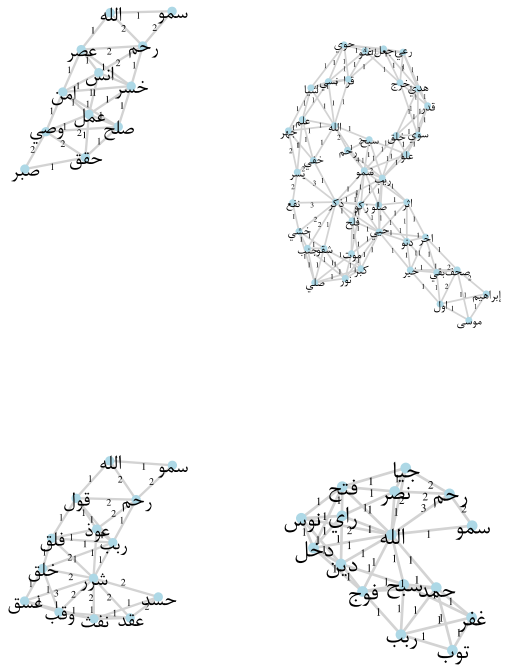
در بخش ۱-۱-۲ اشاره شد که خوشه‌بندی برای فهم ساختار داده نیز به کار می‌رود. خوشه‌های به دست‌آمده از سوره‌ها در این پژوهش می‌تواند برای تحلیل به متخصص قرآنی سپرده شود. روش پیشنهادی تحلیل این است که متخصص برای هر خوشه با توجه به اعضایش وجه اشتراکشان را بیابد و در نهایت اساس غیر ظاهری خوشه‌بندی به دست‌آمده را تشخیص دهد.

۱-۴- جنبه نوآوری بحث

برای بازنمایی بردار یک متن روش‌های مختلفی استفاده شده‌اند. *روسیو* و همکاران در مقاله‌ای [۲۳] از یک رویکرد مبتنی بر گراف برای رده‌بندی متون استفاده کرده‌اند. آن‌ها برای هر متن پیش‌پردازش‌شده از مجموعه‌ی متون، گرافی تشکیل می‌دهند که گره‌های آن واحد متن بوده و بین هر دو واحد متنی که در یک فاصله‌ی تعیین‌شده در متن آمده باشند، یالی وجود دارد. در مرحله‌ی بعد زیرگراف‌های پرتکرار در مجموعه‌ی گراف‌ها استخراج شده و به عنوان ویژگی در بازنمایی برداری متن استفاده می‌شوند. برای هر متن یک بردار دودویی تولید می‌شود که هر درایه از آن در صورت وجود ویژگی مربوطه در متن مربوطه برابر یک خواهد بود. در مرحله‌ی آخر الگوریتم **SVM** بر بردارهای متن اعمال می‌گردد تا هر متن را به یک دسته‌ی از پیش مشخص‌شده منتسب کند. نویسندگان در این مقاله برای کاهش تعداد زیرگراف‌های پرتکرار از تبدیل هر گراف متن به گراف هسته‌ی اصلی استفاده نموده‌اند.

1 Expectation Maximization
2 Diffusion Frèchet Function

در پژوهش جاری نیز رویکرد گرافی مشابهی را برای بازنمایی برداری متن به کار برده‌ایم. در این مقاله به جای زیرگراف‌های پرتکرار، از زیرگراف‌های پرتکرار بسته استفاده می‌شود. تعداد ویژگی‌ها با استفاده از زیرگراف‌های پرتکرار بسته نسبت به زیرگراف‌های پرتکرار به طور کلی کمتر بوده یا مساوی با آن است و در مورد قرآن بسیار کمتر می‌شود. خوشه‌بندی با این رویکرد برای سوره‌های قرآن قبلاً انجام نشده است. در ادامه روش مورد استفاده توضیح داده شده و نتایج حاصل گزارش شده‌اند.



شکل ۲: گراف سوره‌های «اعلیٰ»، «عصر»، «نصر» و«فلق» با استفاده از پنجره‌ای به طول ۴، به ترتیب از راست به چپ و بالا به پایین.

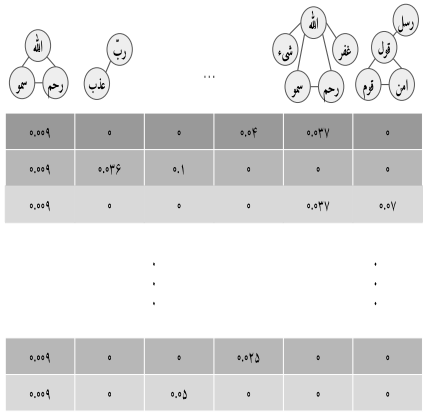
Figure 2: Graphs of surahs "Al-A'la", "Al-'Asr", "An-Nasr" and "Al-Falaq",from right to left and top to bottom, respectively, obtained using a window of size 4.

۲- روش پیشنهادی

شکل ۱ مراحل انجام کار را نشان می‌دهد. همان‌گونه که در شکل ۱ پیداست، در این پژوهش ابتدا از روی متن پیش‌پردازش‌شده‌ی قرآن، برای هر سوره گرافی ساخته می‌شود. در مرحله‌ی بعد بردار هر سوره با توجه به گراف آن تشکیل می‌گردد و پس از آن خوشه‌بندی بر بردار سوره‌ها اعمال می‌شود. خوشه‌های حاصل متشکل از سوره‌های قرآن می‌تواند توسط متخصص قرآنی تحلیل شود و تفسیر خوشه ها به عنوان خروجی به دست آید. در این پژوهش تنها مراحل داخل کادر مستطیلی پیاده‌سازی شده و مرحله‌ی بررسی خوشه‌ها توسط متخصص قرآنی انجام نشده‌است. در ادامه‌ی این بخش مراحل انجام کار با جزئیات بیشتر توضیح داده خواهد شد.

۲-۱- تشکیل گراف

در ابتدا با لغزاندن یک پنجره روی متن پیش‌پردازش‌شده‌ی هر سوره گرافی تشکیل می‌شود. در هر یک از این گراف‌ها هر واحد متن یک گره است و بین هر دو واحد متنی که درون یک پنجره قرار می‌گیرند یالی برقرار می‌شود. وزن هر یال تعداد هم‌وقوعی واحدهای متنی دو سر آن در یک پنجره را در آن سوره نشان می‌دهد. بنابراین گراف‌های ساخته‌شده برای هر سوره بدون جهت و وزن‌دار هستند. در این شیوه از تشکیل گراف، برچسب هر گره یکتاست؛ به عبارت دیگر هر واحد متنی تنها با یک گره در گراف متناظر است. در عین حال برچسب گره متناظر با یک واحد متنی در گراف‌های مختلف یکسان است. همچنین به دلیل استفاده از پنجره، گراف هر سوره هم‌بند خواهد بود. زیرگراف‌های پرتکراری که در این پژوهش استخراج خواهند شد نیز همبند هستند. در این پژوهش یال بازگشتی(یال از یک گره به خودش) درنظر گرفته نشده است. ملاک تعیین پرتکرار بودن یک زیرگراف، تعداد سوره‌ای است که آن زیرگراف در گراف متناظر با آن‌ها وجود دارد. در تعیین پرتکراری، یک زیرگراف حداکثر یک‌بار در هر سوره شمرده می‌شود، اما در برداری کردن سوره‌ها تعداد تکرار یک زیرگراف به صورت کمینه‌ی وزن یال‌های آن زیرگراف در گراف یک سوره در نظر گرفته می‌شود.



۰	۰.۰۳۷	۰.۰۴	۰	۰	۰.۰۵۹
۰	۰	۰	۰.۱	۰.۰۳۶	۰.۰۵۹
۰.۰۷	۰.۰۳۷	۰	۰	۰	۰.۰۵۹
.	.			.	
.	.			.	
.	.			.	
۰	۰	۰.۰۴۵	۰	۰	۰.۰۵۹
۰	۰	۰	۰.۰۵	۰	۰.۰۵۹

شکل ۳: نمونه‌ی ماتریس داده. این ماتریس از نظر ساختار و روش تشکیل، منطبق با روش برداری کردن سوره‌هاست اما اعداد درون ماتریس مثالی بوده و ممکن است در بردار واقعی سوره‌ها متفاوت باشد.

Figure 3: An Example of data matrix. This matrix, in terms of structure and method of formation, is consistent with the method of surah vectorization, but the numbers inside the matrix are exemplary and may differ in the actual vector of the surahs.

۲-۲- بازنمایی برداری

در مجموعه‌ی گراف‌های ساخته‌شده، زیرگراف‌هایی هستند که در گراف تعداد زیادی از سوره‌ها حضور دارند. نحوه‌ی تشکیل گراف سوره به گونه‌ای است که اگر دو کلمه‌ی «الف» و «ب» باهم در جایی از یک سوره در فاصله‌ی کمی قرار داشته باشند و در جای دیگری دو کلمه‌ی «ب» و «ج»، آن‌گاه بین «الف» و «ب» و نیز بین «ب» و «ج» یالی در گراف وجود دارد. حال فرض کنید هم‌وقوعی «الف» و «ب» و نیز «ب» و «ج» در جاهای مختلفی از تعداد زیادی از سوره‌های دیگر هم تکرار شده و در گراف سوره‌ها به همراه تعدادی یال دیگر زیرگراف پرتکراری تشکیل‌دهند؛ هر کدام از زیرگراف‌های پرتکرار در مجموعه‌ی سوره‌ها ممکن است نماینده‌ی یک ویژگی لفظی، معنایی یا حتی یک اشاره‌ی رمزی در قرآن باشد. در این‌جا هر سوره را با تعداد تکرار نسبی زیرمجموعه‌ای از زیرگراف‌های پرتکرار در آن به عنوان ویژگی‌هایش، توصیف می‌نماییم. به این منظور حد آستانه‌ای برای پرتکرار محسوب‌شدن یک زیرگراف(یا به

عبارتی کمینه‌پشتیبان که به اختصار در ادامه آن را **msp**[1] می‌نامیم) تعیین شده و زیرگراف‌های پرتکرار در مجموعه‌ی سوره‌ها استخراج می‌گردد. چنانچه یک سوره یک زیرگراف پرتکرار مانند «د» را داشته باشد، همه‌ی زیرگراف‌های آن زیرگراف را هم دارد که اگر در مجموعه‌ی سوره‌ها تعداد حضور آن‌ها با تعداد حضور «د» برابر باشد، یعنی آن‌ها هیچ جا مستقل از «د» نیامده‌اند و با دانستن مقدار ویژگی «د» در بردار هر سوره مقادیر ویژگی‌های متناظر با زیرگراف‌های «د» را نیز می‌دانیم، در نتیجه ویژگی‌های اضافی هستند. بنابراین زیرگراف‌های اضافی از مجموعه‌ی زیرگراف‌های پرتکرار، حذف می‌شوند. اعضای مجموعه‌ی باقیمانده یا همان زیرگراف‌های پرتکرار بسته که به اختصار **CFSG**[2] نامیده می‌شوند، ویژگی‌های سازنده‌ی بردار سوره‌ها خواهند بود. با استفاده از زیرگراف‌های پرتکرار بسته به جای همه‌ی زیرگراف‌های پرتکرار، تعداد ویژگی‌ها کاهش قابل توجهی پیدا می‌کند.

شکل ۲ نمونه‌ی گراف چند سوره را نشان می‌دهد. همان‌گونه که در شکل ۲ مشاهده می‌شود، به خاطر استفاده از پنجره در هر گراف، بین هر جفت گره حداقل یک مسیر وجود دارد. کلمه‌ی «اسم» و «ربّ» در اولین آیه از سوره‌ی « اعلیٰ» با کلمه‌ی «سبّح» و در پانزدهمین آیه با کلمه‌ی «فذکّر» در یک پنجره آمده‌اند که در نتیجه‌ی آن، گره‌های متناظر با ریشه‌های «سمو»، «ربب»، «سبح» و «ذکر» با هم در گراف آن سوره، یک زیرگراف تشکیل داده‌اند. زیرگراف مثلثی «سمو»، «الله» و «رحم» به خاطر وجود آیه‌ی شریفه‌ی «بسم الله الرّحمٰن الرّحیم» در گراف ۱۱۳ سوره وجود دارد. همچنین زیرگراف با دو یال متصل‌کننده‌ی«سمو» و «الله» و نیز «الله» و «رحم» در همه‌ی ۱۱۴ سوره تکرار شده و پرتکرارترین زیرگراف محسوب می‌شود. چنانچه تکرار دو زیرگراف مذکور با هم برابر بود، تنها زیرگراف بالادستی یعنی زیرگراف اولی در مجموعه‌ی زیرگراف‌های پرتکرار بسته قرار می‌گرفت و دیگری حذف می‌شد.

بعد از به دست‌آمدن مجموعه‌ی زیرگراف‌های پرتکرار بسته، ماتریس سوره‌ها تشکیل می‌گرد. شکل ۳ یک نمونه ماتریس سوره‌ها را نشان می‌دهد. هر سطر از این ماتریس، بردار متناظر با یک سوره را نمایش می‌دهد. هر ستون متناظر با یک **CFSG** است که در شکل بالای آن ستون نشان داده شده است. مقدار هر ویژگی مطابق رابطه(۱) محاسبه می‌گردد. A ماتریس سوره‌هاست و A_{ij} مقدار سطر i از ستون j را نشان می‌دهد. W_m وزن یال m و E_{ij} یال‌های زیرگراف j در گراف سوره‌ی i می‌باشد. در رابطه(۱)، صورت کسر نماینده‌ی تعداد تکرار **CFSG** متناظر با ستون j در گراف سوره i و مخرج کسر تعداد تکرار آن **CFSG** در بین همه‌ی سوره‌هاست. تعداد تکرار یک زیرگراف در هر گراف، برابر با کمینه وزن یال‌هایش در آن گراف درنظر گرفته می‌شود. با این روش مجموع هر ستون از این ماتریس برابر با یک خواهد بود.

$$A_{ij}=\frac{\min\limits_{m\in E_{ij}} W_m}{\sum\limits_{k=1}^{n} \min\limits_{m\in V_{kj}} W_m} \qquad (۱)$$

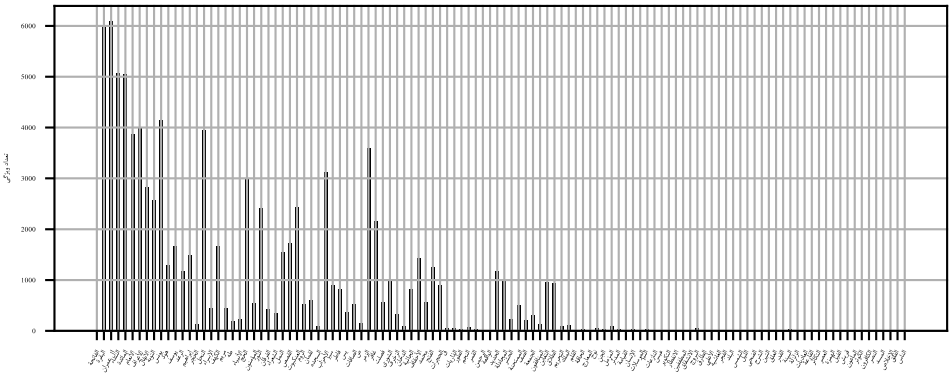
از آن‌جا که ویژگی‌ها زیاد(بسته به حد آستانه‌ی پرتکراربودن یک زیرگراف، در این پژوهش تا چند هزار ویژگی)، و دارای اشتراکات فراوان هستند، الگوریتم خوشه‌بندی یک‌بار هم پس از اعمال یک روش کاهش بعد روی ماتریس داده اجرا می‌شود.

۲-۳- خوشه‌بندی

در مرحله‌ی پایانی خوشه‌بندی بردار سوره‌ها به صورت ناهم‌پوشان[3] انجام می‌شود؛ یعنی در نهایت هر سوره‌ای که به این مرحله راه‌یافته به یک و فقط یک خوشه منتسب می‌گردد. چون طول سوره‌ها بسیار متفاوت است، برای مقایسه‌ی نمونه‌های کاهش بعد نیافته معیار شباهتی تعریف شده که نسبت تعداد **CFSG** های موجود مشترک در دو سوره به حداکثر تعداد **CFSG** های موجود مشترک ممکن بین دو سوره را نشان می‌دهد:

$$Similarity(a,b)=(a.b)/min(a_1,b_1) \qquad (۲)$$

در رابطه(۲)، **a** و **b** بردار تبدیل‌شده‌ی دو سوره است که در آن به جای مقادیر بیش‌تر از صفر بردار اصلی سوره مقدار یک گذاشته شده است. عملگر . عمل ضرب نقطه‌ای را انجام می‌دهد و حاصل آن برای دو بردار دودویی ورودی، مجموع تعداد یک‌های مشترک در هر دو می‌باشد. همچنین $\mathbf{a_1}$,$\mathbf{b_1}$ به ترتیب تعداد یک‌های بردارهای **a** و **b** را نشان می‌دهند.



شکل ۴: تعداد زیرگراف‌های پرتکرار بسته(CFSG) در بردار هر سوره برای msp= ۱۱, ws= ۲ . محور افقی شماره‌ی هر سوره و محور عمودی تعداد CFSG را نشان می‌دهد.

Figure 4: Number of closed frequently sub-graphs (CFSG) in the vector of each surah for msp = 11, ws = 2. The horizontal and vertical axes show each surah number and number of CFSGs, respectively.

1 Minimum Support
2 Closed Frequent Sub-Graph

۳ هر نمونه تنها به یک خوشه منتسب خواهد شد.

۳- آزمایش‌های تجربی

۳-۱- داده‌ها

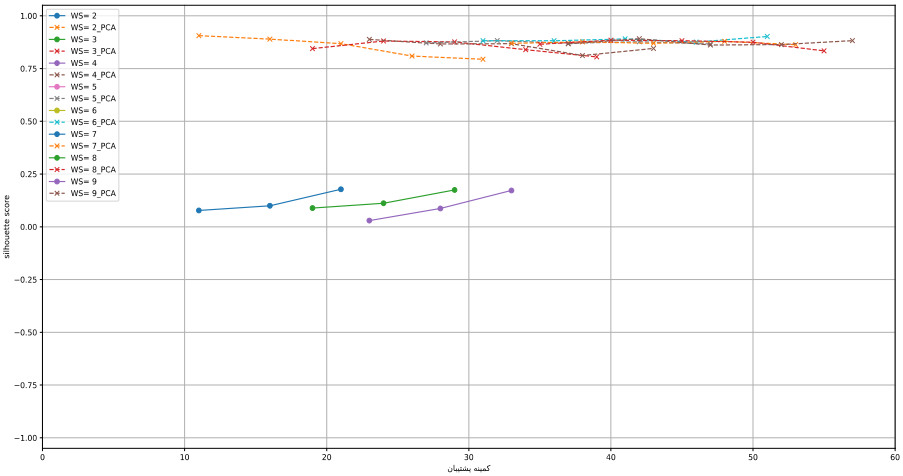
متن پیش‌پردازش‌شده‌ی ورودی سوره‌ها بر اساس یک پیکره‌ی موجود[۱]، ریشه‌ی هر کلمه[۲] (در صورت وجود) است. همچنین برخی کلمات فاقد ریشه مانند اسامی خاص[۳] نیز به آن اضافه شده‌اند. در پیکره مطلوب بود کلماتی چون «کانَ» و «حیث» با توجه به نحوه‌ی تشکیل گراف و استفاده از هم‌وقوعی در یک پنجره، به دلیل عمومی بودن، از متن سوره‌ها حذف شوند. بنابراین «کانَ» و خانواده‌اش[۴]، «حیث»، «أنَىٰ»، «کُل» و «کیف» از داده‌ی ورودی حذف شدند[۵،۶]. لازم به ذکر است که آیه‌ی شریفه‌ی «بسم الله الرّحمٰن الرّحیم» از ابتدای سوره‌ها حذف نشده است.

با توجه به متن ورودی در گراف هر سوره همه‌ی کلمات هم‌ریشه تنها یک گره متناظر با ریشه‌ی مشترک دارند. در این نوع بازنمایی، کلمات هم‌ریشه با هم مرتبط در نظر گرفته می‌شوند. این ارتباط در مورد کلماتی مانند اسم و تنها زیرگراف‌های پرتکرار مورد استفاده قرار خواهند گرفت، تکرار کم هم‌وقوعی با ریشه‌های دیگر منجر به این می‌شود که در برخی موارد ریشه‌ی مشترک نماینده‌ی برخی و نه همه‌ی کلمات هم‌ریشه باشد. جایگزینی ریشه‌ی کلمات به جای خود کلمات علاوه بر کاهش تعداد گره‌ها در گراف هر سوره باعث می‌شود تعداد وقوع زیرگراف‌ها افزایش یابد.

۳-۲- اجرا

کد برنامه[۷] به زبان پایتون نوشته شده و بر روی کامپیوتری با ۱۲ **Gb** حافظه اجرا شده است. زیرگراف‌های پرتکرار با استفاده از نرم‌افزار **gSpan** به دست می‌آید. حداقل تعداد یال مجاز برای یک زیرگراف پرتکرار برابر با یک درنظر گرفته شده است. همچنین برای کاهش بعد کتاب‌خانه‌ی **scikit-learn** به کار رفته است.

برای خوشه‌بندی سلسله‌مراتبی نیز از کتاب‌خانه‌ی **scipy** استفاده شده است. معیار شباهت بین خوشه‌ای از روش «پیوند میانگین» محاسبه می‌شود؛ یعنی برای بررسی



شکل ۵: تأثیر مقدار کمینه‌پشتیبان بر مقدار معیار ارزیابی کیفیت خوشه‌بندی؛ نیم‌رخ(silhouette).

Figure 5: The effect of minimum support value on the result of clustering quality evaluation measure; silhouette score.

سماوات که ظاهراً معنای مشترکی ندارند در حد حروف اصلی مشترک و در مورد کلماتی که صورت‌های صرفی یک ریشه‌اند به صورت اشتراک معنایی وجود دارد. از آن‌جا که

۱ نسخه‌ی ۰/۴ پیکره‌ی عربی قرآنی (دوکس، ۲۰۱۱) قابل دسترسی از سایت corpus.quran.com

۲ حروف اصلی سازنده‌ی کلمه که معمولا سه یا چهار حرف است؛ مثل سمو و علم که به ترتیب ریشه‌ی اسم و عالم محسوب می‌شوند.

۳ اسامی خاص مشخص‌شده در پیکره با اندکی اصلاحات چه ریشه‌دار چه فاقد ریشه به صورت جداگانه در نظر گرفته‌شدند.

۴ أصبحَ / أضحی / ظلّ / باتَ / أمسی / مازالَ / مابَرِحَ / ما انفکَّ / مافَتِئَ / مادامَ / صارَ / لیسَ

۵ مواردی از هم‌ریشه‌ها با معنای غیرعام، حذف نشدند. برای مثال "کَوَنَ" آن‌جا که کلمه‌ی متناظرش قید مکان بود، حذف نشد. همچنین ریشه‌ی "کلل" در "کلالة" و کَلّ" به معنای سربار حذف نشد.

۶ با وجود این حذف و اضافات هنوز هم داده‌ی ورودی دارای کلماتی مثل "عند" بوده و ممکن است لازم باشد در آینده از پیکره حذف شوند. البته از آن‌جا که پرتکراری مانند صافی، واحدهای متنی اضافی را حذف می‌کند، در مورد واحدهای کم‌تکرارتر ممکن است تلاش برای حذفشان صرفه‌ی زمانی نداشته باشد.

خوب‌بودن ادغام یک خوشه‌ی تشکیل‌شده با دیگری در یک مرحله، مقدار میانگین فواصل دو به دوی اعضای هر خوشه با اعضای خوشه‌ی دیگر محاسبه می‌گردد.[۸]

برای تعیین کیفیت خوشه‌های به‌دست‌آمده از معیار نیم‌رخ[۹] استفاده شده است. مقدار این معیار بین ۱- و ۱ بوده و هرچه مقدار معیار نیم‌رخ به یک نزدیک‌تر باشد، مطلوب‌تر است. نیم‌رخ برای هر نمونه از رابطه (۳) محاسبه می‌شود. مقدار معیار نیم‌رخ برای خوشه‌بندی برابر میانگین مقدار نیم‌رخ تمامی نمونه‌هاست [۲۴].

$$S(i)=\frac{b(i)-a(i)}{max(a(i),b(i))} \qquad (۳)$$

۷ کد از آدرس https://github.com/mmtIr/Clustering-Quranic-Surahs قابل دسترسی است.

۸ همچنین روش یافتن خوشه‌ها برابر maxclust تعیین شده است. در این روش حد آستانه‌ای پیدا می‌کنیم که هیچ دو نمونه‌ی اصلی که با هم در یک خوشه قرار گرفته‌اند فاصله‌ای بیشتر از آن نداشته باشند و در عین حال تعداد خوشه‌ها بیشتر از یک تعداد مورد نظر تعیین‌شده نشود.

9 Silhouette score

در رابطه (۳)، **a(i)** و **b(i)** به ترتیب برای نمونه‌ی **i** ام که به یک خوشه منتسب شده است، میانگین فاصله درون خوشه و میانگین فاصله تا نزدیک‌ترین خوشه را نشان می‌دهند. در خوشه‌بندی سلسله‌مراتبی، ادغام تا رسیدن به یک خوشه انجام می‌شود. برای تعیین سطح برش در اجراهای با اعمال کاهش بعد، مقدار نیم‌رخ برای خوشه‌بندی‌های به دست‌آمده بین ۲ تا ۲۰ خوشه محاسبه می‌گردد؛ سپس بهترین آن به عنوان نتیجه در نظر گرفته می‌شود. برای اجراهای بدون کاهش بعد، تمامی نتایج ۲ خوشه‌ای بوده و همان گزارش شده است[1]. لازم به ذکر است در پیاده‌سازی از هر مجموعه بردار یکسان پیش از خوشه‌بندی تنها یکی نگه‌داشته شده و بقیه حذف شده‌اند و در محاسبه‌ی معیار نیم‌رخ نیز درنظر گرفته نشده‌اند.

جدول ۱: بهترین نتایج به دست‌آمده از نظر معیار نیم‌رخ برای هر اندازه پنجره

Table 1: The best results according to silhouette score for each window size.

تعداد خوشه	مقدار معیار نیم‌رخ (sil)	تعداد بعد پس از کاهش (f)	تعداد بعد پیش از کاهش	مقدار کمینه‌پشتیبان (msp)	اندازه پنجره (ws)
8	0.91	2	7023	11	2
14	0.88	2	1116	24	3
9	0.89	2	9733	23	4
4	0.88	2	1301	32	5
2	0.90	2	28	51	6
5	0.88	2	75	48	7
2	0.88	2	1114	40	8
2	0.89	2	1217	42	9

۳-۳- گزارش و تحلیل نتایج

با توجه به وابستگی مسئله به پارامترهای اندازه‌ی کمینه‌پشتیبان(**msp**)، تعداد بعد پس از کاهش(**f**[2])، اندازه‌ی پنجره(**ws**[3])ـ و معیار شباهت استفاده‌شده، روش برای هر اندازه‌پنجره و معیار شباهت بین نمونه‌ای تعیین‌شده و با مقادیرمتفاوت کمینه‌ی پشتیبان، با اعمال کاهش بعد به تعداد مختلف و بدون آن انجام شده‌است. در این بخش ابتدا نتایج به صورت کلی و مقایسه‌ای گزارش شده و سپس به بررسی بهترین نتایج به دست‌آمده پرداخته‌شده است.

شکل ۴، به عنوان نمونه تعداد **CFSG** در بردار هر سوره در اجرای با کمینه‌پشتیبان ۱۱ و اندازه پنجره‌ی ۲ را نشان می‌دهد. بدیهی است تعداد **CFSG** ممکن در هر سوره به طول سوره بستگی دارد. البته لزوماً بیشتر بودن تعداد واحد متن موجود در متن پیش‌پردازش‌شده بین دو سوره به معنی تعداد **CFSG** بیشتر حاضر در بردار سوره نیست؛ برای مثال سوره‌ی «الحجرات» دارای ۱۸ واحد متن کمتر از سوره‌ی «ق» است، اما در حدود ۸۰۰ تا **CFSG** بیشتر از آن دارد. چنین اختلافی ممکن است مثلا به این دلیل باشد که در سوره با تعداد واحد متنی بیشتر، تنوع هم‌وقوعی واحدهای متنی به نسبت، بیشتر بوده، بنابراین یال‌های بیشتری وزن کمتری داشته و تعداد **CFSG** کمتری دارد.

شکل ۵، نمودار مقایسه‌ای مقادیر معیارنیم‌رخ به دست‌آمده در اجراهای مختلف با اندازه پنجره‌های متفاوت و با تغییر ۵ واحد، ۵ واحد کمینه‌پشتیبان را نشان می‌دهد. برای بردارهای به دست‌آمده، در اجراهای بدون کاهش بعد معیار شباهت رابطه(۲) و در اجراهای با کاهش بعد، معیار شباهت کسینوسی استفاده شده است. در تمامی نتایج به‌دست‌آمده بدون اعمال کاهش بعد، همه‌ی سوره‌ها در یک خوشه و سوره‌ی توبه در خوشه‌ی دیگری قرار می‌گیرد. نمودارهای مربوط به این اجراها در شکل ۵، با خط پر نمایش داده‌شده‌اند. همان‌گونه که مشاهده می‌شود، با افزایش کمینه‌پشتیبان به اجراهایی می‌رسیم که با پارامترهای تنظیم‌شده، بردارها همگی به یک خوشه منتسب شده‌اند و در نتیجه نقطه‌ای برای نمایندگی ندارند.

مطابق شکل ۵، بدون استفاده از کاهش بعد، برای اندازه‌پنجره‌های ۲و ۳ و ۴ با افزایش کمینه پشتیبان که به کاهش تعداد ویژگی بردارها منجر می‌شود، مقدار معیار نیم‌رخ افزایش می‌یابد.

نمودارهای نقطه‌چین‌شده مربوط به اجراهای با اعمال کاهش بعد می‌باشند. هر نقطه از نمودارهای نقطه‌چین‌شده، مربوط به بهترین مقدار نیم‌رخ در بین اجراهای با اندازه‌ی پنجره‌ی مربوطه و تعداد ابعاد کاهش‌یافته‌ی مختلف است. اجراهای با اندازه پنجره‌های بیشتر با توجه به محدودیت حافظه از اندازه پشتیبان‌های بالاتری شروع می‌شوند. در نمودارهای نقطه‌چین از یک کمینه‌پشتیبان به بعد، به دلیل این‌که تعداد ویژگی بردارها کم(کمتر از ۲۰) است، دیگر کاهش بعد اعمال نشده و نمودار متوقف می‌شود. تغییر مقدار پشتیبان، کمتر از ۰.۲ در میزان معیار نیم‌رخ نتایج مربوط به یک اندازه پنجره اختلاف ایجاد کرده است.

جدول ۲: نتایج خوشه‌بندی سوره‌ها برای اجرا با پارامترهای msp= ۴۸, ws= ۷ و f= ۲.

Table 2: The results of surahs clustering for the execution with parameters msp= 48, ws= 7, f= 2.

شماره خوشه	اعضای خوشه	تعداد سوره‌ی مکی-مدنی[4]
۱	الفاتحة، الحجر، مریم، الفرقان، القصص، السّجده، یس، الصّافات، ص، فصّلت، الزّخرف، الدّخان، محمد، ق، الذّاریات، الطّور، النّجم، القمر، الرّحمٰن، الواقعة، الممتحنة، المنافقون، الطّلاق، التّحریم، الملک، القلم، الحاقة، المعارج، نوح، الجن، المزّمّل، المدّثّر، القیامة، الإنسان، المرسلات، النّبإ، النّازعات، عبس، التّکویر، الإنفطار، المطفّفین، الإنشقاق، البروج، الطّارق، الأعلیٰ، الغاشیة، الفجر، الشّمس، اللّیل، العلق، البیّنة، الزّلزلة، العادیات، العصر، الفیل، الماعون، الکوثر، الکافرون، النّصر، الإخلاص، الفلق	۵۱ سوره‌ی مکی و ۱۰ سوره‌ی مدنی
۵	البقرة، الأنعام، الأعراف، یونس، هود، یوسف، الرّعد، إبراهیم، الکهف، الحج، المؤمنون، النّمل، العنکبوت، الرّوم، الأحزاب، سبإ، فاطر، الشّوریٰ، الجاثیة، الأحقاف، الحدید، الحشر، الصّف، التّغابن	۱۶ سوره‌ی مکی و ۸ سوره‌ی مدنی
۴	آل عمران، النّساء، المائدة، النّحل، النّور، الزّمر، غافر، الفتح،	۳ سوره‌ی

۱ از آن‌جا که در تابع مورد استفاده مقدار بیشینه تعداد خوشه‌ی مورد نظر قابل تنظیم بوده، خروجی اجراهای بدون کاهش بعد برای تعداد حداکثر ۲ تا ۲۰ خوشه، یک خوشه‌ای یا دو خوشه‌ای به‌دست‌آمده است.

2 Feature

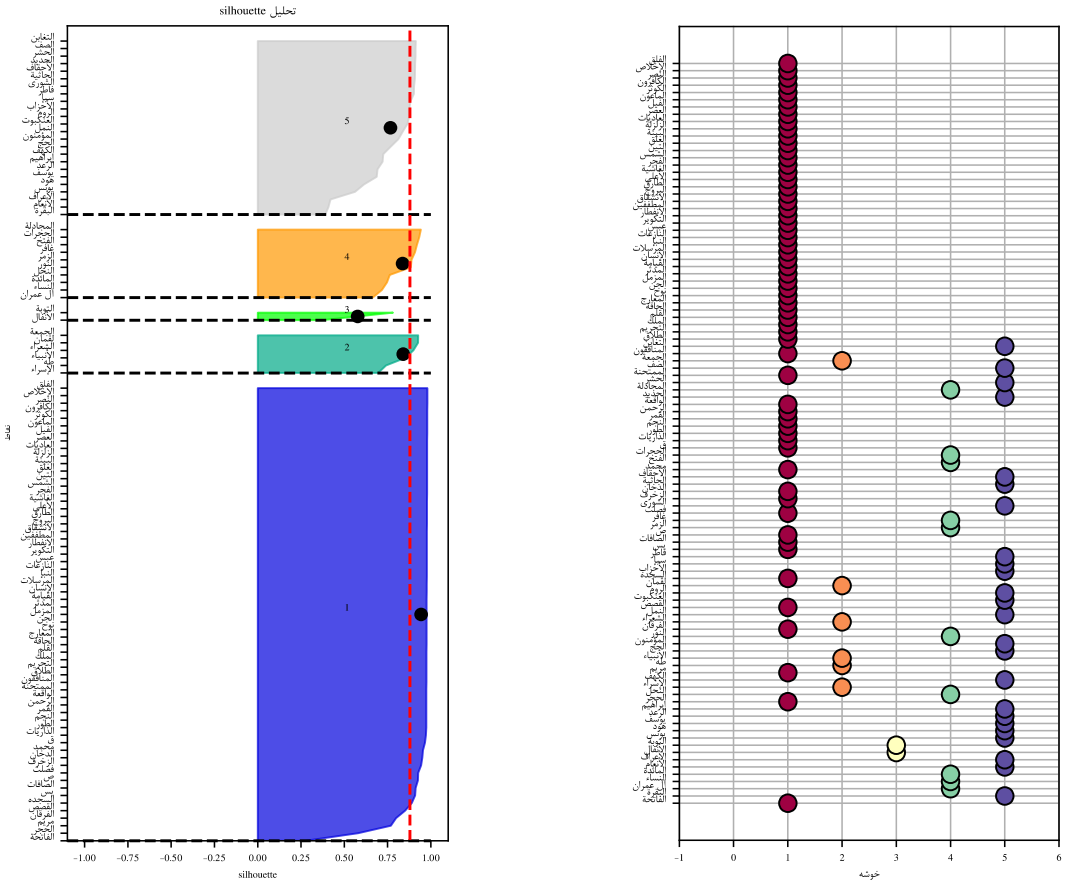
3 Window size

۴ بر اساس رده‌بندی مکی-مدنی مشهور ابتدای مصاحف شریف.

	الحجرات، المجادلة	مکی و ۷ سوره‌ی مدنی
۲	الإسراء، طه، الأنبياء، الشّعراء، لقمان، الجمعة	۵ سوره‌ی مکی و ۱ سوره‌ی مدنی
۳	الأنفال، التّوبة	۲ سوره‌ی مدنی

برای هر اندازه پنجره نمودار تغییرات مقادیر نیم‌رخ با تعداد متفاوت ویژگی پس از کاهش بعد نیز تولید می‌گردد. این نمودارها تقریباً نزولی بوده و بهترین مقدار نیم‌رخ برای هر اندازه پنجره در کمترین تعداد بعد ثانویه یعنی ۲ به دست آمده است. این مسئله می‌تواند به علت اشتراکات زیاد گره‌ها و یال‌ها بین زیرگراف‌های پرتکرار بسته باشد.

جدول ۱، مقادیر تعیین‌شده برای پارامترها در اجراهای دارای بهترین نتایج به دست‌آمده از نظر معیار نیم‌رخ را برای هر اندازه‌پنجره نشان می‌دهد. بهترین مقادیر نیم‌رخ به‌دست‌آمده در تمامی اندازه پنجره‌ها به هم نزدیک و در بازه‌ی ۰/۸۸ تا ۰/۹۱ قرار دارد. بهترین مقدار نیم‌رخ به



(ب)
(b)

(الف)
(a)

شکل ۶: نتایج خوشه‌بندی برای اجرای با پارامترهای ۷ =ws، ۴۸ =msp و ۲ =f. نمودار (الف) عضویت سوره‌ها در خوشه‌ها به ترتیب شماره‌ی سوره‌ها. محور افقی و محور عمودی به ترتیب نشان‌دهنده‌ی شماره‌ی خوشه و نام سوره‌هاست. نمودار (ب) مقادیر نیم‌رخ هر نمونه. نقاط مشکی میانگین مقدار نیم‌رخ اعضای هر خوشه و خط نقطه‌چین قرمز مقدار نیم‌رخ خوشه‌بندی را نشان می‌دهد.

Figure 6: Clustering results for the execution with parameters msp = 48, ws = 7 and f = 2. (a) membership of the surahs in clusters in the order of the number of surahs. The horizontal and vertical axes show the cluster number and name of surahs, respectively. (b) silhouette scores for each sample. Each black dot indicates the average silhouette score of the members of the corresponding cluster, and the red dotted line indicates the clustering silhouette score.

دست‌آمده (۰/۹۱) مربوط به اجرای با اندازه‌پنجره‌ی ۲، با مقدار کمینه‌پشتیبان ۱۱ است. از آن‌جا که روش خوشه‌بندی، سلسله‌مراتبی است برای اجراهای با بیش‌تر از دو خوشه از جدول ۱، خوشه‌بندی تا سطح دو خوشه‌ای نیز تولیدشد و هشت خوشه‌بندی جدول، دو به دو با هم مقایسه شدند. برای مثال دو اجرای دارای بالاترین مقادیر نیم‌رخ در ۲۱ سوره اختلاف دارند. همچنین نتایج خوشه‌بندی اجراهای اندازه پنجره‌ی ۶ و ۸ در ۶ سوره‌ی «الإسراء»، «الأنبیاء»،«الجمعة»، «الأحزاب»، «التّوبة» و «طه» متفاوتند. چنان‌چه ادغام خوشه‌ها در الگوریتم سلسله مراتبی در اجراهای با اندازه پنجره‌های ۷و ۸ ادامه می‌یافت، نتایج تنها در خوشه‌بندی سوره‌ی «لقمان» متفاوت بودند.[۱] نتایج مقایسه‌ها نشان می‌دهد در مجموعه‌ی اجراهای جدول ۱، خوشه‌بندی‌ها در انتساب ۱ تا ۲۱ سوره اختلاف دارند که هریک زیرمجموعه‌ای از ۲۳ سوره‌ی «الکهف»، «الأنفال»، «القصص»، «الجاثیة»، «التّوبة»، «فاطر»، «الشّوری»، «الحجرات»، «الرّوم»، «الفتح»، «لقمان»، «سبإ»، «الجمعة»، «التّغابن»، «المجادلة»، «المؤمنون»، «الحدید»، «الحشر»، «الصّف»، «طه»، «الإسراء»، «الأحزاب» و «الأنبیاء» می‌باشند.

از بین اجراهای جدول ۱، اجرا با بیشترین مقدار نیم‌رخ، دارای ۸ خوشه، ۷۰۲۳ تا **CFSG** و ۱۶۳ گره سازنده‌ی **CFSG** است. از آن‌جا که متخصص انسانی احتمالا در تحلیل خوشه‌بندی با تعداد خوشه‌ی بیش‌تر و تعداد **CFSG** کمتر موفق‌تر خواهد بود، در بخش ۳-۳-۱ نتایج حاصل از اجرای با اندازه پنجره‌ی ۷ با مقدار نیم‌رخ ۰/۸۸ که ۱۷ گره سازنده‌ی **CFSG** دارد و در آن ۵ خوشه به دست‌آمده، توضیح داده خواهد شد.

۳-۳-۱- خوشه‌بندی منتخب بر اساس معیار نیم‌رخ

در اجرای با اندازه پنجره‌ی ۷ با توجه به مقدار بالای کمینه‌ی پشتیبان (۴۸)، بردارها تنها دارای ۷۵ ویژگی اولیه بوده‌اند که با کاهش بعد به ۲ ویژگی کاهش یافته‌اند. این ویژگی‌ها زیرمجموعه‌هایی از ۱۷ واحد متنی «علم»، «قول»، «ربب»، «سمو»، «الله»، «رحم»، «ارض»، «قوم»، «بین»، «شیء»، «کفر»، «امن»، «اتی»، «عمل»، «رسل»، «اخر» و «رأی» هستند. جدول ۲، خوشه‌بندی به دست‌آمده در اجرای با اندازه پنجره‌ی ۷ را نمایش می‌دهد.

در خوشه‌بندی حاصل بردار سوره‌ی «عبس» با «القدر»، «القارعة»، «التّکاثر»، «الهمزة»، «قریش»، «المسد»، «البلد»، «اللّیل»، «الضّحی» و «الشّرح» و همچنین بردار سوره‌ی «الفلق» با «النّاس» یکسان بوده، بنابراین همان‌گونه که قبل‌تر در بخش ۳-۲ اشاره شد، پیش از خوشه‌بندی از هر مجموعه بردار یکسان یکی نگه داشته شده، بقیه حذف شده‌اند و در محاسبه‌ی مقدار نیم‌رخ نیز محسوب نبوده‌اند. جدول ۲، همچنین تعداد سوره‌های مکی و مدنی در هر خوشه را نشان می‌دهد. اگرچه سوره‌های با بردار یکسان از نظر مکی-مدنی بودن یک‌دست و همگی مکی هستند، اما در ترکیب مکی مدنی سوره‌ها در خوشه‌های مختلف، نظم خاصی دیده نشده و برای یافتن وجه ارتباط این خوشه‌ها لازم است جنبه‌ی دیگری جز زمان نزول را بررسی نمود.

با روش خوشه‌بندی مورد استفاده، باهم قرارگرفتن سوره‌های درون یک خوشه تنها به خاطر ویژگی‌های مشترک بین همه‌ی سوره‌های عضو آن نیست و سایر ویژگی‌ها که در این سوره‌ها وجود داشته‌اند نیز مؤثر بوده‌است.

شکل ۶، نتایج بهترین اجرا با اندازه پنجره‌ی ۷ را از نظر مقدار معیار نیم‌رخ و پراکندگی اعضای خوشه‌ها در قرآن نشان می‌دهد. مقدار نیم‌رخ نهایی برای ارزیابی کیفیت یک خوشه‌بندی گزارش می‌شود. این مقدار برابر میانگین نیم‌رخ نمونه‌هاست. در نمودار مقادیر نیم‌رخ، مقدار نیم‌رخ به دست‌آمده برای هر سوره، به تفکیک خوشه‌ها نشان داده می‌شود. با توجه به این نمودار می‌توان در مورد تناسب یک سوره با یک خوشه و نیز خوب‌بودن یک خوشه از نظر معیار نیم‌رخ اظهار نظر کرد. در حالت کلی، برای یک نمونه مقدار نیم‌رخ ۱ به معنی انتساب به خوشه‌ی درست و نیم‌رخ ۱- به معنی انتساب به خوشه‌ی غلط درنظرگرفته می‌شود. همان‌گونه که در شکل ۶- الف پیداست، مقدار نیم‌رخ تمامی نمونه‌ها مثبت است؛ یعنی فاصله‌ی درون خوشه‌ای نمونه از فاصله‌ی بین خوشه‌ای کمتر بوده است. همچنین میانگین تمامی خوشه‌ها به جز خوشه‌ی دوتایی سوره‌های« الأنفال» و «التّوبه»، بیشتر از ۰/۷۵ می‌باشد. خوشه‌ی اول دارای بالاترین مقدار میانگین نیم‌رخ خوشه‌هاست. کم‌ترین (۰/۲۸) و بیشترین (۰/۹۸) مقدار نیم‌رخ نمونه‌ها در این خوشه‌ی۶۱ عضوی قرار دارد. همان‌گونه که در شکل ۶-الف مشاهده می‌شود، به جز دو سوره‌ی «الفاتحة» و «الحجر» سایر اعضای خوشه ۱ دارای نیم‌رخ بالاتر از ۰/۷۵ هستند و میانگین نیم‌رخ‌ها در این خوشه ۰/۹۴ است. مطابق شکل ۶-ب سوره‌های انتهای قرآن به همراه تعدادی سوره‌ی دیگر مانند سوره‌ی «القصص» در این خوشه قرار داشته و باقی سوره‌ها در چهار خوشه‌ی دیگر پراکنده‌اند.

شکل ۷ نمودار سلسله مراتبی ادغام سوره‌ها تا رسیدن به دو خوشه را برای اجرای با اندازه پنجره‌ی ۷ نشان می‌دهد. در روند تحلیل خوشه‌ها با استفاده از این نمودار چنان‌چه در حدس زدن وجه اشتراک کلی موفق نباشیم، می‌توانیم تحلیل با توجه به ترتیب ادغام در نمودار سلسله مراتبی مورد نظر را نیز امتحان کنیم.

از نظر محتوایی، سوره‌های موجود در خوشه‌های به‌دست‌آمده به هم نزدیک است. خوشه‌ی ۱ حاوی اصول اعتقادی توحید و معاد می‌باشد. در خوشه‌ی ۴ و ۵ پرداختن به مسائل فقهی و حکومت‌داری پررنگ‌تر است. سوره‌ی بقره در خوشه‌ی ۵ بعد از استقرار حکومت اسلامی نازل شده و حاوی مسائل اجتماعی است. خوشه‌ی ۴ را می‌توان ترکیبی از موضوعات اخلاقی و فقه دانست. در این خوشه مسائل فقهی جزئی‌تر مثل احکام ارث در سوره‌های نور و نساء، دیده می‌شود. سوره‌های خوشه‌ی ۳، انفال و توبه، حاوی آیات نبرد و جهاد می‌باشد. در خوشه‌ی۲ نیز داستان انبیاء و موعظه‌های حکمی ایشان نمود بیشتری دارد. لازم به ذکر است این تحلیل در محدوده‌ی دانش نویسندگان و یک مقاله‌ی حوزه‌ی علوم رایانه بیان شده و لازم است توسط خبره‌ی قرآنی تأیید شود. نتایج به دست‌آمده با هدف یافتن ارتباط سوره‌های هر خوشه می‌تواند در اختیار متخصص قرآنی قرار گیرد. این ارتباط ممکن است از نوع موضوعات، وقایع بیان‌شده، هدف غایی سوره، زمان نزول یا دیگر انواع ممکن ارتباطی برای سوره‌ها باشد، مثلا ممکن است موضوع مشترک سوره‌های یک خوشه انفاق و برای خوشه‌ی دیگر امتحانات الهی تشخیص داده شود. این کار می‌تواند با توجه به مبنای خوشه‌بندی انجام شود و متخصص در صورت نیاز با رعایت اصول منطقی خوشه‌هایی را با هم ادغام نموده یا تقسیم نماید.

برای تحلیل خوشه‌ها می‌توان به روش زیر عمل کرد:

---

۱ البته تمام موارد اختلافی گزارش شده در مقاله بدون احتساب سوره‌هایی است که در مرحله‌ی آماده‌سازی بردارها از خوشه‌بندی یک اجرا حذف شده بودند، اما در دیگری به صورت مستقل (قبلاً در ۳-۲ اشاره شد از بین سوره‌هایی که بردارشان یکسان است، تنها یکی نگه داشته می‌شود و بقیه به صورت مستقل درون ماتریس داده قرار ندارند)، حضور داشته‌اند.

* از خوشه‌های با تعداد اعضای کمتر شروع شود.
* ابتدا با توجه به آشنایی با سوره‌ها سعی شود تا وجه اشتراک بین سوره‌ها حدس زده شود.
*اگر تعداد زیرگراف‌های پرتکرار بسته(**CFSG**) یا حتی تعداد گره‌های سازنده‌شان، کم بود، به آن‌ها نیز توجه شود. ممکن است ترکیبات مختلفشان در یافتن وجه اشتراک کمک کند. البته مطلوب یافتن وجه اشتراکی جز مبنای اولیه‌ی خوشه‌بندی است، هرچند ممکن است یافتن مبنای خوشه‌بندی نیز منجر به یافته‌ی جدیدی شود.
* چنان‌چه در حدس زدن وجه اشتراک کلی موفقیت حاصل نشد، تحلیل با توجه به ترتیب ادغام در نمودار سلسله مراتبی مورد نظر نیز می‌تواند امتحان شود. مثلا برای یافتن وجه اشتراک بین سوره‌های خوشه‌ی شماره ۲، ابتدا چند وجه اشتراک معنایی یا وجه اشتراک قرآنی دیگری را برای سوره‌های «الجمعه» و« الأنبیاء» بیابیم بعد با اضافه کردن سوره‌ی «الشّعراء» وجوهی که در این سوره‌ها نیست را حذف کنیم، به همین ترتیب برای سوره‌های «لقمان» و «الإسراء» و سپس سوره‌ی «طه» پیش رفته تا در نهایت بتوان از بین مجموعه وجوه اشتراکی اولیه برای خوشه‌ی شماره ۲، یک یا چندتا را انتخاب نمود.
* اگر وجه اشتراکی پیدا شد که در بیشتر سوره‌های خوشه وجود داشت، اما در بقیه صادق نبود، به نمودار مقادیر نیم‌رخ مراجعه شود؛ اگر سوره‌هایی که آن وجه اشتراک را ندارند، مقدار نیم‌رخ کمتری داشته باشند، احتمال اینکه واقعا عضو آن خوشه نباشند و بشود آن‌ها کنار گذاشت بیشتر است.

ادامه می‌یافت، اختلاف در خوشه‌بندی ۱ تا ۲۱ سوره را نشان می‌دهد.

لازم به ذکر است با توجه به وابستگی نتایج حاصل به پارامترهای مختلف در مسأله و معیارهای شباهت درنظر گرفته‌شده، خوشه‌بندی‌های به‌دست‌آمده قطعی نیست. همچنین کیفیت خوشه‌بندی بر اساس شباهت سوره‌ها در فضای ویژگی درنظر گرفته‌شده گزارش شده است. در زمان نگارش مقاله به دلیل وجود پژوهش‌های اندک مشابه و عدم ارائه‌ی مناسب معیار کمی برای اندازه‌گیری کیفیت خوشه‌بندی در کارهای مشابه، نتایج با سایر پژوهش‌ها مقایسه نشده است. برای ارزیابی معنادار بودن، این نتایج می‌تواند در اختیار متخصص انسانی قرار داده شود.

۵- فهرست منابع

[۱] اصلانی، اکرم، اسماعیلی، مهدی. «یافتن الگوهای مکرّر در قرآن کریم به‌کمک روش‌های متن‌کاوی»، پردازش علائم و داده‌ها،۱۵ (۳)، ۸۹-۱۰۰، ۱۳۹۷. https://jsdp.rcisp.ac.ir/. [دسترسی در ۱۹ اردیبهشت ۱۴۰۳].

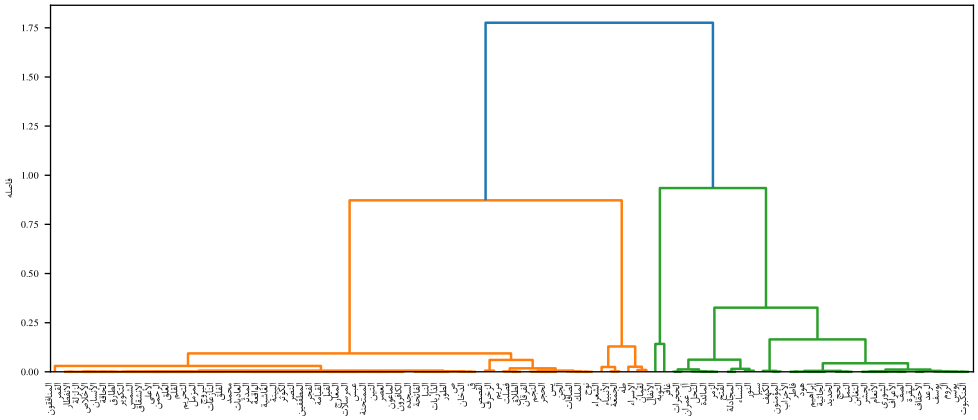
A. Aslani A, M. Esmaeili, "Finding Frequent Patterns in Holy Quran UsingText Mining," JSDP, vol. 15, no.3, pp. 89-100, 2018. Available: https://jsdp.rcisp.ac.ir/. [Accessed: May. 10, 2024].

[۲] صادق‌زمانی، فهیمه، ضرغامی، محمدحسین، «تعیین ساختار روابط بین اعضای خانواده بر اساس ویژگیهای شخصیتی»، *اندازه‌گیری تربیتی*، شماره ۳۳، ۲۰۸-۱۸۹، ۱۳۹۷. http://journals.atu.ac.ir. [دسترسی در ۵ اسفند۹۹].

F. Sadegh-Zamani and M. H. Zarghami, "Determining structure of relations between family members based on personality characteristics," Quarterly of Educational Measurement, no. 33, pp. 189-208, 2018. Available: http://journals.atu.ac.ir. [Accesed: Feb. 24, 2021].

[۳] صوفی، محسن، علی‌احمدی، علی‌رضا، علی‌احمدی، حسین، مینایی، بهروز، «خوشه‌بندی سوره‌های قرآن با تکنیک‌های داده‌کاوی»، *رهیافت‌هایی در علوم قرآن و حدیث*، دوره ۵۰، ۱۲۰-۱۰۳، ۱۳۹۷. https://jquran.um.ac.ir. [دسترسی در ۵ اسفند۹۹].

M. Sufi, A. R. Ali-Ahmadi, H. Ali-Ahmadi, B. Minaei-Bidgoli, "Clustering of Qur'anic Surahs



شکل ۷: نمودار خوشه‌بندی سلسله‌مراتبی به دست‌آمده برای اجرای با پارامترهای ۷ =ws ,۴۸ =msp و ۲ =f.

Figure 7: The hierarchical clustering dendrogram obtained for the execution with the parameters msp = 48, ws = 7 and f = 2.

۴- نتیجه‌گیری

در این پژوهش، یک رویکرد مبتنی بر گراف برای خوشه‌بندی سوره‌های قرآن به کار گرفته شده است. در این روش پس از استخراج زیرگراف‌های پرتکرار از گراف سوره‌ها، بردار سوره‌ها با زیرگراف‌های پرتکرار بسته به عنوان ویژگی تشکیل شده و خوشه‌بندی پس از اعمال کاهش بعد انجام می‌شود. نتایج به دست‌آمده نشان می‌دهد با اعمال کاهش بعد، بهترین مقدار نیم‌رخ در خوشه‌بندی‌های حاصل از مقادیر متفاوت کمینه پشتیبان در اندازه پنجره‌های مختلف در بازه‌ای بین ۰/۸۸ تا ۰/۹۱ نزدیک به هم بوده و لزوما افزایش کمینه پشتیبان یا کوچک‌تر بودن اندازه پنجره منجر به افزایش کیفیت خوشه‌بندی نمی‌شود. مقایسه‌ی دوبه دوی نتایج بهترین خوشه‌بندی‌های حاصل از هر اندازه پنجره، چنانچه خوشه‌بندی سلسله مراتبی تا رسیدن به دو خوشه

[13] ***M. H. Elahimanesh, B. Minaei-Bidgoli, M. J. Gholami, and H. Juzi, "An Introduction to Noor Corpus and its Language Model." First International Conference on Persian language Processing(ICPLP), Semnan university, 2012. Available: researchgate.net. [Accessed: Feb. 25, 2021].***

[14] **H. Veeramani, S. Thapa and U. Naseem , "LowResContextQA at Qur'an QA 2023 Shared Task: Temporal and Sequential Representation Augmented Question Answering Span Detection in Arabic," In Proceedings of ArabicNLP 2023, pp. 708-713, 2023. Available: https://aclanthology.org/2023.arabicnlp-1.78. [Accessed: May. 10, 2024].**

[15] ***A. Lim, "The berth planning problem," perations research letters, vol. 22, pp. 105-110, March 1998. Available: https://www.sciencedirect.com/ [Accessed: Feb. 25, 2021].***

[16] **R. Malhas and T. Elsayed, "Arabic machine reading comprehension on the Holy Qur'an using CL-AraBERT," Information Processing & Management, Vol. 59, no. 6, 2022. Available: https://doi.org/10.1016/j.ipm.2022.103068. [Accessed: May. 10, 2024].**

[17] **G. Mediamer, "Semantic Feature Analysis for Multi-Label Text Classification on Topics of the Al-Quran Verses," Journal of Information Processing Systems, vol. 20, no.1, 2024. Available: https://jips-k.org/digital-library/2024/20/1/1. [Accessed: May. 10, 2024].**

[18] **Y. Mellah, I. Touahri, Z. Kaddari, Z. Haja, J. Berrich and T. Bouchentouf , "LARSA22 at Qur'an QA 2022: text-to-text transformer for finding answers to questions from Qur'an," In Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, pp. 112-119, 2022. Available: https://aclanthology.org/2022.osact-1.13. [Accessed: May. 10, 2024].**

[19] **M. Mohammed, S. Amin and MM. Aref , "An english islamic articles dataset (eiad) for developing an islambot question answering chatbot," In 2022 5th International Conference on Computing and Informatics (ICCI), pp. 303-309, 2022. Available: https://ieeexplore.ieee.org/abstract/document/9756122/. [Accessed: May. 10, 2024].**

[20] ***H. Moisl, "Sura Length and Lexical Probability Estimation in Cluster Analysis of the Qur'an," ACM Transactions on Asian Language Information Processing Using Data Mining Techniques", New Approaches in Quran and Hadith Studies, vol. 50, pp. 103-120, 2018-2019. Available: https://jquran.um.ac.ir. [Accesed: Feb. 24, 2021].***

[۴] قلی‌زاده، بهروز، *ساختمان‌های گسسته*، چاپ سی و یکم، تهـران، مؤسسـه انتشـارات علمی دانشـگاه شریف، چاپ۳۱، ۱۳۹۱.

B. Gholizadeh, Discrete Mathematics, Tehran: Sharif University Press, 2012-2013.

[۵] هاشمی رفسنجانی، علی‌اکـبر، و جمعی از محققـان مرکز فرهنگ و معـارف قـرآن، تفسـیر راهنمـا، قم: بوستان کتاب قم، چاپ سوم، ۱۳۷۹.

A. A. Hashemi-Rafsanjani, and a group of researchers from Quranic Science and Culture Center, Tafsir Rahnama, Qom: Bustan Ketab Publisher, 2000-2001.

[6] ***E. Aftab and MK. Malik, "eRock at Qur'an QA 2022: contemporary deep neural networks for Qur'an based reading comprehension question answers," In Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection 2022. pp. 96-103, 2022. Available: https://aclanthology.org/2022.osact-1.11.pdf [Accessed: May. 10, 2024].***

[7] ***M. E. Aktas, and E. Akbas, "Text classification via network topology: A case study on the Holy Quran," In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), 16 Dec 2019, 1557-1562. Available: http://www.math.uco.edu. [Accessed: Feb. 24, 2021].***

[8] ***E. Alpaydin, Introduction to machine learning, 2nd ed. Massachusetts: Massachusetts Institutes of Technology, 2010. Available: https://books.google.com [Accessed: Feb. 24, 2021].***

[9] ***E. Atwell, Habash Nizar, Louw Bill, B. Abu Shawar, T. McEnery, W. Zaghouani, and M. El-Haj, "Understanding the Quran: A new grand challenge for computer science and artificial intelligence," ACM-BCS Visions of Computer Science 2010, 2010. Available: https://eprints.lancs.ac.uk. [Accessed: Feb. 25, 2021].***

[10] ***C. M. Bishop, Pattern Recognition and Machine Learning, New York: Springer-Verlag, 2006. Available: http://users.isr.ist.utl.pt/. [Accessed: Feb. 25, 2021].***

[11] ***K. Dukes, and N. Habash, "Morphological Annotation of Quranic Arabic," Lrec, 2010. Available: http://citeseerx.ist.psu.edu/. [Accessed: Feb. 25, 2021].***

[12] ***K. Dukes, "Quranic Arabic Corpus," May. 1, 2011. [online]. Available: http://corpus.quran.com/. [Accessed: Feb. 24, 2021]***

[29] *I. Takigawa, H. Mamitsuka, "Efficiently mining δ-tolerance closed frequent subgraphs," Machine Learning, vol. 82, pp. 95-121. Available: https://link.springer.com. [Accessed: Feb. 25, 2021].*

[30] *P. N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, Introduction to data mining, second edition, New York: Pearson Education, 2018.*

[31] *N. Thabet, "Understanding the thematic structure of the Qur'an: an exploratory .multivariate approach," In Proceedings of the ACL Student Research Workshop, 2005, pp. 7-12. Available: https://www.aclweb.org/. [Accessed: Feb. 25, 2021].*

[32] *X. Yan, and J. Han, "gspan: Graph-based substructure pattern mining," 2002 IEEE International Conference on Data Mining,2002, pp. 721-724. Available: https://sites.cs.ucsb.edu/. [Accessed: Feb. 25, 2021].*

[33] *R. Zafarani, M. A. Abbasi, and H. Liu, Social Media Mining: An Introduction, London: Cambridge University Press, 2014. Available: http://citeseerx.ist.psu.edu/. [Accessed: Feb. 25, 2021].*

بهروز مینایی بیدگلی: دانش آموخته دانشگاه ایالتی میشیگان آمریکا در رشته علوم و مهندسی کامپیوتر با تخصص هوش مصنوعی و داده‌کاوی است. او در حال حاضر عضو هیأت علمی و دانشیار دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت و رئیس دانشکده مهندسی کامپیوتر است. او سرپرستی گروه پژوهشی فناوری‌های بازی‌های رایانه‌ای و نیز آزمایشگاه داده‌کاوی را به عهده دارد. محاسبات نرم، یادگیری ماشین، بازی‌های رایانه‌ای، داده‌کاوی، متن‌کاوی، و پردازش زبان طبیعی، زمینه‌های پژوهشی مورد علاقه ایشان است. نشانی رایانامه‌ی ایشان عبارت است از:

b_minaei@iust.ac.ir

مریم‌سادات متقی دانش‌آموخته‌ی رشته‌ی مهندسی کامپیوتر مقطع کارشناسی دانشگاه شاهد و مقطع کارشناسی ارشد رشته‌ی قرآن‌کاوی رایانشی پژوهشکده‌ی اعجاز قرآن دانشگاه شهیدبهشتی تهران می‌باشد. او هم‌اکنون دانشجوی دکترای هوش مصنوعی دانشگاه شهیدبهشتی است. زمینه‌ی پژوهشی وی قرآن‌کاوی رایانشی و پردازش متن می‌باشد. نشانی رایانامه‌ی ایشان عبارت است از:

m.motaghi88@chmail.ir

(TALIP), vol. 8, pp. 1-19, 2009. Available: https://eprints.ncl.ac.uk/. [Accessed: Feb. 25, 2021].

[21] *A. B. Muhammad, "Annotation of conceptual co-reference and text mining the Qur'an,"Ph.D. dissertation, Dept. school of computing, Leeds Univ., UK, 2012. Available: http://etheses.whiterose.ac.uk/. [Accessed: Feb. 25, 2021].*

[22] *C. Nicolini, C. Bordier, and A. Bifone, "Community detection in weighted brain connectivity networks beyond the resolution limit," Neuroimage, vol. 146, pp. 28-39, 2017. Available: researchgate.net. [Accessed: Feb. 25, 2021].*

[23] *F. Rousseau, E. Kiagias, and M. Vazirgiannis, "Text categorization as a graph classification problem," In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26-31 July 2015, pp. 1702–1712. Available: https://www.aclweb.org/. [Accessed: Feb. 25, 2021].*

[24] *P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Journal of computational and applied mathematics, vol. 20, pp. 53-65, 1987. Available: https://www.sciencedirect.com/. [Accessed: Feb. 25, 2021].*

[25] *A. B. M. Sharaf, and E. Atwell, "QurAna: Corpus of the Quran annotated with Pronominal Anaphora," LREC, 2012, pp. 130-137. Available: http://citeseerx.ist.psu.edu/. [Accessed: Feb. 25, 2021].*

[26] *A. B. M. Sharaf, and E. Atwell, "QurSim: A corpus for evaluation of relatedness in short texts," LREC, 2012, pp. 2295-2302. Available: http://textminingthequran.com/. [Accessed: Feb. 25, 2021].*

[27] *S. A. Shirkhorshidi, S. Aghabozorgi Saeed, and T. Y. Wah, "A comparison study on similarity and dissimilarity measures in clustering continuous data," PloS one, vol. 10, 2015. Available: researchgate.net. [Accessed: Feb. 25, 2021].*

[28] *M. A. Siddiqui, S. M. Faraz, and S. A. Sattar, "Discovering the thematic structure of the Quran using probabilistic topic model," 2013 In 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, IEEE, 2013, pp. 234-239. Available: researchgate.net. [Accessed: Feb. 25, 2021].*