POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Evaluation of posts by bioinformatics code developers on stack overflow platform: topic modeling and community detection

Biyoenformatik alanındaki kod geliştiricilerin stack overflow platformunda paylaştıkları soruların değerlendirilmesi: konu modelleme ve topluluk tespiti

***Yazar(lar) (Author(s))**: Gülbahar Merve ŞILBIR[1]*

ORCID[1]: 0000-0003-0321-7259

To cite to this article: Şılbır G. M., "Evaluation of posts by bioinformatics code developers on stack overflow platform: topic modeling and community detection", ***Journal of Polytechnic***, *(*): *, (*).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Şılbır G. M., "Evaluation of posts by bioinformatics code developers on stack overflow platform: topic modeling and community detection", ***Politeknik Dergisi***, *(*): *, (*).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1587995

Evaluation of Posts by Bioinformatics Code Developers on Stack Overflow Platform: Topic Modeling and Community Detection

Highlights

- *The main areas of bioinformatics developers' Genomic Research and Analysis.*
- *Topics included gene expression, protein interaction prediction, and genomic data management.*
- *The presentation identified seven bioinformatics communities and described 100 central items.*

Graphical Abstract

In this study, posts of bioinformatics code developers on the Stack Overflow platform between March 2017 and 2024 were analyzed using LDA topic modeling and the Louvain community finding algorithm.

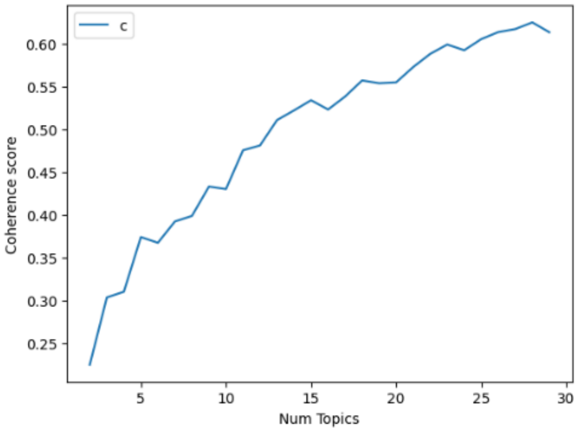


Figure. Graphic Abstract

Aim

In this study, we present an analysis of the posts shared on the Stack Overflow website within the field of bioinformatics. We examine the types of questions, the tools and methods discussed, and the trends and patterns noted in the bioinformatics-related discussions on this platform.

Design & Methodology

*In this study, we analyzed the posts shared about bioinformatics on the Stack Overflow platform using LDA topic modeling and the Louvain community fin*ding algorithm.

Originality

The originality of this study is to gain a deeper understanding of the challenges faced by code developers in bioinformatics. It is anticipated that the findings will identify areas of knowledge that are currently lacking in the field, thereby guiding future research efforts based on these insights.

Findings

Our finding revealed that bioinformatics developers' questions focused on 28 topics in four main categories. We also presented that topics in bioinformatics consist of seven communities and the trends of these communities and the relationship between the 100 most central words.

Conclusion

Based on the results we obtained from this study, the problems that bioinformatics developers have encountered over time have been revealed with topic modeling and community detection.

Declaration of Ethical Standards

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Evaluation of Posts by Bioinformatics Code Developers on Stack Overflow Platform: Topic Modeling and Community Detection

Araştırma Makalesi / Research Article

Gülbahar Merve ŞILBIR[1*]

[1]Trabzon University, Trabzon, Türkiye

(Geliş/Received : 19.11.2024 ; Kabul/Accepted : 17.04.2025 ; Erken Görünüm/Early View : 03.05.2025)

ABSTRACT

Developers are key to managing, storing and analysing the growing biological data. Platforms like Stack Overflow help identify current trends in the field. In this study, we present an analysis of the posts shared on the Stack Overflow website within the field of bioinformatics. We analyzed the posts shared about bioinformatics on the Stack Overflow platform using LDA topic modeling and the Louvain community finding algorithm. Our finding revealed that bioinformatics developers' questions focused on 28 topics in four main categories. We found that the most popular topics were "Gene Expression and Function", "Protein Interaction Prediction", "Gene and Protein Structure Analysis", "Sample Analysis in Network Problems", and "Genomic Data Management". Besides, we also presented that topics in bioinformatics consist of seven communities and the trends of these communities and the relationship between the 100 most central words. Our finding also revealed that the topics that code developers are most interested in in the field of bioinformatics are "next generation sequencing", "genome", "gene", "phylogeny", "proteins", and "sequence". Based on the results we obtained from this study, the problems that bioinformatics developers have encountered over time have been revealed with topic modeling and community detection.

Keywords: Bioinformatics, bioinformatics topics, topic modeling, community detection, stack overflow.

Biyoenformatik Alanındaki Kod Geliştiricilerin Stack Overflow Platformunda Paylaştıkları Soruların Değerlendirilmesi: Konu Modelleme ve Topluluk Tespiti

ÖZ

Kod geliştiriciler, artan biyolojik verileri yönetmek, depolamak ve analiz etmek için anahtar konumdadır. Stack Overflow gibi platformlar, geliştiriciler için alandaki mevcut eğilimleri belirlemeye yardımcı olan tartışma platformlarıdır. Bu çalışmada, biyoenformatik alanında Stack Overflow web sitesinde paylaşılan gönderilerin bir analizini sunuyoruz. LDA konu modellemesi ve Louvain topluluk bulma algoritmasını kullanarak Stack Overflow platformunda biyoenformatik hakkında paylaşılan gönderileri analiz ettik. Bulgularımız, biyoenformatik geliştiricilerinin sorularının dört ana kategoride 28 konuya odaklandığını ortaya koydu. En popüler konuların "Gen İfadesi ve İşlevi", "Protein Etkileşim Tahmini", "Gen ve Protein Yapısı Analizi", "Ağ Sorunlarında Örnek Analizi" ve "Genomik Veri Yönetimi" olduğunu bulduk. Ayrıca, biyoenformatikteki konuların yedi topluluktan oluştuğunu ve bu toplulukların eğilimlerini ve en merkezi 100 kelime arasındaki ilişkiyi de sunduk. Bulgularımız ayrıca biyoenformatik alanında kod geliştiricilerinin en çok ilgi duyduğu konuların "yeni nesil dizileme", "genom", "gen", "filogeni", "proteinler" ve "sekans" olduğunu ortaya koydu. Bu çalışmadan elde ettiğimiz sonuçlara dayanarak, biyoenformatik kod geliştiricilerinin zaman içinde karşılaştığı sorunlar konu modelleme ve topluluk tespiti ile ortaya konmuştur.

Anahtar Kelimeler: Biyoinformatik, biyoinformatik konuları, konu modelleme, topluluk tespiti, stack overflow.

1. INTRODUCTION

Bioinformatics integrates computational techniques and algorithms with biological data analysis, playing a crucial role in the field [1]. This interdisciplinary approach has led to progress in various areas of life sciences, such as genomics, proteomics, and molecular medicine [2]. The explosion of biological data from high-throughput sequencing technologies has created an immediate requirement for efficient data management, storage, and analysis methods [3, 4].

Bioinformatics encompasses a wide range of studies, including the analysis of genetic data, protein structures, systems biology, and personalized medicine [5]. In recent years, there has been great interest in developing advanced algorithms for collecting and processing high-volume biological data [6]. The integration of artificial intelligence and machine learning techniques has revolutionized the discovery of genetic variations and biomarkers, leading to significant advancements in biomedical research and clinical applications [7]. In addition, these technological advances have led to the development of new approaches in big data analytics and computational biology, further advancing bioinformatics studies [8-10]. In this context, bioinformatics continues to play a critical role in biological research and medical applications thanks to rapidly developing data analysis tools.

*Sorumlu Yazar (Corresponding Author)
e-posta : gmervecakmak@trabzon.edu.tr

Stack Overflow, a widely used platform for developers, has become a valuable resource for the bioinformatics community. The discussions and problem-solving threads on this website offer valuable insights into the real-world challenges encountered by bioinformatics practitioners and the corresponding strategies utilized to tackle these challenges [11]. To understand the topics discussed on Stack Overflow and analyze the textual data obtained from this platform to determine trends. The analyses also uncover significant implications for the research field [12].

Questions posed on knowledge-sharing platforms like Stack Overflow serve as a valuable data source for identifying knowledge gaps in bioinformatics and the practical challenges developers face. Topic modeling methods can be employed to uncover the semantic relationships formed by the rapidly growing body of textual data in bioinformatics. Latent Dirichlet Allocation (LDA), a widely used technique in topic modeling, is a text mining method designed to uncover hidden thematic structures within large datasets [13]. Topic modeling algorithms identify patterns within textual data by generating topics and using these topics to associate documents with similar content [14]. This approach to analyzing textual data is utilized across various disciplines, including social sciences and computer science [15].

Another method for identifying semantic relationships within textual data in bioinformatics is the use of community detection algorithms. Community detection algorithms are extensively used in fields such as social network analysis and bioinformatics, and in recent years, they have also become a key tool for uncovering semantic relationships in textual data [16, 17]. These algorithms identify clusters of words or concepts within a text to reveal its underlying semantic structures. Graph-based approaches, particularly the Louvain algorithm, are highly effective in detecting communities within networks that represent textual data, and are employed to analyze semantic closeness and divergence in texts [18]. For example, these algorithms are utilized to group posts on similar topics in social media data or to recognize common themes in scientific articles [19]. As a result, community detection algorithms offer powerful methods for analyzing the complex structure of textual data and identifying meaningful relationships within the data [20].

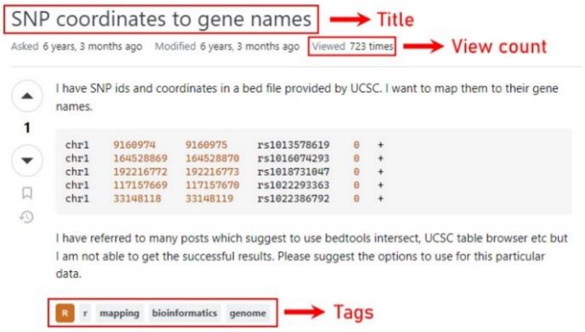
When examining literature studies, it was found that topic modeling methods were used to identify trends in bioinformatics. Hahn et al. (2017) analyzed trends in bioinformatics literature using the LDA model for topic modeling and identified subfields of this field that have attracted more attention in recent years. This study has revealed that topics such as "big-data", "next generation sequencing", and "cancer" are rapidly developing areas in bioinformatics [15]. Similarly, Youssef and Rich (2018) used the LDA model to analyze scientific research in bioinformatics and examined the topics that received more attention. As a result of the study, "cancer biology" and "clinical informatics" were identified as the most popular research categories, while "next-generation sequencing", "metagenomics", "PPIs", and "microbiome analysis" were identified as developing research areas [21]. Ebrahimi et al. (2023) utilized the LDA and TF-IDF models to analyze bioinformatics literature submitted to the Scopus Citation Database by Iranian researchers. Their findings indicated that the main topics included "molecular modeling," "gene expression," "biomarker," "coronavirus," "immunoinformatics," "cancer bioinformatics," and "systems biology" [22]. These studies highlight the effectiveness of LDA and similar topic modeling techniques in identifying bioinformatics research trends.

In this research paper, we present an analysis of the posts shared on the Stack Overflow website within the field of bioinformatics. We examine the types of questions posed, the tools and methods discussed, and the trends and patterns noted in the bioinformatics-related discussions on this platform. In this study, we aim to determine the topics that users talk about by analyzing the questions asked on Stack Overflow about "Bioinformatics" using the topic modeling technique LDA. In addition, we aim to identify communities by analyzing the tags associated with these questions using the Louvain algorithm. Our goal is to determine the popularity of the tags used in bioinformatics and to analyze the distribution of these tags by year. The anticipated results are expected to provide a deeper understanding of the challenges faced by code developers in bioinformatics, identify knowledge gaps in the field, and guide future research endeavors based on these insights. Additionally, it is believed that this study will improve the understanding of software development processes in bioinformatics and the challenges encountered, particularly through the analysis of data from platforms like Stack Overflow..

2. MATERIAL and METHOD

2.1. Data Collection and Extraction

In this study, a total of 22,335 shares in the field of bioinformatics were downloaded on the Stack Overflow platform until March 2024 (Access date: 30.03.2024). A sample image and structure of the downloaded data is given in Figure 1. Figure 1 illustrates, "title" indicates the title of the question sent by users, "tags" indicates the keywords associated with the question, and "view count" indicates the number of times the question was viewed by users. In this study, we used the LDA algorithm in the "title" analysis and the Louvain algorithm in the "tags" analysis. Data lines with empty "title" and "tags" data were removed from the data and the analysis was performed using the remaining 6,093 data. The total number of analyzed words is 56,449 for "title" and 16,594 for "tags".



SNP coordinates to gene names

Asked 6 years, 3 months ago Modified 6 years, 3 months ago Viewed 723 times

1

I have SNP ids and coordinates in a bed file provided by UCSC. I want to map them to their gene names.

```
chr1    9160974     9160975     rs1013578619    0   +
chr1    164528869   164528870   rs1016074293    0   +
chr1    192216772   192216773   rs1018731047    0   +
chr1    117157669   117157670   rs1022293363    0   +
chr1    33148118    33148119    rs1022386792    0   +
```

I have referred to many posts which suggest to use bedtools intersect, UCSC table browser etc but I am not able to get the successful results. Please suggest the options to use for this particular data.

R r mapping bioinformatics genome

Figure 1. Stack Overflow data

2.2. Data Preprocessing

In data pre-processing operations, punctuation marks, HTML tags, and characters other than alphabetic characters (meaningless characters, web links, and numbers) were removed from the data. All letters are converted to lower case to avoid differences between words. Data consisting of sentences is separated into words by the tokenization process. Ineffective words that do not disrupt the semantic integrity of the data were removed from the data. For this purpose, previously determined English stopwords were extracted from the data using the Python nltk library. Lemmatization involves identifying word roots and reducing words while preserving meaning. Finally, bigrams and trigrams operations were performed using the n-gram method to detect the consecutive use of words in pairs or triplets.

Table 1. Parameter values for LDA

Parameter	Brief Explanation	Value
num_topics	Number of topics to be obtained from the documents	[2-30]
random_state	Random seed determinant that ensures the same results are reproduced	100
update_every	Determines how often model parameters are updated	1
chunksize	Number of documents to be used in each training iteration	1000
passes	Total number of training iterations	5
α	Coefficient that determines document-subheading density. A high value of α finds more headings, while a low value of α finds fewer subheadings.	symmetric
β	Coefficient determining word-subheading density. A high β value indicates that a heading consists of a large number of words, while a low β value indicates that a heading consists of fewer words.	symmetric
per_word_topics	If the model is desired to be found as a list of topics sorted in descending order of the most probable topics for each word, the value True is given.	True

2.3. Topic Modeling Latent Dirichlet Allocation (LDA)

Topic modeling is an unsupervised method that enables the discovery of latent semantic structures known as topics in textual data [23, 24]. In this study, Latent Dirichlet Allocation (LDA) method, one of the topic modeling methods, was used. LDA is an unsupervised topic modeling method that detects topics in textual data in a way that best provides maximum likelihood or posterior probability [13, 25].

LDA aims to perform an optimization over the probabilities *P*(word | topic) and *P*(topic | document). Topic modeling is performed by randomly matching a predetermined number of K topics with documents. The topic that best represents each document is determined and learning is achieved by determining the best words that represent each topic. In the first step, topics and documents are randomly matched, and then various statistics are examined to calculate the meaningfulness of these assignments. At this stage, local statistics are calculated for which words are associated with each topic in each document, and the rate at which each word is associated with the topics in all documents is calculated with global statistics. The graphical representation of LDA is illustrated in Figure 2.

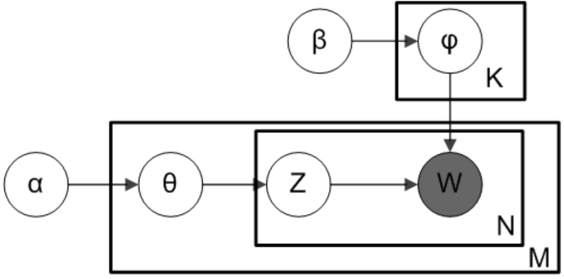


Figure 2. LDA model architecture

As illustrated in Figure 2, M is the total number of documents, K is the number of hidden topics, α and β are Dirichlet parameters. Θ shows the probability of finding the topics in the document, and φ shows the distribution of words in the topics. Z represents the topics assigned to each word. The LDA algorithm assumes that a word belongs to a topic and a document belongs to at least one topic. A document may also belong to more than one subject. Due to this multiplicity, the Dirichlet distribution Dirt(α) is needed and at this point the Dirichlet parameter α is used. Therefore, the value of the Dirichlet parameter α directly affects the result of LDA. The selection of these parameters is very important for the efficient operation of the algorithm, topic selection and distribution of topics.

In LDA, the number of topics K is determined by the coherence value [26]. The coherence value measures the similarity between words. It is very important to model the system with an appropriate number of topics. Therefore, the K value with the highest value among the coherence values calculated for certain topic numbers is selected as the topic number of the model.

2.3.1. Parameter settings

In this study, the LDA model was developed on the Google Colobaroty Pro+ platform using the Gensim library in the Python programming language. In order to find the model with the appropriate coherence value, models were created between 2 and 30 topics and their coherence values were compared. Accordingly, the parameters used to obtain the highest coherence score are presented in Table 1 with their explanations.

As demonstrated in Table 1, α and β parameter values with the highest coherence scores are "symmetric". According to these parameter values, the optimum number of topics was found to be 28. Accordingly, the coherence score distribution in determining the optimum number of topics is illustrated in the graph in Figure 3.

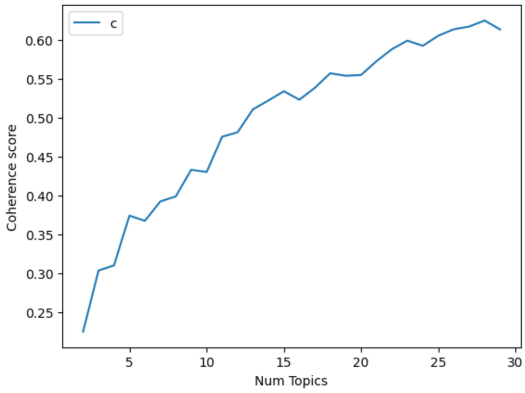


Figure 3. Num topics and coherence score

2.4. Louvain Community Detection Algorithm

A group of entities in a dataset that are more closely related to each other is called a community [27]. The intense interaction between entities can be measured by similarity or distance. Community detection aims to uncover groups of nodes that are densely related to each other. In this study, the Louvain algorithm was used to detect communities of textual data through network analysis. Louvain algorithm is used in many application areas due to its fast convergence properties, high modularity and hierarchical partitioning [28, 29]. The Louvain algorithm, one of the unsupervised algorithms, is suitable for use in large networks thanks to its ability to recreate communities.

The Louvain algorithm is a hierarchical method. This algorithm partially uses greedy optimization, and includes an additional step of rebuilding communities to improve usability across the wider network. Each node is initially considered as a community. The modularity gain, ΔQ, for each node is calculated by adding the node to its neighboring communities. The node is added to the community with the highest modularity gain and removed from the old community it belongs to, or remains in the old community in case there is no gain. Louvain applies this procedure for each node in successive iterative steps until no improvement occurs. Indicates the end of step one, when there is no improvement and the iterations are stopped. Step 2 involves creating a new network using the communities found at the end of step 1. In the new network, connections within a community are shown as self-loops, and connections between communities are shown as weighted connections. Then step 1 is applied to the newly established network again and this process continues until the community structure remains stable.

Equation 1 shows the mathematical equation of the Louvain algorithm. In this equation, ΔQ represents the modularity value, $\sum in$ represents the total value of weights in the collection C, and $\sum tot$ represents the total value of weights of edges connected to vertices in C. In addition, the total weight of the edges connected to vertex i is expressed as ki, the total value of the weights of the edges between vertex i and community C is expressed as $ki{,}in$ and finally the total value of the weights of all the edges in the network is expressed as m.

$$\Delta Q = \left[\frac{\sum_{in} + 2k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{\text{in}}}{2m} - \left(\frac{\sum_{\text{tot}}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \quad (1)$$

In this study, the communities detected with the Louvain algorithm and the centrally determined words are presented in the Findings section with a network graph. In addition, in order to determine the popularity of the detected communities in the field of bioinformatics, the distribution of the communities by year and number of views was analyzed and given in graphical form.

3. RESULTS AND DISCUSSION

In this study, topic modeling and community detection were performed using the "title" and "tag" data of posts related to bioinformatics on the Stack Overflow platform. LDA model was developed for topic modeling and topics were determined after determining the optimum parameter values. According to the data obtained from Stack Overflow platform, we found that there are 28 ideal topics related to bioinformatics and 10 descriptive words for each topic. The keywords, topic names and rates related to the identified topics are presented in detail in Table 2.

As demonstrated in Table 2, descriptive keywords were determined for each topic and given as the ratio of the topic to which each document was assigned in the entire corpus. The ratios given in Table 2 are listed in decreasing order according to their percentages. Accordingly, the 5 most important topics in the field of bioinformatics were found to be "Script Mapping", "Genomic Data Management", "Generating Mutation Formats", "Genome Assembly Methods", "Human Metagenomic Analysis". When the ratios were examined, the 5 topics with the least importance level were found to be "Mapping and Removing Patterns", "Gene Expression and Function", "Gene Set Enrichment Analysis (GSEA) Test and Analysis", "Genetic Residue Analysis", "Error Handling in Data Models".

The graphical visualization of the topics was performed using the pyLDAvis tool for the distribution of 28 topics detected with the LDA model. In the graph created with pyLDAvis, the keywords belonging to the topics were sorted in decreasing order according to the topic-specific probability (λ = 1) and are illustrated in Figure 4. Principal Component Analysis (PCA) was performed for the 28 topics identified with LDA and are shown as PC1 and PC2 in Figure 4. The top 30 words related to the topic are given in the display of the topics. Accordingly, the blue area given in the bar graph indicates the frequency over the entire corpus, and the red area indicates the frequency over the documents where the selected topic was detected. As an example, the words belonging to Topic 1 are illustrated in Figure 4. The graphs related to other topics are presented in Appendix 1.

As a result of the LDA model developed in this study, it was seen that a wide variety of topics in the field of bioinformatics were asked by bioinformatics developers. In order to understand the basic knowledge areas of researchers who do programming/code development in the field of bioinformatics, the topics discovered with the LDA model were divided into categories. Accordingly, the topics determined with the LDA model are given according to categories and rates (Table 3).

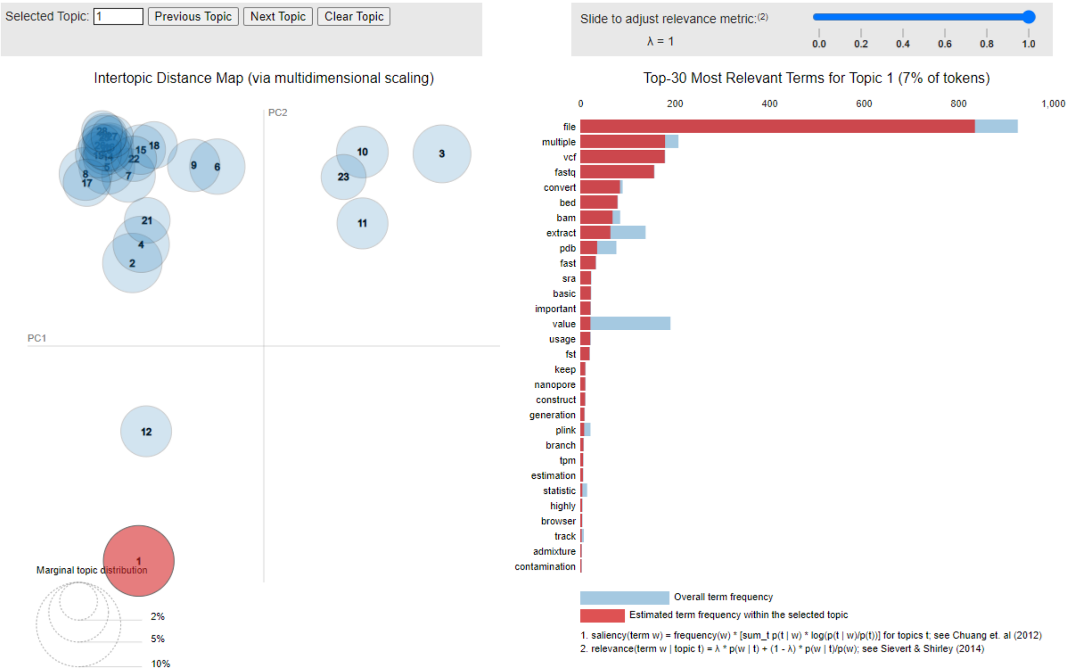
As demonstrated in Table 3, the prominent titles in the field of bioinformatics are "Genomic Research and Analysis", "Sequence and Alignment Tools", "Computational Biology and Bioinformatics Techniques" and "Programming and Data Management" in order of priority according to the total rate.

In the field of bioinformatics, we detected 7 communities according to the community analysis with the Louvain algorithm using the "tags" data of the posts on the Stack Overflow platform. The communities and the keywords that make up the tags are presented in Table 4.

The prominent keywords in each community are presented in Table 4. In order to create a more understandable framework about the popularity of these prominent communities in the field of bioinformatics, the data obtained from Stack Overflow is summarized in descending order according to the "view count" variable and illustrated in Figure 5.

Table 2. Topics and topic information determined by LDA

Topic Number	Words and weights	Topic Name	Topic Rate
1	0.091*"script" + 0.080*"low" + 0.065*"mapping" + 0.054*"error" + 0.052*"interpret" + 0.049*"value" + 0.031*"bash_script" + 0.030*"plant" + 0.028*"report" + 0.027*"total"	Script Mapping	%7
2	0.121*"reference" + 0.113*"create" + 0.085*"file" + 0.068*"site" + 0.058*"index" + 0.052*"filter" + 0.052*"genome" + 0.047*"time" + 0.043*"know" + 0.032*"memory"	Genomic Data Management	%5
3	0.169*"generate" + 0.134*"format" + 0.093*"mutation" + 0.079*"long" + 0.052*"transcript" + 0.036*"common" + 0.033*"quality" + 0.031*"uniprot" + 0.013*"refseq" + 0.011*"form"	Generating Mutation Formats	%4.7
4	0.169*"genome" + 0.097*"assembly" + 0.090*"gene" + 0.078*"method" + 0.073*"name" + 0.037*"merge" + 0.035*"genbank" + 0.034*"organism" + 0.025*"particular" + 0.025*"obtain"	Genome Assembly Methods	%4.4
5	0.096*"column" + 0.086*"design" + 0.054*"enrichment" + 0.053*"match" + 0.049*"motif" + 0.045*"metagenomic" + 0.040*"human" + 0.036*"solve" + 0.034*"row" + 0.033*"reproduce"	Human Metagenomic Analysis	%4.4
6	0.212*"result" + 0.089*"bacterial" + 0.084*"annotation" + 0.056*"table" + 0.043*"genome" + 0.041*"search" + 0.037*"block" + 0.034*"value" + 0.026*"style" + 0.026*"clinical"	Bacterial Genome Value Analysis	%4.3
7	0.201*"get" + 0.100*"database" + 0.065*"try" + 0.051*"genomic" + 0.046*"effect" + 0.045*"coordinate" + 0.044*"blast" + 0.028*"public" + 0.028*"conservation" + 0.025*"con"	Public Genomic Databases	%4
8	0.231*"protein" + 0.111*"calculate" + 0.063*"predict" + 0.056*"cluster" + 0.045*"bioinformatic" + 0.040*"visualize" + 0.035*"software" + 0.033*"snakemake" + 0.030*"nucleotide" + 0.028*"interaction"	Protein Interaction Prediction	%3.9
9	0.198*"variant" + 0.095*"specific" + 0.087*"identify" + 0.070*"genotype" + 0.054*"code" + 0.047*"loop" + 0.038*"measure" + 0.036*"help" + 0.029*"final" + 0.020*"annotate"	Identifying Genotype Variants	%3.9
10	0.071*"string" + 0.062*"order" + 0.049*"core" + 0.047*"status" + 0.046*"retrieve" + 0.046*"book" + 0.045*"locus" + 0.027*"pymol" + 0.026*"record" + 0.022*"phase"	String Database and Genomic Locus	%3.8
11	0.179*"plot" + 0.078*"python" + 0.077*"group" + 0.074*"package" + 0.064*"deseq" + 0.049*"show" + 0.037*"object" + 0.034*"feature" + 0.031*"selection" + 0.020*"copy"	Feature Selection and Plotting	%3.6
12	0.172*"cell" + 0.113*"single" + 0.109*"sample" + 0.062*"build" + 0.057*"extract" + 0.056*"download" + 0.049*"compare" + 0.032*"data" + 0.027*"normal" + 0.027*"type"	Extracting and Analyzing Single Cell Samples	%3.6
13	0.096*"structure" + 0.091*"gene" + 0.090*"protein" + 0.074*"text" + 0.049*"parse" + 0.048*"pdb" + 0.046*"information" + 0.041*"express" + 0.029*"additional" + 0.023*"symbol"	Gene and Protein Structure Analysis	%3.6
14	0.343*"analysis" + 0.058*"input" + 0.056*"fail" + 0.046*"perform" + 0.039*"high" + 0.029*"set" + 0.029*"significance" + 0.027*"understand" + 0.024*"polish" + 0.018*"methylation"	Analysis of High-Performance Data	%3.5
15	0.104*"network" + 0.101*"position" + 0.053*"comparison" + 0.048*"sample" + 0.045*"problem" + 0.042*"deletion" + 0.041*"visualisation" + 0.037*"contact" + 0.029*"normalization" + 0.025*"replicate"	Sample Analysis in Network Problems	%3.5
16	0.094*"region" + 0.090*"set" + 0.081*"gene" + 0.080*"list" + 0.050*"coverage" + 0.036*"peptide" + 0.036*"seq" + 0.033*"calculation" + 0.031*"start" + 0.030*"similarity"	Peptide Sequences Coverage Analysis	%3.3
17	0.219*"output" + 0.098*"score" + 0.081*"issue" + 0.053*"large" + 0.038*"homer" + 0.038*"salmon" + 0.036*"aligner" + 0.027*"scale" + 0.026*"short" + 0.024*"imputation"	Large Scale Sequence Alignment Analysis	%3.1
18	0.393*"sequence" + 0.093*"alignment" + 0.092*"tool" + 0.040*"link" + 0.037*"tree" + 0.031*"correlation" + 0.027*"prediction" + 0.026*"chain" + 0.019*"compare" + 0.019*"multiple"	Sequence Alignment Tools	%3.1
19	0.126*"matrix" + 0.085*"location" + 0.077*"length" + 0.074*"amino_acid" + 0.051*"rsid" + 0.047*"dump" + 0.043*"allele" + 0.038*"quantify" + 0.026*"open" + 0.024*"include"	Amino Acid Matrix Analysis	%3
20	0.129*"number" + 0.069*"header" + 0.062*"module" + 0.049*"change" + 0.049*"way" + 0.043*"study" + 0.039*"threshold" + 0.033*"variable" + 0.033*"determine" + 0.031*"upset"	Determining Changes in Thresholds	%2.9
21	0.270*"different" + 0.108*"base" + 0.085*"add" + 0.047*"contain" + '0.037*"check" + 0.030*"chromosome" + 0.029*"question" + 0.026*"post" + 0.024*"omic" + 0.023*"performance"	Different Base Combinations	%2.9
22	0.346*"datum" + 0.087*"way" + 0.072*"good" + 0.051*"look" + 0.045*"miss" + 0.042*"synthetic" + 0.029*"repeat" + 0.020*"chip" + 0.020*"algorithm" + 0.014*"access"	Synthetic Data and Algorithms	%2.9
23	0.408*"file" + 0.088*"multiple" + 0.087*"vcf" + 0.076*"fastq" + 0.041*"convert" + 0.038*"bed" + 0.033*"bam" + 0.031*"extract" + 0.017*"pdb" '+ 0.016*"fast"	File Format Conversion	%2.8
24	0.234*"read" + 0.103*"give" + 0.076*"count" + 0.072*"map" + 0.042*"remove" + 0.038*"pair" + 0.035*"split" + 0.034*"unique" + 0.029*"pattern" + 0.025*"exact"	Mapping and Removing Patterns	%2.7
25	0.294*"gene" + 0.107*"expression" + 0.068*"function" + 0.043*"classification" + 0.039*"size" + 0.032*"specific" + 0.029*"label" + 0.028*"limma" + 0.025*"differential" + 0.023*"insertion"	Gene Expression and Function	%2.7
26	0.184*"run" + 0.153*"test" + 0.074*"gsea" + 0.069*"call" + 0.053*"produce" + 0.048*"difference" + 0.037*"profile" + 0.031*"support" + 0.031*"soft" + 0.017*"bcftool"	Gene Set Enrichment Analysis (GSEA) Test and Analysis	%2.5
27	0.260*"find" + 0.060*"population" + 0.058*"distance" + 0.057*"residue" + 0.045*"line" + 0.039*"installation" + 0.038*"pipeline" + 0.035*"command" + 0.028*"return" + 0.027*"genetic"	Genetic Residue Analysis	%2.5
28	0.211*"error" + 0.128*"dataset" + 0.071*"mean" + 0.062*"process" + 0.052*"model" + 0.051*"work" + 0.029*"category" + 0.024*"distinguish" + 0.023*"properly" + 0.023*"gsva"	Error Handling in Data Models	%2.3



Selected Topic: 1

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 4. Graphical representation of the developed LDA model in pyLDAvis

Table 3. Categorization of topics determined according to the LDA model

Category	Topic Name	Topic Rate	Total Rate
Genomic Research and Analysis	Genome Assembly Methods	%4.4	%28,7
	Human Metagenomic Analysis	%4.4	
	Bacterial Genome Value Analysis	%4.3	
	Public Genomic Databases	%4	
	Identifying Genotype Variants	%3.9	
	Gene Expression and Function	%2.7	
	Gene Set Enrichment Analysis (GSEA) Test and Analysis	%2.5	
	Genetic Residue Analysis	%2.5	
Sequence and Alignment Tools	Extracting and Analyzing Single Cell Samples	%3.6	%24,5
	Peptide Sequences Coverage Analysis	%3.3	
	Large Scale Sequence Alignment Analysis	%3.1	
	Sequence Alignment Tools	%3.1	
	Amino Acid Matrix Analysis	%3	
	Different Base Combinations	%2.9	
	File Format Conversion	%2.8	
	Mapping and Removing Patterns	%2.7	
Computational Biology and Bioinformatics Techniques	Protein Interaction Prediction	%3.9	%24,8
	String Database and Genomic Locus	%3.8	
	Feature Selection and Plotting	%3.6	
	Gene and Protein Structure Analysis	%3.6	
	Analysis of High-Performance Data	%3.5	
	Sample Analysis in Network Problems	%3.5	
	Synthetic Data and Algorithms	%2.9	
Programming and Data Management	Script Mapping	%7	%21,9
	Genomic Data Management	%5	
	Generating Mutation Formats	%4.7	
	Determining Changes in Thresholds	%2.9	
	Error Handling in Data Models	%2.3	

Table 4. Communities detected with the Louvain algorithm

Community Number	Keywords
1	['hashing', 'k-mer', 'nanopore', 'genome', 'positive-selection', 'ngs', 'simulated-data', 'assembly', 'de-bruijn-graphs', 'long-reads', 'genome-sequencing', 'canu', 'star', 'minion', 'barcode', 'adapter', 'trimming', 'metagenome', 'qc', 'quality-control', 'pacbio', 'scaffold', 'crispr', 'khmer', 'rna-alignment', 'rna-splicing', 'albacore', 'illumina', 'fast5', 'gmap', 'errors', 'microbiology', 'bacteria', 'data-preprocessing', '1d2-reads', 'blasr', 'minknow', 'porcine', 'base-calling', 'seqtk', 'cutadapt', 'genome-index', 'galaxy', 'read-correction', 'pcr', 'library', 'demultiplexing', 'paralogs', 'fastqc', 'microbial-genomics', 'microbiome', '16rrna', 'ecology', 'mice']
2	['hgp', 'phylogenetics', 'docking', 'fasta', 'sequence-alignment', 'codon', 'phylogeny', 'blat', 'blast', 'data-management', 'nucleotide-models', 'linux', 'protein-structure', 'motifs', 'python', 'pdb', 'shell', 'benchmarking', 'hidden-markov-models', 'sequence-analysis', 'text', 'api', 'rna-structure', 'covid-19', 'sars-cov-2', 'sequence-homology', 'quantification', 'taxonomy', 'thermodynamics', 'interactions', 'perl', '3d-structure', 'literature-search', 'homework', 'domains', 'virulign', 'awk', 'coverage', 'parsing', 'mtdna', 'heteroplasmy', 'protein-protein-interaction', 'blastp', 'structural-biology', 'looper', 'yaml', 'systems-biology', 'mauve', 'primer', 'software-usage', 'virus', 'pssm', 'psi-blast', 'duplications', 'bitscore', 'makeblastdb', 'bash', 'orf', 'bugs', 'orthologues', 'ete3', 'mrbayes', 'membrane-protein', 'beast', 'genometools', 'nomenclature', 'pymol', 'entrez', 'efetch', 'sequence', 'rosetta', 'pyranges', 'population-genetics', 'loop', 'epidemiology', 'coronavirus', 'molecular-clock', 'aws', 'synonymous-mutations', 'bioedit', 'phangorn', 'matrix', 'godon', 'molecular-dynamics', 'gromacs', 'lipid-bilayer', 'orthofinder', 'grep', 'e-utilities', 'seqio', 'multiple-sequence-alignment', 'ggtree', 'rdkit', 'biopython', 'substitution-model', 'haplotypenetwork', 'curl', 'seqkit', 'multi-fasta', 'parallel', 'command-line', 'segmasker', 'sed', 'populations', 'ragtag', 'for-loop', 'tsv', 'hiv', 'dna', 'whole-genome', 'substitution', 'amino-acids', 'distance', 'ancient-dna', 'dss', 'differential-methylation', 'alphafold', 'iqtree', 'brap', 'dnanexus', 'biobank', 'zsh', 'cpu', 'ksnp4', 'peptide']
3	['genotyping', 'gatk', 'snp', 'structural-variation', 'variant-calling', 'genomics', 'cancer', 'vcf', 'variation', 'snv', 'wgs', 'array-cgh', 'exome', 'gwas', 'imputation', 'liftover', 'terminology', 'indel', 'isoform', 'filtering', 'radseq', 'workflow-management', 'human', 'molecular-genetics', 'plink', 'cnv', 'mutations', 'somatic', 'non-coding', 'haplotypes', 'ld', 'vcftools', 'maf', 'copy-number', 'snp-chip', 'adam', 'pedigree', 'docker', 'wdl', 'genetics', 'phasing', 'figure-explanation', 'point-mutation', 'bcftools', 'circos', 'maftools', 'text-mining', 'nextflow', '1000genomes', 'allele-frequency', 'snpeff', 'tabix', 'chromosomes', 'qtls', 'variants', 'logistic-regression', 'hg19', 'deep-learning', 'polygenic-risk-score', 'pytorch', 'multi-allelic', 'wes', 'pathogenicity', 'variant-filtration', 'pbwt', 'tools', 'phred-scores', 'gnomad', 'rsid', 'dictionary', 'admixture', 'singularity', 'fst', 'mutect2', 'annovar', 'troubleshooting', 'dbnsfp']
4	['proteins', 'transcriptome', 'rna-seq', 'normalization', 'r', 'fpkm', 'microarray', 'gse', 'bioconductor', 'deseq2', 'differential-expression', 'modelling', 'statistics', 'networks', 'cell-line', 'computation', 'ebseq', 'rsem', 'scrnaseq', 'clustering', 'hts', 'flow-cytometry', 'subset', 'edger', 'methylation', 'rna', 'pathway', '10x-genomics', 'visualization', 'best-practice', 'data-download', 'combat', 'strandedness', 'gsea', 'biostrings', 'go', 'go-enrichment', 'single-cell', 'geoquery', 'spike-in', 'matlab', 'software-quality', 'gviz', 'ggplot2', 'salmon', 'pca', 'scran', 'linear-regression', 'ribosomal', 'batch-effects', 'heatmap', 'ercc', 'merge', 'cellranger', 'tximport', 'tx2gene', 'figure-reproduction', 'upsetr', 'noise', 'gene-expression', 'public-dataset', 'seurat', 'groupgo', 'clusterprofiler', 'mass-spectrometry', 'limma', 'umap', 'genomicranges', 'chromplot', 'tdtomato', 'secondary-structure', 'proteomics', 'cummerbund', 'design', 'fusions', 'rcurl', 'tpm', 'tmm', 'umi', 'geo', 'rpy2', 'features', 'splitseq', 'correlation', 'subread', 'featurecounts', 'sva', 'rstudio', 'wgcna', 'peptide-shaker', 'tmt', 'fold-change', 'maxquant', 'kallisto', 'stringtie', 'java', 'genomefeatures', 'p-values', 'sbml', 'cibersort', 'monocle', 'statsmodels', 'complexheatmap', 'gtex', 'prosite', 'coexpression', 'outlier', 'chi-square', 'log-likelihood', 'epigenetics', 'meta-analysis', 'qpcr', 'phyloseq', 'graphs', 'stringdb', 'sctransform', 'trinity', '10x', 'tsne', 'hdf5', 'cytoscape', 'unsupervised-learning', 'deconvolution', 't-test', 'cibersortx', 'wormbase', 'dimensionality-reduction', 'diffusion-map', 'biostatistics', 'inference', 'non-parametric', 'lme4']
5	['algorithms', 'reproducibility', 'git', 'machine-learning', 'chip-seq', 'peak-calling', 'homer', 'deeptools', 'encode', 'multi-omics', 'transcriptome-regulation', 'software-installation', 'ucsc', 'atac-seq', 'scipy', 'snakemake', 'slurm', 'trna', 'trnascan-se', 'hmmer', 'bioconda', 'sketch', 'macs2', 'jupyter', 'conda', 'anaconda', 'literature', 'busco', 'transcription-factors', 'windows', 'hpc', 'tensorflow', 'nmf', 'github', 'megax', 'distance-matrix', 'pairwise', 'neighbor-joining', 'pdist']
6	['drugs', 'ontology', 'sequence-annotation', 'ensembl', 'refseq', 'gencode', 'biomart', 'kegg', 'database', 'data-retrieval', 'public-databases', 'repeat', 'repeat-elements', 'transposable-elements', 'gene', 'conversion', 'identifiers', 'gene-ontology', 'data-mining', 'centromere', 'telomere', 'uniprot', 'webservice', 'edirect', 'functional-annotation', 'ratt', 'embl', 'repeatmasker', 'sbol', 'blastn', 'computational-biochemistry', 'nt', 'publishing', 'cheminformatics', 'small-molecules', 'icd-codes', 'vep', 'eggnog', 'plants', 'pangenome', 'grch37', 'tblastn', 'tsa', 'mane']
7	['human-genome', 'storage', 'file-formats', 'fastq', 'sam', 'bwa', 'read-mapping', 'reads', 'bed', 'format-conversion', 'reference-genome', 'samtools', 'software-recommendation', 'mpileup', 'pysam', 'impute2', 'htslib', 'bedtools', 'sambamba', 'gff3', 'gtf', 'bigwig', 'bioawk', 'cigar', 'genome-browser', 'mapq', 'mappability', 'text-processing', 'c++', 'interval', 'exon', 'igv', 'markduplicate', 'validation', 'sratoolkit', 'picard', 'genbank', 'emboss', 'seqret', 'cram', 'base-clipping', 'pybedtools', 'minimap2', 'hi-c', 'gff', 'multithreading', 'bowtie2', 'rseqc', 'chromosome-capture', 'indexing', 'pandas', 'cloud', 'gffutils', 'bgzip', 'phase', 'fastq-dump', 'dataframe', 'egf', 'sra', 'paf', 'json', 'ncbi', 'parameters', 'seaborn']

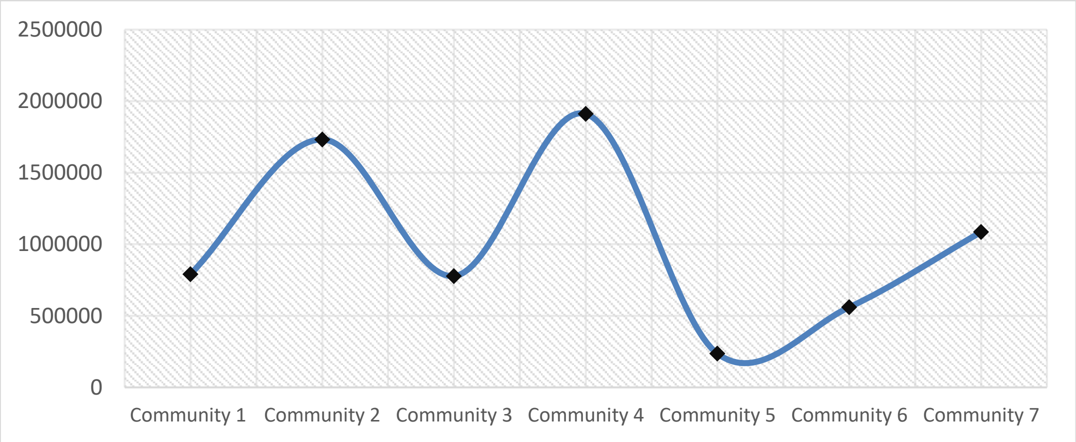


Figure 5. Most viewed communities by users

As illustrated in Figure 5, when the distribution of communities is examined, it is determined that Community 2 and Community 4 are more prominent compared to other communities. When we look at the distribution of communities over the years, we observe that Community 4 was prominent in 2017, 2018, and 2019, Community 2 in 2020, both Community 2 and Community 4 in 2021, Community 4 in 2022 and 2023, and finally Community 4 until March 2024 (Figure 6).

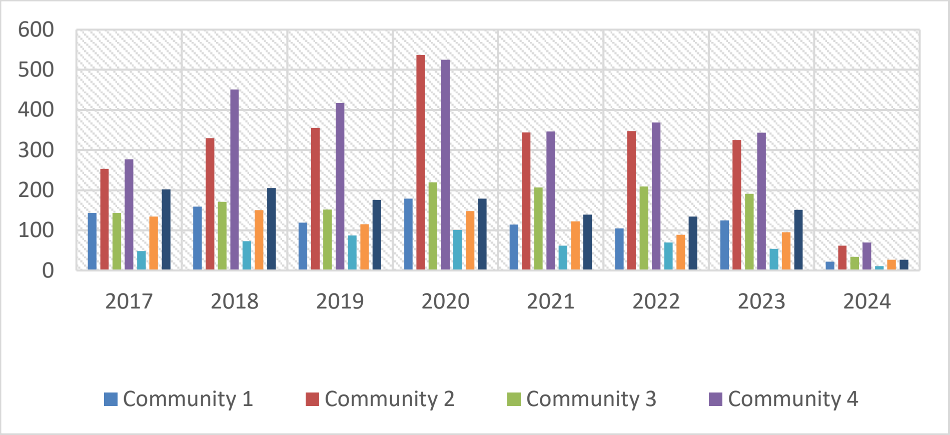


Figure 6. Distribution of communities by year

The analysis was carried out by taking into account the number of views in determining the degree centrality weights of the keywords related to the communities given in Table 4. Accordingly, the 100 most central words in the field of bioinformatics and their degree centrality values are presented in Table 5.

As demonstrated in Table 5, the first five prominent words in the field of bioinformatics on the Stack Overflow platform were determined to be “Python”, “R”, “phylogenetics”, “rna-seq”, and “sequence-alignment”. For these 100 central words, a network graph was created by adjusting the node sizes according to the centrality measurements (Figure 7).

Table 5. The top 100 most central words for communities

Keyword	Degree Centrality	Keyword	Degree Centrality	Keyword	Degree Centrality
python	0,4757	differential-expression	0,1517	quality-control	0,1011
R	0,4551	illumina	0,1517	chip-seq	0,0993
phylogenetics	0,4494	metagenome	0,1479	networks	0,0993
rna-seq	0,4157	read-mapping	0,1442	coverage	0,0993
sequence-alignment	0,3577	protein-structure	0,1442	long-reads	0,0974
ngs	0,3202	file-formats	0,1423	homework	0,0955
genome	0,2903	reference-genome	0,1423	proteomics	0,0955
gene	0,2734	clustering	0,1423	seurat	0,0936
phylogeny	0,2697	sars-cov-2	0,1404	perl	0,0918
fasta	0,2640	format-conversion	0,1348	algorithms	0,0899
sequence-annotation	0,2603	bed	0,1330	motifs	0,0899
sam	0,2566	wgs	0,1273	identifiers	0,0880
proteins	0,2547	ensembl	0,1255	plink	0,0861
genomics	0,2509	nanopore	0,1236	virus	0,0861
bioconductor	0,2472	machine-learning	0,1217	ncbi	0,0861
sequence-analysis	0,2434	bacteria	0,1199	methylation	0,0843
statistics	0,2360	mutations	0,1199	pathway	0,0843
vcf	0,2247	cancer	0,1180	cnv	0,0843
database	0,2191	gatk	0,1161	software-usage	0,0843
snp	0,2172	modelling	0,1161	gtf	0,0824
scrnaseq	0,2060	filtering	0,1161	awk	0,0824
fastq	0,2004	data-preprocessing	0,1161	pacbio	0,0805
public-databases	0,1985	reads	0,1142	taxonomy	0,0805
assembly	0,1929	data-download	0,1142	public-dataset	0,0805
variant-calling	0,1910	microarray	0,1124	literature-search	0,0787
bash	0,1835	bedtools	0,1124	multiple-sequence-alignment	0,0787
software-recommendation	0,1760	deseq2	0,1105	genotyping	0,0768
covid-19	0,1760	pdb	0,1105	genbank	0,0768
rna	0,1742	software-installation	0,1086		
blast	0,1685	errors	0,1086		
single-cell	0,1685	human-genome	0,1067		
gene-expression	0,1685	genetics	0,1067		
transcriptome	0,1610	gwas	0,1049		
visualization	0,1610	k-mer	0,1030		
samtools	0,1592	normalization	0,1030		
linux	0,1592	variation	0,1030		

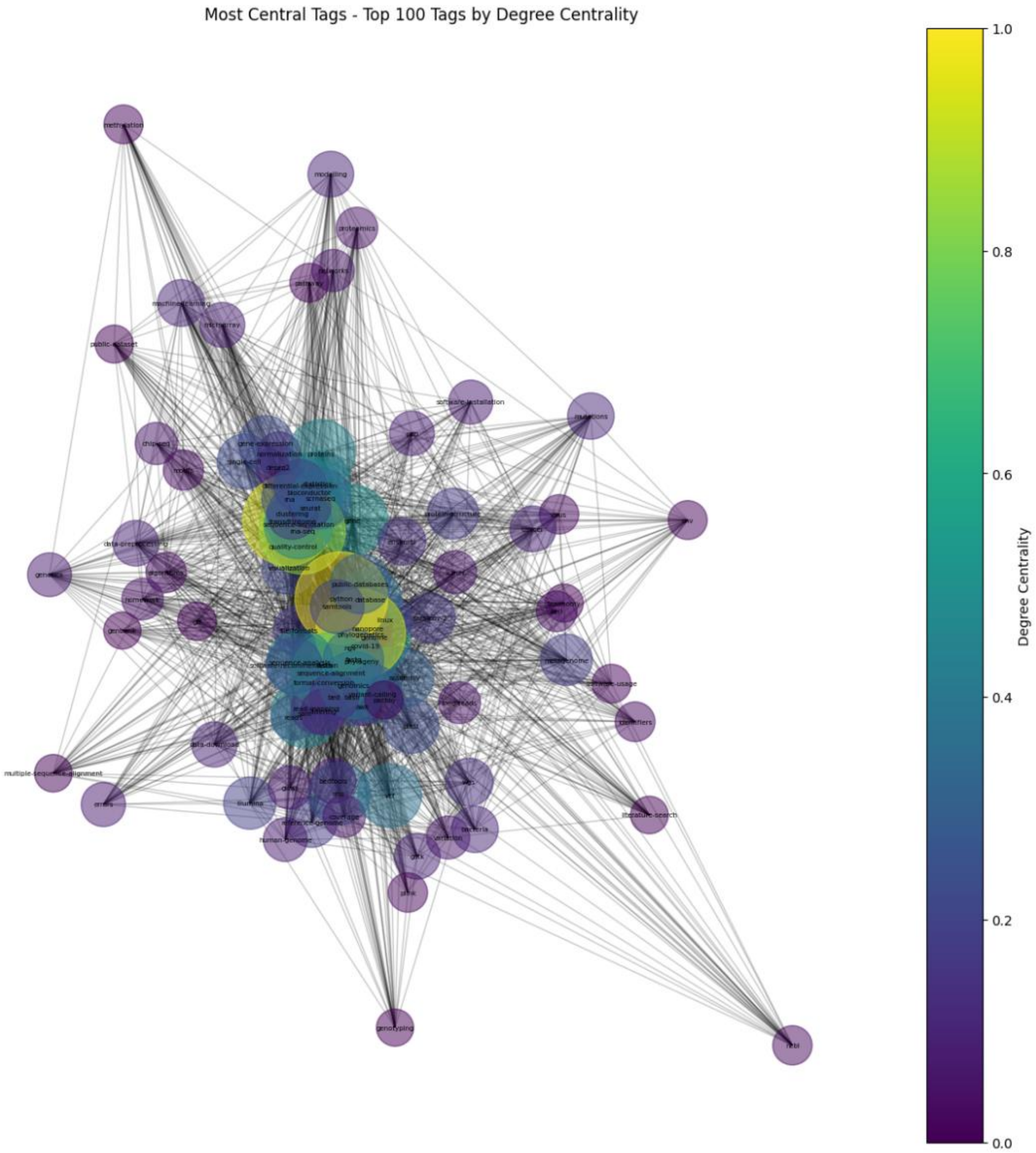


Figure 7. Network graph of the 100 most central words

As illustrated in Figure 7, the yellow filled circles are the most central words and the large sized circles have more viewing numbers. The lines between the words given in the graph show the relationship between the words and other words.

Developers in the field of bioinformatics focus on algorithms, artificial intelligence, machine learning, data mining techniques to analyze biological data and model biological processes [30, 31]. Developers often use languages such as Python, R, C, and Java to create workflows to make sense of various biological data [32]. This enables the discovery of new findings in bioinformatics research and the development of more precise treatment strategies in clinical applications [33-35].

According to the LDA analysis conducted in this study, we found that the questions most frequently asked by developers working in the field of bioinformatics were 28 separate topics. According to the findings obtained as a result of the analysis, we found that the topics were distributed under four broad basic categories: “Genomic Research and Analysis”, “Sequence and Alignment Tools”, “Computational Biology and Bioinformatics Techniques”, and “Programming and Data Management”. The topics obtained as a result of the analysis are presented in Table 6, compared with other studies in the literature.

As demonstrated in Table 6, the topics of “Gene Expression and Function”, “Protein Interaction Prediction”, “Gene and Protein Structure Analysis”, “Sample Analysis in Network Problems”, and “Genomic Data Management” are found in at least three more studies. Accordingly, we can say that these five topics are the most prominent topics in the field of bioinformatics. In addition, the topics of “Genome Assembly Methods”, “Large Scale Sequence Alignment Analysis”, “Sequence Alignment Tools”, “Feature Selection and Plotting”, and “Synthetic Data and Algorithms” are found in at least two studies.

Table 6. Comparison of the findings of this study with the literature

Topic Name	[15]	[21]	[22]	[36]
Genome Assembly Methods	X		X	
Human Metagenomic Analysis		X		
Bacterial Genome Value Analysis		X	X	
Public Genomic Databases	X			
Identifying Genotype Variants	X			
Gene Expression and Function	X		X	X
Gene Set Enrichment Analysis (GSEA) Test and Analysis				
Genetic Residue Analysis				X
Extracting and Analyzing Single Cell Samples		X		
Peptide Sequences Coverage Analysis		X		
Large Scale Sequence Alignment Analysis	X	X		
Sequence Alignment Tools	X	X		
Amino Acid Matrix Analysis				
Different Base Combinations				
File Format Conversion				
Mapping and Removing Patterns				
Protein Interaction Prediction	X	X	X	X
String Database and Genomic Locus				
Feature Selection and Plotting	X		X	
Gene and Protein Structure Analysis	X	X	X	X
Analysis of High-Performance Data	X			
Sample Analysis in Network Problems		X	X	X
Synthetic Data and Algorithms			X	X
Script Mapping				
Genomic Data Management	X		X	X
Generating Mutation Formats			X	
Determining Changes in Thresholds				
Error Handling in Data Models				

On the other hand, it is seen that the topics of “Public Genomic Databases”, “Identifying Genotype Variants”, “Genetic Residue Analysis”, “Extracting and Analyzing Single Cell Samples”, “Peptide Sequences Coverage Analysis”, “Analysis of High-Performance Data”, and “Generating Mutation Formats” were only mentioned in one study. Unlike other studies, we saw that the topics of “GSEA Test and Analysis”, “Amino Acid Matrix Analysis”, “Different Base Combinations”, “File Format Conversion”, “Mapping and Removing Patterns”, “String Database and Genomic Locus”, “Script Mapping”, “Determining Changes in Thresholds”, and “Error Handling in Data Models” were only included in our current study. Accordingly, we can say that these nine topics we identified represent the problems and difficulties that bioinformatics developers have encountered in recent years. In addition, unlike other studies and topics, we also revealed the areas in which program/code developers in the field of bioinformatics have technical difficulties.

In this study, we obtained important findings about seven communities found according to the community analysis with the Louvain algorithm and their time-dependent trends (Table 4, Figure 5, and Figure 6). Accordingly, we found that the 2nd and 4th communities depicted the main focus of developers in the field of bioinformatics. Accordingly, we found that the most popular tags within the scope of the 4th community were “proteins”, “transcriptome”, “rna-seq”, and “microarray”. In similar studies in the literature, we also found that these words define the areas researched for bioinformatics [15, 21, 22, 36]. The fact that these topics are prominent both in scientific research and in the questions asked by developers shows the importance of these topics in the field of bioinformatics, but also indicates that there are still unsolved problems in these areas.

In this study, according to the analysis of the most central words used in the bioinformatics field on the Stack Overflow platform, we found that the first five prominent words were “Python”, “R”, “phylogenetics”, “rna-seq”, and “sequence-alignment”. We also encountered the Python language in the words related to the “Feature Selection and Plotting” topic. These findings show us that the programming languages used by code developers in the bioinformatics field are Python and R in order of priority. These programming languages are frequently used platforms by developers due to both their practical use and their bioinformatics-specific libraries [32]. On the other hand, we also found that the prominent operating system for bioinformatics developers is “linux” (Table 5). We also saw that questions were asked about the file format for biological data. Among these, we found that the most prominent file formats are “fasta”, “sam”, “bed”, “gtf”, “vcf”, “fastq”, and “pdb”. It has been observed that questions about reading and processing these file formats in the programming language used by the developers have come to the fore. Programming languages, operating systems, and file formats show us the environments that developers use technically.

According to the findings obtained in this study, it was determined that the topics that code developers in the field of bioinformatics are most interested in are “next generation sequencing”, “genome”, “gene”, “phylogeny”, “proteins”, and “sequence”. These topics are also widely researched in scientific research [15, 37, 38]. In particular, it is seen that methods such as artificial intelligence and machine learning, deep learning, and natural language processing are used in terms of model development in these mentioned topics [39-44]. According to the findings obtained in this study, it was determined that machine learning-based models were used by bioinformatics code developers and they asked questions about the algorithms (Table 5). Therefore, this finding obtained within the scope of this study is a guiding finding for code developers who want to conduct scientific research and develop models in the field of bioinformatics.

6. CONCLUSION

The Stack Overflow platform and similar discussion platforms are very effective platforms for code developers to find solutions to their problems. The analysis of the data collected regarding the posts shared on these platforms can guide both researchers and practitioners in solving problems. Within the framework of the results obtained from the findings of this study, it can help both researchers who want to conduct scientific research in the field of bioinformatics and code developers who are looking for a solution to a bioinformatics problem to get to know the field of bioinformatics and develop a perspective on how they can contribute to the field. With this study, 28 basic topics and the first 100 most central words in the field of bioinformatics were identified. These findings can provide a guide for researchers to identify the problem in the field of bioinformatics and to focus on which point of the field they will focus on. Therefore, researchers can conduct research by prioritizing the most viewed topics and problems awaiting solutions based on the findings we obtained as a result of this study.

Practitioners can contribute to the development and innovation of the field by creating useful tools and applications to solve the dominant problems of code developers in the field of bioinformatics that our findings reveal. Bioinformatics code developer candidates who are new to the field can make a career in these fields by considering which areas have talent gaps and which topics and tools are popular. For example, the most viewed "Python", "R", "Perl", "Bioconductor", and "Biopython" by users provide important perspectives to practitioners on which development tools they should focus on. More supportive libraries, guidelines, auxiliary tools, or guide documents can be prepared for such development tools that are widely used by practitioners.

In this study, posts in the field of bioinformatics on the Stack Overflow platform between 2017-2024 (March) were analyzed using LDA topic modeling and the Louvain community finding algorithm. The findings of our study revealed the questions frequently asked by code developers in the field of bioinformatics under four main categories: "Genomic Research and Analysis", "Sequence and Alignment Tools", "Computational Biology and Bioinformatics Techniques", and "Programming and Data Management". We found that the most popular topics were "Gene Expression and Function", "Protein Interaction Prediction", "Gene and Protein Structure Analysis", "Sample Analysis in Network Problems", and "Genomic Data Management". We found that the topics "GSEA Test and Analysis", "Amino Acid Matrix Analysis", "Different Base Combinations", "File Format Conversion", "Mapping and Removing Patterns", "String Database and Genomic Locus", "Script Mapping", "Determining Changes in Thresholds", and "Error Handling in Data Models" represent the problems and challenges that bioinformatics developers have encountered in recent years. Our findings also showed that the development platforms most used by code developers are "Python", "R", "Perl", "Bioconductor", and "Biopython". We found that the topics that code developers are most interested in in the field of bioinformatics are "next generation sequencing", "genome", "gene", "phylogeny", "proteins", and "sequence".

In this study, only posts shared on the Stack Overflow platform were analyzed. Researchers who want to conduct a different research on this subject can analyze the data on a different platform (Kaggle, GitHub, etc.) with the methodology used in this study and compare it with the findings of this study. In this way, the results of this research can be further expanded.

DECLARATION OF ETHICAL STANDARDS

The author(s) of this manuscript declare that the materials and methods used in their studies do not require ethics committee approval and/or legal-specific permission.

AUTHORS' CONTRIBUTIONS

Gülbahar Merve ŞILBIR: Contributed to conceptualisation, methodology, writing, reviewing and editing. Performed the experiments and analyse the results.

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] Ramsden J., "Bioinformatics: An Introduction", ***Springer Nature***, fourth ed., Switzerland, (2023).

[2] Rastogi S.C., Rastogi P., Mendiratta N., "Bioinformatics: Methods and Applications-Genomics, Proteomics and Drug Discovery", ***PHI Learning Pvt. Ltd.***, fifth ed., Delhi, (2022).

[3] Satam H., Joshi K., Mangrolia U., Waghoo S., Zaidi G., Rawool S., Thakare R.P., Banday S., Mishra A.K., Das G., Malonia S.K., "Next-generation sequencing technology: current trends and advancements", ***Biology***, 12(7);997, (2023).

[4] Kitsou K., Katzourakis A., Magiorkinis G., "Limitations of current high-throughput sequencing technologies lead to biased expression estimates of endogenous retroviral elements", ***NAR Genomics and Bioinformatics***, 6(3), (2024).

[5] Lesk A., "Introduction to Bioinformatics", ***Oxford University Press***, fifth ed., United Kingdom, (2019).

[6] Hie B., Peters J., Nyquist S.K., Shalek A.K., Berger B., Bryson B.D., "Computational methods for single-cell RNA sequencing", ***Annual Review of Biomedical Data Science***, *3*(1);339-364, (2020).

[7] Topol E., "Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again", ***Basic Books***, United States, (2019).

[8] Tolani P., Gupta S., Yadav K., Aggarwal S., Yadav A.K., "Big data, integrative omics and network biology", ***Advances in Protein Chemistry and Structural Biology***, 127;127-160, (2021).

[9] Kashyap H., Ahmed H.A., Hoque N., Roy S., Bhattacharyya D.K., "Big data analytics in bioinformatics: architectures, techniques, tools and issues", ***Network Modeling Analysis in Health Informatics and Bioinformatics***, 5;1-28, (2016).

[10] Chen C., Wu Y., Li J., Wang X., Zeng Z., Xu J., Liu Y., Feng J., Chen H., He Y., Xia R., "TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining", ***Molecular Plant***, 16(11);1733-1742, (2023).

[11] Ahmed S.S., Wang S., Tian Y., Chen T.H.P., Zhang H., "Studying and recommending information highlighting in Stack Overflow answers", ***Information and Software Technology***, 172;107478, (2024).

[12] Gürcan F., Özyurt Ö., "Identification of trend topics discussed in Stackoverflow posts by word frequency analysis", ***GUSTIJ***, 11(2); 357-368, (2021).

[13] Blei D.M., Ng A.Y., Jordan M.I., "Latent dirichlet allocation", ***Journal of Machine Learning Research,*** 3; 993-1022, (2003).

[14] Alghamdi R., Alfalqi K., "A survey of topic modeling in text mining", ***Int. J. Adv. Comput. Sci. Appl.***, 6(1), (2015).

[15] Hahn A., Mohanty S.D., Manda P., "What's hot and what's not? Exploring trends in bioinformatics literature using topic modeling and keyword analysis", ***in: Proceedings of ISBRA,*** pp. 279-290, (2017).

[16] Fortunato S., Hric D., "Community detection in networks: A user guide", ***Physics Reports***, 659; 1-44, (2016).

[17] Rossetti G., Pappalardo L., Rinzivillo S., "A novel approach to evaluate community detection algorithms on ground truth", ***in: Proceedings of Complex Networks VII***, pp. 133-144, (2016).

[18] Blondel V.D., Guillaume J.L., Lambiotte R., Lefebvre E., "Fast unfolding of communities in large networks", ***J. Stat. Mech.***, P10008, (2008).

[19] Malliaros F.D., Vazirgiannis M., "Clustering and community detection in directed networks: A survey" ***Physics Reports***, 533(4); 95-142, (2013).

[20] Su X., Xue S., Liu F., Wu J., Yang J., Zhou C., Hu W., Paris C., Nepal S., Jin D., Sheng Q.Z., Yu P.S., "A comprehensive survey on community detection with deep learning", ***IEEE Transactions on Neural Networks and Learning Systems,*** 35(4); 4682-4702, (2024).

[21] Youssef A., Rich A., "Exploring trends and themes in bioinformatics literature using topic modeling and temporal analysis", ***in: 2018 IEEE Long Island Systems, Applications and Technology Conference***, pp. 1-6, (2018).

[22] Ebrahimi F., Dehghani M., Makkizadeh F., "Analysis of persian bioinformatics research with topic modeling", ***BioMed Research International***, 3728131, (2023).

[23] Papadimitriou C.H., Tamaki H., Raghavan P., Vempala S., "Latent semantic indexing: A probabilistic analysis", ***in: Proceedings of ACM Sigact-Sigmod-Sigart***, pp. 159-168, (1998).

[24] Blei D.M., "Probabilistic topic models", ***Communications of the ACM***, 55(4);77-84, (2012).

[25] Sayed A.H., "In Inference and Learning from Data: Inference", ***Cambridge University Press***, United Kingdom, (2023).

[26] Röder M., Both A., Hinneburg A., "Exploring the space of topic coherence measures", ***in: Proceedings of WSDM,*** pp. 399-408, (2015).

[27] Ma T., Liu Q., Cao J., Tian Y., Al-Dhelaan A., Al-Rodhaan M., "LGIEM: Global and local node influence based community detection", ***Future Generation Computer Systems***, 105;533-546, (2020).

[28] Orman G., Labatut V., Cherifi H., "Qualitative comparison of community detection algorithms", ***in: Proceedings of DICTAP,*** pp. 265-279, (2011).

[29] Salha-Galvan G., Lutzeyer J.F., Dasoulas G., Hennequin R., Vazirgiannis M., "Modularity-aware graph autoencoders for joint community detection and link prediction", ***Neural Networks***, 153;474-495, (2022).

[30] Langmead B., Salzberg S.L., "Fast gapped-read alignment with Bowtie 2", ***Nature Methods***, 9(4);357-359, (2012).

[31] Libbrecht M.W., Noble W.S., Machine learning applications in genetics and genomics", ***Nature Reviews Genetics***, 16(6);321-332, (2015).

[32] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Bruncher M., Perrot M., Duchesnay E., "Scikit-learn: Machine learning in Python", ***Journal of Machine Learning Research,*** 12;2825-2830, (2011).

[33] Cibulskis K., Lawrence M.S., Carter S.L., Sivachenko A., Jaffe D., Sougnez C., Gabriel S., Meyerson M., Lander E.S., Getz G., "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples", ***Nature Biotechnology***, 31(3); 213-219, (2013).

[34] Friedman A.A., Letai A., Fisher D.E., Flaherty K.T., "Precision medicine for cancer with next-generation functional diagnostics", ***Nature Reviews Cancer***, 15(12);747-756, (2015).

[35] Van Dijk D., Sharma R., Nainys J., Yim K., Kathail P., Carr A.J., Burdziak C., Moon K.R., Chaffer C.L., Pattabiraman D., Bierie B., Mazutis L., Wolf G., Krishnaswamy S., Pe'er D., "Recovering gene interactions from single-cell data using data diffusion", ***Cell***, 174(3);716-729, (2018).

[36] Gurcan F., Çağıltay N.E., "Exploratory analysis of topic interests and their evolution in bioinformatics research using semantic text mining and probabilistic topic modeling", ***IEEE Access***, 31480-31493, (2022).

[37] Qian X.B., Chen T., Xu Y.P., Chen L., Sun F.X., Lu M.P., Liu Y.X., "A guide to human microbiome research: study design, sample collection, and bioinformatics analysis", ***Chinese Medical Journal***, 133(15);1844-1855, (2020).

[38] Pereira R., Oliveira J., Sousa M., "Bioinformatics and computational tools for next-generation sequencing

analysis in clinical genetics", ***Journal of Clinical Medicine***, 9(1);132, (2020).

[39] Auslander N., Gussow A.B., Koonin E.V., "Incorporating machine learning into established bioinformatics frameworks", ***International Journal of Molecular Sciences,*** 22(6);2903, (2021).

[40] Shi J., "Machine learning and bioinformatics approaches for classification and clinical detection of bevacizumab responsive glioblastoma subtypes based on miRNA expression", ***Scientific Reports***, 12(1);8685, (2022).

[41] Wang L., Deng C., Wu Z., Zhu K., Yang Z., "Bioinformatics and machine learning were used to validate glutamine metabolism-related genes and immunotherapy in osteoporosis patients", ***Journal of Orthopaedic Surgery and Research,*** 18(1); 685, (2023).

[42] Barun M.N., Önder E., "Unlocking the multidisciplinary potential of data science: Insights from apriori analysis", ***Politeknik Dergisi***, 1-1, (2024).

[43] Akalın F., Yumuşak N., "Classification of exon and ıntron regions on dna sequences with hybrid use of SBERT and ANFIS approaches", ***Politeknik Dergisi***, 27(3), 1043-1053, (2024).

[44] Tokdemir G., "Using text mining for research trends in empirical software engineering", ***Politeknik Dergisi***, 24(3), 1227-1235, (2021).