Жамият ва инновациялар –
Общество и инновации –
Society and innovations
Journal home page:
https://inscience.uz/index.php/socinov/index

Using the Jaccard similarity method for recommendation system of books

Khabibulla MADATOV[1], Sapura SATTAROVA[2]

Urgench State University

ARTICLE INFO

Article history:
Received December 2023
Received in revised form
15 December 2023
Accepted 20 January 2024
Available online
25 February 2024

Keywords:
corpus,
token,
similarity of texts,
NLTK,
Jaccard algorithm,
set,
intersection of sets,
union of sets.

ABSTRACT

The main goal of pedagogy is to educate the young generation to become mature, knowledgeable and well-rounded individuals in all respects. In this regard, one of the main tasks of the education system is to form a culture of reading among young people, to provide them with textbooks and works of art suitable for their age and intellectual potential. But only if young readers read books suitable for their intellectual potential based on their age characteristics, their knowledge, spirituality, outlook and other positive aspects will develop. If the students do not read the works according to their potential, the reader will not be able to absorb fully the contents of the work he has read, the information in the book will "weight" him. As a result, the reader's desire to read begins to fade. Readers should not read literature that is shallow in content, incompatible with our national spirituality and values, moral standards, and may have a negative impact on the education of young people. Therefore, it is necessary to create a system of recommending works suitable for the intellectual potential of readers. This article examines the application of the Jaccard similarity method to the creation of appropriate reading lists for high school students. For this, a corpus is created on the basis of high-class literature textbooks, and this corpus is compared with literary works. Books with the highest similarity results are recommended for reading. The problem was fully solved on the basis of literature textbooks of 5th-11th grade students and works of art in the Uzbek language.

DOI: https://doi.org/10.47689/2181-1415-vol5-iss1-pp59-69

[1] Candidate of physical and mathematical sciences. Urgench State University. E-mail: habi1972@mail.ru
[2] Teacher, Urgench State University. Urgench, Uzbekistan. E-mail: sprsattarova@gmail.com

Жамият ва инновациялар – Общество и инновации –
Society and innovations
Issue – 5 № 1 (2024) / ISSN 2181-1415

Kitob tavsiya qilish tizimini yaratishda jakkard o‘xshashlik usulidan foydalanish

Kalit so‘zlar:
korpus,
token,
matnlar o‘xshashligi,
NLTK,
Jaccard algoritmi,
to‘plam,
to‘plamlar kesishmasi,
to‘plamlar birlashmasi.

ANNOTATSIYA

Ushbu maqolada Jaccard o‘xshashlik usulini yuqori sinf o‘quvchilari mutolaa qilishi uchun mos keladigan adabiyotlar ro‘yxatini tuzishga tatbiqi haqida fikrlar bayon etiladi.Qo‘yilgan masala 5-11-sinf o‘quvchilarining adabiyot darsliklari va o‘zbek tilidagi badiiy asarlar asosida to‘la tahlil qilingan.

Использование метода жаккардового подобия для создания системы рекомендаций книг

Ключевые слова:
корпус,
токен,
сходство текстов,
НЛТК,
алгоритм Жаккара,
множество,
пересечение множеств,
объединение множеств.

АННОТАЦИЯ

Основная цель педагогики – воспитать молодое поколение зрелыми, знающими и всесторонне развитыми личностями. В связи с этим одной из главных задач системы образования является: формирование культуры чтения среди молодежи, обеспечение ее учебниками и произведениями искусства, соответствующими ее возрасту и интеллектуальному потенциалу. Но только если юные читатели будут читать книги, соответствующие их интеллектуальному потенциалу и учитывающие их возрастные особенности, их знания, духовность, мировоззрение и другие положительные стороны будут развиваться. Если учащиеся будут читать произведения, не соответствующие их возможностям, они не смогут полностью усвоить содержание прочитанного произведения, и информация в книге будет их «утяжелять». В результате у читателя может угаснуть желание читать. Читателям не следует выбирать литературу, которая поверхностна по содержанию, несовместима с нашей национальной духовностью и ценностями, нравственными нормами и способна оказать негативное влияние на образование молодежи. Поэтому необходимо создать систему рекомендации произведений, соответствующую интеллектуальному потенциалу читателей. В данной статье рассматривается применение метода подобия Жаккара для создания подходящих списков чтения для старшеклассников. Для этого на основе учебников литературы высокого класса создается корпус и этот корпус сравнивается с литературными произведениями. Книги с наибольшим результатом сходства рекомендуются к прочтению. Проблема была успешно решена на основе учебников литературы для учащихся 5-11 классов и произведений искусства на узбекском языке.

INTRODUCTION.

Based on the idea of the President of the Republic of Uzbekistan that "New Uzbekistan starts from the threshold of the school, from the educational system", large-scale reforms are being implemented. In particular, in order to strengthen the legal basis for the development and support of reading culture in the Republic, 5 decisions of the President of the Republic of Uzbekistan and the Government were adopted in 2017–2020, and the development of reading culture was determined as a priority of the state policy.

Solving many problems in the field of natural language processing is based on methods for determining the similarity of texts. Today, text similarity algorithms are the main solution for plagiarism detection, document classification, data retrieval, summarization and a number of other tasks. This algorithm was created by the Swiss scientist Jaccard Paul in 1901, and with its help, issues of similarity between various documents and text files are being determined. Jaccard similarity is actively used in ecology, geobotany, molecular biology, bioinformatics, genomics, informatics and other fields. The Jaccard formula has long been the standard solution for similarity problems. Jaccard algorithm is easy to understand, it works without vectorization and not based on cosine similarity. Using Jaccard similarity, it is possible to determine the similarity between simple texts, sets of numbers, as well as complex types of text files. Jaccard similarity (also called Jaccard similarity coefficient or Jaccard index) is one of the algorithms used to determine the similarity between two sets. It tokenizes words and compares them through collections. It can be used to measure the similarity between two objects, for example two text files. In Python programming, Jaccard similarity is mainly used to measure the similarity between two sets or two asymmetric binary vectors. Mathematically, calculating Jaccard similarity simply takes the ratio of set intersection to set union.

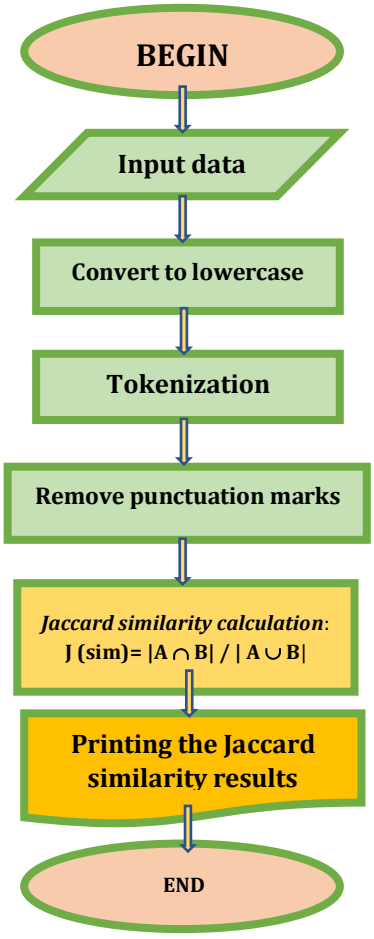
LITERATURE REVIEW

In this section, we discuss the work done on text similarity. Uzbek and foreign researchers are conducting a number of scientific researches on text processing in the field of natural language processing. Because the creation of modern applications related to natural language processing, conducting scientific research will undoubtedly be an important factor in the development of any low-resource language. In today's era of rapid development of computer technologies and the Internet, any user is faced with text processing processes such as searching for textual information, categorizing them, comparing and processing texts. One of the biggest challenges, especially when working with a large number of documents, is finding information that matches your interest, determining the degree of similarity between two works, and creating a glossary of large volumes of books. These problems can be easily solved by methods of determining the similarity of texts. In this paper [1], two new simple but effective similarity models are developed considering all user rating vectors to classify relevant neighborhoods and generate recommendations in less computational time. The source [2] proposes a new class of tests for homogeneity of two independent polynomial samples. Their tests are a natural extension of tests based on Jaccard's dissimilarity index, and the authors study the asymptotic powers of these tests. The authors of this paper [3] proposed a method to measure the similarity between words by using the Jaccard coefficient. Technically, they developed the Jaccard similarity measure with the Prolog programming language to compare the similarity between datasets. The authors of this paper [4] propose a min-max hash method, which cuts the hashing time in half, but it has a slightly smaller difference in pairwise Jaccard similarity estimation. In addition, the min-max hash estimator only

involves checking pairwise equality, so it is well suited for approximate nearest-neighbor searches. The authors of this paper [5] use three new similarity measures called Jaccard vector similarity, Jaccard cross-correlation, and the inner product of Jaccard Frobenius covariances for early motion detection by associating past features with future features. The purpose of this study [6] was to find the optimal value similarity for text mining. They used the Jaccard similarity method, a combination of Jaccard similarity, cosine similarity, and Jaccard similarity and cosine similarity. By combining the two similarities, it was achieved to increase the similarity value of the two names. The results of this study are that the cosine similarity method gave the best value of closeness or similarity compared to the Jaccard similarity and the combination of the two. The goal of the project proposed by the authors of this article [7] was to create a tool for analyzing large amounts of data related to large-scale social networks on the Internet. In particular, the project suggested creating a Map Reduce program to calculate the Jaccard similarity coefficient based on shared page changes among Wikipedia users. The program was then generalized to compute the Jaccard similarity between objects in any arbitrary column of the dataset co-occurring with another arbitrary column. The program was implemented in Java with the MapReduce programming technique. The authors of this article [8] determined the similarity of Uzbek texts using Jaccard and cosine similarity methods. The authors of the article written by Uzbek researchers [8, 9, 10] developed a cosine similarity detection algorithm based on TF-IDF for texts in the Uzbek language.

RESEARCH METHODOLOGY.

Jaccard similarity detection algorithm



3.2. Description of Jaccard similarity

The Jaccard similarity index is a measure of similarity between two data sets. Developed by Paul Jaccard, the index ranges from 0 to 1. The closer to 1, the more similar the two data sets are. If two data sets have exactly the same elements, their Jaccard similarity index is 1. Conversely, if they have no members in common, their similarity is 0.

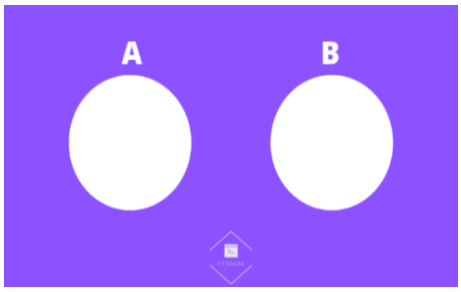


Fig. 1: Given data sets

The following examples show how to calculate the Jaccard similarity index for different data sets. Let us be given 2 sets A and B. (Fig. 1) The Jaccard similarity (or Jaccard index) of these sets is defined as formula (1)

$$J = |A \cap B| / |A \cup B| = |A \cap B| / |A| + |B| - |A \cup B| \quad (1)$$

We divide this formula into two components:

1.Intersection of sets. It calculates the embedded intersection between A and B, shown by the yellow area in the infographic below. (Fig. 2)

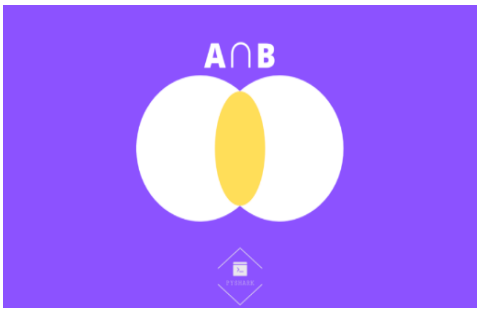


Fig. 2: Intersection of sets

2.Combination of sets. The denominator is actually a built-in combination of A and B, shown in yellow in the figure below. (fig-3)

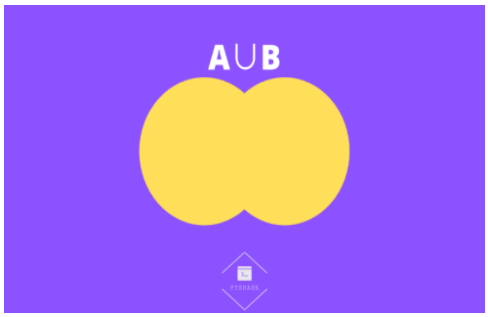


Fig. 3: Combination of sets

Using the Jaccard similarity formula, we can see that the similarity statistic is the ratio of the two visualizations above, where:

• If both sets are identical, for example A=1, 2, 3 and B=1, 2, 3, then their ***Jaccard similarity = 1.***

• If sets A and B have no elements in common, say A=1, 2, 3 and B=4, 5, 6, then ***their Jaccard similarity = 0***

• If sets A and B have elements in common, for example, A=1, 2, 3 and B=3, 4, 5, then their ***Jaccad similarity is 0≤ J (A, B) ≤1***. will have some value in the interval.

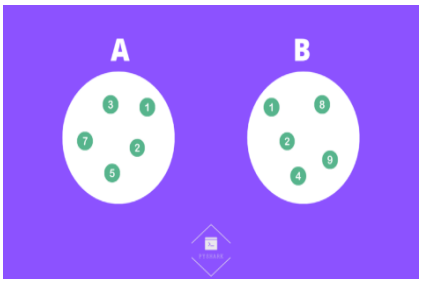


Fig. 4: Given set

3.3 Jaccard similarity calculation.

Let's look at two sets (fig-4)

A = {1, 2, 3, 5, 7}

B = {1, 2, 4, 8, 9}

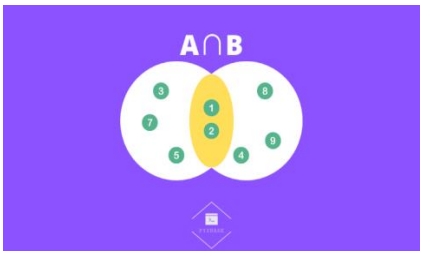


Fig. 5: Intersection of sets

Step 1:

As a first step, we need to find the intersection of sets A and B:

In this case: The intersection of sets is A∩B= {1,2}

Step 2:

In the second step, the union of sets A and B is found**.** Combination of sets:

A∪B= {1, 2, 3, 5, 7, 4, 8, 9}

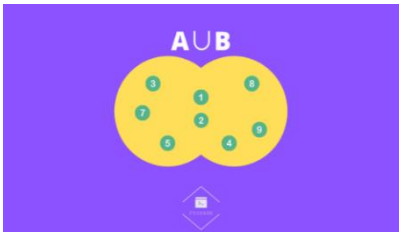


Fig. 6: Combination of sets

Step 3:

And the last step – the ratio of the number of elements in the intersection of the set to the number of elements of their union is obtained.

$$J(A,B) = \frac{n(A \cap B)}{n(A)+n(B)-n(A \cap B)} \quad (2)$$

From formula (2), we get the following result

$$J(A,B) = \frac{n(A \cap B)}{n(A)+n(B)-n(A \cap B)} = \frac{2}{5+5-2} = \frac{2}{8} = \frac{1}{4} = 0.25$$

RESULTS

The Python programming language uses the NLTK library to process text through the Jaccard algorithm, and it is performed in the following steps. The algorithm in Figure 6 is used to calculate it.

Example 1: Below, we determine the similarity between sets of numbers using Jaccard similarity

A = {0, 1, 2, 5, 6, 8, 9}, B = {0, 2, 3, 4, 5, 7, 9}

To calculate the Jaccard similarity between them, we first find the ratio of the intersection of the two sets to their union.

- The same number of elements in both sets: {0, 2, 5, 9} = 4
- Elements in both sets: {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} = 10
- Jaccard similarity: 4 / 10 = 0.4

So, The result of Jaccard similarity index is 0.4.

Example 2: We determine the similarity of the following 2 sets:

C = {0, 1, 2, 3, 4, 5}, D = {6, 7, 8, 9, 10}

To calculate the Jaccard similarity between them, we first find the ratio of the intersection of the two sets to their union.

- Same number of elements in both sets: {} = 0
- Elements in both sets: {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10} = 11
- **Thus, the result** of the Jaccard similarity: 0 / 11 = 0

The Jaccard similarity index turned out to be 0. This indicates that the two data sets have no common elements.

Example 3: Jaccard similarity for words

We can determine the Jaccard similarity index for a data set that contains characters as opposed to numbers.

E = {‘maktab’, ‘kitob’, ‘daftar’, ‘qalam’} E ={'school', 'book', 'notebook', 'pen'}

F = {‘qalam’, ‘ustoz’, ‘doska’, ’parta’} F ={'pen', 'teacher', 'blackboard', 'desk'}

To calculate the Jaccard similarity between them, we first find the ratio of the intersection of the two sets to their union.

- Same number of elements in both sets: {'qalam)'} = 1
- Elements in both sets: {‘maktab’, ‘kitob’, ‘daftar’, ‘qalam’, ‘ustoz’, ‘doska’, ‘parta’} = 7
- **Thus, the result** of the Jaccard similarity: 1 / 7= 0.142857

Example 4: We have already understood the calculation of Jaccard similarity from the above examples. Now, with the help of this similarity, we will determine the similarity of texts in Uzbek language. Alisher Navoi's "Lison-ut-Tair" and Farididdin Attar's "Mantiq-ut-Tair" were selected for this purpose. We have already understood the

calculation of Jaccard similarity from the above examples. Now, with the help of this similarity, we will determine the similarity of the texts in the Uzbek language. For this purpose, Alisher Navoi's "Lison-ut-tair" and Farididdin Attor's "Mantiq-ut-tair" were selected. The code written by the authors in Python program for calculating Jaccard similarity is also presented in detail.

```
import nltk
import string
import re
# Open for reading from a file in text mode
f1 = open("Navoiy.txt", "rt")
data1 = f1.read()
f2 = open("Attor.txt", "rt")
data2 = f2.read()
#Tokenization process
tokens1= nltk.word_tokenize(data1)
tokens2= nltk.word_tokenize(data2)
# Print tokens
print(tokens1)
print(tokens2)
#Determining the number of tokens
k1 = len(tokens1)
k2 = len(tokens2)
print('number of tokens in file 1 =',k1)
print( 'number of tokens in file 2=',k2)
A = set(tokens1)
B = set(tokens2)
def Jaccard_similarity(A, B):
 nominator = A.intersection(B)
 denominator = A.union(B)
 similarity = len(nominator)/len(denominator)
 return similarity
similarity = Jaccard_similarity(A, B)
print('1 st file: Lison-ut- Tayr; File 2: Mantiq-ut Tayr, jac.sim: = ',similarity)
```

The result is as follows:

```
================RESTART:D:\PYTHON\Jaccard-disser.py =============
number of tokens in file 1= 48817
number of tokens in file 2= 65047
1 st file:Lison-ut Tayr ; 2nd file:Mantiq-ut Tayr,jac. sim: = 0.22059697942680462
```

Table 1

List of sources that we investigated in this article (This list consists of school literature textbooks and works of fiction in various genres)

№	File name	Source name	Number of tokens	Number of unique words
1.	5.txt	Literature textbook for grade 5	67 168	20 667
2.	6.txt	Literature textbook for grade 6.	61 664	19 577
3.	7.txt	Literature textbook for grade 7	66 325	20 905
4.	8.txt	Literature textbook for grade 8	81 209	22 555
5.	9.txt	Literature textbook for grade 9	79 310	24 110
6.	10.txt	Literature textbook for grade 10	75 938	22 985
7.	11.txt	Literature textbook for grade 11	78 706	24 505
8.	lison.txt	Alisher Navoi's "Lison-ut-Tair".	48456	12845
9.	mantiq.txt	Farididdin Attar's "Mantiq-ut-Tair	64666	17943
10.	feruz.txt	A collection of ghazals by Muhammad Rahimkhan Feruz "Ne bolldi yorim kelmadi".	7216	2477
11.	shaytan.txt	Tahir Malik's novel "Shaytanat".	124837	30258
12.	garri.txt	Harry Potter by Joanna Kathleen,	406772	34494

Table 2

Comparison of school literature textbooks:
(In the process of defining file similarities, after comparing the text from the corresponding stage of the school bases with it, their Jaccard similarity was equal to 1, and in the other cases, the result was $0<Jac(sim)\leq 1$)

№	Files	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt
1.	**5.txt**	1.0	0.201	0.212	0.212	0.198	0.197	0.194
2.	**6.txt**	0.201	1.0	0.200	0.198	0.201	0.200	0.193
3.	**7.txt**	0.212	0.200	1.0	0.190	0.203	0.199	0.195
4.	**8.txt**	0.212	0.198	0.190	1.0	0.198	0.196	0.191
5.	**9.txt**	0.198	0.201	0.203	0.198	1.0	0.204	0.202
6.	**10.txt**	0.197	0.200	0.199	0.196	0.204	1.0	0.205
7.	**11.txt**	0.194	0.193	0.195	0.191	0.202	0.205	1.0

According to the final results, according to the sources listed in Table 1, the Jaccard similarity algorithm of the Uzbek language texts was achieved using the Python programming language, and the results listed in Tables 2-3 were achieved. From the results, we can see that the same source similarity is equal to 1, while the Jaccard similarity is equal to $0<Jac(sim)\leq 1$ in other cases. The essence of our article is that, using

the algorithm shown in Figure 1, we determined the Jaccard similarity of works of different genres recommended to schoolchildren of grades 5-11. Based on the maximum similarity result, we can conclude which class the material of the given genre corresponds to. Using this algorithm to create a list of books for a young readers' review gives effective results.

Table 3

Final results for book recommendation

№	Files	lison.txt	mantiq.txt	Feruz.txt	Shaytan.txt	Garri.txt
1.	**5.txt**	0.156	0.163	0.028	0.167	0.156
2.	**6.txt**	0.146	0.146	0.029	0.172	0.163
3.	**7.txt**	0.147	0.146	0.032	0.162	0.153
4.	**8.txt**	0.169	0.173	0.031	0.155	0.152
5.	**9.txt**	0.149	0.155	0.035	0.164	0.157
6.	**10.txt**	0.147	0.150	0.031	0.166	0.154
7.	**11.txt**	0.143	0.149	0.032	0.161	0.153

From this table, we can see that all the links except the feruz.txt file are suitable for students of school textbooks. Feruz. txt collection of ghazals, since it consists of classical vocabulary, we cannot recommend it for school students.

DISCUSSION

In this article, the importance of text similarity algorithms, their fields of application, the work carried out by Uzbek and foreign scientists in this regard, and the issue of applying the Jacquard similarity algorithm to texts in the Uzbek language were considered. The computational algorithm and Python code program for the Jaccard similarity method are clearly and simply explained by the authors. The main purpose of the article was to use the jacquard similarity method to recommend a list of books that match the intellectual potential of schoolchildren. The problem was completely solved in the case of the educational corpus consisting of school textbooks and various works of art. For this purpose, a corpus of 5-11th grade literature textbooks and several works written in Uzbek language was created. They are divided into tokens, a list of unique words is defined, and the similarity of texts is determined using Python program code based on the created algorithm. The software code presented in this article can be used to calculate the degree of similarity of any text document or literary source.

CONCLUSION

The main goal of pedagogy is to educate the young generation to become mature, knowledgeable, well-rounded individuals in all respects. From this point of view, one of the main tasks of the education system is to form a culture of reading among young people, to provide them with textbooks and works of art suitable for their age and intellectual potential [10]. The creation of terminological dictionaries [11] for the books that young people read will have an effective effect on their development to become perfect people who meet the requirements of the time. He should also create a model of

the universe and the solar system [12] in various programming environments, and regularly work on himself outside of class, reading works of art. Using the method proposed by the authors of this article, it is possible to develop the reading culture of students by creating a system of recommending works of art that match the intellectual potential of students.

In this article, a method based on the Jaccard algorithm for the similarity of texts in the Uzbek language using the Python program was considered. To calculate this similarity, the similarity between numbers, words and works was calculated. The method considered in the article is interpreted in a completely new way. Using the algorithm proposed by the authors, it is proposed to create a system that recommends suitable books for young readers. We believe that the information presented in the article will be a useful resource for literary critics analyzing works of art, students studying natural language processing, and any researcher

REFERENCES:

1. Bag S., Kumar S. K., Tiwari M. K. An efficient recommendation generation using relevant Jaccard similarity //Information Sciences. – 2019. – T. 483. – C. 53-64.

2. Ivchenko G. I., Honov S. A. On the Jaccard similarity test //Journal of Mathematical Sciences. – 1998. – T. 88. – C. 789-794.

3. Niwattanakul S. et al. Using of Jaccard coefficient for keywords similarity //Proceedings of the international multiconference of engineers and computer scientists. – 2013. – T. 1. – №. 6. – C. 380-384.

4. Ji J. et al. Min-max hash for Jaccard similarity //2013 IEEE 13th International Conference on Data Mining. – IEEE, 2013. – C. 301-309.

5. Fernando B., Herath S. Anticipating human actions by correlating past with the future with Jaccard similarity measures //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2021. – C. 13224-13233.

6. Zahrotun L. Comparison Jaccard similarity, cosine similarity and combined both of the data clustering with shared nearest neighbor method //Computer Engineering and Applications Journal. – 2016. – T. 5. – №. 1. – C. 11.

7. Bank J., Cole B. Calculating the Jaccard similarity coefficient with map reduce for entity pairs in Wikipedia //Wikipedia Similarity Team. – 2008. – T. 1. – C. 94.

8. S. M. X.Madatov, S.Matlatipov Kosinus o‘xshashlik va uning o‘zbek tili matnlariga tatbiqi haqida //O‘zMU xabarlari. – 2016. – T. 2. – №. 1.

9. Madatov K., Matlatipov S., Aripov M. Uzbek text's correspondence with the educational potential of pupils: a case study of the School corpus //arXiv preprint arXiv:2303.00465. – 2023.

10. Akhmedovich K. M., Beknazarovna S. S. METHODS OF CHECKING THE GIVEN LITERATURE ON THE INTELLECTUAL POTENTIAL OF SCHOOLCHILDREN

11. S.B.Sattarova, F.X.Bekchanova, A.K.Shermetov TERMINOLOGIK LUG‘AT YARATISH TEXNOLOGIYASI VA UNING TA’LIM TIZIMIDAGI AHAMIYATI //Academic research in educational sciences. – 2023. – T. 4. – №. 5. – C. 422-434.

12. Khodjinazarovna B. F., Kamaliddinovich S. A., Beknazarovna S. S. VISUALIZING THE SOLAR SYSTEM USING PYTHON AND ITS IMPORTANCE IN EDUCATION //International journal of advanced research in education, technology and management. – 2023. – T. 2. – №. 6.