УДК 81.322:003.26:004.056.55 DOI: 10.65599/III6781

ДАР БОРАИ АЛГОРИТМИ ТАҲИЯИ МАҶМУИ ХУСУСИИ N-ГРАММАҲО АЗ УНСУРҲОИ МАТН

М.Ҳ. Ғафуров, Иброҳими Ю.

Донишгоҳи техникии Тоҷикистон ба номи академик М.С.Осимӣ

Дар мақолаи мазкур раванди таҳлил ва коркарди иттилооти матнӣ тавассути таҳияи маҷмуи хусусии n-граммаҳо (униграммаҳо, биграммаҳо ва триграммаҳо) аз унсурҳои матн шарҳ дода шудааст. Усули мазкур дар соҳаҳои коркарди додаҳои забони табиӣ (NLP), моделсозии масъалаҳои соҳаи лингвистика ва дар самти криптография барои бадалсозии объектҳои матнӣ аҳамияти калидӣ дорад. Дар мақола алгоритмҳои ҷудокунии ин унсурҳо пешниҳод карда шудаанд, ки дар онҳо: униграммаҳо - ҳар як аломати матн ҳамчун воҳиди алоҳида, биграммаҳо ва триграммаҳо - мутаносибан ба гурӯҳҳои ду ё сеҳарфа тақсим кардани матн мебошад. Дар ҳолати норасоии аломатҳо дар охири объекти матнӣ символи махсуси «ҷои ҳеҷ» - (ø) илова карда мешавад. Барои татбиқи амалии ин алгоритмҳо забони барномасозии C# истифода шуда, намунаҳои матни барномаи он оварда шудаанд. Қисмати муҳими кор ба таҳияи маҷмуи хусусии n-граммаҳои беназир бахшида шудааст. Қайд карда мешавад, ки усули беҳтарин барои ин кор истифодаи «хеш-маҷмуъҳо» (HashSet) мебошад, зеро он аз лиҳози суръати иҷро ва соддагии татбиқ самараноктар аст.

Калидвожаҳо: *матн, алгоритм, униграмма, биграмма, триграмма, алифбои беназири хусусӣ.*

ОБ АЛГОРИТМЕ ФОРМИРОВАНИЯ ЧАСТНОГО МНОЖЕСТВА N-ГРАММ ИЗ ЭЛЕМЕНТОВ ТЕКСТА

М.Х. Гафуров, Иброхими Ю.

В данной статье рассматривается процесс анализа и обработки текстовой информации посредством формирования частного множества n-грамм (униграмм, биграмм и триграмм) из элементов текста. Данный метод имеет ключевое значение в таких областях, как обработка естественного языка (NLP), моделирование лингвистических задач и криптография для шифрования текстовых объектов. В статье представлены алгоритмы разделения этих элементов, в которых: униграммы - каждый символ текста как отдельная единица, биграммы и триграммы - делятся на группы по два или три символа соответственно. В случае нехватки символов для создания n-грамм в конце текстового объекта добавляется специальный символ «нулевого места» - (ø). Для практической реализации этих алгоритмов используется язык программирования C#, приводятся примеры его программного текста. Значительная часть работы посвящена созданию множество уникальных n-грамм. Отмечается, что наилучшим методом для этого является использование «хеш-множеств» (HashSet), поскольку он более эффективен с точки зрения скорости выполнения и простоты реализации.

Ключевые слова: *текст, алгоритм, униграмма, биграмма, триграмма, частный уникальный алфавит.*

ON THE ALGORITHM FOR DEVELOPING A PRIVATE SET OF N-GRAMS FROM TEXT ELEMENTS

M.H. Gafurov, Ibrohimi Yu.

This article describes the process of analyzing and processing textual information by developing a private set of n-grams (unigrams, bigrams, and trigrams) from text elements. This method is of key importance in fields such as Natural Language Processing (NLP), linguistic modeling, and cryptography for the transformation of text objects. The article presents algorithms for dividing these elements, in which unigrams - each text character as a separate unit - while bigrams and trigrams are divided into groups of two or three characters, respectively. If there are not enough characters to create n-grams, a special "zero place" character (ø) is added to the end of the text object. The C# programming language is used for the practical implementation of these algorithms, and examples of its program code are provided. A significant portion of the work is devoted to the creation of sets of unique n-grams. It is noted that the best method for this is the use of "hash sets" (HashSet), since it is more efficient in terms of execution speed and ease of implementation.

Keywords: *text, algorithm, unigram, bigram, trigram, private unique alphabet.*

Муқаддима

Таҳлили иттилооти матнӣ яке аз масъалаҳои муҳим дар соҳаи коркарди забони табиӣ (NLP-Natural Language Processing), моделсозии масъалаҳои забоншиносӣ, криптография ва криптотаҳлил мебошад. Дар корҳои [1, 2] усулҳои таҳияи алифбои беназир дар самти бадалсозии иттилоот, дар [3] раванди бадалсозии объекти матнӣ бо истифодаи символҳои забон, дар [4] татбиқи биграммаҳо дар бадалсозии объекти матнӣ дар [5, 6] татбиқи триграммаҳо дар бадалсозии иттиллоот, дар [7] масъалаи басомади ҳарфҳо ва дар [8] басомади биграммаҳо дар забони тоҷикӣ омӯхта шудаанд. Дар раванди тадқиқот дар самти муайянкунӣ ва ҷудокунии n-граммаҳо корҳои дар боло овардашударо истифода мекунем. Барои ин ва дигар масъалаҳои соҳаи криптография тадқиқи усулҳои муайянкунӣ ва ҷудокунии n-граммаҳо муҳим аст.

Алгоритмҳои ҳалли масъала

Равиши заминавии тақсимкунии матнҳо ба n-граммаҳо аз пай дар пай ба n символ ҷудо кардани онҳо мебошад, ки ҷои холӣ (пробел) ҳамчун символи алоҳида ба қайд гирифта мешавад. Барои ҷудокунӣ ва коркарди иттилооти матнӣ асосан аз униграмм, биграмм ва триграмм (n=1, 2, 3) истифода мекунанд. Униграммаҳо (n=1) дар шакли воҳиди хурдтарини матн тасвир мешавад ва замина барои сохтани моделҳои мураккаб, аз он ҷумла биграммаҳо (n=2), триграммаҳо (n=3) ва n-граммаҳо ба ҳисоб меравад.

Алгоритми ҷудокунии n-граммаҳо аз қисматҳои зерин иборат аст:

1. Объекти матнӣ дода, интихоб ё сохта мешавад;
2. Барои сохтани n-граммаҳо аз тарзҳои зерин истифода мекунанд:

а) объекти додашударо аз аввал сар карда бо баинобат гирифтани ҳамаи нишонаҳои аломатҳои оддӣ, махсус ва рақамҳо ба унсурҳои аз n-символ иборатбуда (n-граммаҳои тарафи чап) тақсим мекунанд. Агар дар охир барои ба n-грамма табдил додан символ кифоя набошад, ба он нишонаи мавқеи ҳеҷро (ø) ҳамроҳ мекунанд;

b) аз символҳои ҳарфӣ омода кардани n-граммаҳо кифоя аст, ки аломатҳои оддӣ, махсус ва рақамҳоро ба символҳои ҳарфии дар объект набуда иваз кунанд, яъне пешакӣ калид-ивазкунандаи аломатҳоро ба символҳо месозанд, ки он чунин намудро дорад:

$$K = \{a_i \rightarrow s_j \; ; \; i,j = 1,2,3,\ldots,n \},$$

ки a_i – оддӣ, махсус ва ё рақамҳо, s_j – символҳои ҳарфии ихтиёрии дар объект набуда;

Дар ҳолати n>1 будан (биграммаҳо, триграммаҳо, ...) дарозии матни додашуда санҷида шуда, ҳангоми ба n-граммаи пурра дар охири матн норасоӣ пайдо гардидан дар охири матн миқдори зарурии символи ҷои ҳеҷ (ø) илова карда мешавад.

3. Бо мақсади дар оянда истифода кардани маҷмӯи ҳамаи n-граммаҳои дар объект буда, аз онҳо маҷмуи n-граммаҳои беназирро месозанд.

А. Униграмма

Бигзор матни T ки аз пайдарпайии аломатҳои оддӣ, махсус ва рақамҳо иборат аст, дода шудааст:

$$T = c_1 c_2 c_3 \ldots c_m$$

дар инҷо c_i-символи i-ум, m-дарозии матн мебошад.

Маҷмуи хусусии униграммаҳои матни додашударо ба таври зерин ифода кардан мумкин аст:

$$U = \{u_i\}_{i=1}^{m}, u_i = c_i$$

Барои тақсими матни додашуда ба униграммаҳо амалҳои зерин иҷро карда мешаванд, ки барномаи он дар забони барномасозии C# таҳия карда шудааст ва қисматҳои асосии онро меорем:

1. Дохилкунии матни T

```
string text = richTextBox1.Text;
```

2. Аз хотираи фаврӣ ҷудо кардани блоки дохилкунии символҳо ва унсурҳо $U = [\,]$

```
string[] unigrams = new string[text.Length];
```

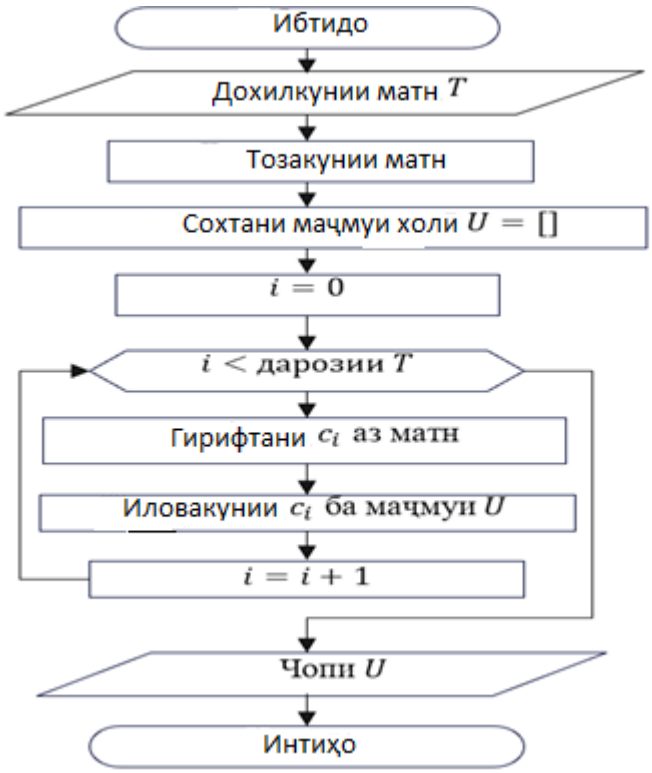
3. Барои ҳар як символи матн, аломатҳои махсус ва рақамҳо, ки ҳамчун c_i ба қайд гирифта мешавад, аз матни T иловакунии c_i ба маҷмуи U - униграммаҳо

```
for(int i = 0;i< text.Length; i++)
{
   unigrams[i] = text[i].ToString();
}
```

4. Сохтани маҷмуи символҳои беназир ва хориҷкунии банди 3

```
string[] uniqueUnigrams = unigrams.Distinct().ToArray();
for (int i = 0; i < uniqueUnigrams.Length; i++)
{
   listBox2.Items.Add(i + ". " + uniqueUnigrams[i]);
}
```

Блок-схемаи ҷудокунии униграммаҳоро аз матни додашуда дар шакли зерин тавсиф менамоем.



Расми 1 – Блок-схемаи ҷудокунии униграммаҳоро аз матни додашуда

В. Биграмма – пайдарпайии аз ду символи дар матн бударо меноманд. Он яке аз ҳолати хусусии n-грамма ($n = 2$) мебошад. Дар раванди бадалсозии объекти матнӣ биграммаҳои тарафи чапро истифода мекунанд, ки он дар ҷараёни бадалкунӣ ва аксбадалкунӣ мувофиқи мақсад аст.

Бигузор объекти матнии T ки аз n символҳо, аломатҳои имлоӣ, фаннӣ, махсус ва рақамҳо иборат аст, дода шудааст:

$$T = b_1 b_2 b_3 \dots b_n$$

дар инҷо b_i-символи i-ум, n-дарозии матн мебошад.

Ҳангоми ҷудокунии матн ба биграммаҳо шарти асосии он ҷуфт будани миқдори символҳо дар объекти додашуда мебошад. Агар миқдори символҳо тоқ бошад, дар охири объект илова кардани символи ҷои ҳеҷ (ø) ба мақсад мувофиқ аст.

Тақсим кардани объект ба биграммаҳо марҳилаҳои зеринро дарбар мегирад:

1. Дохилкунии матн

```
string text = richTextBox1.Text;
```

2. Муайянкунии дарозии объекти матнӣ $n = |T|$ ва дар ҳолати зарурат ба он илова кардани символи ҷои ҳеҷ (ø)

```
while (text.Length % 2 != 0)
{
text += "ø";
}
```

3. Агар n тоқ бошад $T' = T\text{ø}$ ва $n' = n + 1$, вагарна $T' = T$ ва $n' = n$, ки матни барнома дар банди 2 оварда шудааст.
4. Муайянкунии миқдори биграммаҳо $N = \frac{n'}{2}$

```
string[] bigramms = new string[text.Length / 2];
```

5. Сохтани маҷмуи биграммаҳои объекти додашуда $B_i = (T'_{2i-1}, T'_{2i}), i = 1, 2, \dots, N$

```
for (int i = 0; i < text.Length; i += 2)
```

```
{
string bigramm = text[i].ToString() + text[i + 1].ToString();
bigramms[i / 2] = bigramm;
}
```

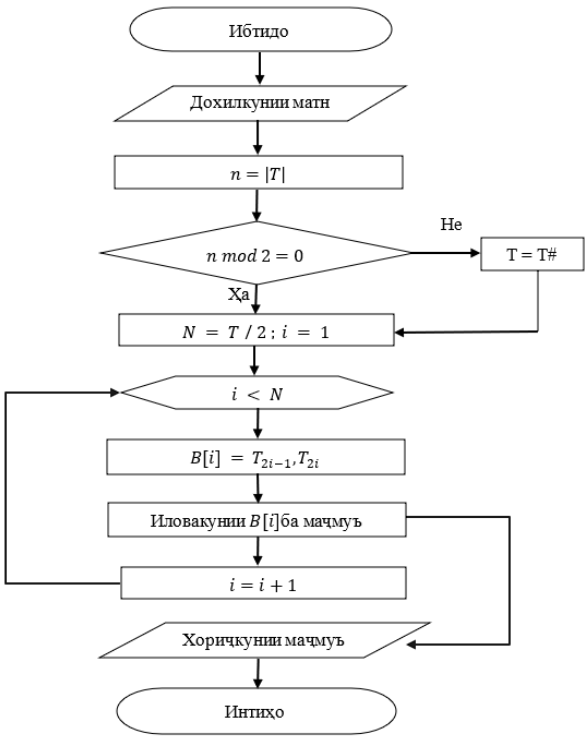
6. Натиҷаи ниҳоӣ дар намуди $Bigrams(T') = \{B_1B_2, \dots . B_N\}$ ва дар асоси он сохтани биграммаҳои беназир

```
string[] uniqueBigrams = new HashSet<string>(bigramms).ToArray();
```

7. Хориҷкунии маҷмуи биграммаҳои беназир

```
for(int i = 0; i < uniqueBigrams.Length; i++)
{
listBox2.Items.Add(i+". "+uniqueBigrams[i]);
}
```

Блок-схемаи ҷудокунии объекти матнира бо истифодаи биграммаҳо дар шакли зерин тавсиф кардан мумкин аст.



Расми 2 - Блок-схемаи ҷудокунии объекти матнира бо истифодаи биграммаҳо

С. Триграмма – пайдарпайии аз се символи дар матн буда мебошад. Он ҳолати хусусии n-грамма ҳангоми $n = 3$ будан аст.

Бигузор объекти матнӣ дар намуди зерин дода шудааст:

$$S = t_1t_2 \dots t_n,$$

ки дар инҷо S – объекти матнӣ, t_i-символи i-уми матн, $n = |S|$-дарозии матн мебошад.

Тавре, ки маълум аст, ҳар як триграмма аз се символ, аломатҳои имлоӣ, фаннӣ, махсус ва рақамҳо иборат буда метавонад. Дар бадалсозии объект триграммаҳои тарафи чапро истифода мекунанд. Пас бо назардошти шарти $n\ mod\ 3 \neq 0$ триграммаи охирин метавонад пурра набошад. Барои бартараф кардани ҳолати мазкур амали пуркуниро истифода мекунанд. Дар чунин ҳолат ба матн илова кардани 1 ё 2 символи ҷои ҳеҷ (ø) барои пурра кардани триграммаи охирин зарур мебошад. Агар $n\ mod\ 3 = 2$ бошад, пас 1 символи ҷои ҳеҷ (ø) ва агар $n\ mod\ 3 = 1$ бошад, пас 2 символи ҷои ҳеҷ (ø) дар охири матн илова мекунанд. Дар натиҷа объекти матнии S намуди $S = S'$-ро мегирад. Агар дар охири матн символи ҷои ҳеҷ (ø) илова карда нашавад, пас $S' = S$ ва $n' = n$ аст.

Барои сохтани маҷмуи хусусии триграммаҳои объекти матнии додашуда аввал миқдори триграммаҳое, ки аз ҳамин объект ҳосил кардан имконпазир аст, бо истифода аз формулаи $k = \frac{n'}{3}$ муайян мекунем. Дар инҷо n'-миқдори ҳамаи символҳо дар матни васеъкардашудаи S' ва 3 дарозии ҳар як триграмма мебошад. Миқдори триграммаҳо ба k баробар аст. Барои объекти додашударо ба триграммаҳо тақсим кардан формулаи зеринро истифода мекунем:

$$T_j = s'_{3(j-1)+1}s'_{3(j-1)+2}s'_{3(j-1)+3}$$

дар инҷо $j = 1, 2, \dots, k$.

Дар натиҷа маҷмуи хусусии триграммаҳои ҳосилшуда чунин намудро мегирад: $Trigrams(S') = \{T_1, T_2, \dots, T_k\}$

Тақсим кардани объект ба триграммаҳо марҳилаҳои зеринро дарбар мегирад:

1. Дохилкунии матн

```
string text = richTextBox1.Text;
```

2. Дар ҳолати зарурат барои триграммаи охиринро сохтан ба матн илова кардани символи ҷои ҳеҷ (ø)

```
while (text.Length % 3 != 0)
{
  text += "ø";
}
```

3. Муайянкунии миқдори триграммаҳо $k = \frac{n'}{3}$

```
string[] trigramms = new string[text.Length / 3];
```

4. Сохтани маҷмуи триграммаҳои объекти додашуда $T_j = s'_{3(j-1)+1}s'_{3(j-1)+2}s'_{3(j-1)+3}$

```
for (int i = 0; i < text.Length; i += 3)
 {
   string trigramm = text[i].ToString() + text[i + 1].ToString() + text[i + 2].ToString();
   trigramms[i / 3] = trigramm;
 }
```

5. Натиҷаи ниҳоӣ дар намуди $Trigrams(S') = \{T_1, T_2, \dots, T_k\}$ ва дар асоси он сохтани триграммаҳои беназир

```
 string[] uniqueTrigrams = new HashSet<string>(trigramms).ToArray();
```

6. Хориҷкунии маҷмуи тригаммаҳои беназир

```
for (int i = 0; i < uniqueTrigrams.Length; i++)
{
  listBox2.Items.Add(i + 1+ ". " + uniqueTrigrams[i]);
}
```

Таҳияи n-граммаҳои беназир

Барои истифода бурдани n-граммаҳо дар бадалсозии объекти додашуда аз сохтани маҷмуи хусусии n-граммаҳои беназири матнӣ истифода мекунанд, ки дар он символҳо, аломатҳои имлоӣ, фаннӣ, махсус ва рақамҳо такрорнашаванда мебошанд. n-граммаҳои беназирро асосан бо яке аз 3 тарзи зерин ҳосил кардан имкон дорад:

1. Усули хеш-ҷадвалҳо ё хеш-маҷмуъҳо (HashSet): Шакли махсуси маҷмуъҳо, ки бо истифодаи имкониятҳои махсуси забонҳои барномасозӣ унсурҳоро ба маҷмуъ дар ҳолате илова мекунад, агар он беназир бошад. Дар ин ҳолат унсур пеш аз ба маҷмуъ илова шудан бо истифодаи усули хеширонӣ санҷида мешавад ва агар ин унсур дар маҷмуъ ҷой надошта бошад, пас он ба маҷмуъ илова мегардад.

2. Усули санҷиши пайдарпай: Ҳар як унсур бо унсурҳои дар маҷмуъбуда муқоиса мешаванд ва агар унсури ҷорӣ дар маҷмуъ набошад, пас он илова карда мешавад.

3. Усули ҳисобкунии басомади вохурӣ: Аввал миқдори такроршавии унсур дар маҷмуъ санҷида мешавад. Пас барои ҷудокунии унсурҳои беназир ҳамон унсурҳое гирифта мешаванд, ки басомади такроршавии онҳо ба 1 баробар бошад.

Дар байни усулҳои номбаршуда аз лиҳози тадбиққунӣ усули хеш-ҷадвалҳо ё хеш-маҷмуъҳо нисбатан содда, қулай ва суръати баланди иҷроиш дорад. Ҳамаи забонҳои барномасозии муосир воситаҳои махсуси барномавии кор бо хеш-маҷмуъҳоро доранд ва имконият медиҳанд, ки ба осонӣ маҷмуи унсурҳои беназир таҳия карда шаванд. Дар барномаи сохташуда, ки қисмҳои он дар боло нишон дода шуд, барои сохтани n-граммаҳои беназир усули хеш-маҷмуъҳо истифода гардидааст.

Хулоса

Тибқи блок-схема ва алгоритмҳои овардашуда имконият фароҳам меояд, ки аз объекти матнии додашуда бо осонӣ маҷмуи хусусии n-грамма (униграмма, биграмма, триграмма)-ҳо муайян ва таҳия карда шаванд. Барои татбиқи n-граммаҳо дар самти бадалсозии иттилоот ҳамчунин ҷудокунии n-граммаҳои беназир зарур аст, ки барои ҳалли ин масъала истифода кардани усули хеш-ҷадвалҳо ё хеш-маҷмуъҳо ба мақсад мувофиқ мебошад.

Муқарриз: Ғуломсафдаров А.Ғ. – н.и.т., дотсенти кафедраи барномасозӣ ва зеҳни сунъии Донишгоҳи технологии Тоҷикистон.

Адабиёт

1. Ғафуров, М. Х. Об одном способе разработки уникальных вариантов алфавита шифрования / М. Х. Гафуров, А. А. Косимов, А. Абдукарим // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2022. – № 1(57). – С. 47-50. – EDN JUKVIS.
2. Ғафуров М. Ҳ. Тарзи сохтани алифбои уникалии бадалсозӣ / М. Ҳ. Ғафуров, А. А. Косимов // Вестник ПИТТУ имени академика М.С. Осими. 2022. №1 (22). С. 16-22.
3. Ғафуров, М. Ҳ. Бадалсозии объекти матнӣ бо истифодаи символҳои забон / М. Ҳ. Ғафуров // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2020. – No. 4(52). – P. 31-35. – EDN SMXVKZ.
4. Ғафуров М. Ҳ. Татбиқи биграммаҳо дар бадалсозии объект бо истифодаи калиди дукарата / М. Ҳ. Ғафуров, А. А. Қосимов // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. 2023. №3 (63). С. 43-46.
5. Гафуров М. Х. Шифрование объекта с использованием триграмм и двойного ключа / М.Х. Гафуров, А. А. Косимов, С. Исфандиёри // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. 2023. №4(64). С. 45-49.
6. Гафуров, М. Х. Применение биграмм и триграмм при шифровании объекта с использованием квадрата Полибея / М. Х. Гафуров // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2024. – № 1(65). – С. 72-75. – EDN LUNXVJ.
7. Усманов З. Д. Частотность букв таджикской литературы / З. Д. Усманов, А. А. Косимов // Доклады Академии наук Республики Таджикистан. – 2015. – Т. 58, № 2. – С. 112-115.
8. Усманов З. Д. Частотность биграмм таджикской литературы / З. Д. Усманов, А. А. Косимов // Доклады Академии наук Республики Таджикистан. 2016. Т. 59. № 1-2. С. 28-32.
9. Ғафуров, М. Ҳ. Дар бораи як тарзи бадалсозии объект бо истифодаи калиди дукарата / М. Ҳ. Ғафуров // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2023. – No. 1(61). – P. 38-41. – EDN BQCOID.

МАЪЛУМОТ ДАР БОРАИ МУАЛЛИФОН - СВЕДЕНИЯ ОБ АВТОРАХ – INFORMATION ABOUT THE AUTHORS

TJ	RU	EN
Ғафуров Миршафи Ҳамитович	Гафуров Миршафи Хамитович	Gafurov Mirshafi Khamitovich
н.и.т., дотсент	к.т.н., доцент	Candidate of technical sciences, associate professor
Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ	Таджикский технический университет имени академика М.С. Осими	Tajik technical university named after academician M.S. Osimi
E-mail: mirugaf56@gmail.com		
TJ	RU	EN
Иброҳими Юсуф	Иброхими Юсуф	Ibrohimi Yusuf
Докторанти PhD кафедраи САИ	Докторант PhD кафедры АСУ	PhD student in ASM Department
Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ	Таджикский технический университет имени академика М.С. Осими	Tajik technical university named after academician M.S. Osimi
E-mail: ibrohimi-yusuf@mail.ru		