

◇◇◇◇◇◇◇◇◇◇ **МЕТОДИКА ОБУЧЕНИЯ** ◇◇◇◇◇◇◇◇◇◇

УДК 378.018.43

**Автоматические рекомендации видеоматериалов
к уроку в онлайн-школе с использованием
нейролингвистического программирования**

Смирнов Д.А.*

Московский авиационный институт

(национальный исследовательский университет)

(ФГБОУ ВО МАИ (НИУ)), г. Москва, Российская Федерация

ORCID: <https://orcid.org/0000-0002-7092-2612>

e-mail: daniil.smirnov2311@yandex.ru

Сологуб Г.Б.**

Московский авиационный институт

(национальный исследовательский университет)

(ФГБОУ ВО МАИ (НИУ)), г. Москва, Российская Федерация

ORCID: <https://orcid.org/0000-0002-5657-4826>

e-mail: glebsologub@ya.ru

В статье изложен подход к автоматизации подбора видеоматериалов к текстовым слайдам на уроках английского языка в онлайн-школе путем векторизации текста слайда и субтитров видео при помощи меры TF-IDF и максимизации косинусной меры сходства этих векторных представлений.

Ключевые слова: автоматизация преподавания, дистанционное обучение, сходство текстов.

Для цитаты:

Смирнов Д.А., Сологуб Г.Б. Автоматические рекомендации видеоматериалов к уроку в онлайн-школе с использованием нейролингвистического программирования // Моделирование и анализ данных. 2020. Том 10. № 2. С. 102–109. DOI: [10.17759/mda.2020100208](https://doi.org/10.17759/mda.2020100208)

***Смирнов Даниил Алексеевич**, студент бакалавриата института «Информационные технологии и прикладная математика» Московского авиационного института (национального исследовательского университета), г. Москва, Россия, ORCID: <https://orcid.org/0000-0002-7092-2612>, e-mail: daniil.smirnov2311@yandex.ru

****Сологуб Глеб Борисович**, кандидат физико-математических наук, доцент кафедры математической кибернетики института «Информационные технологии и прикладная математика» Московского авиационного института (национального исследовательского университета), г. Москва, Россия, ORCID: <https://orcid.org/0000-0002-5657-4826>, e-mail: glebsologub@ya.ru

1. ВВЕДЕНИЕ

В настоящее время дистанционное обучение является достаточно удобной, а в ряде ситуаций и единственно доступной формой обучения.

В частности, распространено проведение индивидуальных уроков по английскому языку с преподавателем в онлайн-школах. Такие уроки представляют из себя видеоконференцию между учеником и преподавателем и проводятся, как правило, на специальной платформе, которая содержит и позволяет демонстрировать во время урока учебный контент различных типов, в частности, текстовые слайды и видеоматериалы. При подготовке урока преподаватель тратит время и силы на подбор релевантных видеоматериалов к слайдам урока по выбранной теме.

2. ПОСТАНОВКА ЗАДАЧИ

В качестве исходных данных дан набор текстовых слайдов и коллекция субтитров видеоматериалов. В системе для преподавателя это выглядит следующим образом:

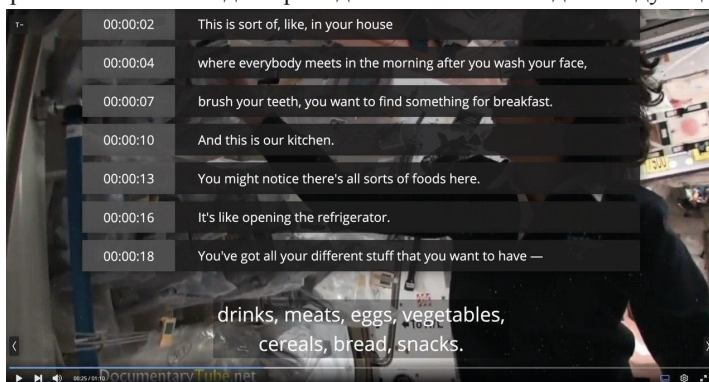


Рис. 1. Пример субтитров в системе для преподавателя

Listen again and complete sentences 1-6

00:00 / 02:58

- 1 Superfoods are good for you because they have lots of .
- 2 The most important thing is to eat healthy food every .
- 3 In the past, astronauts ate a type of food pill when they were in .
- 4 Food pills might become more .
- 5 If you like chocolate ice cream, but your friend likes strawberry, you can eat the same ice cream and it will taste for both of you.
- 6 The technology might not replace normal drinks and food, but it may become in the future.

Рис.2. Пример слайда в системе для преподавателя



Исходный код субтитров и слайдов выглядит следующим образом.

```
In [6]: df.content[0]
Out[6]: ' I have five speaking tips for you whenever you're in public. The first is you have to know your content. Winging it rarely looks good. So when friends call or text me saying, "I have a big presentation. I'm so nervous! What do I do?" I say, "How well do you know your content?" My second tip is be authentic. How many times have you sat through a meeting or watched a speaker, where they just went in this, like, presentation mode? They looked uncomfortable, maybe they looked a little nervous, their eyes are all over the place and their body deceives them. Just be authentic! Be yourself! Guess what? We're human. We relate to other humans. So just be yourself, be likeable, be relatable. The third tips I have is eye contact. Well, let me give you a little tip. Three to five seconds is really good eye contact. Now, if you actually try it, it will feel like eternity. But I promise you that it makes the impression of confidence, authority and like you really do know your stuff. And hopefully, you do. The fourth tip has to do with these. What on earth do we do with our hands when we're talking? I'm always surprised that at weddings, when people ask, "Hey, so what do you do?" And I tell them my business, they're like, "What do I do with my hands?" Well, here is some really practical advice. Just relax them. My fifth practical tip has to do with those filler words. You know which ones I'm talking about? The pesky ones, like "uhm", "ah", "and", "so", "like", "well" – they're riddled throughout our presentations. Why do we do that? So those are my five very practical tips to speaking with more confidence.'
```

Рис.3. Пример исходного кода субтитров

Фактически, исходный код субтитров представляет собой текст на естественном языке с минимальным использованием служебных символов.

```
In [6]: slides[2]
Out[6]: '\n2018-04-11 21:44:28.867","2018-04-11 21:26:47.199","1523482007087","1493376238359","0","2016-12-17 15:39:40","f","57060.000000","1","17613.000000","t","","1237.000000","6","1","7","Family","2018-04-11 20:47:25","homework k","f","<vim-content-section>\n\n <vim-content-section-content><vim-image-set>\n\n <vim-image resource-id=\n237011"></vim-image>\n\n <vim-image resource-id=\n237021"></vim-image-set></vim-content-section-content>\n\n</vim-content-section>\n\n<vim-content-section>\n\n <vim-content-section-title>Read the text again and complete the answers to the questions</vim-content-section-title>\n\n <vim-content-section-content><h3>My family</h3>\n\n <p>Hi! My name's Ian Haig and I'm from Glasgow in Scotland. I'm not English &ndash; I'm Scottish! I have a big family. My father's name is Gordon and he's 50. My mother is 45. Her name's Anna and she's very nice.</p>\n\n <p>My mother's Scottish but her mother and father are Italian. They're from Milan. I have three sisters. Their names are Rosie, Jenny, and Valeria. Valeria is an Italian name. Rosie is 26. She's tall and beautiful, and she's married. Her husband's name is Tom and he's very tall. Jenny and Valeria are 24, but Jenny is tall with short hair and Valeria is short with long hair.</p>\n\n <p>We have a big, new house and an old car. It's slow and cheap, and very small! But it's OK! My mother and father are very short!</p>\n\n <p>Me? Well, I'm 21, I'm not tall, and I'm slim with short hair. I have a girlfriend but I'm not married. Her name is Lucy and she's very beautiful. She's tall with short, dark hair. She's English but she's nice!</p>\n\n <vim-example> What is Ian's surname?\n\n <br> It's <vim-input value=\nHaig1" disabled=\ntrue1"></vim-input>.\n\n </vim-example>\n\n <ol>\n\n <li> Where is Ian from?\n\n <br> It's from <vim-input-answers><vim-input-item>Glasgow</vim-input-item><vim-input-item>Glasgow in Scotland</vim-input-item></li>\n\n <li> What is Ian's father's name?\n\n <br> It's <vim-input-answers><vim-input-item>Gordon</vim-input-item></li>\n\n <li> How old is Ian's mother?\n\n <br> It's <vim-input-answers><vim-input-item>45</vim-input-item></li>\n\n <li> Where is Anna's mother from?\n\n <br> She's from <vim-input-answers><vim-input-item>Italy</vim-input-item><vim-input-item>Milan</vim-input-item></li>\n\n <li> What are Jenny's sisters' names?\n\n <br> Their names are <vim-input-answers><vim-input-item>Rosie1, Valeria</vim-input-item><vim-input-item>Rosie and Valeria</vim-input-item></li>\n\n <li> What is Rosie's husband's name?\n\n <br> His name is <vim-input-answers><vim-input-item>Tom</vim-input-item></li>\n\n <li> What are Jenny's sister's names?\n\n <br> Their names are <vim-input-answers><vim-input-item>Rosie1, Valeria</vim-input-item><vim-input-item>Rosie and Valeria</vim-input-item></li>\n\n <li> What is Rosie's husband's name?\n\n <br> His name is <vim-input-answers><vim-input-item>Tom</vim-input-item></li>\n\n <li> Is their car fast and expensive?\n\n <br> No, it's <vim-input-answers><vim-input-item>slow and cheap</vim-input-item><vim-input-item>slow, cheap</vim-input-item></li>\n\n <li> Where is Ian's girlfriend from?\n\n <br> She's from <vim-input-answers><vim-input-item>England</vim-input-item></li>\n\n </ol></vim-content-section-content>\n\n","","t","",,""
```

Рис.4. Пример исходного кода слайда

Исходный код слайда наоборот содержит много служебной информации, включая теги разметки, но внутри тегов находятся предложения на естественном языке, которые составляют текстовую часть слайда.

Требуется сформировать алгоритм автоматической рекомендации видеоматериалов, упорядоченных по значениям, для каждого слайда, выбранного преподавателем. Под рекомендацией понимается подборка видеоматериалов учебного контента. В результате работы алгоритма должна быть получена автоматическая сформированная подборка видеоматериалов для каждого слайда учебного контента.

3. МЕТОД РЕШЕНИЯ

Для автоматического подбора видеоматериалов к уроку будем сравнивать тексты слайдов и субтитры видео и находить среди них похожие.

Сформулируем алгоритм автоматических рекомендаций.

1. Провести предобработку исходных данных путём очистки их от тегов разметки и прочей служебной информации до состояния текстов на естественном языке.
2. Лемматизировать (привести к нормальным формам [1]) слова в текстах.
3. Представить тексты в виде числовых векторов.
4. Вычислить меру сходства каждого текста слайда с каждым текстом субтитров видео.
5. Рекомендовать для каждого слайда 5 видео с наиболее высокой мерой сходства текста субтитров с текстом данного слайда.

Реализуем алгоритм на языке программирования Python. Для предобработки исходных данных хорошо подходит библиотека для вычисления регулярных выражений `re` [2]. Остальные шаги будем осуществлять при помощи библиотеки методов машинного обучения `scikit-learn` и библиотеки инструментов нейролингвистического программирования `nlTK`. В частности, лемматизацию выполним при помощи встроенных функций библиотеки `nlTK`.

```
In [13]: l = lemmatize_sentence(df1.content[0])
         df_lem_subs.content[0]

Out[13]: "i have five speak tip for you whenever you in public the first be you have to know your content wing it rarely look good so when friend call or text me say `` i have a big presentation i so nervous what do i do `` i say `` how we ll do you know your content `` my second tip be be authentic how many time have you sit through a meeting or watch a speaker where they just go in this like presentation mode they look uncomfortable maybe they look a little nervous s their eye be all over the place and their body deceive them just be authentic be yourself guess what we human we relate to other human so just be yourself be likeable be relatable the third tip i have be eye contact well let me give you a little tip three to five second be really good eye contact now if you actually try it it will feel like eternity but i promise you that it make the impression of confidence authority and like you really do know your stuff and hopefully you do the fourth tip have to do with these what on earth do we do with our hand when we talk i always surprise that at wedding when people ask `` hey so what do you do `` and i tell them my business they like `` what do i do with my hand `` well here be some really practical advice just relax them my fifth practical tip have to do with those filler word you know which one i talk about the pesky one like `` umh `` `` ah `` `` and `` `` so `` `` like `` 'well '`` - they riddle throughout our presentation why do we do that so those be my five very practical tip to speak with more confidence"
```

Рис.5. Пример исходного кода субтитров после предобработки и лемматизации

```
In [22]: l = lemmatize_sentence(df_slides.slide_content[0])
         for j in l:
             if j == "." or j == "!" or j == "?" or j == ",":
                 l.remove(j)
             sub = ' '.join(l)
         sub

Out[22]: "read the text again and complete the answer to the question my family hi my name 's ian haig and i 'm from glasgow in scotland i 'm not english & ndash ; i 'm scottish i have a big family my father 's name be gordon and he 's 50 m y mother be 45 her name 's anna and she 's very nice my mother 's scottish but her mother and father be italian the y 're from milan i have three sister their name be rosie jenny and valeria valeria be an italian name rosie be 26 s he 's tall and beautiful and she 's married her husband 's name be tom and he 's very tall jenny and valeria be 24 but jenny be tall with short hair and valeria be short with long hair we have a big new house and an old car it 's slow and cheap and very small but it 's ok my mother and father be very short me well i 'm 21 i 'm not tall and i 'm slm with short hair i have a girlfriend but i 'm not marry her name be lucy and she 's very beautiful she 's tall with short dark hair she 's english but she 's nice what be ian 's surname it 's where be ian from he 's from glasgow glasgow in scotland what be ian 's father 's name it 's gordon how old be ian 's mother she 's 45 forty-five where be anna 's mother from she 's from italy milan milan italy what be jenny 's sister ' name their name be rosie valeria rosie and valeria valeria rosie valeria and rosie what be rosie 's husband 's name hi name be tom be their car fast and expensive no it 's slow and cheap slow cheap slow and cheap and very small slow cheap small where be ian 's girlfriend from she 's from england"
```

Рис.6. Пример исходного кода слайда после предобработки и лемматизации

Для векторизации текстов слайдов и субтитров видео будем использовать статистическую меру *TF-IDF* (от англ. *TF* – term frequency, частотность терминов, *IDF* – inverse document frequency, обратная частотность документов).



Эта мера используется в качестве простого и удобного способа оценки важности слова в тексте, являющемся частью коллекции текстов [3]. Будем использовать её, чтобы оценить важность каждого слова для текста слайда или субтитров видео относительно остальных текстов и составить вектор из таких оценок для каждого текста.

Мера *TF-IDF* вычисляется как произведение величин *TF* и *IDF*.

TF – это величина, которая показывает относительную частоту встречаемости данного слова в данном тексте и вычисляется по формуле:

$$TF(t) = \frac{n_i}{\sum_{k=1}^m n_k},$$

где t – слово, частотность которого вычисляется; n_i – количество раз, когда слово t встретилось в тексте, $\sum_{k=1}^m n_k$ – количество всех слов в тексте, m – количество уникальных слов.

IDF – это величина, обратная частоте встречаемости слова в текстах коллекции субтитров, которая вычисляется по формуле:

$$IDF = \log \frac{|D|}{|d_i \supseteq t|},$$

где $|D|$ – количество текстов в коллекции; $|d_i \supseteq t|$ – количество текстов, в которых встречается слово t .

Каждому тексту слайдов и субтитров видео сопоставим вектор оценок *TF-IDF* входящих в него слов, вычисленных при помощи встроенного метода из библиотеки *scikit-learn*.

```
In [42]: get_vectors(df_lem_subs.content[0])
Out[42]: array([[ 1,  1,  1,  1,  1,  1,  5,  1,  1,  2,  1, 13,  1,  1,  1,  1,
  1,  2,  2,  2,  1, 15,  1,  1,  3,  1,  1,  1,  1,  3,  1,  1,
  1,  1,  1,  2,  1,  2,  7,  1,  1,  1,  2,  2,  1,  1,  2,  4,
  4,  4,  1,  6,  1,  2,  3,  1,  1,  1,  2,  1,  1,  1,  5,  2,
  1,  1,  1,  2,  2,  1,  2,  1,  1,  1,  1,  3,  3,  1,  1,  1,
  3,  1,  1,  1,  1,  2,  2,  1,  6,  1,  2,  1,  1,  1,  2,  1,
  1,  3,  6,  2,  3,  1,  5,  1,  1,  2,  1,  1,  1,  1,  7,  6,
  1,  1,  1,  1,  1,  5,  1,  4,  5,  3,  1,  1,  1,  1,  1,  1,
  5,  1, 12,  3,  2]])
```

Рис.7. Результат векторизации текста субтитров

```
In [40]: get_vectors(slide)
Out[40]: array([[ 1,  1,  1,  1,  1,  1,  5,  1,  1,  2,  1, 13,  1,  1,  1,  1,
  1,  2,  2,  2,  1, 15,  1,  1,  3,  1,  1,  1,  1,  3,  1,  1,
  1,  1,  1,  2,  1,  2,  7,  1,  1,  1,  2,  2,  1,  1,  2,  4,
  4,  4,  1,  6,  1,  2,  3,  1,  1,  1,  2,  1,  1,  1,  5,  2,
  1,  1,  1,  2,  2,  1,  2,  1,  1,  1,  1,  3,  3,  1,  1,  1,
  5,  3,  1,  1,  1,  1,  2,  2,  1,  6,  1,  2,  1,  1,  1,  2,
  1,  1,  3,  6,  2,  3,  1,  5,  1,  1,  2,  1,  1,  1,  1,  7,
  6,  1,  1,  1,  1,  1,  5,  1,  4,  5,  3,  1,  1,  1,  1,  1,
  1,  5,  1, 12,  3,  2]])
```

Рис.8. Результат векторизации текста слайда



Для вычисления сходства текстов по полученным векторам будем использовать меру косинусного сходства, которая вычисляется как косинус угла между этими векторами по формуле:

$$\text{similarity} = \cos \alpha = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

где $A \cdot B$ – скалярное произведение; $|A||B|$ – произведение длин векторов.

Чем больше косинус, тем меньше угол между векторами, а значит и меньше различие между текстами, которые мы сравниваем [1]. Для каждой пары текста слайдов и текста субтитров видео вычислим меру косинусного сходства при помощи встроеного метода из библиотеки scikit-learn.

```
In [48]: get_cosine_sim(slide, df_lem_subs.content[0])  
Out[48]: array([[1.          , 0.37487662],  
              [0.37487662, 1.          ]])
```

Рис.9. Пример расчета меры косинусного сходства для векторов

Далее для каждого заданного преподавателем слайда будем рекомендовать 5 видеоматериалов, сходство текста субтитров которых с текстом этого слайда максимально. Видеоматериалы предъявляются преподавателю в соответствии с убыванием меры сходства.

```
In [32]: lst[0]  
Out[32]: " Look at Robert. What is he doing? He's playing soccer. What is he wearing? He's wearing shorts and a T-shirt. Is it hot? Yes, it is. Look, it's Ben. He is playing in the snow. Is he wearing shorts and a T-shirt? No, it's cold. He's wearing a hat, a coat, a sweater, gloves and boots. What is this girl doing? She's running. Is she wearing a coat and a hat? No, it's warm. Is she wearing shorts and a T-shirt? No, it's raining. She's wearing pants and a sweat shirt. What are you wearing? Is it cold? Are you wearing a coat and a hat? Is it warm? Are you wearing shorts and a T-shirt?"  
  
In [33]: df_slides.slide_content[0]  
Out[33]: " Read the text again and complete the answers to the questions My family Hi! My name's Ian Haig and I'm from Glasgow in Scotland. I'm not English &ndash; I'm Scottish! I have a big family. My father's name is Gordon and he's 50. My mother is 45. Her name's Anna and she's very nice. My mother's Scottish but her mother and father are Italian. They're from Milan. I have three sisters. Their names are Rosie , Jenny , and Valeria. Valeria is an Italian name. Rosie is 26. She's tall and beautiful , and she's married. Her husband's name is Tom and he's very tall. Jenny and Valeria are 24 , but Jenny is tall with short hair and Valeria is short with long hair. We have a big , new house and an old car. It's slow and cheap , and very small! But it's OK! My mother and father are very short! Me? Well , I'm 21 , I'm not tall , and I'm slim with short hair. I have a girlfriend but I'm not married. Her name is Lucy and she's very beautiful. She's tall with short , dark hair. She's English but she's nice! What is Ian's surname? It's . Where is Ian from? He's from Glasgow Glasgow in Scotland . What is Ian's father's name? It's Gordon . How old is Ian's mother? She's 45 forty-five . Where is Anna's mother from? She's from Italy Milan Milan , Italy . What are Jenny's sisters' names? Their names are Rosie , Valeria Rosie and Valeria Valeria , Rosie Valeria and Rosie . What is Rosie's husband's name? His name is Tom . Is their car fast and expensive? No , it's slow and cheap slow , cheap and slow and cheap , and very small slow , cheap , small . Where is Ian's girlfriend from? She's from England . "  
  
In [ ]:
```

Рис.10. Пример текста субтитров видеоматериалов, рекомендованных к тексту слайда



- 1 Superfoods are good for you because they have lots of .
- 2 The most important thing is to eat healthy food every .
- 3 In the past, astronauts ate a type of food pill when they were in .
- 4 Food pills might become more .
- 5 If you like chocolate ice cream, but your friend likes strawberry, you can eat the same ice cream and it will taste for both of you.
- 6 The technology might not replace normal drinks and food, but it may become in the future.

Related videos

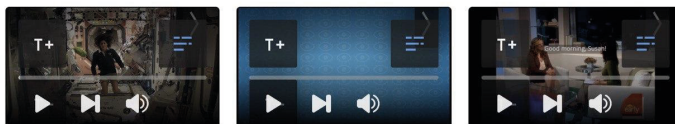


Рис.11. Пример результата работы программы в системе для преподавателя

4. ЗАКЛЮЧЕНИЕ

Поставлена задача построения автоматических рекомендаций видеоматериалов к слайдам урока в онлайн-школе английского языка. Сформулирован алгоритм решения этой задачи на основе методов нейролингвистического программирования путем векторизации текстов слайдов и текстов субтитров видео при помощи меры *TF-IDF* и максимизации косинусной меры сходства этих векторных представлений. Описана реализация этого алгоритма на языке Python с применением свободно-распространяемых библиотек подпрограмм.

Предложенный алгоритм может быть использован для построения автоматических рекомендаций любого контента, имеющего текстовое представление.

Литература

1. *Daniel Jurafsky, James H. Martin. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second Edition. Pearson Education International. – 2009. – 1024 pp.*
2. *Фридл Дж. Регулярные выражения. – СПб.: «Питер», 2001. – 352 с. – ISBN 5–318–00056–8.*
3. *Jones K.S. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. – MCB University Press. – 2004. – Vol. 60, no. 5. – P. 493–502. – ISSN 0022–0418.*



Automatic Recommendation of Video for Online School Lesson Using Neuro-Linguistic Programming

Daniil A. Smirnov*

Moscow Aviation Institute
(National Research University), Moscow, Russia
ORCID: <https://orcid.org/0000-0002-7092-2612>
e-mail: daniil.smirnov2311@yandex.ru

Gleb B. Sologub**

Moscow Aviation Institute
(National Research University), Moscow, Russia
ORCID: <https://orcid.org/0000-0002-5657-4826>
e-mail: glebsologub@ya.ru

The article describes an approach to automating the matching of video materials to text slides in English classes in an online school by vectorizing slide text and video subtitles using the TF-IDF measure and maximizing the cosine similarity measure of these vector representations.

Keywords: teaching automation, distance learning, text similarity.

For citation:

Smirnov D.A., Sologub G.B. Automatic Recommendation of Video for Online School Lesson Using Neuro-Linguistic Programming. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2020. Vol. 10, no. 2, pp. 102–109. DOI:10.17759/mda.2020100208 (In Russ., abstr. in Engl.).

References

1. Daniel Jurafsky, James H. Martin. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second Edition. *Pearson Education International*, 2009. 1024 pp.
2. Friedl, John. *Regulyarnye vyrazheniya = Regular Expressions*. *Piter*, 2001. 352 pp. ISBN 5-318-00056-8.
3. Jones K.S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*. *MCB University Press*, 2004. Vol. 60, no. 5, pp. 493–502. ISSN 0022-0418.

***Daniil A. Smirnov**, undergraduate student of the Institute of Information Technology and Applied Mathematics, Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: <https://orcid.org/0000-0002-7092-2612>, e-mail: daniil.smirnov2311@yandex.ru

****Gleb B. Sologub**, candidate of physical and mathematical sciences, associate professor of the Department of Mathematical Cybernetics, Institute of Information Technologies and Applied Mathematics, Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: <https://orcid.org/0000-0002-5657-4826>, e-mail: glebsologub@ya.ru