Розглянуто особливості застосування технологій NLP, Information Retrieval, SEO та Web-mining для визначення стійких словосполучень при ідентифікації ключових слів в опрацюванні Web-ресурсів. Лінгвостатистичний аналіз природомовного тексту використовує переваги контент-моніторінгу на основі методів NLP для ідентифікації стійких словосполучень. Квантитативний аналіз стійких словосполучень використано для визначення степеня приналежності множині ключових слів. Запропоновано метод визначення стійких словосполучень при ідентифікації ключових слів україномовного контенту

Ключові слова: стійке словосполучення, NLP, Information Retrieval, SEO, Web-mining, статистичний лінгвістичний аналіз, квантитативна лінгвістика, рубрикація

Рассмотрены особенности применения технологий NLP, Information Retrieval, SEO и Web-mining для определения устойчивых словосочетаний при идентификации ключевых слов в разработке Web-ресурсов. Лингвостатистический анализ естественноязыкового текста использует преимущества контент-мониторинга на основе методов NLP для идентификации устойчивых словосочетаний. Квантитативный анализ устойчивых словосочетаний использован для определения степени принадлежности множеству ключевых слов. Предложен метод определения устойчивых словосочетаний при идентификации ключевых слов украиноязычного контента

Ключевые слова: устойчивое словосочетание, NLP, Information Retrieval, SEO, Web-mining, статистический лингвистический анализ, квантитативная лингвистика, рубрикация

UDC 004.89
DOI: 10.15587/1729-4061.2018.126009

ANALYSIS OF STATISTICAL METHODS FOR STABLE COMBINATIONS DETERMINATION OF KEYWORDS IDENTIFICATION

V. Lytvyn
Doctor of Technical Sciences, Professor*
E-mail: vasyl.v.lytvyn@lpnu.ua
V. Vysotska
PhD, Associate Professor*
E-mail: victoria.a.vysotska@lpnu.ua
D. Uhryn
PhD, Associate Professor
Department of Information Systems
Chernivtsi Faculty of National Technical University
«Kharkiv Polytechnic Institute»
Holovna str., 203A, Chernivtsi, Ukraine, 58000
E-mail: ugrund38@gmail.com
M. Hrendus
Assistant*
E-mail: mhirnyak@ukr.net
O. Naum
Assistant
Department of Information Systems and Technologies
Drohobych Ivan Franko State Pedagogical University
I. Franko str., 24, Drohobych, Ukraine, 82100
E-mail: oleh.naum@gmail.com
*Department of Information Systems and Networks
Lviv Polytechnic National University
S. Bandery str., 12, Lviv, Ukraine, 79013

1. Introduction

In modern intellectual systems of linguistic nature, it is important to determine effectively stable word combinations for identifying a set of keywords while processing Web resources [1]. An optimally appropriate set of stable word combinations is used for information retrieval (IR), SEO and Web-mining technologies, natural language processing (NLP) and automatic machine translation. It is also essential to identify the content by using specific natural language texts and rubrics as well as reflexive and automatic analysis of comments on published products. A new direction is the automatic processing of texts while integrating data from various sources of different fields, including Internet tourism [2]. Stable word combinations are used in algorithms for correct tokenization, compiling dictionaries (lexicography), automatic translation, learning foreign languages (міцний чай – strong tea ↔ міцний сон – fast sleep), and distinguishing terminology [3].

Analysis of stable word combinations is used for identifying relevant content, indexing in IR, tokenization, content categorization, creating a search image of some content, and constructing thematic ontologies [4]. Usually, this work is the prerogative of the person who is the moderator of Web-resources [5]. Automating the process of extracting data or knowledge of natural language content using NLP methods greatly reduces the time and the amount of Web resources to obtain the desired result [6]. The use of methods of artificial intelligence (AI) in linguistic processing of natural language texts is usually effective after qualitative

© V. Lytvyn, V. Vysotska, D. Uhryn, M. Hrendus, O. Naum, 2018

morphological and syntactic parsing of these texts [7]. If for English texts these questions are easily solved by a simple parser and using the Porter algorithm, then for Slavic languages, including Ukrainian texts, it is not so easy [8]. Therefore, there appears the problem of choosing an optimal statistical method for determining stable word combinations for identifying keywords in the development of Ukrainian-language Web resources [9].

The use of knowledge engineering for effective NLP improves the quality of the results of research on texts [10]. This entails developing new NLP approaches and techniques, including the automatic determination of stable word combinations for identifying keywords when processing Web resources [11]. With the proliferation of Internet services and their introduction into everyday life of every ordinary person, there appears redundancy of information as an IR result. The so-called informational noise negatively affects both the Internet business and the irritability of a regular user of these services. The daily IR results are the following: Google>8 billion pages, Yandex>600 million pages, and 2.5 million sites [12]. Therefore, the qualitative and optimal definition of stable word combinations as a set of keywords in Ukrainian and English texts will significantly reduce the time for receiving relevant content search results in response to user queries.

2. Literature review and problem statement

Modern NLP methods are increasingly used not only in AI and computer linguistics, but in Internet environments, especially in the IR direction (Fig. 1). Today in IR, it implies not only representation, storage, organization and access to information elements. It also focuses on the needs of regular and potential users in on-line information and the emphasis on finding important relevant content (and not data, Fig. 2) [13].

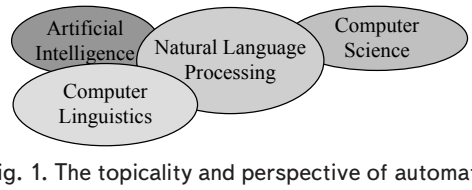


Fig. 1. The topicality and perspective of automatic processing of texts

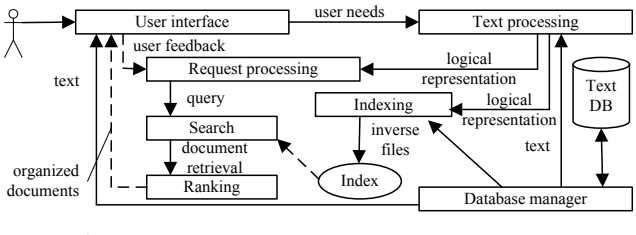


Fig. 2. The process of information search according to Baeza-Yates & Ribeiro-Neto (1999)

The main models and methods of IR are indexing, the Boolean model, the vector model and the evaluation of the search quality [14]. The average complexities of a direct search (Brute Force, O(n+m)) and a complex search (Dboyer-Moore, O(n/m)) [14] were experimentally tested long ago. The effectiveness of the indexing method is directly proportional to the effectiveness of the process of creating a document or content search image (logical representation). This, in turn, affects the efficiency of presenting relevant information on the Internet [14].

$$Dictionary\begin{cases} Content & \Rightarrow 2\rightarrow 4\rightarrow 8\rightarrow 16\rightarrow 32\rightarrow 64\rightarrow 128, \\ NLP & \Rightarrow 1\rightarrow 2\rightarrow 3\rightarrow 5\rightarrow 8\rightarrow 13\rightarrow 21, \\ Web\ mining & \underbrace{\Rightarrow 13\rightarrow 16.}_{Postings} \end{cases}$$

The effectiveness of any NLP method depends directly on the quality of the prior processing of the text content. This, in turn, depends on extracting and/or receiving text (HTML, PDF...) [15]; coding and language [16]; breakdown into words and sentences (tokenization); elimination of stop words [17]; and stemming as determining the word form [18]. Tokenization is a process of demarcating and classifying sections of a series of input characters for the desired content:

– dates, numbers (13/03/2014, 1415);

– adverbs (Ukr. нарешті, зазвичай, відтоді, потім, наприклад);

– introductory words (іншими словами, в підсумок скажемо, між іншим);

– prepositions (напередодні, незважаючи на);

– particles (все ж таки, немов би, немов як, до того ж, ніби то як);

– verbose tokens (collocations, Улан-Уде, Нью-Йорк, Іван Іванович);

– boundaries of sentences ("І. І. Іванов приїхав в м. Львів минулої зими.").

The resulting tokens are then subjected to a different form of processing. The process is considered as a subtask for analysing input data [19].

Stop-words (or noise words) are words that do not carry a meaningful load, so their usefulness and role during searching are not significant. A text, in its turn, is an unstructured set of meaningful words ("bag of words"), where stop-words belong to the functional parts of speech, that is, they are prepositions, conjunctions, particles (а, га, ай, ау, ах, ба, без, поблизу, брр, зась, ніби, б, бути, в, ви, ваш, поблизу, вглиб, до того ж, уздовж, адже, замість, замість, поза, усередині, як, біля, навколо, геть,...) [20].

An effective IR model is highly dependent on the quality and method: presentation of text files and content [21]; setting informational needs (queries) of users; estimation of the proximity between the query and the document [22].

The Boolean model of IR considers content as a plurality of words (terms) and a query as a Boolean expression: "(кішка OR пес) AND корм"; "птаха ANDNOT військовий" [14]. Processing a query in this model is the operation on sets that correspond to words (terms) (Table 1).

Table 1

An example of the Boolean model (keywords in articles found as to [23])

BM/Article	[24]	[25]	[26]	[27]	[28]	[29]
Content	1	1	0	0	0	1
NLP	1	1	0	1	0	0
Web-mining	1	1	0	1	1	1
IR	0	1	0	0	0	0
SEO	1	0	0	0	0	0
Web resources	1	0	1	1	1	1
Keywords	1	0	1	1	1	0

The advantages of the Boolean model are simplicity and convenience for those who are familiar with logical operators. The disadvantage is that this model is too "contrast-based" (in terms of both content submission and its relevance).

The Vector model presents IR content and query as vectors in the space of words (terms), where the vector component is the meaning of a word for the document (query). The model uses a measure of proximity (ranking). This is the cosine of the angle between the vectors (Fig. 3):

$$sim(\bar{d},\bar{q}) = \frac{\sum d_i \cdot q_i}{|\bar{d}||\bar{q}|},$$

where d_i is the weight of a term in the content (frequency of use in the content/collection); q_i is the weight of the term i in the query. A well-known example of the vector model is the approach *TF×IDF* (*TF* is term frequency, *IDF* is inverse document frequency). The basic version of *TF×IDF* [14] is

$$tf_{ij} = \frac{f_{ij}}{\max_k f_{kj}}, \quad idf_i = \log\frac{N}{n_i}, \quad w_{ij} = tf_{ij} \cdot idf_i.$$

A modified version of *TF×IDF* [14] is

$$TFIDF_D(l) = \beta + (1-\beta) \cdot tf_D(l) \cdot idf_D(l),$$

$$tf_D(l) \frac{freq_D(l)}{freq_D(l) + 0.5 + 1.5 \cdot \frac{dl_D}{avg\_dl}},$$

$$idf(l) = \frac{\log\left(\frac{|c|+0.5}{df(l)}\right)}{\log(|c|+1)},$$

where avg_dl is the average length of a document, and c is the size of the collection β=0...1.

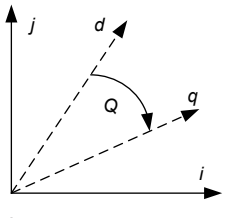


Fig. 3. The vector pattern of IR

The advantage of a vector model is the effectiveness of processing primary static collections. It also involves partial coincidence. The disadvantage of the model is that it is easily attacked (spammed) and does not work well on short texts. However, the Web is an uncontrolled collection (Fig. 4), large volumes of content, its various formats, variety (language, themes, etc.), high competition (spam), present clicks and links (PageRank). Therefore, the two previous IR models do not resolve the problem of the quality of searching for relevant content. The basis of the quality assessment of IR is the notion of relevance (compliance with the information needs) of the content sought. It necessarily contains the following signs (features): precision ($p = a/b$), recall ($r = a/c$), and F-measures ($F = (p+r)/2pr$), where a represents the relevant signs in the answer, b denotes all the signs in the answer, and c is all relevant signs (Fig. 5).

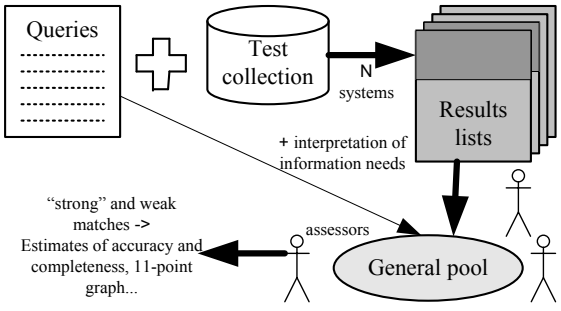


Fig. 4. The method of general information search pool

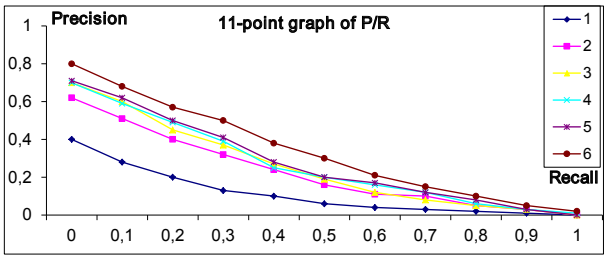


Fig. 5. The 11-point graph of P/R information retrieval results

The main initiators of the IR evaluation methods are (Fig. 4, 5) the following: TREC (Text Retrieval Evaluation Conference, trec.nist.gov) and CLEF (Cross-Language Evaluation Forum, www.clef-campaign.org).

3. The aim and objectives of the study

The aim of the work is to analyse statistical methods for developing an optimal approach to determining stable word combinations in identifying keywords when developing and processing Ukrainian-language Web-resources based on the technology of computational linguistics.

To achieve this aim, the following tasks are set and done:

– to develop a method for determining stable word combinations while identifying keywords in Ukrainian-language texts based on the analysis of lexical speech coefficients in standard content fragmentation;

– to devise a formal approach to designing content monitoring software to determine stable word combinations when identifying keywords in Ukrainian texts based on Web Mining and NLP;

– to obtain and analyse the results of experimental testing of the proposed content-monitoring method for determining stable word combinations in identifying keywords in Ukrainian-language scientific texts on technical matter.

4. The method for determining stable word combinations when identifying keywords for text content

The method for determining stable word combinations consists of the following phases: morphological analysis (MA), syntactic analysis (SA), keyword selection, and stability analysis of word combinations from a multitude of keywords.

Fig. 6 lists the main steps in determining stable word combinations when identifying keywords for text content.

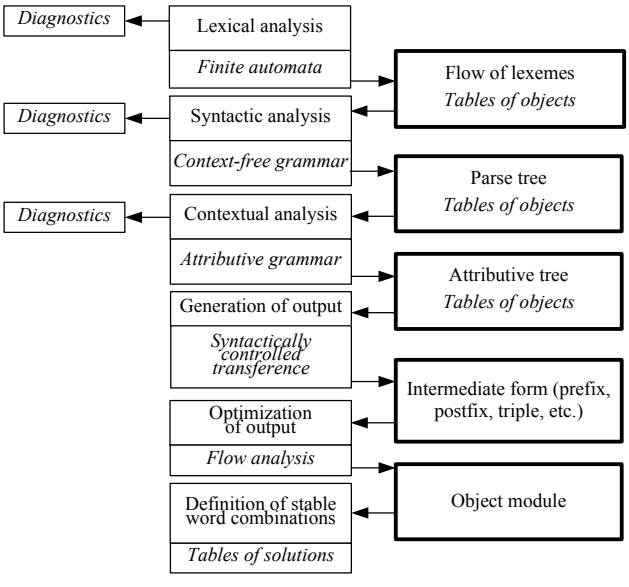


Fig. 6. A flowchart of linguistic analysis of Ukrainian-language texts to identify stable word combinations as keywords

Stage 1. The aim of the MA is to define the keyword equivalence classes in IR [30]. MA methods for identifying keywords are procedural, tabular, and statistical stemming or their various combinations. One of the well-known MA algorithms is the Porter stemmer [31], which is the stemming algorithm published by Martin Porter in 1980. The original version of the Stemmer is for English and written in BCPL. Stemming is the process of reducing the word to the stem by rejecting it auxiliary parts, such as an inflexion or suffix. For Ukrainian texts in MA, it is best to use combinations of approaches such as procedural, tabular and statistical stemming [32]. In the procedural approach of MA, the emphasis is placed on analysing words of the dictionaries of stems and full forms dictionaries (FFDs). Then the MA algorithm consists of three main stages: the search in the FFD, the selection of the stem and the search for the stem in the dictionary. Examples of the tabular approach are *вовка* → *вовк* (masc., animate, sg., [gen.|mean.]); *не* → *не* (particle); *годуй* → *годувати* (imperf., imperat., sg.); *в* → *в* (preposition). An example of the model for a Ukrainian word change is *лев* masc. 1*b (animal); *лев* masc. 1*a (currency); *стриже* imperf. 8*b (-г-); *гостьова* femin. 4a (п). The basis of most of machine MA of the Ukrainian language is a tree or a finite state automaton (Fig. 7) [33].

The statistical stemmer is based on the probability of determining the stem of the word, for example: словниками → словник → словник-ами → ник-ами; сокирами → сокира → сокир-ами → ир-ами; літаючого → літати → літ-аючого → іт-аючого; літаючого → літаючий → літаюч-ого → юч-ого. The main rule is one vowel in the stem of the word. The types of words are determined by the forms of their inflexions (Fig. 8).

Features of the algorithm. The algorithm works with separate words, so the context in which the word is used is unknown. Other unavailable categories of linguistics are the word structure (root, suffix, etc.) and the part of speech (noun, adjective, etc.). We currently have the following techniques for analysing words:

– the ending in removed from the word, for example, the ending *увати* transfers the word *критикувати* into *критик*;

– the word has a stable ending: the words with this ending are left unchanged, for example, *ск* and the invariable words *блиск*, *тиск*, *обеліск*, etc.;

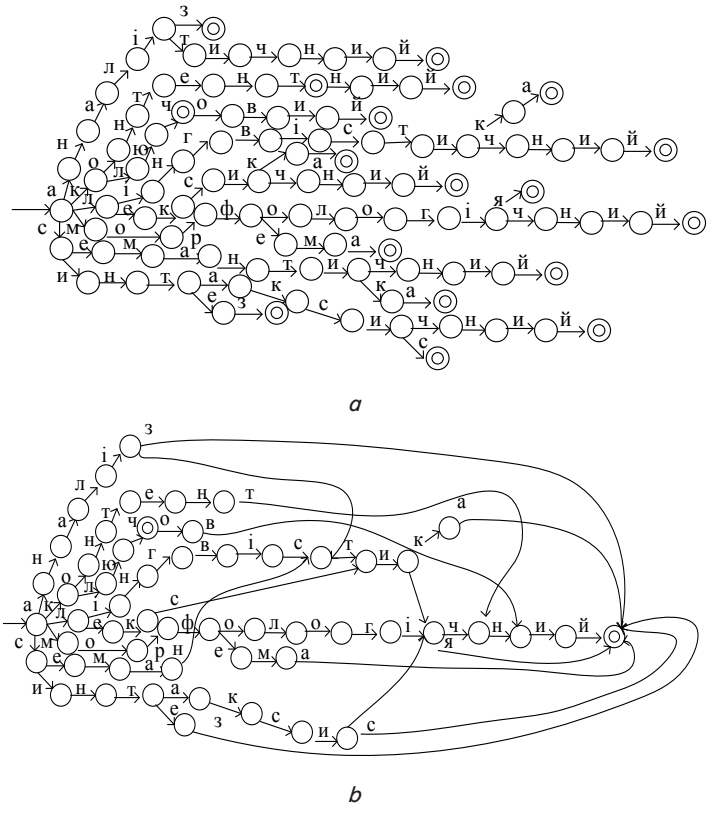
– the word changes the ending, but this rule applies to words in which certain letters drop out (*ядро* and *ядер*, where the ending *ер* changes to *р*) or change (*чоловік* and *чоловіче*, where *к* changes into *ч*);

– the word corresponds to a stable expression: this is an attempt to combine several rules into one complex, for example, in the code there are expressions similar to (ов)*у ва(в|вши|вшись|ла|ло|ли|ння|нні|нням|нню|ти|вся|всь|лись|лися|тись|тися);

– the word does not change during its stemming, but there is an exception to the rules: it is necessary to maintain a dictionary of exception words, for example, *віче*, *наче*;

– the word changes with stemming, but it is also an exception: it is necessary to keep in the dictionary at once two forms of the word (original and schematised), for example, the word *відер* should change to *відр*, although other words ending in *ер* are not stemmed so (*авіадиспетчер*, *вітер*, *гравер*, etc.);

– the short words remain unchanged: the functional parts of speech (prepositions, conjunctions, particles) are usually very short words and ignored by the algorithm (words up to 2 letters inclusive).



a

b

Fig. 7. Methods for storing MA results: *a* is a tree and *b* is the Finite State Automata, FSA

```
var $ADJECTIVE =
'/(ими|ій|ий|а|е|ова|ове|ів|є|їй|єє|еє|я|ім|ем|им|ім|их|іх|ою|йми|іми|у|ю|ого|ому|ої)$/';
//http://uk.wikipedia.org/wiki/Прикметник + http://wapedia.mobi/uk/Прикметник
 var $PARTICIPLE = '/(ий|ого|ому|им|ім|а|ій|у|ою|ій|і|их|йми|их)$/';
//http://uk.wikipedia.org/wiki/Дієприкметник
 var $VERB = '/(сь|ся|ив|ать|ять|у|ю|ав|али|учи|ячи|вши|ши|е|ме|ати|яти|є)$/';
//http://uk.wikipedia.org/wiki/Дієслово
 var $NOUN =
'/(а|ев|ов|е|ями|ами|еи|и|ей|ой|ий|й|иям|ям|ием|ем|ам|ом|о|у|ах|иях|ях|ы|ь|ию|ью|ю|
ия|ья|я|і|ові|ї|ею|єю|ою|є|еві|ем|єм|ів|їв|\'ю)$/';
//http://uk.wikipedia.org/wiki/Іменник
```

Fig. 8. Definition of the word type by the inflexion form

All these techniques are used for groups that generate and illustrate the rules of stemming. However, this greatly complicates the search algorithm for keywords. First, it is necessary to take into account the widespread endings (not the traditional inflexions, as part of the word), that is, the sequence of letters in which a word ends. Tables 2, 3 contain endings of words from 1 to 4 letters in length. Five or more letters are not given, as there are few such words (for 5 maximum *йтесь* (6,837), for 6 (4,656), etc.). This is a peculiar map for the stemming project. For the effectiveness of the search algorithm, it is necessary to construct a static tree of endings and to cover all branches of the tree [34]. The level of the tree detail varies within 500–600 words with a common ending.

Table 3

A static tree of endings the total proportion of which is less than 1 %

р (2,709)	ч (959)	г (636)	п (341)	щ (110)
н (2,531)	с (914)	з (581)	б (281)	ц (34)
д (1,038)	л (754)	ж (353)	ф (214)	ґ (4)

Stage 2. Syntax represents the rules of combining words in correct expressions such as word combinations and sentences [35]. The task of a SA (syntactic analyser, parser) is to construct the syntactic structure of an input sentence [36]. The aspects of implementing the SA are dictionaries (data on individual units of language); formal rules and interaction with adjacent levels of processing (MA, semantic analysis). Often, the SA uses the Context-free grammar (CFG) rules: <N, T, X, R>, where N is the set of nonterminal characters, T is the set of terminal characters $(N \cap T = \varnothing)$, X is the axiom $(X \in N)$, R is the set of rules of transformation (substitutions) of type $Y \to \alpha$, where $Y \in N$, α is a list of terminal and non-terminal characters. An example of the CFG:

$$N = \{S, NP, PP, V, N, A\},$$

$$T = \{\textit{система, рубрикувати, україномовний, контент, за, ключовий, слово}\}, S.$$

Table 2

A static table of common endings in the Ukrainian language

я (164,062)	тися (10,379)	мось (20,536)	али (10,666)	ному (19,112)	ові (17,191)	а (68,134)	их (31,127)
ся (148,160)	лися (10,338)	лось (10,231)	ними (19,089)	о (90,454)	сті (8,731)	на (21,328)	ах (20,023)
ня (9,765)	теся (19,103)	тись (10,366)	м (119,779)	мо (33,568)	ості (7,636)	ла (17,945)	ях (9,855)
ося (30,769)	лася (10,230)	лись (10,337)	т (2,980)	го (31,445)	ю (80,877)	ка (11,029)	них (19,092)
ься (25,211)	ь (151,355)	тесь (19,105)	ім (31,343)	ло (17,238)	ою (39,616)	істю (7,598)	ї (34,702)
ися (21,940)	сь (111,459)	лась (10,229)	им (31,166)	ймо (11,229)	ню (10,075)	й (77,109)	ої (31,421)
еся (19,105)	ть (33,055)	ість (7,606)	ам (20,154)	ємо (11,136)	ною (20,280)	ій (33,241)	ної (19,098)
шся (11,775)	ось (30,788)	и (123,402)	ом (17,018)	ого (31,389)	кою (7,497)	ий (31,136)	в (32,681)
ася (10,235)	ись (22,656)	ми (62,080)	ям (15,717)	ало (10,465)	нню (9,054)	ала (10,610)	ів (15,898)
вся (10,076)	есь (19,114)	ти (20,025)	нім (19,333)	ного (19,090)	стю (7,648)	е (66,988)	ав (10,547)
юся (8,044)	ась (10,239)	ли (17,711)	ним (19,093)	і (90,275)	у (94,504)	те (32,651)	ш (19,163)
ння (9,001)	всь (10,016)	ими (31,121)	ням (9,434)	ні (31,679)	му (35,023)	не (20,257)	єш (11,138)
мося (20,532)	сть (7,688)	ами (20,106)	нням (8,975)	ві (22,543)	ну (23,125)	йте (11,230)	є (11,466)
лося (10,233)	юсь (8,047)	ями (9,844)	ку (11,624)	ті (12,596)	ній (19,549)	єте (11,137)	к (7,299)
ться (25,036)	ють (11,222)	ати (10,819)	ому (31,585)	нні (9,909)	ний (19,042)	х (61,506)	

$R = \{S \to NP\,VP, S \to NP\,VP\,PP, NP \to A\,N, PP \to P\,NP, VP \to V\,NP,$
$NP \to$ система, $V \to$ рубрикувати, $A \to$ україномовний,
$A \to$ ключовий, $N \to$ контент, $N \to$ слово, $P \to$ за$\}$.

The disadvantage of using the CFG is the periodic appearance of ambiguity in the SA, for example, "Система рубрикує україномовний контент за ключовими словами / The system categorizes Ukrainian-language content by keywords" (Fig. 9).

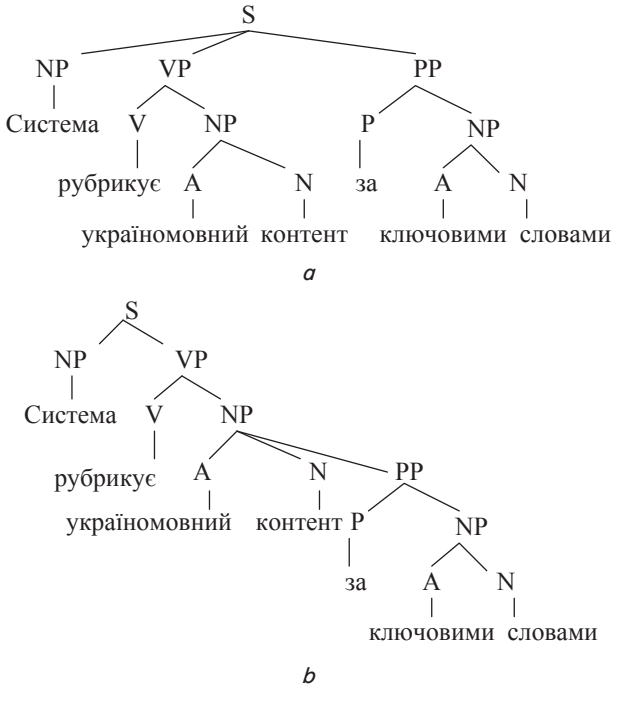


Fig. 9. The CFG ambiguity: *a* — example 1; *b* — example 2

The examples of known SA systems for English language tests are "Machinese Word combination Tagger" [37] and VISL [38]. There is no available online information resource for the SA of Ukrainian-language texts. We will analyse the results of the SA of an English text through these resources through the example of such a sentence set: "The train went up the track out of sight, around one of the hills of burnt timber. Nick sat down on the bundle of canvas and the baggage that I had got out of the door of the baggage car." "Machinese Word combination Tagger" is a text analyser that processes base forms and component structures. It also recognizes the "part of speech" classes (noun, adjective, verb, pronoun, etc.) and generates a micro-indicative syntax of a word combination, marks fragments or brackets noun word combinations (Fig. 10).

The Connexor Machinese Tokenizer is a set of components of the program that performs the basic tasks of text analysis at a very high speed and provides relevant word information for bulk programs. The Machinese Tokenizer splits the text into clear words and provides possible forms and classes for the words (Fig. 11). The first column displays the position of a token in the text (the calculation in characters); in the following column, there is information on the length of the token; the third column is for the text form, and the other columns contain the main form(s) and the tag denoting the part of speech (PRON=pronoun, V=verb, DET=determinant, or N=noun). If a word has several meanings, the analysis also includes several columns of the main parts of speech.

The Ontology Matcher Demo uses metadata to identify ontology objects in the text (Fig. 12). The program corresponds to the concepts in the Finnish general ontology with approximately 28,000 notions in each language. The found notions of ontology are given in the text below as a reference. With the cursor over a word, there appears the notion to which the word refers.

Text	Baseform	Phrase syntax and part-of-speech
The	the	premodifier, determiner
train	train	nominal head, noun, single-word noun phrase
went	go	main verb, indicative past
on	on	adverbial head, adverb
up	up	preposed marker, preposition
the	the	premodifier, determiner
track	track	nominal head, noun, single-word noun phrase
out	out	adverbial head, adverb
of	of	preposed marker, preposition
sight	sight	nominal head, noun, single-word noun phrase
,	,	
around	around	preposed marker, preposition
one	one	nominal head, pro-nominal
of	of	postmodifier, preposition

Fig. 10. English Machinese Word combination Tagger 4.9.1 analysis

0	4	This	this	PRON		
5	2	is	be	V		
8	1	a	a	DET		
10	4	test	test	N	test	V

Fig. 11. The Machinese Tokenizer

"The train went on up the track out of sight, around one of the hills of burnt timber. Nick sat down on the bundle of canvas and bedding the baggage man had pitched out of the door of the baggage car."

Fig. 12. The Ontology Matcher

Fig. 13, 14 show the result of SA through the VISL information resource

For SA of Ukrainian-language texts, such information resources do not exist [39–42]. Moreover, the SA process itself is rather cumbersome for Ukrainian-language content [43–46]. Let us consider the example of the input sentence: "Він зробив це так незручно, що зачепив образок мого ангела, який висів на дубовій спинці ліжка, і що вбита муха впала мені прямо на голову" ("He made it so uncomfortable that he touched the image of my angel that hung on the oak-bed backboard, and that the killed fly fell to my head").

Its SA example with using pre-syntax is shown in Fig. 15.

Parsing by chunks is a breakdown of sentences into non-intersecting word combinations [47–51], i. e. (flat structure)≠complete parsing, for example, (the boy (with the hat)) ←→ (the boy) with (the hat)).

Tree structure

Enter English text to parse:

The train went on up the track out of sight, around one of the hills of burnt timber.

Parse and Show | Export and Download | Reset

Visualization: Vertical ▼ Notational convention default ▼

```
&lt;ß&gt;
&lt;s&gt;

SOURCE: Running text
1. The train went on up the track out of sight, around one of the hills of burnt timber.
A1
STA:cl(fcl)
,
.
|-S:g(np)
| |-D:art('the' S/P)    The
| |-H:n('train' S NOM)  train
|-P:g(vp)
| |-H:v('go' IMPF)      went
| |-D:adv('on') on
|-A:g(pp)
| |-H:prp('up') up
| |-D:g(np)
|   |-D:art('the' S/P)  the
|   |-H:n('track' S NOM)        track
|-A:g(pp)
  |-H:prp('out_of')     out_of
  |-D:n('sight' S NOM)  sight
|-D:g(pp)
  |-H:prp('around')     around
  |-D:g(np)
    |-H:num('one' &lt;card&gt; S)       one
    |-D:g(pp)
      |-H:prp('of')     of
      |-D:g(np)
```

Fig. 13. The structure of a tree on the VISL information resource

```
<ß>
The [the] <*> <def> ART S/P @>N #1->2
train [train] <DA:tog> <Vground> <def> <nhead> N S NOM @SUBJ> #2->3
went [go] <DA:gå> <move> <mv> V IMPF @FS-STA #3->0
on [on] <DA:på> ADV @MV< #4->3
up [up] <DA:op=ad> PRP @<SA #5->3
the [the] <def> ART S/P @>N #6->7
track [track] <DA:spor> <Lpath> <sem-l> <def> <nhead> N S NOM @P< #7->5
out of [out=of] <complex> <DA:ude=af> PRP @<ADVL #8->3
sight [sight] <DA:sigt> <percep-w> <Labs> <idf> <nhead> N S NOM @P< #9->8
, [.] PU @PU #10->0
around [around] <insertion> <DA:omkring> PRP @>A #11->0
one [one] <fr:78> <f:3664212> <card> NUM S @P< #12->11
of [of] <DA:af> <np-close> PRP @N< #13->12
the [the] <def> ART S/P @>N #14->15
hills [hill] <DA:høj> <Lmountain> <def> <nhead> N P NOM @P< #15->13
of [of] <DA:af> <np-close> PRP @N< #16->15
burnt [burnt] <DA:brændt> <SYN:cooked> <SYN:destroyed> <jppl> <tempered-2>
<SYN:treated> ADJ POS @>N #17->18
timber [timber] <DA:tømmer> <mat> <idf> <nhead> N S NOM @P< #18->16
. [.] PU @PU #19->0
</ß>
```

Fig. 14. The result of SA through the VISL information resource

Частина речення: (*він зробив це так незручно,*)
--- він[1](дієслово)зробив[2](кого)це[3]
зробив[2] (як) так [4]
зробив[2](предикатив) незручно[5]
зробив[2] (як) незручно[5]
Частина речення: (*що зачепив образок мого ангела,*)
--- образок[9] (дієслово) зачепив[8](кого)що[7]
образок[9](який) мого[10]
{і[20]} ангела[11](якого) мого[10]
Частина речення: (*який висів на спинці ліжка,*)
{образок[9]}(який)який[13] (дієслово) висів[14](прийменник)на[15](чому) спинці[17](якій дубовій[16] спинці[17](чого) ліжка[18]
Частина речення: (*і*)
{образок[9]}і[20]
Частина речення: (* що вбита муха впала мені прямо на голову.*)
--- муха[23] (дієслово) впала[24](кому) мені[25] (прийменник)на[27](кого) голову[28] на[27](кого) голову[28]
впала[24](прийменник)на[27]
впала[24](як)прямо[26]
муха[23] (яка) вбита[22]
{і[20]} що[21]
--- мені[25] (прийменник)на[27]
незвязн: він[1], муха[23],
==в реченні слів всього: 25, слів незв'язно: 2, із них прийменників:0, час опрацювання: 0.050с.
Він[1] зробив[2] це[3] так[4] незручно[5] ,[6] що[7] зачепив[8] образок[9] мого[10] ангела[11] ,[12] який[13] висів[14] на[15] дубовій[16] спинці[17] ліжка[18] ,[19] і[20] що[21] вбита[22] **муха[23]** впала[24] мені[25] прямо[26] на[27] голову[28] .[29]

Fig. 15. The result of the SA of the Ukrainian-language sentence

5. Results of studying the definition of stable word combinations when identifying keywords for text content

To isolate stable word combinations in the analysed texts and to conduct a comparative analysis, we will use 4 different methods: FREQ (frequency+morphological patterns, that is, direct counting of the number of words) [52]; t-test [53]; statistics χ^2 [54]; LR as a likelihood ration [55]. A collocation is a word combination that has features of a syntactically and semantically integral unit [56]. In it, the choice of one component is based on the context, and the choice of another depends on the choice of the first element [57]. For example, *ставити умови* (to set conditions): the choice of the verb *ставити* (to set) is determined by tradition and depends on the noun *умови*; with the noun *пропозицію* (suggestion, proposal), there will be another verb – *вносити* (to make). This concerns a limited (selective) combining of words: word combinationologisms, idioms, proper names, and brandnames. A collocation also usually includes components of toponyms, anthroponyms, and other frequently used naming conventions (for example, *супермаркет «Метро»* (*Metro* supermarket), *завод «Електрон»* (*Electron* factory)) [58]. Other names for the same phenomenon are stable (set) or word combinationological units and N-grams. Examples of collocations are the following:

– *грати роль, мати значення, впливати, справляти враження*;

– *засоби масової..., зброя масової..., вищий навчальний...;*

– *глибокий старець ↔ поверхневий/мілкий невеликий юнак*;

– *міцний чай ↔ сильний чай*;

– *кока-кола, Microsoft Windows*;

– *Гола Пристань, Володимир Волинський, Нью Йорк, Стів Джобс.*

1. The FREQ method is a direct calculation of the frequency of using pairs (triples) of words. For example, FREQ for the sentence "*В літературі описано декілька підходів до автоматичного виділення стійких словосполучень.*" → *в літературі*; *літературі описано*; *описано декілька*; *декілька підходів*; *підходів до*; *до автоматичного*; *автоматичного виділення*; *виділення стійких*; *стійких словосполучень*. Unfortunately, as a result of applying this method to large volumes of text, we receive information noise due to the high frequency of function words. The method also requires taking into account the frequency of use and the patterns of word combinations. An example of morphology rules in FREQ is as follows:

A N: *турецький гамбіт* (Turkish gambit), *перша похідна* (first derivative), *інформаційний ресурс* (information resource);

N N_G: *контент аналіз* (content analysis), *баланс інтересів* (balance of interests), *контент-комерція* (content commerce), *контент моніторинг* (content monitoring);

N Pr N: *трава у дворі* (grass in the yard), *дрова на траві* (firewood on the grass).

2. The t-test method consists in checking statistical hypotheses and using the statistical model of MA:

– H_0: words found accidentally;

– $P(w^1w^2)=P(w^1)P(w^2)$;

– taking into account not only pairs but also the frequency of using separate words (those that make up a pair);

– $t=\dfrac{\overline{x}-\mu}{\sqrt{\dfrac{s^2}{N}}}$,

where $\overline{x}$ is the empirical average, μ is the theoretical average, s^2 is the empirical variance, and N is the size of the empirical sample.

The method is not quite correct for the language, but it helps get results in practice – for example, the frequency of the occurrence of the stable word combination *контент аналіз* (content analysis) in [14] with P(контент)=28/1368 and P(аналіз)=38/1368 is

$$H_0: p=P(\text{контент аналіз})=P(\text{контент})P(\text{аналіз})=$$
$$=0.20468\cdot 0.27778\approx 5.69\cdot 10^{-4}.$$

In the Bernoulli scheme

$$s^2=p(1-p)\approx p$$

with

$$\overline{x}=18/1368$$

and

$$t=\frac{\overline{x}-\mu}{\sqrt{\dfrac{s^2}{N}}}=\frac{0.013158-5.69\cdot 10^{-4}}{\sqrt{\dfrac{5.69\cdot 10^{-4}}{1{,}368}}}\approx 19.52816.$$

3. The Pearson χ^2 method is applied to tables of 2×2 (Table 4). In the calculations, normality is not expected.

Table 4

An example of using the Pearson χ^2 method

w_i	w_1=контент	w_1≠контент
w_2=аналіз	18 (content analysis)	20 (e. g., statistical analysis)
w_2≠аналіз	10 (including content monitoring)	1320 (including statistical monitoring)

For example,

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})} = \frac{1368(18\cdot 1320 - 20\cdot 10)^2}{(18+20)(18+10)(20+1{,}320)(10+1{,}320)} \approx 400.44106.$$

4. The LR method is used to calculate the hypotheses ($p_1 >> p_2$)

$$H_1 : P(w^2 \mid w^1) = p = P(w^2 \mid \neg w^1),$$

$$H_2 : P(w^2 \mid w^1) = p_1 \neq p_2 = P(w^2 \mid \neg w^1),$$

where

$$p = \frac{c_2}{N}; \quad p_1 = \frac{c_{12}}{c_1};$$

$$p_2 = \frac{c_2 - c_{12}}{N - c_1}.$$

Then, using a binomial distribution

$$b(m,n,p) = C_n^m p^m (1-p)^{n-m},$$

we obtain the relation of likelihood LR:

$$L(H_1) = b(c_{12}, c_1, p) b(c_2 - c_{12}, N - c_1, p),$$

$$L(H_2) = b(c_{12}, c_1, p_1) b(c_2 - c_{12}, N - c_1, p_2),$$

$$\log\lambda = L(H_1) \Big/ L(H_2),$$

where $-2\log\lambda$ in the asymptotics is distributed as χ^2, i. e.,

$$L(H_1) = C_{c_1}^{c_{12}} \cdot p^{c_{12}} \cdot (1-p)^{c_1 - c_{12}} \cdot C_{N-c_1}^{c_2 - c_{12}} p^{c_2 - c_{12}} \cdot (1-p)^{N - c_1 - c_2 + c_{12}} = \frac{c_1!}{c_{12}! \cdot (c_1 - c_{12})!} \cdot p^{c_{12}} \cdot (1-p)^{c_1 - c_{12}} \cdot \frac{(N-c_1)!}{(c_2 - c_{12})! \cdot (N - c_1 - (c_2 - c_{12}))!} \times \times p^{c_2 - c_{12}} \cdot (1-p)^{N - c_1 - c_2 + c_{12}},$$

$$L(H_2) = C_{c_1}^{c_{12}} \cdot p_1^{c_{12}} \cdot (1-p_1)^{c_1 - c_{12}} \cdot C_{N-c_1}^{c_2 - c_{12}} \cdot p_2^{c_2 - c_{12}} \cdot (1-p_2)^{N - c_1 - c_2 + c_{12}} = \frac{c_1!}{c_{12}! \cdot (c_1 - c_{12})!} \cdot p_1^{c_{12}} \cdot (1-p_1)^{c_1 - c_{12}} \cdot \frac{(N-c_1)!}{(c_2 - c_{12})! \cdot (N - c_1 - (c_2 - c_{12}))!} \times \times p_2^{c_2 - c_{12}} \cdot (1-p_2)^{N - c_1 - c_2 + c_{12}},$$

Then

$$\log\lambda = L(H_1) \Big/ L(H_2) = = \frac{p^{c_{12}} \cdot (1-p)^{c_1 - c_{12}} \cdot p^{c_2 - c_{12}} \cdot (1-p)^{N - c_1 - c_2 + c_{12}}}{p_1^{c_{12}} \cdot (1-p_1)^{c_1 - c_{12}} \cdot p_2^{c_2 - c_{12}} \cdot (1-p_2)^{N - c_1 - c_2 + c_{12}}}.$$

For example, with $c_1 = 28$, $c_{12} = 18$ and $c_2 = 38$,

$$\log\lambda = L(H_1) \Big/ L(H_2) = = \frac{\left(\frac{38}{1{,}368}\right)^{18} \cdot \left(1 - \frac{38}{1{,}368}\right)^{28-18} \cdot \left(\frac{38}{1{,}368}\right)^{38-18} \cdot \left(1 - \frac{38}{1{,}368}\right)^{1{,}368-28-38+18}}{\left(\frac{18}{28}\right)^{18} \cdot \left(1 - \frac{18}{28}\right)^{28-18} \cdot \left(\frac{38-18}{1{,}368-28}\right)^{38-18} \cdot \left(1 - \frac{38-18}{1{,}368-28}\right)^{1{,}368-28-38+18}} \approx \approx 4.53355 \cdot 10^{-23}.$$

In order to choose the optimal statistical method for determining stable word combinations, it is necessary to analyse a Ukrainian language text based on the stems of words without taking into account their inflexions. It will greatly improve the accuracy of the result.

6. Discussion of the research results on identifying stable word combinations for keyword identification

An experiment of distinguishing terms was carried out on 3 technical articles [1–3] written in two languages – Ukrainian and English. The template for the experiment contained the following: [Adj+N], [Diyeprykm.+N], [N+N, Gen.], [N+N, Abl.], [N+'–'+N]. The experiment included 6 methods for determining the keywords: manually by authors (A); via the system Victana.lviv.ua [23], according to Zipf's law (B); by FREQ (C); by t-test (D); by LR (F); by χ^2 (G). The analysis of the 3 articles [1–3] was conducted in Ukrainian and the results were translated into English (Tables 5, 6). The keywords in bold are those that occurred in the results of applying all the methods, the italicized keywords are only those obtained through the B–G methods, and the underlined keywords are those in the methods A and C–G. While conducting linguistic analysis for compiling alphanumeric dictionaries of two words, the following features and algorithms were used:

– bigrams were formed within the punctuation marks (if there was at least one punctuation mark between the words, these words were not considered as a bigram);

– an alphanumeric dictionary of two-word combinations was formed on the basis of stems, that is, the bigrams *контений аналіз* and *контентного аналізу* were considered as one and the same bigram;

– in the analysis of the inflexions of the analysed words, verbs were not taken into account when forming the alphanumeric dictionary of bigrams (verbs were considered as punctuation marks);

– before the linguistic analysis of the texts, all stop words (particles, adverbs, conjunctions) and pronouns (they were also considered as punctuation marks) were excluded.

The statistical methods make it possible to take into account the use of separate words. The peculiarities that are associated with using the methods for different volumes of data and probability ranges (better than the t-test for larger p, where normality is violated; the likelihood ratio is better approximated with χ^2 than tables 2×2 for small volumes). They are often used not for the acceptance/rejection of hypotheses but for the ranking of candidate word combinations.

Table 5

The list of frequency index for stable word combinations in articles [1–3]

No.	Author's	as to [23]	FREQ, t-test	LR	χ^2
1	2	3	4	5	6
Q	A	B	C, D	F	G
In work [1] in Ukrainian					
1	**Стиль автора**	Стоп-слово	Відносна частота	*Коефіцієнт кореляції*	*Коефіцієнт кореляції*
2	Статистичний аналіз	Метод визначення	*Коефіцієнт кореляції*	Відносна частота	Відносна частота
3	Лінгвістичний аналіз	**Визначення стилю**	**Стиль автора**	*Частота появи*	*Частота появи*
4	Квантитативна лінгвістика	**Стиль автора**	**Визначення стилю**	Стопове слово	Авторська атрибуція
5	Авторська атрибуція	Аналіз уривку	Стопове слово	Україномовний текст	**Стиль автора**
6	**Визначення стилю**	*Частота появи*	Україномовний текст	**Стиль автора**	Україномовний текст
7	Україномовні тексти	*Автор тексту*	*Частота появи*	Поява слова	Стопове слово
8	Технологія лінгвометрії	Уривок тексту	Авторська атрибуція	Авторська атрибуція	**Визначення стилю**
9	Технологія стилеметрії	*Коефіцієнт кореляції*	Поява слова	**Визначення стилю**	Поява слова
10	Технологія глоттохронології	Дослідження тексту	*Автор тексту*	Слова уривку	Слова уривку
In work [2] in Ukrainian					
1	**Web Mining**	**Ключове слово**	**Ключове слово**	*Текстовий контент*	*Текстовий контент*
2	Контент-моніторінг	**Контент-аналіз**	*Текстовий контент*	**Ключове слово**	Тематичний словник
3	**Ключові слова**	Визначена системою	**Web Mining**	Тематичний словник	**Ключове слово**
4	**Контент-аналіз**	*Формування системою*	Тематичний словник	*Слова контенту*	*Слова контенту*
5	Стеммер Портера	**Web Mining**	Визначення слів	Ключове словосполучення	*Множина слів*
6	Лінгвістичний аналіз	*Слова контенту*	Ключове словосполучення	Визначення слів	*Формування системою*
7	Метод визначення	*Текстовий контент*	*Слова контенту*	*Формування системою*	**Web Mining**
8	Визначення слів	Аналіз статистики	*Множина слів*	**Web Mining**	Визначення слів
9	Слов'янськомовні тексти	Ключове словосполучення	*Формування системою*	*Слова контенту*	*Слова контенту*
10	Технологія NLP	*Множина слів*	**Контент-аналіз**	Контент-моніторінг	Контент-моніторінг
In work [3] in Ukrainian					
1	Інформаційний ресурс	**Контент-аналіз**	Психологічний стан	Психологічна особистість	Психологічна особистість
2	**Контент-аналіз**	Стоп- слово	Психологічна особистість	Психологічний стан	Психологічний стан
3	Лінгвістичний аналіз	Тематичний словник	**Контент-аналіз**	Формування зрізу	Формування зрізу
4	Морфологічний аналіз	Пости користувача	Марковане слово	*Стан особистості*	Зріз стану
5	**Соціальна мережа**	Повідомлення користувача	Психологічний зріз	Марковане слово	Марковане слово
6	Формування зрізу	Користувач мережі	*Стан особистості*	Психологічний зріз	**Контент-аналіз**
7	Зріз розуміння	*Стан особистості*	Формування зрізу	**Контент-аналіз**	Психологічний зріз
8	Розуміння особистості	Аналізована особистість	Зріз стану	Зріз стану	*Стан особистості*
9	Україномовні тексти	**Соціальна мережа**	Зріз особистості	Аналізована особистість	**Соціальна мережа**
10	Big-Five	Диспозиції особистості	**Соціальна мережа**	**Соціальна мережа**	Аналізована особистість
In work [1] in English					
1	**Style of the author**	*Reference fragment*	*Reference fragment*	Words fragment	Words fragment
2	Statistical analysis	**Author's style**	Words fragment	*Reference fragment*	*Reference fragment*
3	Linguistic analysis	*Author's text*	*Syntactic words*	*Stop words*	Recognition author
4	Quantitative linguistics	*Syntactic words*	Frequency fragment	Swadesh list	*Stop words*
5	Author's attribution	*Stop words*	Swadesh list	Recognition author	Swadesh list
6	Recognition of style	Formatted fragments	*Stop words*	*Syntactic words*	*Syntactic words*
7	Ukrainian texts	*Anchor words*	**Author style**	Frequency fragment	Frequency fragment
8	Linguometry technology	Author's language	Recognition author	*Author's text*	*Author's text*

Continuation of Table 5

1	2	3	4	5	6
9	Stylemetry technology	Method of anchor	*Author's text*	*Anchor words*	**Author style**
10	Glottochronology technology	Frequency dictionary	*Anchor words*	**Author style**	*Anchor words*
In work [2] in English					
1	Web Mining	*Text content*	*Text content*	Web mining	Web mining
2	**Content monitoring**	**Content analysis**	Web mining	*Text content*	*Text content*
3	**Content analysis**	Analysis of statistics	Keywords text	*Keywords content*	*Keywords content*
4	Porter stemmer	Defined systematically	Keywords defined	Keywords text	Analysis text
5	Linguistic analysis	*Stop word*	Analysis text	Keywords defined	Keywords text
6	Determining the keywords	Potential keywords	*Keywords content*	*Stop word*	Keywords defined
7	Slavic language	**Content monitoring**	**Content monitoring**	Analysis text	*Stop word*
8	Slavic texts	*Author's keywords*	**Content analysis**	*Author's keywords*	**Content monitoring**
9	Method for determining	*Keywords content*	*Stop word*	**Content monitoring**	**Content analysis**
10	Web technology	Direct word	*Author's keywords*	**Content analysis**	*Author's keywords*
In work [3] in English					
1	Information resource	**Content analysis**	**Content analysis**	Psychological personality	**Content analysis**
2	**Content analysis**	*Psychological state*	Psychological personality	*Psychological state*	Psychological personality
3	Linguistic analysis	Personality analysis	*Psychological state*	**Content analysis**	*Psychological state*
4	Morphological analysis	Personality disposition	Social networks	Based analysis	Based analysis
5	Social network	Psychological analysis	Marked words	State personality	Psychological base
6	Status of personality	Personality model	State personality	Psychological base	State personality
7	Personality understanding	Stop words	Based analysis	Social networks	Social networks
8	Formation of the status	Psychological disposition	Psychological base	Marked words	Psychological base
9	Stop words	Content monitoring	State based	State based	Marked words
10	Method of formation	Social network	Based content	Psychological base	State based

Table 6

Differences of the methods according to the rating list of 100 stable word combinations

Q	A	B	C	D	F	G	A	B	C	D	F	G	A	B	C	D	F	G
For the Ukrainian articles [1–3]																		
A	1	0.23	0.47	0.35	0.27	0.21	1	0.27	0.51	0.39	0.31	0.25	1	0.25	0.49	0.36	0.29	0.23
B	0.23	1	0.63	0.61	0.52	0.43	0.27	1	0.65	0.63	0.57	0.47	0.25	1	0.64	0.62	0.55	0.45
C	0.47	0.63	1	0.93	0.17	0.71	0.51	0.65	1	0.94	0.25	0.73	0.49	0.64	1	0.93	0.21	0.72
D	0.35	0.61	0.93	1	0.19	0.75	0.39	0.63	0.94	1	0.26	0.77	0.36	0.62	0.93	1	0.22	0.76
F	0.27	0.52	0.17	0.19	1	0.26	0.31	0.57	0.25	0.26	1	0.39	0.29	0.55	0.21	0.22	1	0.33
G	0.21	0.43	0.71	0.75	0.26	1	0.25	0.47	0.73	0.77	0.39	1	0.23	0.45	0.72	0.76	0.33	1
For the English articles [1–3]																		
A	1	0.27	0.51	0.47	0.31	0.27	1	0.31	0.55	0.51	0.35	0.31	1	0.29	0.53	0.49	0.33	0.29
B	0.27	1	0.66	0.64	0.55	0.47	0.31	1	0.69	0.67	0.59	0.49	0.29	1	0.68	0.65	0.57	0.48
C	0.51	0.66	1	0.95	0.23	0.76	0.55	0.69	1	0.96	0.27	0.77	0.53	0.68	1	0.95	0.24	0.75
D	0.47	0.64	0.95	1	0.21	0.79	0.51	0.67	0.96	1	0.29	0.81	0.49	0.65	0.95	1	0.25	0.78
F	0.31	0.55	0.23	0.21	1	0.31	0.35	0.59	0.27	0.29	1	0.41	0.33	0.57	0.24	0.25	1	0.37
G	0.27	0.47	0.76	0.79	0.31	1	0.31	0.49	0.77	0.81	0.41	1	0.29	0.48	0.75	0.78	0.37	1

To compare the results, we used the Google-based library word2vec, which has proven itself as an alternative of *TF×IDF* (A_1 in Table 7 according to the template ['bigram', number of uses]). We also used the built-in methods to search for word combinations in Python. However, for these datasets, it did not work effectively, because for high-quality work, it needs a huge corpus [58]. The most interesting thing is that the system allows doing it after transferring each word from the corpus to a space whose dimension is specified by the user, for example, ['*king*' + '*woman*' – '*man*' = '*queen*'].

Table 7

Differences of other methods by ranking the frequency of occurrence of stable word combinations is articles [1–3]

Method	Language	Article [1]	Article [2]	Article [3]
1	2	3	4	5
A_1	UA	(‘контент_моніторінгу’, 13)	(‘тематичного_словника’, 11) (‘слов_янськомовних’, 10)	(‘психологічного_стану’, 16) (‘формування_зрізу’, 12) (‘sfx_a’, 12) (‘структурну_схему’, 7) (‘відкритість_досвіду’, 6) (‘зрізу_психологічного’, 2)
	ENG	(‘swadesh_list’, 18) (‘based_on’, 15)	(‘based_on’, 20) (‘slavic_language’, 15) (‘author_s’, 13)	(‘based_on’, 35) (‘psychological_state’, 26) (‘social_networks’, 22) (‘his_her’, 11) (‘following_structural’, 8) (‘big_five’, 7) (‘let_us’, 7) (‘structural_scheme’, 4)
A_2	UA	((‘службових’, ‘слів’), 32) ((‘стопових’, ‘слів’), 24) ((‘визначення’, ‘визначення’), 23) ((‘стилю’, ‘стилю’), 22) ((‘слів’, ‘слів’), 22) ((‘списку’, ‘сводеша’), 20) ((‘в’, ‘уривку’), 19) ((‘опорних’, ‘слів’), 18) ((‘стилю’, ‘автора’), 17) ((‘автора’, ‘автора’), 17)	((‘ключових’, ‘слів’), 72) ((‘текстового’, ‘контенту’), 21) ((‘на’, ‘етапі’), 17) ((‘визначення’, ‘ключових’), 16) ((‘крок’, ‘1’), 16) ((‘крок’, ‘2’), 16) ((‘web’, ‘mining’), 15) ((‘слів’, ‘в’), 14) ((‘тематичного’, ‘словника’), 11) ((‘для’, ‘визначення’), 10)	((‘на’, ‘основі’), 21) ((‘психологічного’, ‘стану’), 18) ((‘контент’, ‘аналізу’), 16) ((‘маркованих’, ‘слів’), 15) ((‘зрізу’, ‘психологічного’), 14) ((‘стану’, ‘особистості’), 14) ((‘формування’, ‘зрізу’), 12) ((‘особистості’, ‘на’), 12) ((‘sfx’, ‘a’), 12) ((‘основі’, ‘контент’), 11)
	ENG	((‘of’, ‘the’), 107) ((‘author’, ‘s’), 52) ((‘of’, ‘a’), 51) ((‘in’, ‘the’), 46) ((‘the’, ‘author’), 45) ((‘reference’, ‘fragment’), 31) ((‘analysis’, ‘of’), 24) ((‘words’, ‘in’), 22) ((‘to’, ‘the’), 21) ((‘the’, ‘method’), 21)	((‘of’, ‘the’), 134) ((‘in’, ‘the’), 61) ((‘by’, ‘the’), 45) ((‘analysis’, ‘of’), 39) ((‘of’, ‘a’), 31) ((‘the’, ‘text’), 30) ((‘the’, ‘system’), 30) ((‘to’, ‘the’), 29) ((‘of’, ‘keywords’), 28) ((‘text’, ‘content’), 27)	((‘of’, ‘the’), 134) ((‘is’, ‘the’), 117) ((‘the’, ‘content’), 45) ((‘of’, ‘a’), 43) ((‘analysis’, ‘of’), 37) ((‘based’, ‘on’), 35) ((‘on’, ‘the’), 34) ((‘in’, ‘the’), 33) ((‘content’, ‘analysis’), 30) ((‘the’, ‘process’), 27)
A_3	UA	((‘слів’, ‘слів’), 88) ((‘стилю’, ‘автора’), 68) ((‘службових’, ‘слів’), 63) ((‘визначення’, ‘стилю’), 61) ((‘списку’, ‘сводеша’), 56) ((‘стопових’, ‘слів’), 48) ((‘визначення’, ‘автора’), 45) ((‘авторського’, ‘мовлення’), 33) ((‘опорних’, ‘слів’), 31) ((‘стилю’, ‘стилю’), 30)	((‘ключових’, ‘слів’), 74) ((‘слів’, ‘в’), 24) ((‘web’, ‘mining’), 22) ((‘текстового’, ‘контенту’), 21) ((‘на’, ‘2’), 20) ((‘визначення’, ‘ключових’), 19) ((‘ключових’, ‘в’), 19) ((‘визначення’, ‘слів’), 18) ((‘слів’, ‘для’), 18) ((‘на’, ‘крок’), 18)	((‘на’, ‘основі’), 21) ((‘психологічного’, ‘стану’), 18) ((‘психологічного’, ‘особистості’), 17) ((‘контент’, ‘аналізу’), 16) ((‘стану’, ‘особистості’), 15) ((‘маркованих’, ‘слів’), 15) ((‘зрізу’, ‘психологічного’), 14) ((‘зрізу’, ‘стану’), 14) ((‘зрізу’, ‘особистості’), 14) ((‘особистості’, ‘на’), 14)
	ENG	((‘of’, ‘the’), 186) ((‘the’, ‘of’), 169) ((‘of’, ‘of’), 152) ((‘of’, ‘a’), 81) ((‘the’, ‘the’), 75) ((‘the’, ‘author’), 66) ((‘and’, ‘of’), 63) ((‘in’, ‘the’), 57) ((‘of’, ‘author’), 57) ((‘of’, ‘words’), 55)	((‘of’, ‘the’), 258) ((‘the’, ‘of’), 235) ((‘of’, ‘of’), 137) ((‘the’, ‘the’), 122) ((‘of’, ‘keywords’), 72) ((‘in’, ‘the’), 71) ((‘a’, ‘of’), 70) ((‘and’, ‘of’), 69) ((‘by’, ‘the’), 64) ((‘of’, ‘content’), 63)	((‘the’, ‘of’), 304) ((‘of’, ‘the’), 243) ((‘the’, ‘the’), 168) ((‘of’, ‘of’), 162) ((‘is’, ‘the’), 154) ((‘of’, ‘a’), 91) ((‘the’, ‘is’), 76) ((‘the’, ‘content’), 71) ((‘is’, ‘of’), 61) ((‘and’, ‘the’), 57)
A_4		((‘слів’, ‘слів’), 88) ((‘стилю’, ‘автора’), 68) ((‘службових’, ‘слів’), 63) ((‘визначення’, ‘стилю’), 61) ((‘списку’, ‘сводеша’), 56) ((‘стопових’, ‘слів’), 48) ((‘визначення’, ‘автора’), 45) ((‘авторського’, ‘мовлення’), 33) ((‘опорних’, ‘слів’), 31) ((‘стилю’, ‘стилю’), 30)	((‘text’, ‘content’), 30) ((‘web’, ‘mining’), 24) ((‘keywords’, ‘text’), 23) ((‘keywords’, ‘defined’), 22) ((‘stage’, ‘1’), 20) ((‘analysis’, ‘text’), 18) ((‘step’, ‘2’), 18) ((‘keywords’, ‘content’), 17) ((‘content’, ‘monitoring’), 17) ((‘step’, ‘1’), 17)	((‘на’, ‘основі’), 21) ((‘психологічного’, ‘стану’), 18) ((‘психологічного’, ‘особистості’), 17) ((‘контент’, ‘аналізу’), 16) ((‘стану’, ‘особистості’), 15) ((‘маркованих’, ‘слів’), 15) ((‘зрізу’, ‘психологічного’), 14) ((‘зрізу’, ‘стану’), 14) ((‘зрізу’, ‘особистості’), 14) ((‘особистості’, ‘на’), 14)

Continuation of Table 7

1	2	3	4	5
		(('fragment', 'fragment'), 37) (('reference', 'fragment'), 35) (('words', 'fragment'), 25) (('syntactic', 'words'), 21) (('frequency', 'fragment'), 19) (('swadesh', 'list'), 19) (('stop', 'words'), 18) (('author', 'style'), 17) (('fragment', '3'), 17) (('recognition', 'author'), 16)	(('ключових', 'слів'), 74) (('слів', 'в'), 24) (('web', 'mining'), 22) (('текстового', 'контенту'), 21) (('на', '2'), 20) (('визначення', 'ключових'), 19) (('ключових', 'в'), 19) (('визначення', 'слів'), 18) (('слів', 'для'), 18) (('на', 'крок'), 18)	(('content', 'analysis'), 40) (('psychological', 'personality'), 27) (('psychological', 'state'), 26) (('social', 'networks'), 22) (('marked', 'words'), 21) (('state', 'personality'), 20) (('based', 'analysis'), 19) (('psychological', 'based'), 18) (('state', 'based'), 18) (('based', 'content'), 18)

After the transference into a space of some dimension, each word becomes a vector, so words can form basic relational operations of addition, subtraction, multiplication, etc. Besides, let us consider the analysis through the bigrams (A_2 in Table 7) and the skipgrams (A_3 in Table 7). The results are better than those obtained through word2vec, which means that it is the best way to analyse skipgrams with a value of 3 and also to eliminate stop words in English (A_4 in Table 7). However, these results are far enough from the ones listed in Table 5. The outcome is worse due to the failure to identify punctuation marks and the use of stop words in linguistic analysis as content units of speech.

7. Conclusions

1. The study has developed a method for determining stable word combinations while identifying keywords of text content in standard passages of an author's text. For this purpose, the well-known statistical methods for determining stable word combinations when identifying keywords of text content were analysed. The factors influencing the quality of identifying stable word combinations were determined during the pre-linguistic elaboration of these texts. A comparative analysis of the corresponding methods was carried out on the basis of the obtained results. The developed method consists in using Zipf's law in the formation of stable word combinations as keywords, taking into account the following rules of a preliminary linguistic processing of the text:

– removing all word stops; bigrams are formed only within the limits of punctuation marks; the verb and the pronoun are to be considered punctuation marks;

– verbs are determined by their inflexions; bigrams are formed on the basis of stems without taking into account inflexions;

– adjectives are identified by their inflexions, and it is assumed that adjectives should occupy only the first place in the bigrams of Ukrainian texts.

This allowed taking into account the peculiarities of constructing keywords in the Ukrainian language, regardless of the inflexions within the word combinations. Also, the results obtained were closer to the number of keywords identified by the authors. This increases 1.4 times the degree of relevancy of the analysed content.

2. A program set has been developed to identify stable word combinations as keywords. An approach has been suggested for devising linguistic content analysis software to determine stable word combinations while identifying keywords of Ukrainian and English text-based contents. The peculiarity of the approach is that the linguistic statistical analysis of lexical units is adapted to the peculiarities of Ukrainian-language and English-language words/texts.

The developed information system, which is based on identified stable word combinations, helps convey more accurately the analysed content in accordance with the author' idea about it. This can produce a more accurate search result for the user and can better render the opinion of the author about the content under analysis.

3. The results of the experimental testing of the proposed method of content analysis of English and Ukrainian texts for determining stable word combinations when identifying the keywords of technical texts have been verified.

The developed method conveys the content of the analysed text by the identified keywords in the form of stable word combinations more accurately than other known resources. Further experimental research requires approbation of the proposed method for determining stable word combinations in other categories of texts – scientific, humanitarian, belletristic, journalistic, etc.

References

1. Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology / Lytvyn V., Vysotska V., Pukach P., Bobyk I., Uhryn D. // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 4, Issue 2 (88). P. 10–19. doi: 10.15587/1729-4061.2017.107512
2. Development of a method for determining the keywords in the slavic language texts based on the technology of web mining / Lytvyn V., Vysotska V., Pukach P., Brodyak O., Ugryn D. // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 2, Issue 2 (86). P. 14–23. doi: 10.15587/1729-4061.2017.98750
3. The method of formation of the status of personality understanding based on the content analysis / Lytvyn V., Pukach P., Bobyk I., Vysotska V. // Eastern-European Journal of Enterprise Technologies. 2016. Vol. 5, Issue 2 (83). P. 4–12. doi: 10.15587/1729-4061.2016.77174
4. Mobasher B. Data mining for web personalization // The adaptive web. 2007. P. 90–135. doi: 10.1007/978-3-540-72079-9_3
5. Dinucă C. E., Ciobanu D. Web Content Mining // Annals of the University of Petro ani. Economics. 2012. Vol. 12, Issue 1. P. 85–92.

6. Xu G., Zhang Y., Li L. Web content mining // Web Mining and Social Networking. 2011. P. 71–87. doi: 10.1007/978-1-4419-7735-9_4
7. Khomytska I., Teslyuk V. The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level // Advances in Intelligent Systems and Computing. 2017. Vol. 512. P. 149–163. doi: 10.1007/978-3-319-45991-2_10
8. Khomytska I., Teslyuk V. Specifics of phonostatistical structure of the scientific style in English style system // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: 10.1109/stc-csit.2016.7589887
9. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i komp'yuternaya lingvistika / Bol'shakova E., Klyshinskiy E., Lande D., Noskov A., Peskova O., Yagunova E. Moscow: MIEM, 2011. 272 p.
10. Anisimov A., Marchenko A. Sistema obrabotki tekstov na estestvennom yazyke // Iskusstvennyy intellekt. 2002. Issue 4. P. 157–163.
11. Perebyinis V. Matematychna linhvistyka. Ukrainska mova. Kyiv, 2000. P. 287–302.
12. Buk S. Osnovy statystychnoi lingvistyky. Lviv, 2008. 124 p.
13. Perebyinis V. Statystychni metody dlia linhvistiv. Vinnytsia, 2013. 176 p.
14. Braslavskiy P. I. Intellektual'nye informacionnye sistemy. URL: http://www.kansas.ru/ai2006/
15. Lande D., Zhyhalo V. Pidkhid do rishennia problem poshuku dvomovnoho plahiatu // Problemy informatyzatsii ta upravlinnia. 2008. Issue 2 (24). P. 125–129.
16. Varfolomeev A. Psihosemantika slova i lingvostatistika teksta. Kaliningrad, 2000. 37 p.
17. Sushko S., Fomychova L., Barsukov Ye. Chastoty povtoriuvanosti bukv i bihram u vidkrytykh tekstakh ukrainskoiu movoiu // Ukrainian Information Security Research Journal. 2010. Vol. 12, Issue 3 (48). doi: 10.18372/2410-7840.12.1968
18. Kognitivnaya stilometriya: k postanovke problemy. URL: http://www.manekin.narod.ru/hist/styl.htm
19. Kocherhan M. Vstup do movoznavstva. Kyiv, 2005.
20. Rodionova E. Metody atribucii hudozhestvennyh tekstov // Strukturnaya i prikladnaya lingvistika. 2008. Issue 7. P. 118–127.
21. Meshcheryakov R. V., Vasyukov N. S. Modeli opredeleniya avtorstva teksta. URL: http://db.biysk.secna.ru/conference/conference.conference.doc_download?id_thesis_dl=427
22. Morozov N. A. Lingvisticheskie spektry. URL: http://www.textology.ru/library/book.aspx?bookId=1&textId=3
23. Victana. URL: http://victana.lviv.ua/index.php/kliuchovi-slova
24. Method of Integration and Content Management of the Information Resources Network / Kanishcheva O., Vysotska V., Chyrun L., Gozhyj A. // Advances in Intelligent Systems and Computing. 2017. Vol. 689. P. 204–216. doi: 10.1007/978-3-319-70581-1_14
25. Information resources processing using linguistic analysis of textual content / Su J., Vysotska V., Sachenko A., Lytvyn V., Burov Y. // 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). 2017. doi: 10.1109/idaacs.2017.8095038
26. The risk management modelling in multi project environment / Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: 10.1109/stc-csit.2017.8098730
27. Peculiarities of content forming and analysis in internet newspaper covering music news / Korobchinsky M., Chyrun L., Chyrun L., Vysotska V. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: 10.1109/stc-csit.2017.8098735
28. Intellectual system design for content formation / Naum O., Chyrun L., Vysotska V., Kanishcheva O. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: 10.1109/stc-csit.2017.8098753
29. The Contextual Search Method Based on Domain Thesaurus / Lytvyn V., Vysotska V., Burov Y., Veres O., Rishnyak I. // Advances in Intelligent Systems and Computing. 2017. Vol. 689. P. 310–319. doi: 10.1007/978-3-319-70581-1_22
30. Marchenko O. Modeliuvannia semantychnoho kontekstu pry analizi tekstiv na pryrodniy movi // Visnyk Kyivskoho universytetu. 2006. Issue 3. P. 230–235.
31. Jivani A. G. A Comparative Study of Stemming Algorithms // Int. J. Comp. Tech. Appl. 2011. Vol. 2, Issue 6. P. 1930–1938.
32. Using Structural Topic Modeling to Detect Events and Cluster Twitter Users in the Ukrainian Crisis / Mishler A., Crabb E. S., Paletz S., Hefright B., Golonka E. // Communications in Computer and Information Science. 2015. Vol. 528. P. 639–644. doi: 10.1007/978-3-319-21380-4_108
33. Rodionova E. Metody atribucii hudozhestvennyh tekstov // Strukturnaya i prikladnaya lingvistika. 2008. Issue 7. P. 118–127.
34. Bubleinyk L. Osoblyvosti khudozhnoho movlennia. Lutsk, 2000. 179 p.
35. Kowalska K., Cai D., Wade S. Sentiment Analysis of Polish Texts // International Journal of Computer and Communication Engineering. 2012. Vol. 1, Issue 1. P. 39–42. doi: 10.7763/ijcce.2012.v1.12
36. Kotsyba N. The current state of work on the Polish-Ukrainian Parallel Corpus (PolUKR) // Organization and Development of Digital Lexical Resources. 2009. P. 55–60.
37. Machinese Phrase Tagger. URL: http://www.connexor.com
38. VISL. URL: http://visl.sdu.dk
39. Classification Methods of Text Documents Using Ontology Based Approach / Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. // Advances in Intelligent Systems and Computing. 2017. Vol. 512. P. 229–240. doi: 10.1007/978-3-319-45991-2_15

40. Vysotska V. Linguistic analysis of textual commercial content for information resources processing // 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET). 2016. doi: 10.1109/tcset.2016.7452160
41. Vysotska V., Chyrun L., Chyrun L. Information technology of processing information resources in electronic content commerce systems // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: 10.1109/stc-csit.2016.7589909
42. Vysotska V., Chyrun L., Chyrun L. The commercial content digest formation and distributional process // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: 10.1109/stc-csit.2016.7589902
43. Content linguistic analysis methods for textual documents classification / Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: 10.1109/stc-csit.2016.7589903
44. Lytvyn V., Vysotska V. Designing architecture of electronic content commerce system // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). 2015. doi: 10.1109/stc-csit.2015.7325446
45. Vysotska V., Chyrun L. Analysis features of information resources processing // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). 2015. doi: 10.1109/stc-csit.2015.7325448
46. Application of sentence parsing for determining keywords in Ukrainian texts / Vasyl L., Victoria V., Dmytro D., Roman H., Zoriana R. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: 10.1109/stc-csit.2017.8098797
47. Maksymiv O., Rak T., Peleshko D. Video-based Flame Detection using LBP-based Descriptor: Influences of Classifiers Variety on Detection Efficiency // International Journal of Intelligent Systems and Applications. 2017. Vol. 9, Issue 2. P. 42–48. doi: 10.5815/ijisa.2017.02.06
48. Peleshko D., Rak T., Izonin I. Image Superresolution via Divergence Matrix and Automatic Detection of Crossover // International Journal of Intelligent Systems and Applications. 2016. Vol. 8, Issue 12. P. 1–8. doi: 10.5815/ijisa.2016.12.01
49. The results of software complex OPTAN use for modeling and optimization of standard engineering processes of printed circuit boards manufacturing / Bazylyk O., Taradaha P., Nadobko O., Chyrun L., Shestakevych T. // 2012 11th International Conference on "Modern Problems of Radio Engineering, Telecommunications and Computer Science" (TCSET). 2012. P. 107–108.
50. The software complex development for modeling and optimizing of processes of radio-engineering equipment quality providing at the stage of manufacture / Bondariev A., Kiselychnyk M., Nadobko O., Nedostup L., Chyrun L., Shestakevych T. // TCSET'2012. 2012. P. 159.
51. Riznyk V. Multi-modular Optimum Coding Systems Based on Remarkable Geometric Properties of Space // Advances in Intelligent Systems and Computing. 2017. Vol. 512. P. 129–148. doi: 10.1007/978-3-319-45991-2_9
52. Development and Implementation of the Technical Accident Prevention Subsystem for the Smart Home System / Teslyuk V., Beregovskyi V., Denysyuk P., Teslyuk T., Lozynskyi A. // International Journal of Intelligent Systems and Applications. 2018. Vol. 10, Issue 1. P. 1–8. doi: 10.5815/ijisa.2018.01.01
53. Basyuk T. The main reasons of attendance falling of internet resource // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). 2015. doi: 10.1109/stc-csit.2015.7325440
54. Pasichnyk V., Shestakevych T. The model of data analysis of the psychophysiological survey results // Advances in Intelligent Systems and Computing. 2017. Vol. 512. P. 271–281. doi: 10.1007/978-3-319-45991-2_18
55. Zhezhnych P., Markiv O. Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism Documentation Objects // Advances in Intelligent Systems and Computing. 2018. Vol. 689. P. 656–667. doi: 10.1007/978-3-319-70581-1_45
56. Burov E. Complex ontology management using task models // International Journal of Knowledge-based and Intelligent Engineering Systems. 2014. Vol. 18, Issue 2. P. 111–120. doi: 10.3233/kes-140291
57. Smart Data Integration by Goal Driven Ontology Learning / Chen J., Dosyn D., Lytvyn V., Sachenko A. // Advances in Big Data. 2016. P. 283–292. doi: 10.1007/978-3-319-47898-2_29
58. Google – word2vec. URL: https://github.com/danielfrg/word2vec/blob/master/examples/word2vec.ipynb