UDC 81`32
doi 10.17072/2073-6681-2022-2-18-26

Topic Modelling in Computer Security Discourse: a Case Study of Whitepaper Publications and News Feeds

Ekaterina V. Isaeva
Head of the Department of English for Professional Communication
Perm State University
15, Bukireva st., Perm, 614990, Russian Federation. ekaterinaisae@psu.ru

SPIN-code: 4468-9991
ORCID: https://orcid.org/0000-0003-1048-7492
ResearcherID: O-6777-2015
Scopus Author ID: 57204498718

Submitted 02 Mar 2022
Revised 30 Apr 2022
Accepted 15 May 2022

For citation
Isaeva E. V. Topic Modelling in Computer Security Discourse: a Case Study of Whitepaper Publications and News Feeds. *Vestnik Permskogo universiteta. Rossiyskaya i zarubezhnaya filologiya* [Perm University Herald. Russian and Foreign Philology], 2022, vol. 14, issue 2, pp. 18–26. doi 10.17072/2073-6681-2022-2-18-26 (In Eng.)

Abstract. Up-to-date information plays a crucial role in modern linguistic research. For this reason, computational linguistic methods, including those aided with analytical and machine-learning tools, are attracting growing attention. Some of their applications in cognitive-discursive linguistics are keyword extraction, topic modelling, and content analysis. Text-mining tools facilitate time-consuming linguistic work and add to the results' reliability and greater statistical precision by processing a significantly larger data volume. Most studies, however, have overlooked interference of socially significant but context-irrelevant (e.g. political) information into a specialized discourse by focusing mainly on one data format. The current study, aimed at topic modelling, has been carried out on the computer security discourse. We have implemented the project on the KNIME analytical platform. The model enables comparison between topics extracted from published articles and date-specific RSS news feeds. The study provides important insights into infodemiology and political incidental news exposure occurring in computer-security-oriented RSS feeds on the Kaspersky website but untraceable in the papers published on the same website in a PDF format. The results reported here provide further evidence for the need to consider the hypercontext of professional communication and employ real-time data in solving similar problems within cognitive-discursive linguistics.

Our contribution to the development of cognitive-discursive linguistics is the method for comparing topics within one discourse, taking into account near-real-time data. For computational linguistics, the significance of our work lies in describing a new application of the topic extraction workflow freely available on the KNIME hub.

Key words: topic modelling; computer security discourse; KNIME; infodemiology; political incidental news exposure; content analysis; RSS feeds; cognitive-discursive linguistics.

Introduction

The Internet is generally regarded as the leading information repository and plays a crucial role in news transfer. One of the methods for targeted news tracking and coping with an overwhelming amount of information is RSS feeds, "which are XML documents that provide users with new, frequently updated news content automatically and allow users to subscribe to it" [Gustafson et al. 2008: 232].

According to Y.C. Wu [Wu 2017], large and wide-coverage news corpora are attracting growing attention "in many research domains, including information retrieval, language modelling, question answering, and named entity recognition" because they provide big data for information retrieval and knowledge discovery. A. Chudinov et al. have provided the most in-depth analysis of the political and media discourses. They examined the dominant values determined by national and cultural traditions and the dominant ideology in society in a given period [Chudinov et al. 2021; Mukhametzyanova et al. 2020]. They prove the presence of a potential ideological component of meaning in the semantics of the newspaper titles, highlighting their ability to reflect the era's cultural and historical realities. This makes it possible to attribute the news article titles to artefacts of the period [Mukhametzyanova, Mardieva, Chudinov 2020] and boost the researchers' interest to discover historical and sociological trends through their discursive representation.

The specificity of the news media discourse is featured by A. Photiou et al., who claim that novel and politically oriented content propagates faster due to its broad appeal [Photiou et al. 2021]; thus, it is highly influential in opinion shaping. The power of the news and media discourse to manipulate people's opinions, beliefs, sentiments, and political views triggers interest in currently developing theories of Infodemiology and Political Incidental News Exposure. The former refers to using real-time information across the Internet, mainly social media, in informing public policy and influencing individuals' intentions [Liew, Lee 2021]. The latter is defined as "exposure to information that people encounter without actively seeking for it" [Matthes et al. 2020: 770].

Our paper explores the relationship between topics covered in the computer security texts published on the Kaspersky website[1] in two formats: popular scientific articles and RSS (news) feeds. We discuss the issue of Infodemiology and Political Incidental News Exposure in the computer security discourse triggered by the Russian – Ukraine conflict (February - March 2022). Citizens experience infodemiology and incidental news exposure when they see political information in situations where they use media for other purposes (in our case, updating their knowledge about computer security) than political ones [Matthes et al. 2020]. This work contributes to the research of the ideological component in media discourse and reveals susceptibility to political bias in professional communication.

Computer security discourse has been deeply studied from the cognitive-discursive perspective as permeated by metaphors [Isaeva et al. 2022; Isaeva 2019; Isaeva, Crawford 2019]. A neglected area in this field is the refraction of the computer security professional communication while embedding it into the social and political context, which make up the hyper context.

Cognitive-discursive research on linguistic and extralinguistic peculiarities traditionally relies on a manual categorisation of the language units, conducting opinion polls, surveys, interviews, or focus group discussions. This makes the research time consuming and, consequently, outdated when the study is aimed at highlighting the current social trends. T. Liew and C. Lee suggest "using data from social media platforms […], where researchers can collect near–real-time information that reflects prevailing perspectives and sentiments in the community" [Liew, Lee 2021: 2]. In our project, we follow this hint and address online publications and RSS feeds to obtain up-to-date news occurring on the Kaspersky website in a real-time mode.

This article is organised as follows: after the **Introduction** providing the background for the current study, we describe computational-linguistic **Methods** used in our experiment with a brief overview of other projects relying on similar methods and software. Next, we provide the results of the **Computer security knowledge-mining experiment**, supplying them with our interpretation in the **Discussion** section.

1. Methods

The paper is written on the interface of cognitive-discursive and computational linguistics and is centred around the computer security discourse, content analysis, natural language processing, text mining, and knowledge mining. To reach the goal, we apply modern methods and analytical instruments of computational linguistics and make cognitive discursive inferences.

Data analysis has been implemented within the machine learning principles. G. Ertek and L. Kailas define machine learning as computer algorithms relying on training and capable of identifying patterns, determining insights, and predicting [Ertek, Kailas: 2021]. Unsupervised machine learning "aims to discover patterns or structures in a dataset without considering any target attribute" [Ertek, Kailas 2021: 8-9]. Unsupervised machine learning implemented as text mining, i.e. automated retrieving knowledge from natural language texts, can be used for extracting keywords from various sources and tracking the evolution of a topic over time [Sebestyén, Domokos, Abonyi 2020] or text clustering and reviling different topics in texts united by a common theme and genre [Lee, Lim 2021]. This machine learning technique of topic modelling is focused on identifying critical topics from textual data based on statistical probability and correlations among words. The method reminds traditional linguistic thematic or content analysis. However, unlike thematic analysis, computer-aided topic modelling does not require manual labour to classify textual data, is appropriate for large volumes of texts [Liew, Lee 2021], hence enables efficient data management, greater transparency, and enhanced knowledge [Flores-Ruiz et al. 2021].

For the project discussed in the current paper, the data were collected following the topic modelling method implemented via the Konstanz Information Miner (KNIME), an open-source statistical and data mining platform for data pre-processing, analysing, integration, modelling, and visualisation [Flores-Ruiz et al. 2021]. The software is applicable to data of various types, including natural language texts provided in different formats (text documents, spreadsheets, hyperlinks, and others). Recent developments regarding using KNIME for text-mining purposes have led to a range of multidisciplinary projects. For example, V. N. Dancy-Scott et al. used KNIME to "evaluate the evolving use of HIV-related language in abstracts presented at the IAC from 1989 to 2014" [Dancy-Scott et al. 2018: 1]. They used a text mining module to create a terminology corpus of key HIV terms grouped into expert categories. The team has also applied the Tableau[2] visualisation software to analyse terms' frequencies and visualise the data with line graphs and word clouds [Dancy-Scott et al. 2018].

Another representative example of using KNIME for text mining is the cyberbullying detection model. The holistic multi-dimensional approach "takes into account individual-based, social network-based, episode-based and linguistic content-based cyberbullying features" [Liu, Zavarsky et al. 2019: 404]. The model has been tested on 922 episodes with 59459 comments from Instagram.

A project devoted to the topic modelling following keywords extraction was implemented by Sebestyén et al., who employed the Multi-document summarisation method to extract information from multiple texts written on the same topic. "The purpose of the method is to objectively explore the relationships between documents, identify key topics and compare documents according to the explored set of focal points" [Sebestyén et al. 2020: 2]. Then, the authors used a graph-based approach to extract keywords, generated NGrams of these words, and clustered the words based on the latent Dirichlet allocation (LDA) generative statistical model [Sebestyén et al. 2020].

From the cognitive-discursive linguistic perspective, the data in our project were subjected to content analysis, which, according to E. Budaev, occupies an important place in modern studies of media and рщlitical communication. The method is described as quantitative, ontologically focused on realism and the manifestation of political reality in the text, static representation of the text's semantics, verified by mathematical methods, statistical regularities without digging into the context [Budaev 2017]. In our case, these principles were implemented in the KNIME-aided model of topic extraction and modelling.

2. Results of the computer security knowledge-mining experiment

The experimental design used in the current study refers to computational linguistics. The approach is mainly based on the "Topic Extraction: Optimising the Number of Topics with the Elbow Method" [Dewi, Thiel 2017] and a respective sample model of the workflow provided via the KNIME hub[3].

The model is adapted to extract topics from the papers on computer security provided on the Whitepapers repository[4], and the RSS feeds[5] of the Kaspersky website[6]. To obtain the static data independent of the date the experiment takes place, thus less susceptible to sentiment bias, we provided the first sub-workflow with the papers published online on the Whitepapers repository in a PDF format. This sub-workflow will be used as a reference model, typical of the Kaspersky website publications. For diachronic data with daily updates, thus capturing spontaneous and nonregular deviations from the reference model, we refer to the RSS news feeds present on March 13, 2022.

Thus, the experimental input data on the experiment comprise nine text files in PDF format (the most recently published articles), and six URLs of RSS feeds. The output includes scatter plots indicating the optimal number of clusters per workflow, two tables providing the distribution of the keywords into clusters (one per each sub-workflow), and word clouds representing the keywords in a certain cluster.

To illustrate the algorithm of natural language processing and knowledge mining employing the chosen KNIME model, we briefly describe the process stages. For a deeper analysis of the workflow, one should address the developers' article [Dewi, Thiel 2017]. Fig. 1 gives an overview of the first sub-workflow linked to the PDF files. We use the PDF Parser node to recognise PDF documents. A separate record is created for each file, metadata is extracted, and the full text is recognised using the PDFBox library.

Text pre-processing is started by the Preprocessing node, which has one input port and splits the process into two streams. Stream 1 starts with a vector representation of the document (i.e. Document Vector).

Fig. 2 illustrates how the documents are merged and split into separate words. All the words are recognised as separate units in the PDF texts uploaed. Fig. 3 illustrates the words arranged in the form of a vector.

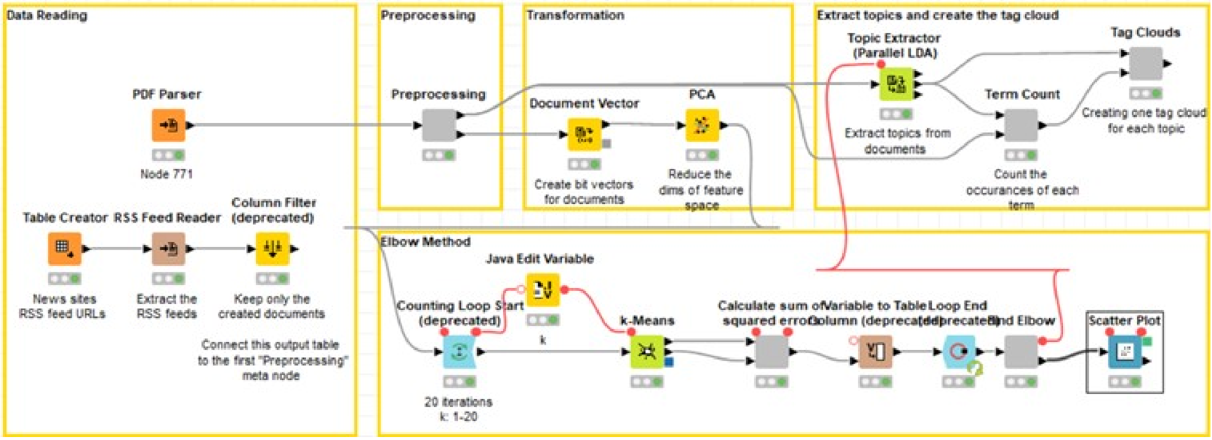
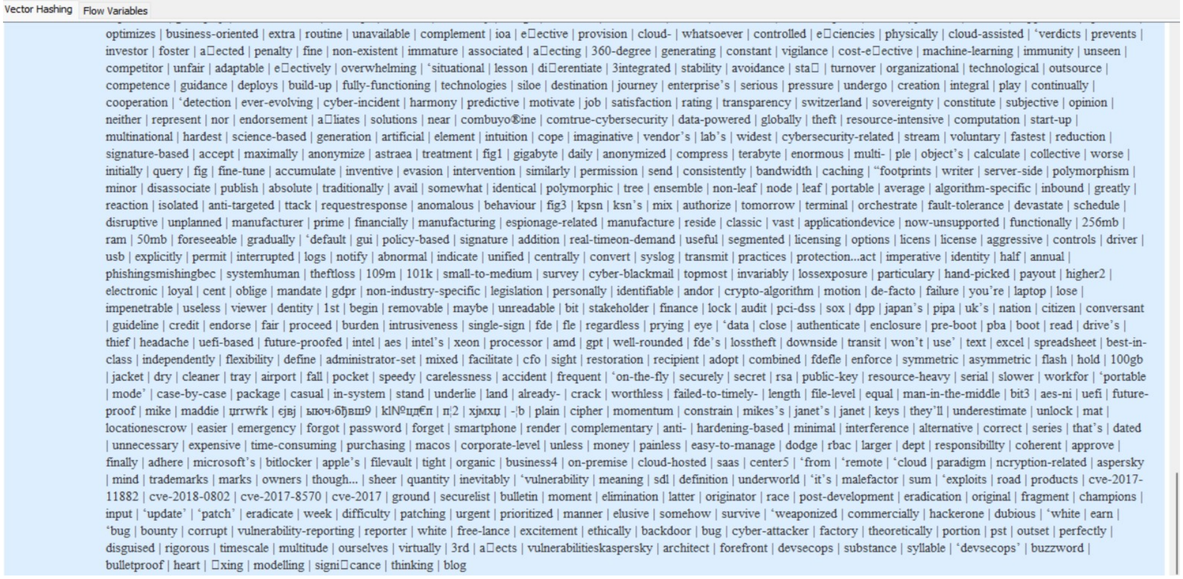
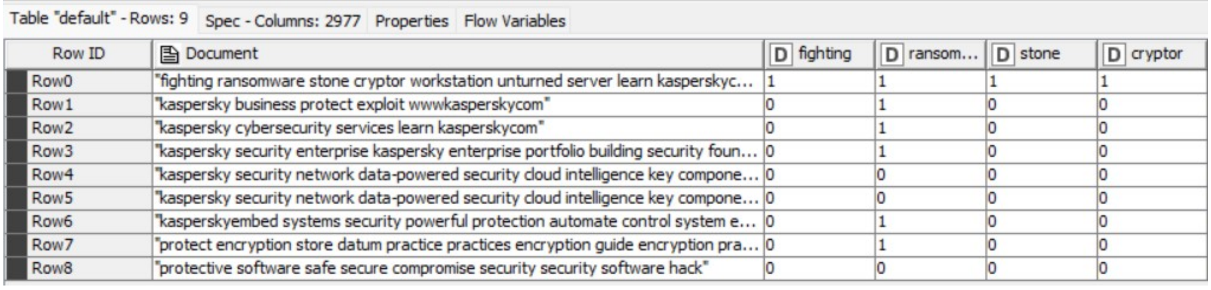

Data Reading
PDF Parser
Node 771
Table Creator
RSS Feed Reader
Column Filter (deprecated)
News sites RSS feed URLs
Extract the RSS feeds
Keep only the created documents
Connect this output table to the first "Preprocessing" meta node
Preprocessing
Preprocessing
Transformation
Document Vector
Create bit vectors for documents
PCA
Reduce the dims of feature space
Extract topics and create the tag cloud
Topic Extractor (Parallel LDA)
Extract topics from documents
Term Count
Count the occurances of each term
Tag Clouds
Creating one tag cloud for each topic
Elbow Method
Java Edit Variable
Counting Loop Start (deprecated)
20 iterations
k: 1-20
k
k-Means
Calculate sum of squared errors
Variable to Table Column (deprecated)
Loop End (deprecated)
Find Elbow
Scatter Plot

Fig. 1. The sub-workflow of knowledge retrieval from PDF files



```
optimizes | business-oriented | extra | routine | unavailable | complement | ioa | e□ective | provision | cloud- | whatsoever | controlled | e□ciencies | physically | cloud-assisted | ‘verdicts | prevents |
investor | foster | a□ected | penalty | fine | non-existent | immature | associated | a□ecting | 360-degree | generating | constant | vigilance | cost-e□ective | machine-learning | immunity | unseen |
competitor | unfair | adaptable | e□ectively | overwhelming | ‘situational | lesson | di□erentiate | 3integrated | stability | avoidance | sta□ | turnover | organizational | technological | outsource |
competence | guidance | deploys | build-up | fully-functioning | technologies | siloe | destination | journey | enterprise’s | serious | pressure | undergo | creation | integral | play | continually |
cooperation | ‘detection | ever-evolving | cyber-incident | harmony | predictive | motivate | job | satisfaction | rating | transparency | switzerland | sovereignty | constitute | subjective | opinion |
neither | represent | nor | endorsement | a□liates | solutions | near | combuyo®ine | comtrue-cybersecurity | data-powered | globally | theft | resource-intensive | computation | start-up |
multinational | hardest | science-based | generation | artificial | element | intuition | cope | imaginative | vendor’s | lab’s | widest | cybersecurity-related | stream | voluntary | fastest | reduction |
signature-based | accept | maximally | anonymize | astraea | treatment | fig1 | gigabyte | daily | anonymized | compress | terabyte | enormous | multi- | ple | object’s | calculate | collective | worse |
initially | query | fig | fine-tune | accumulate | inventive | evasion | intervention | similarly | permission | send | consistently | bandwidth | caching | “footprints | writer | server-side | polymorphism |
minor | disassociate | publish | absolute | traditionally | avail | somewhat | identical | polymorphic | tree | ensemble | non-leaf | node | leaf | portable | average | algorithm-specific | inbound | greatly |
reaction | isolated | anti-targeted | ttack | requestresponse | anomalous | behaviour | fig3 | kpsn | ksn’s | mix | authorize | tomorrow | terminal | orchestrate | fault-tolerance | devastate | schedule |
disruptive | unplanned | manufacturer | prime | financially | manufacturing | espionage-related | manufacture | reside | classic | vast | applicationdevice | now-unsupported | functionally | 256mb |
ram | 50mb | foreseeable | gradually | ‘default | gui | policy-based | signature | addition | real-timeon-demand | useful | segmented | licensing | options | licens | license | aggressive | controls | driver |
usb | explicitly | permit | interrupted | logs | notify | abnormal | indicate | unified | centrally | convert | syslog | transmit | practices | protection...act | imperative | identity | half | annual |
phishingsmishingbec | systemhuman | theftloss | 109m | 101k | small-to-medium | survey | cyber-blackmail | topmost | invariably | lossexposure | particulary | hand-picked | payout | higher2 |
electronic | loyal | cent | oblige | mandate | gdpr | non-industry-specific | legislation | personally | identifiable | andor | crypto-algorithm | motion | de-facto | failure | you’re | laptop | lose |
impenetrable | useless | viewer | dentity | 1st | begin | removable | maybe | unreadable | bit | stakeholder | finance | lock | audit | pci-dss | sox | dpp | japan’s | pipa | uk’s | nation | citizen | conversant
| guideline | credit | endorse | fair | proceed | burden | intrusiveness | single-sign | fde | fle | regardless | prying | eye | ‘data | close | authenticate | enclosure | pre-boot | pba | boot | read | drive’s |
thief | headache | uefi-based | future-proofed | intel | aes | intel’s | xeon | processor | amd | gpt | well-rounded | fde’s | losstheft | downside | transit | won’t | use’ | text | excel | spreadsheet | best-in-
class | independently | flexibility | define | administrator-set | mixed | facilitate | cfo | sight | restoration | recipient | adopt | combined | fdefle | enforce | symmetric | asymmetric | flash | hold | 100gb
| jacket | dry | cleaner | tray | airport | fall | pocket | speedy | carelessness | accident | frequent | ‘on-the-fly | securely | secret | rsa | public-key | resource-heavy | serial | slower | workfor | ‘portable
| mode’ | case-by-case | package | casual | in-system | stand | underlie | land | already- | crack | worthless | failed-to-timely- | length | file-level | equal | man-in-the-middle | bit3 | aes-ni | uefi | future-
proof | mike | maddie | џrrwŕk | єjвj | ьючљбђвш9 | klN№џд€п | п¦2 | хјмхџ | -¦b | plain | cipher | momentum | constrain | mikes’s | janet’s | janet | keys | they’ll | underestimate | unlock | mat |
locationescrow | easier | emergency | forgot | password | forget | smartphone | render | complementary | anti- | hardening-based | minimal | interference | alternative | correct | series | that’s | dated
| unnecessary | expensive | time-consuming | purchasing | macos | corporate-level | unless | money | painless | easy-to-manage | dodge | rbac | larger | dept | responsibillty | coherent | approve |
finally | adhere | microsoft’s | bitlocker | apple’s | filevault | tight | organic | business4 | on-premise | cloud-hosted | saas | center5 | ‘from | ‘remote | ‘cloud | paradigm | ncryption-related | aspersky
| mind | trademarks | marks | owners | though... | sheer | quantity | inevitably | ‘vulnerability | meaning | sdl | definition | underworld | ‘it’s | malefactor | sum | ‘exploits | road | products | cve-2017-
11882 | cve-2018-0802 | cve-2017-8570 | cve-2017 | ground | securelist | bulletin | moment | elimination | latter | originator | race | post-development | eradication | original | fragment | champions |
input | ‘update’ | ‘patch’ | eradicate | week | difficulty | patching | urgent | prioritized | manner | elusive | somehow | survive | ‘weaponized | commercially | hackerone | dubious | ‘white | earn |
‘bug | bounty | corrupt | vulnerability-reporting | reporter | white | free-lance | excitement | ethically | backdoor | bug | cyber-attacker | factory | theoretically | portion | pst | outset | perfectly |
disguised | rigorous | timescale | multitude | ourselves | virtually | 3rd | a□ects | vulnerabilitieskaspersky | architect | forefront | devsecops | substance | syllable | ‘devsecops’ | buzzword |
bulletproof | heart | □xing | modelling | signi□cance | thinking | blog
```

Fig. 2. The documents’ vector hashing



Row ID	Document	D fighting	D ransom…	D stone	D cryptor
Row0	"fighting ransomware stone cryptor workstation unturned server learn kasperskyc…	1	1	1	1
Row1	"kaspersky business protect exploit wwwkasperskycom"	0	1	0	0
Row2	"kaspersky cybersecurity services learn kasperskycom"	0	1	0	0
Row3	"kaspersky security enterprise kaspersky enterprise portfolio building security foun…	0	1	0	0
Row4	"kaspersky security network data-powered security cloud intelligence key compone…	0	0	0	0
Row5	"kaspersky security network data-powered security cloud intelligence key compone…	0	0	0	0
Row6	"kasperskyembed systems security powerful protection automate control system e…	0	1	0	0
Row7	"protect encryption store datum practice practices encryption guide encryption pra…	0	1	0	0
Row8	"protective software safe secure compromise security security software hack"	0	0	0	0

Fig. 3. The document vector representation matrix

As is described by KNIME’s developers, “a document vector is a numeric representation of a text in a matrix where each document represents a row, and each unique term represents a column. The binary encoding shows the absence/presence of a term in the document. The Document Vector node performs this transformation, which is needed, for example, to cluster or perform classification on the text data” [Document Vector Node 2020]. The matrix in Fig. 3 consists of 9 rows (separate documents) and 2977 columns (identified separate word units). If the word is found in the text, the field is assigned 1; if not, – 0.

Next, the PCA node is used. This node performs principal component analysis (PCA) on the given data. The input data is projected from the original feature space into a smaller space with minimal loss of information. After that, we move on to the Elbow method. We start the counter loop, set it to 20 iterations. LoopStart is the node that triggers the loop, executed the specified number of times. At the end of the loop, you need LoopEnd, which collects the results of all loop iterations. All nodes in between are executed as many times as specified in the LoopStart dialogue box.

We use the k-Means node to extract knowledge from the text. This node outputs cluster centres for a predefined number of clusters. K-means performs explicit clustering, assigning the data vector evenly to one cluster. The algorithm terminates when the cluster assignments no longer change. The clustering algorithm uses the Euclidean distance between selected attributes. Next, the root mean square error is calculated, and the Variable to Table Column node is used to extract the variables from the workflow and enter them into the input table. At the end of the cycle, we add a Loop End node. It marks the end of the workflow cycle and collects intermediate results by concatenating the incoming tables line by line (Fig. 4).

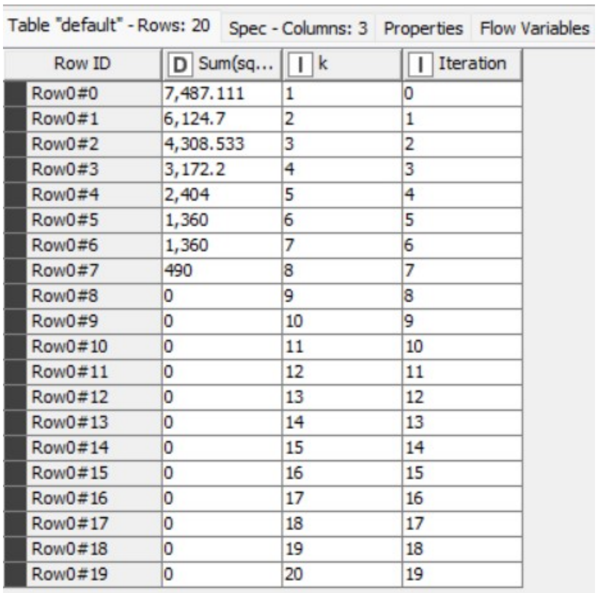


Table "default" - Rows: 20
Spec - Columns: 3
Properties
Flow Variables

Row ID	D Sum(sq...	I k	I Iteration
Row0#0	7,487.111	1	0
Row0#1	6,124.7	2	1
Row0#2	4,308.533	3	2
Row0#3	3,172.2	4	3
Row0#4	2,404	5	4
Row0#5	1,360	6	5
Row0#6	1,360	7	6
Row0#7	490	8	7
Row0#8	0	9	8
Row0#9	0	10	9
Row0#10	0	11	10
Row0#11	0	12	11
Row0#12	0	13	12
Row0#13	0	14	13
Row0#14	0	15	14
Row0#15	0	16	15
Row0#16	0	17	16
Row0#17	0	18	17
Row0#18	0	19	18
Row0#19	0	20	19

Fig. 4. K-Means counter result

The model developers explain the essence of the Elbow method as choosing “the number of clusters at which the SSE decreases abruptly. This produces a so-called ‘elbow’ in the graph” [Dewi, Thiel 2017]. The result of the first sub-flow is a scatter plot showing the number of clusters and their accuracy (Fig. 5), obtained with the Scatter Plot node. According to Fig. 5, the optimal number of clusters is three, i.e. the first drop in the sum of squared errors rate is after the 2nd cluster.

Stream 2 starts with a simple parallel threaded LDA implementation using the Topic Extractor node (Parallel LDA). LDA stands for Latent Dirichlet allocation, a method for identifying the main topics from utterances in an inductive way. Fig. 6 illustrates the distribution of terms into clusters (topics).

As shown in Fig. 6, the software has identified 3 clusters, distributed the keywords into these clusters, and output ten most frequently used terms per each cluster in the table. As a result of the second sub-flow, we get a word cloud for each topic. For instance, Fig. 7 represents a word cloud generated for topic_2. Now we change the input data - instead of PDF files, we use RSS feeds (Fig. 8).

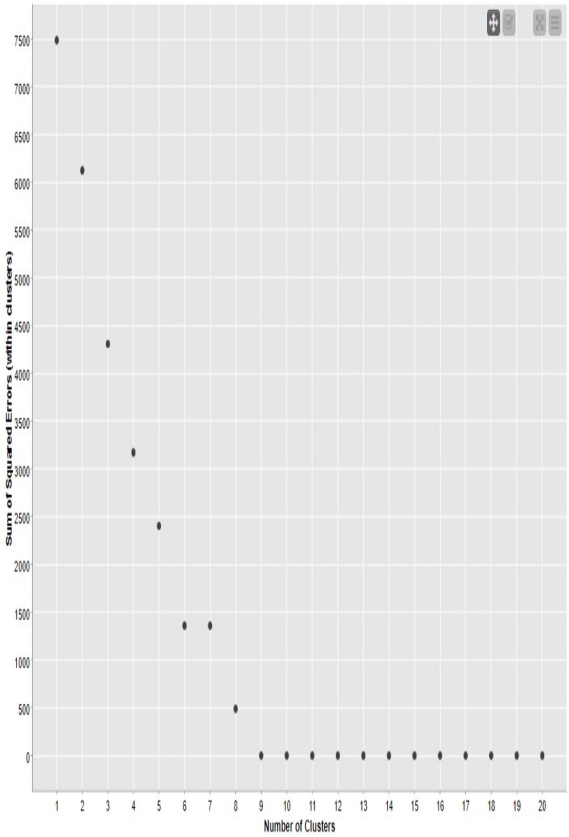


Fig. 5. The number of clusters and definition accuracy (PDF)

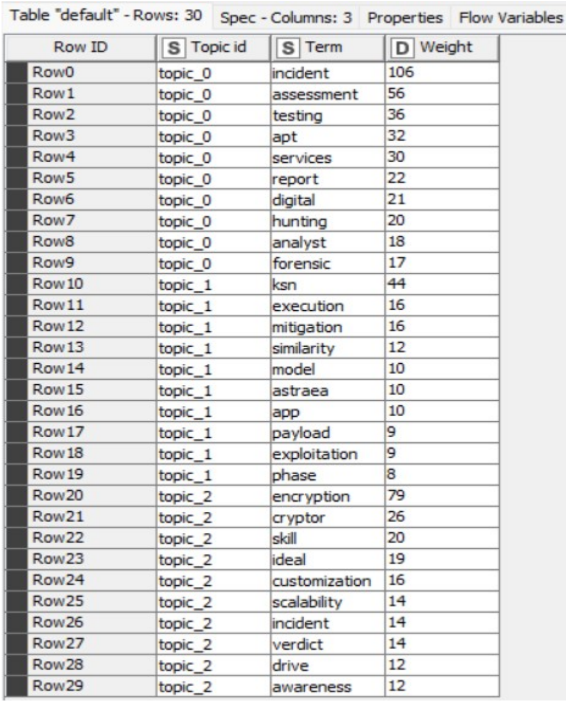


Table "default" - Rows: 30
Spec - Columns: 3
Properties
Flow Variables

Row ID	S Topic id	S Term	D Weight
Row0	topic_0	incident	106
Row1	topic_0	assessment	56
Row2	topic_0	testing	36
Row3	topic_0	apt	32
Row4	topic_0	services	30
Row5	topic_0	report	22
Row6	topic_0	digital	21
Row7	topic_0	hunting	20
Row8	topic_0	analyst	18
Row9	topic_0	forensic	17
Row10	topic_1	ksn	44
Row11	topic_1	execution	16
Row12	topic_1	mitigation	16
Row13	topic_1	similarity	12
Row14	topic_1	model	10
Row15	topic_1	astraea	10
Row16	topic_1	app	10
Row17	topic_1	payload	9
Row18	topic_1	exploitation	9
Row19	topic_1	phase	8
Row20	topic_2	encryption	79
Row21	topic_2	cryptor	26
Row22	topic_2	skill	20
Row23	topic_2	ideal	19
Row24	topic_2	customization	16
Row25	topic_2	scalability	14
Row26	topic_2	incident	14
Row27	topic_2	verdict	14
Row28	topic_2	drive	12
Row29	topic_2	awareness	12

Fig. 6. Topic distribution of terms

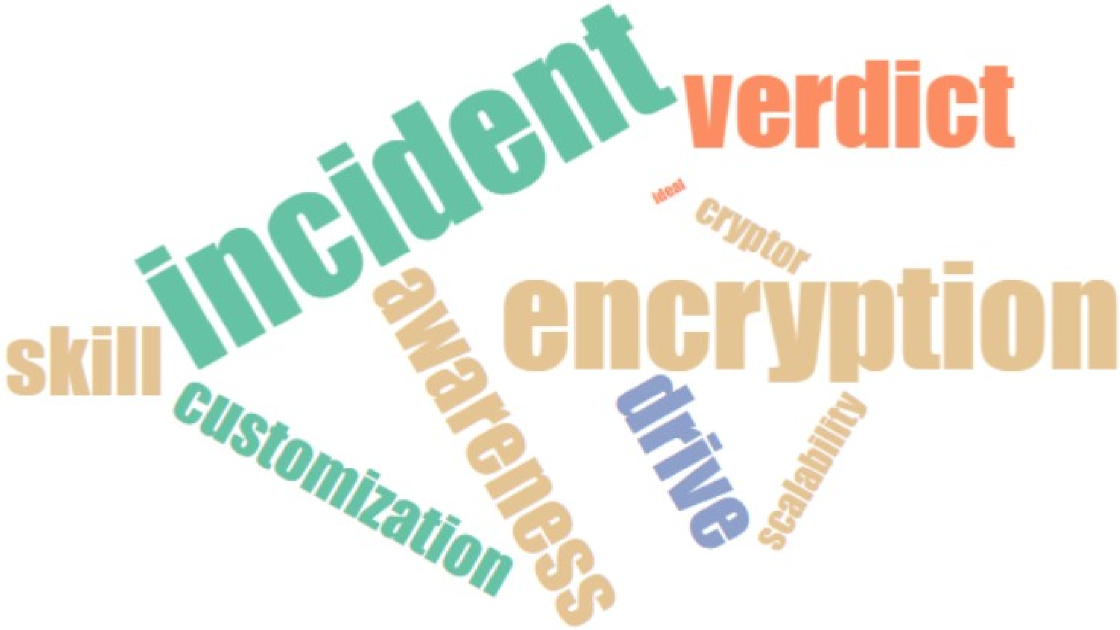


Fig. 7. A sample word cloud obtained from the Whitepapers repository publications (PDF)

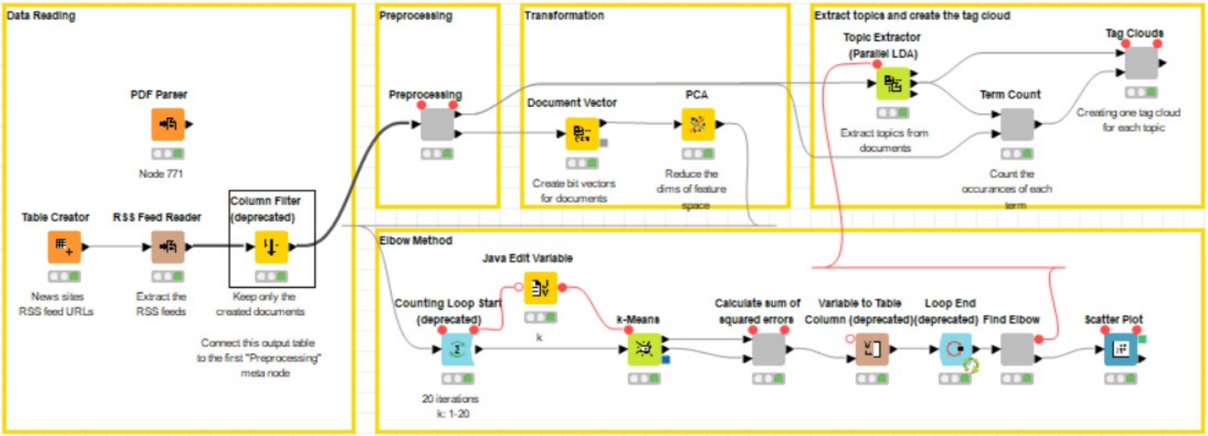
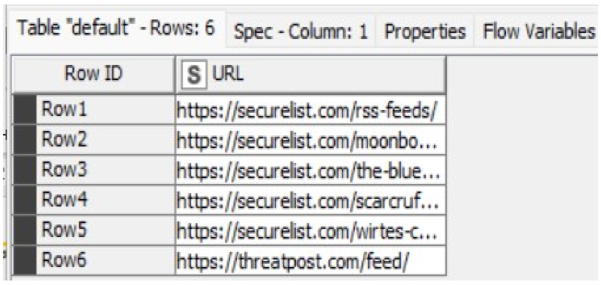

Data Reading
PDF Parser
Node 771
Table Creator
News sites RSS feed URLs
RSS Feed Reader
Extract the RSS feeds
Column Filter (deprecated)
Keep only the created documents
Connect this output table to the first "Preprocessing" meta node
Preprocessing
Preprocessing
Transformation
Document Vector
Create bit vectors for documents
PCA
Reduce the dims of feature space
Extract topics and create the tag cloud
Topic Extractor (Parallel LDA)
Extract topics from documents
Term Count
Count the occurances of each term
Tag Clouds
Creating one tag cloud for each topic
Elbow Method
Java Edit Variable
k
Counting Loop Start (deprecated)
20 iterations k: 1-20
k-Means
Calculate sum of squared errors
Variable to Table Column (deprecated)
Loop End (deprecated)
Find Elbow
Scatter Plot

Fig. 8. A workflow of a data mining of RSS feeds

In the Table Creator node configuration, manually enter 6 RSS threads (Fig. 9).



Row ID	S URL
Row1	https://securelist.com/rss-feeds/
Row2	https://securelist.com/moonbo…
Row3	https://securelist.com/the-blue…
Row4	https://securelist.com/scarcruf…
Row5	https://securelist.com/wirtes-c…
Row6	https://threatpost.com/feed/

Fig. 9. RSS feeds table

Text parsing is implemented through the RSS Feed Reader node. This node parses RSS feeds from the URLs specified in the input table and extracts information such as a title, a description, a publication date, and a link to the article from the feed entries. You can also create a document column, XML or HTTP response code. The document will be made based on the feed entry information, and the XML column displays the XML snippet for the specific feed entry. RSS Feed Parser uses the ROME(1.0) library[7].

Let us apply the Column Filter. This node enables filtering the columns from the input table, with only the desired columns being passed to the output table. In the dialogue box, the columns can be moved between the "Include" and "Exclude" lists. The rest of the steps are similar to those described above.

As a result of the second sub-workflow, an interesting deviation from the standard model has been noted. Fig. 10 illustrates an unexpected occurrence of the keywords *Russia* and *sanction*, clearly falling out of the computer security discourse.

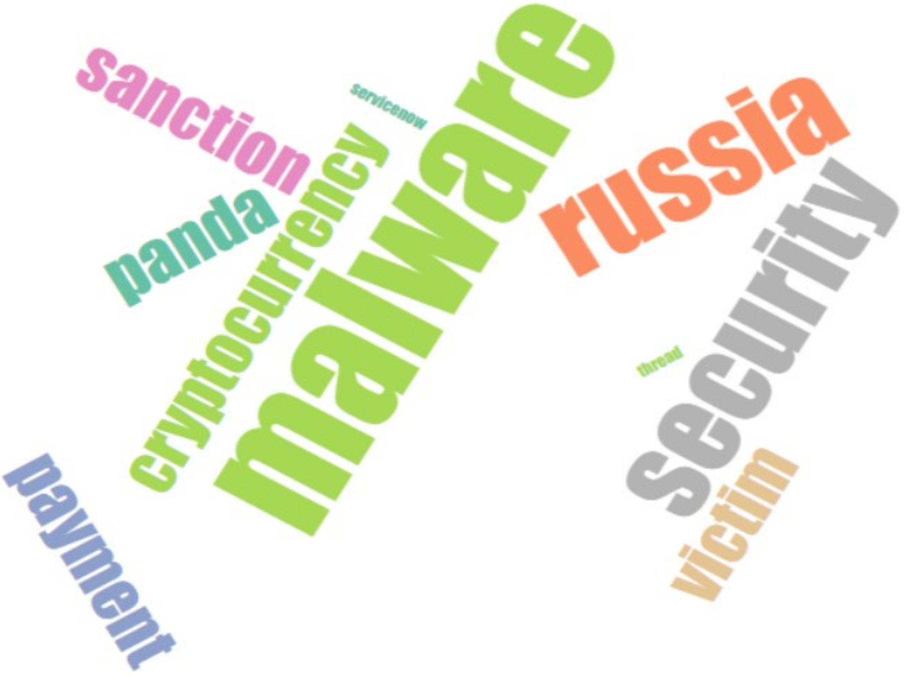


Fig. 10. A word cloud of the computer security RSS feeds' cluster containing nontrivial keywords

3. Discussion

Surprisingly, the results reveal a discrepancy in the topics modelled from the PDF files and RSS feeds retrieved from the Kaspersky website. Both were supposed to be focused on computer security issues. However, as expected, the PDF-based topics are related to software, man-computer interaction, and their safety. At the same time, one of the RSS-generated clusters included *Russia*, *sunction*, and *victim* as its keywords. This finding demonstrates interference of the political content in the computer security discourse. Thus, we may register the case of infodemiology and incidental political news exposure in the media type of professional communication.

This deviation can be explained by the high relevance and novelty of the political discussion of the events of February - March 2022, penetrating all the spheres of international communication, as well as information technology and computer security. We believe that the mismatch between the topics will not be so apparent sometime later when the political discussion loses its novelty. Given the study design and the time the experiment was performed, it was inevitable that the results might differ from those expected in a more stable political period.

Conclusion

This paper has shown how computational linguistics may contribute to getting and processing near-real-time data relevant to cognitive-discursive research. The results obtained using analytical software (KNIME) validate the relevance of the Info demiology and Political Incidental News Exposure theories for professional communication (computer security discourse). These results emphasise the importance of taking into account diachronic data representations when working on such tasks as topic modelling and keyword extraction due to possible episodic interference of extraneous information, novel and particularly urgent for the society in a specific period.

The usefulness of our work lies in testing A. Dewi and K. Thiel's method for topic modelling applied to the task of comparing thematic trends in static (published PDF documents) and regularly updated news feeds, e.g. RSS from Kaspersky website.

The present study has only examined the data current for March 13, 2022. Therefore, we do not make conclusions about a regular interference of novel and socially significant but out-of-context information into professional communication. However, this work has demonstrated that such cased might happen, thus should not be neglected. Future research should consider the potential effects of infodemiology and political incidental news exposure diachronically, comparing samples from different periods.

Endnotes

[1] https://www.kaspersky.com/

[2] https://www.tableau.com/

[3] https://www.knime.com/blog/topic-extraction-optimizing-the-number-of-topics-with-the-elbow-method?

[4] https://www.kaspersky.com/enterprise-security/resources/whitepapers?icid=gl_securelisheader_acq_ona_smm__onl_b2b_securelist_prodmen

[5] https://securelist.com/rss-feeds/

[6] https://www.kaspersky.com/

[7] https://rometools.github.io/rome/ROMEReleases/index.html

References

Budaev E. Metaphors of disease in the Russian press, *XLinguae*. 2021, vol. 10, issue 2, pp. 30–37. doi 10.18355/XL.2017.10.02.03. (In Russ.)

Chudinov A. P., Sergienko N. A., Glushak V. M. Good, Evil, Truth, Lie in Russian, Ukrainian, British, and American linguo-cultures: Results of a psycholinguistic experiment. *Sibirskiy Filologicheskiy Zhurnal* [The Siberian Journal of Philology], 2021, issue 2, pp. 297–311. doi 10.17223/18137083/75/21 (In Russ.)

Dancy-Scott N., Dutcher G. A., Keselman A., Hochstein C., Copty C., Ben-Senia D., Rajan S., Asencio M. G., Choi J. J. Trends in HIV terminology: Text mining and data visualization assessment of international AIDS conference abstracts over 25 years. *JMIR Public Health and Surveillance*, 2018, vol. 4, issue 5. doi 10.2196/PUBLICHEALTH.8552. (In Eng.)

Dewi A., Thiel K. Topic extraction: Optimizing the number of topics with the elbow method. *KNIME*, June 19, 2017. Available at: https://www.knime.com/blog/topic-extraction-optimizing-the-number-of-topics-with-the-elbow-method (accessed 30 Apr 2022). (In Eng.)

Document Vector Node. *KNIMETV*, December 9, 2020. Available at: https://www.youtube.com/watch?v=kLlmCWnknhE (accessed 30 Apr 2022). (In Eng.)

Flores-Ruiz D., Elizondo-Salto A., Barroso-González M. d. l. O. Using social media in tourist sentiment analysis: A case study of Andalusia during the Covid-19 pandemic. *Sustainability,* 2021, vol. 13, issue 7 (3836), pp. 1-19. doi 10.3390/SU13073836. (In Eng.)

Ertek G., Kailas L. Analyzing a decade of wind turbine accident news with topic modeling. *Sustainability,* 2021, vol. 13, issue 12757, pp. 1-34. doi 10.3390/su132212757 (In Eng.)

Isaeva E., Baiburova O., Manzhula O. Anthropomorphism in computer security terminology through the prizm of smart cognitive framing. *Science and Global Challenges of the 21st Century – Science and Technology*. Perm Forum 2021. Lecture Notes in Networks and Systems. 2022, vol. 342, pp. 460–474. doi 10.1007/978-3-030-89477-1_46. (In Eng.)

Isaeva E. V. Metaphor in terminology: Finding tools for efficient professional communication. *Fachsprache*, 2019, vol. 41, special issue 1. doi 10.24989/fs.v41is1.1766. (In Eng.)

Isaeva E. V., Crawford R. Semantic framing of computer viruses: The study of semantic roles' distribution. *Vestnik Permskogo universiteta. Rossiyskaya i zarubezhnaya filologiya* [Perm University Herald. Russian and Foreign Philology], 2019, vol. 11, issue 1, pp. 5–13. doi 10.17072/2073-6681-2019-1-5-13. (In Eng.)

Gustafson N., Pera, M. S., Ng, YK. Generating fuzzy equivalence classes on RSS news articles for retrieving correlated information. In: Gervasi O., Murgante B., Laganà A., Taniar D., Mun Y., Gavrilova M. L. (eds) *Computational Science and Its Applications – ICCSA 2008*. ICCSA 2008. Lecture Notes in Computer Science. 2008. Springer, Berlin, Heidelberg, vol. 5073, pp. 232–247. doi 10.1007/978-3-540-69848-7_20. (In Eng.)

Lee C., Lim C. From technological development to social advance: A review of Industry 4.0 through machine learning. *Technological Forecasting and Social Change*, 2021, vol. 167 (120653). doi 10.1016/J.TECHFORE.2021. 120653. (In Eng.)

Liew T. M., Lee C. S. Examining the utility of social media in Covid-19 vaccination: Unsupervised learning of 672,133 twitter posts. *JMIR Public Health and Surveillance*, 2021, vol. *7*, issue 11, pp. 1–19. doi 10.2196/29789. (In Eng.)

Liu Y., Zavarsky P., Malik Y. Non-linguistic features for cyberbullying detection on a social media platform using machine learning. In: Vaidya, J., Zhang, X., Li, J. (eds) *Cyberspace Safety and Security. CSS 2019.* Lecture Notes in Computer Science, vol. 11982. Springer, Cham, pp. 391–406. doi 10.1007/978-3-030-37337-5_31. (In Eng.)

Matthes J., Nanz A., Stubenvoll M., Heiss R. Processing news on social media. The political incidental news exposure model (PINE). *Journalism*, 2020, vol. 21, issue 8, pp. 1031–1048. doi: 10.1177/1464884920915371. (In Eng.)

Mukhametzyanova L. R., Mardieva L. A., Chudinov A. P. The titles of newspapers and magazines as artifacts of the epoch. *Journal of Research in Applied Linguistics*, 2020, vol. 11, pp. 400–405. doi 10.22055/RALS.2020.16338. (In Eng.)

Photiou A., Nicolaides C., Dhillon P. S. Social status and novelty drove the spread of online information during the early stages of COVID-19. *Scientific Reports*, vol. 11, issue 1 (20098). doi 10.1038/S41598-021-99060-Y. (In Eng.)

Sebestyén V., Domokos E., Abonyi J. Multilayer network based comparative document analysis (MUNCoDA). *MethodsX*, 2020, vol. 7, 100902. doi 10.1016/J.MEX.2020.100902. (In Eng.)

Wu Y. C. Multilingual news extraction via stopword language model scoring. *Journal of Intelligent Information Systems*, 2017, vol. 48, issue 1, pp. 191–213. doi 10.1007/S10844-016-0395-6. (In Eng.)

Тематическое моделирование в дискурсе компьютерной безопасности: исследование на примере публикаций информационных бюллетеней и новостных лент

Екатерина Владимировна Исаева
к. филол. н., доцент, заведующий кафедрой английского языка профессиональной коммуникации
Пермский государственный национальный исследовательский университет
614990, Россия, г. Пермь, ул. Букирева, 15. ekaterinaisae@psu.ru

SPIN-код: 4468-9991
ORCID: https://orcid.org/0000-0003-1048-7492
ResearcherID: O-6777-2015
Scopus Author ID: 57204498718

Статья поступила в редакцию 02.03.2022
Одобрена после рецензирования 30.04.2022
Принята к публикации 15.05.2022

Информация для цитирования
Isaeva E. V. Topic Modelling in Computer Security Discourse: a Case Study of Whitepaper Publications and News Feeds // Вестник Пермского университета. Российская и зарубежная филология. 2022. Т. 14, вып. 2. С. 18–26. doi 10.17072/2073-6681-2022-2-18-26

Аннотация. Актуальная информация играет важную роль в современных лингвистических исследованиях. По этой причине методы компьютерной лингвистики, в том числе с использованием аналитических инструментов и средств машинного обучения, привлекают все большее внимание. Некоторые из них применяются в когнитивно-дискурсивной лингвистике для извлечения ключевых слов, тематического моделирования и контентного анализа. Инструменты для обработки текста облегчают трудоемкую работу лингвиста и повышают надежность и статистическую точность результатов за счет обработки значительно большего объема данных. Большинство исследований, однако, упускают из виду интерференцию социально значимой, но контекстуально не релевантной (например, политической) информации в специализированный дискурс, фокусируясь в основном на каком-то одном формате данных. Настоящее исследование, направленное на тематическое моделирование, выполнено в рамках дискурса компьютерной безопасности. Проект реализован на аналитической платформе KNIME. Разработанная модель позволяет сравнивать темы, извлеченные из опубликованных статей и новостных RSS-лент, привязанных к конкретной дате. Данное исследование позволяет получить важные сведения об инфодемиологии и случайном попадании политических новостей в RSS-ленты сайта Касперского, ориентированные на компьютерную безопасность, которые не прослеживаются в информационных бюллетенях, опубликованных на том же сайте в формате PDF. Представленные в статье результаты служат очередным подтверждением необходимости учитывать гиперконтекст профессиональной коммуникации и оперировать данными реального времени при решении подобных задач в рамках когнитивно-дискурсивной лингвистики. Наш вклад в развитие когнитивно-дискурсивной лингвистики заключается в применении метода сравнения тем в рамках одного дискурса с учетом данных, полученных в режиме реального времени. Для компьютерной лингвистики значимость данной работы заключается в описании нового применения алгоритма извлечения тем, размещенного в свободном доступе на портале KNIME.

Ключевые слова: когнитивно-дискурсивная лингвистика; дискурс компьютерной безопасности; KNIME; инфодемиология; контент-анализ; RSS-ленты; тематическое моделирование.