SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2020
ORGANIZATION AND MANAGEMENT SERIES NO. 148

IDENTIFICATION OF THE LEADING RESEARCH DOMAINS AND GROUPING OF ARTICLES ON THE SMART CITY USING TEXT MINING

Dariusz ZDONEK

Silesian University of Technology in Gliwice, Faculty of Organisation and Management;
dariusz.zdonek@polsl.pl, ORCID: 0000-0002-6190-9643

Purpose: The objective of the paper is to use text mining to identify leading research domains concerning the smart city following an analysis of research articles with a high citation index according to the Web of Science.
Design/methodology/approach: An original method is proposed for analysing academic texts using the R language, tokenisation, lemmatisation, n-grams and correspondence analysis. The author analysed fifty of the most cited articles indexed in the Web of Science from 2014 to 2019.
Findings: The paper presents the advantages and drawbacks of the proposed method of analysing research publications. The assets include automation and repeatability of the analysis of a large number of documents and improved knowledge about links among the articles in terms of research domains. The disadvantage is the loss of information from diagrams and figures. The method identified two leading research domains related to the notion of the smart city, technologies and systems. The analysed publications were categorised by selected keywords.
Research limitations/implications: Future work should include further refinement of the assumptions for the method, analyses of a more significant number of research texts and a narrowing down of the domain of the smart city. It is desirable to consider other functional domains of the city, such as energy, public health, environmental protection or transport.
Practical implications: The proposed method can complement a standard literature analysis regarding the smart city. The leading research domains related to the smart city in the analysed articles were systems and technologies employed to improve how the city operates.
Social implications: Text mining can be employed by various experts focusing on the smart city and constitutes a refreshing complement for other research methods, such as questionnaire surveys, interviews or observations.
Originality/value The publication can be useful for researchers from various fields and managers seeking to create and use simple, useful methods and tools for analysing unstructured text documents for decision-making. The paper proposes a separate text mining analysis of abstracts and whole documents using n-grams. This yielded a more precise list of areas relevant to the smart city. The grouping was done using correspondence analysis of the fifty most cited articles indexed in the Web of Science from 2014 to 2019.

Keywords: smart city, text mining, information and communication technology, scientific papers, research areas.

1. Introduction

The development of the Internet and global digitalisation change how business, industry, entertainment and cities work. This, in turn, affects global societal changes. The increasingly popular notion of the smart city has many definitions in literature (Albino et al., 2015). In simple terms, the smart city is defined as an urban area that uses various types of electronic sensors to collect data and then employs the resulting insight for efficient management of its assets, resources and services (Wikipedia, 2019). Deakin (2013) and A. I. Wear (Deakin and Wear, 2011) define the smart city as one that utilises ICT to meet the demands of its citizens. They believe the community involvement in the process is necessary and list factors that define the smart city.

Interest the smart city has grown dynamically since 2010 and remains constant today. This is apparent from plots on Google Trends (2019) and the numbers of publications in the Web of Science (Fig. 1).

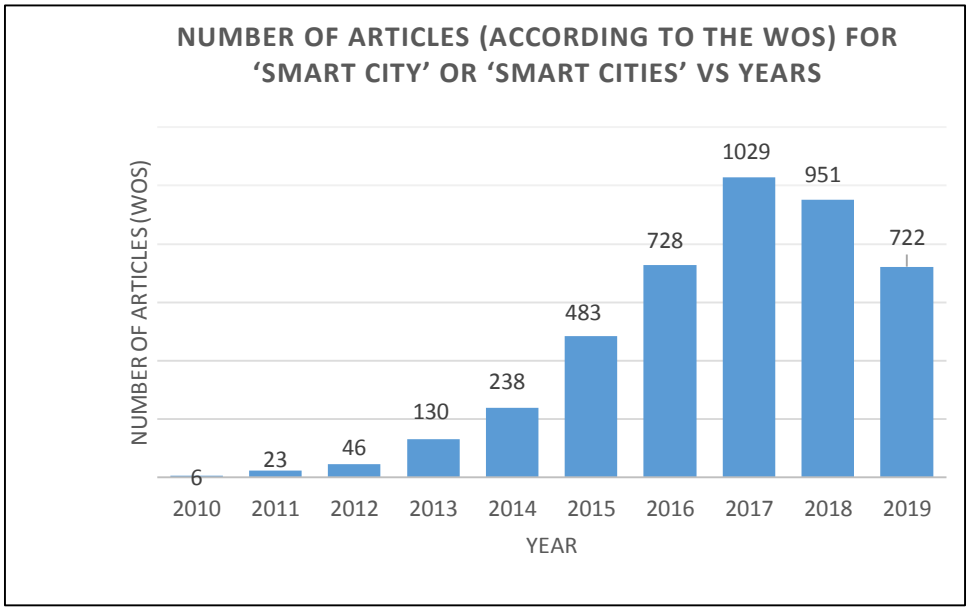

Figure 1. Number of articles (according to the WoS) for 'smart city' or 'smart cities' 2010–2019 (28.11.2019). Source: own work based on https://apps.webofknowledge.com.

Research topics concerning the smart city are growing broader. It is ever more difficult, and mostly even impossible, to keep track of and be familiar with all new publications in this domain, which can focus on technical, social and organisational aspects. Academic efforts yield publications that most often contain structured data, such as tables, and semi-structured data, such as text, pictures or diagrams. In the current era of global digitalisation and Internet prevalence, the amount of data and number of publications is prolific. Hence, there is more and more urgent need for tools and methods that could process and analyse such large amount of data automatically. This applies to structured data in databases and semi-structured textual data. This has put data exploration methods in the spotlight. One of these is text mining, which

searches for unusual patterns in large textual datasets to extract new information from textual resources.

The objective of the paper is to attempt to identify leading research domains concerning the smart city following a text mining analysis of research articles with a high citation index according to the Web of Science.

The author included 50 selected research articles dating from 2014 to 2019, published in two leading journals. Another goal is to find the faults and good points of the application of text mining for identifying research domains concerning the smart city.

2. Text mining for research article analysis

Text mining involves the automated extraction of useful information from textual data. Automation requires programming or appropriate software. According to Suhaib Peerzada, 'text mining is a tool which helps in getting the data cleaned up. Text mining techniques are basically cleaning up unstructured data to be available for text analytics'. He specified several steps to be followed during text mining. These are tokenisation, stemming and lemmatisation, stop words, number and punctuation removal, converting to lowercase, POS tagging, creating text corpus and term-document matrix (Peerzada, 2018). Each step modifies unstructured data to yield as much structured data as possible valuable to the user.

Other authors usually proposed four basic processes for text mining: tokenisation, removal of stop words, stemming and POS tagging (Fan et al., 2006; Vijayarani et al., 2015; Silge and Robinson, 2017). Text mining can be approached differently depending on user needs and the types of analysed documents. Research texts are structured in a particular way and consist of a title, author(s), abstract, keywords, introduction, literature review, method, results, discussion, conclusion, references and acknowledgements. The specific structure depends on the journal and author, as there is no one universal template. Still, some standards are generally accepted and pursued, such as EASE (EASE Guidelines, 2019) or IMRAD (Sollaci, 2004; Mogull, 2017).

Today, it is much easier to gain access to global research publications than a dozen years ago thanks to the Internet and the internationalisation of science. More and more publishing houses offer their resources online for free or for a small charge. This provides access to a considerable number of papers from various disciplines. The author decided to focus on research articles concerning the smart city that have a high citation index.

3. Materials, methods and tools

The paper analyses the most cited papers according to the Web of Science Core Collection from 2014 to 2019, available online. The layout of the articles was found to differ significantly depending on the publisher. The selection of journals was narrowed down to Sustainable Cities and Society and IEEE Access. They were one of the most cited journals on the list, and the author intended to focus on technologies and research problems related to the smart city.

The results were limited to the keywords 'smart city' and 'smart cities' in titles. The analysis involved 50 research articles on the smart city. The list of downloaded articles subjected to text mining can be found in Table 1. The documents were obtained as PDF files. The research process took place from November to December 2019.

Table 1.
Most cited research articles about 'smart city' and 'smart cities' according to the Web of Science published from 2014 to 2019

Id	Article	Times Cited by	
		WoS	Scholar
1	Kylili, A. & Fokaides, P. A. (2015). European smart cities: The role of zero energy buildings. Sustainable cities and society, 15, pp. 86-95.	94	183
2	Silva, B. N., Khan, M. & Han, K. (2018). Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities. Sustainable Cities and Society, 38, pp. 697-713.	60	162
3	Puiu, D., Barnaghi, P., Tönjes, R., Kümper, D., Ali, M. I., Mileo, A., ... & Gao, F. (2016). Citypulse: Large scale data analytics framework for smart cities. IEEE Access, 4, pp. 1086-1108.	59	117
4	Wu, J., Ota, K., Dong, M. & Li, C. (2016). A hierarchical security framework for defending against sophisticated attacks on wireless sensor networks in smart cities. IEEE Access, 4, pp. 416-424.	54	83
5	Mattoni, B., Gugliermetti, F. & Bisegna, F. (2015). A multilevel method to assess and design the renovation and integration of Smart Cities. Sustainable Cities and Society, 15, pp. 105-119.	54	117
6	Pouryazdan, M., Kantarci, B., Soyata, T. & Song, H. (2016). Anchor-assisted and vote-based trustworthiness assurance in smart city crowdsensing. IEEE Access, 4, pp. 529-541.	52	85
7	Mohamed, N., Al-Jaroodi, J., Jawhar, I., Lazarova-Molnar, S. & Mahmoud, S. (2017). SmartCityWare: A service-oriented middleware for cloud and fog enabled smart city services. IEEE Access, 5, pp. 17576-17588.	37	58
8	Liu, J., Xiong, K., Fan, P. & Zhong, Z. (2017). RF energy harvesting wireless powered sensor networks for smart cities. IEEE Access, 5, pp. 9348-9358.	29	37
9	Alvi, A. N., Bouk, S. H., Ahmed, S. H., Yaqub, M. A., Sarkar, M. & Song, H. (2016). BEST-MAC: Bitmap-assisted efficient and scalable TDMA-based WSN MAC protocol for smart cities. IEEE Access, 4, pp. 312-322.	27	46
10	Chen, L. J., Ho, Y. H., Lee, H. C., Wu, H. C., Liu, H. M., Hsieh, H. H., ... & Lung, S. C. C. (2017). An open framework for participatory PM2. 5 monitoring in smart cities. IEEE Access, 5, pp. 14441-14454.	25	41
11	Ullah, R., Faheem, Y. & Kim, B. S. (2017). Energy and congestion-aware routing metric for smart grid AMI networks in smart city. IEEE access, 5, pp. 13799-13810.	22	41
12	Teng, H., Liu, W., Wang, T., Liu, A., Liu, X. & Zhang, S. (2019). A cost-efficient greedy code dissemination scheme through vehicle to sensing devices (V2SD) communication in smart city. IEEE Access, 7, pp. 16675-16694.	21	26

Cont. table 1.

13	Sotres, P., Santana, J. R., Sánchez, L., Lanza, J. & Munoz, L. (2017). Practical lessons from the deployment and management of a smart city Internet-of-Things infrastructure: The smartsantander testbed case. IEEE Access, 5, pp. 14309-14322.	21	38
14	Muhammad, G., Alsulaiman, M., Amin, S. U., Ghoneim, A. & Alhamid, M. F. (2017). A facial-expression monitoring system for improved healthcare in smart cities. IEEE Access, 5, pp. 10871-10881.	20	37
15	Sajjad, M., Khan, S., Jan, Z., Muhammad, K., Moon, H., Kwak, J. T., ... & Mehmood, I. (2016). Leukocytes classification and segmentation in microscopic blood smear: a resource-aware healthcare service in smart cities. IEEE Access, 5, pp. 3475-3489.	19	36
16	Chen, B. W., Ji, W., Jiang, F. & Rho, S. (2015). QoE-enabled big video streaming for large-scale heterogeneous clients and networks in smart cities. IEEE Access, 4, pp. 97-107.	18	23
17	Muhammed, T., Mehmood, R., Albeshri, A. & Katib, I. (2018). UbeHealth: a personalised ubiquitous cloud and edge-enabled networked healthcare system for smart cities. IEEE Access, 6, pp. 32258-32285.	17	45
18	Samani, H. & Zhu, R. (2016). Robotic automated external defibrillator ambulance for emergency medical service in smart cities. IEEE Access, 4, pp. 268-283.	17	27
19	Islam, M. M., Razzaque, M. A., Hassan, M. M., Ismail, W. N. & Song, B. (2017). Mobile cloud-based big healthcare data processing in smart cities. IEEE Access, 5, pp. 11887-11899.	16	41
20	Brisimi, T. S., Cassandras, C. G., Osgood, C., Paschalidis, I. C. & Zhang, Y. (2016). Sensing and classifying roadway obstacles in smart cities: The street bump system. IEEE Access, 4, pp. 1301-1312.	16	26
21	Braun, T., Fung, B. C., Iqbal, F. & Shah, B. (2018). Security and privacy challenges in smart cities. Sustainable cities and society, 39, pp. 499-507.	14	39
22	Massana, J., Pous, C., Burgas, L., Melendez, J. & Colomer, J. (2017). Identifying services for short-term load forecasting using data driven models in a Smart City platform. Sustainable cities and society, 28, pp. 108-117.	14	27
23	Jia, G., Han, G., Jiang, J., Sun, N. & Wang, K. (2015). Dynamic resource partitioning for heterogeneous multi-core-based cloud computing in smart cities. IEEE Access, 4, pp. 108-118.	14	19
24	Raman, R., Sa, P. K., Majhi, B. & Bakshi, S. (2016). Direction estimation for pedestrian monitoring system in smart cities: An HMM based approach. IEEE Access, 4, pp. 5788-5808.	14	19
25	Ianuale, N., Schiavon, D. & Capobianco, E. (2015). Smart cities, big data, and communities: Reasoning from the viewpoint of attractors. IEEE Access, 4, pp. 41-47.	13	23
26	Bonafoni, S., Baldinelli, G. & Verducci, P. (2017). Sustainable strategies for smart cities: Analysis of the town development effect on surface urban heat island through remote sensing methodologies. Sustainable Cities and Society, 29, pp. 211-218.	12	23
27	Cui, L., Xie, G., Qu, Y., Gao, L. & Yang, Y. (2018). Security and privacy in smart cities: Challenges and opportunities. IEEE access, 6, pp. 46134-46145.	10	28
28	Ali, Z., Muhammad, G. & Alhamid, M. F. (2017). An automatic health monitoring system for patients suffering from voice complications in smart cities. IEEE Access, 5, pp. 3900-3908.	10	29
29	Alhussein, M. (2017). Monitoring Parkinson's disease in smart cities. IEEE Access, 5, pp. 19835-19841	9	24
30	Qiu, J., Chai, Y., Liu, Y., Gu, Z., Li, S. & Tian, Z. (2018). Automatic non-taxonomic relation extraction from big data in smart city. IEEE Access, 6, pp. 74854-74864.	8	17
31	Lella, J., Mandla, V. R. & Zhu, X. (2017). Solid waste collection/transport optimisation and vegetation land cover estimation using Geographic Information System (GIS): A case study of a proposed smart-city. Sustainable cities and society, 35, pp. 336-349.	8	16
32	Won, J., Seo, S. H. & Bertino, E. (2017). Certificateless cryptographic protocols for efficient drone-based smart city applications. IEEE Access, 5, pp. 3721-3749.	8	18

Cont. table 1.

33	De Filippi, F., Coscia, C., Boella, G., Antonini, A., Calafiore, A., Cantini, A., ... & Schifanella, C. (2016). MiraMap: A We-government tool for smart peripheries in Smart Cities. IEEE Access, 4, pp. 3824-3843.	8	19
34	Kulkarni, P. & Farnham, T. (2016). Smart city wireless connectivity considerations and cost analysis: Lessons learnt from smart water case studies. IEEE Access, 4, pp. 660-672.	8	17
35	Li, X., Lv, Z., Hijazi, I. H., Jiao, H., Li, L. & Li, K. (2016). Assessment of urban fabric for smart cities. IEEE Access, 4, pp. 373-382.	8	16
36	Yadav, G., Mangla, S. K., Luthra, S. & Rai, D. P. (2019). Developing a sustainable smart city framework for developing economies: An Indian context. Sustainable Cities and Society, 47, p. 101462.	7	14
37	Gohar, M., Muzammal, M. & Rahman, A. U. (2018). SMART TSS: Defining transportation system behavior using big data analytics in smart cities. Sustainable cities and society, 41, pp. 114-119.	6	18
38	Anisetti, M., Ardagna, C., Bellandi, V., Cremonini, M., Frati, F. & Damiani, E. (2018). Privacy-aware Big Data Analytics as a service for public health policies in smart cities. Sustainable cities and society, 39, pp. 68-77.	6	20
39	Usman, M., Asghar, M. R., Ansari, I. S., Granelli, F. & Qaraqe, K. A. (2018). Technologies and solutions for location-based services in smart cities: Past, present, and future. IEEE Access, 6, pp. 22240-22248.	6	10
40	Jain, B., Brar, G., Malhotra, J. & Rani, S. (2017). A novel approach for smart cities in convergence to wireless sensor networks. Sustainable cities and society, 35, pp. 440-448.	6	14
41	Kotevska, O., Kusne, A. G., Samarov, D. V., Lbath, A. & Battou, A. (2017). Dynamic network model for smart city data-loss resilience case study: City-to-city network for crime analytics. IEEE Access, 5, pp. 20524-20535.	5	6
42	Vázquez-Canteli, J. R., Ulyanin, S., Kämpf, J. & Nagy, Z. (2019). Fusing TensorFlow with building energy simulation for intelligent energy management in smart cities. Sustainable cities and society, 45, pp. 243-257.	5	19
43	Vo, N. S., Duong, T. Q., Guizani, M. & Kortun, A. (2018). 5G optimised caching and downlink resource sharing for smart cities. IEEE Access, 6, pp. 31457-31468.	5	24
44	Khan, M., Babar, M., Ahmed, S. H., Shah, S. C. & Han, K. (2017). Smart city designing and planning based on big data analytics. Sustainable cities and society, 35, pp. 271-279.	5	18
45	Deng, Y., Chen, Z., Yao, X., Hassan, S. & Wu, J. (2019). Task scheduling for smart city applications based on multi-server mobile edge computing. IEEE Access, 7, pp. 14410-14421.	4	8
46	Aborokbah, M. M., Al-Mutairi, S., Sangaiah, A. K. & Samuel, O. W. (2018). Adaptive context aware decision computing paradigm for intensive health care delivery in smart cities—A case analysis. Sustainable cities and society, 41, pp. 919-924.	5	15
47	Khan, Z. A. (2018). Using energy-efficient trust management to protect IoT networks for smart cities. Sustainable cities and society, 40, pp. 1-15.	4	9
48	Malik, K. R., Sam, Y., Hussain, M. & Abuarqoub, A. (2018). A methodology for real-time data sustainability in smart city: Towards inferencing and analytics for big-data. Sustainable cities and society, 39, pp. 548-556.	4	15
49	Wan, S., Lu, J., Fan, P. & Letaief, K. B. (2017). To smart city: Public safety network design for emergency. IEEE access, 6, pp. 1451-1460.	4	10
50	Wu, M., Xiong, N. N. & Tan, L. (2018). An intelligent adaptive algorithm for environment parameter estimation in smart cities. IEEE Access, 6, pp. 23325-23337.	4	7

Source: own work based on the Web of Science (https://apps.webofknowledge.com, 2019.11.28).

3.1. Methods

The research consisted of ten procedural stages (Fig. 2).

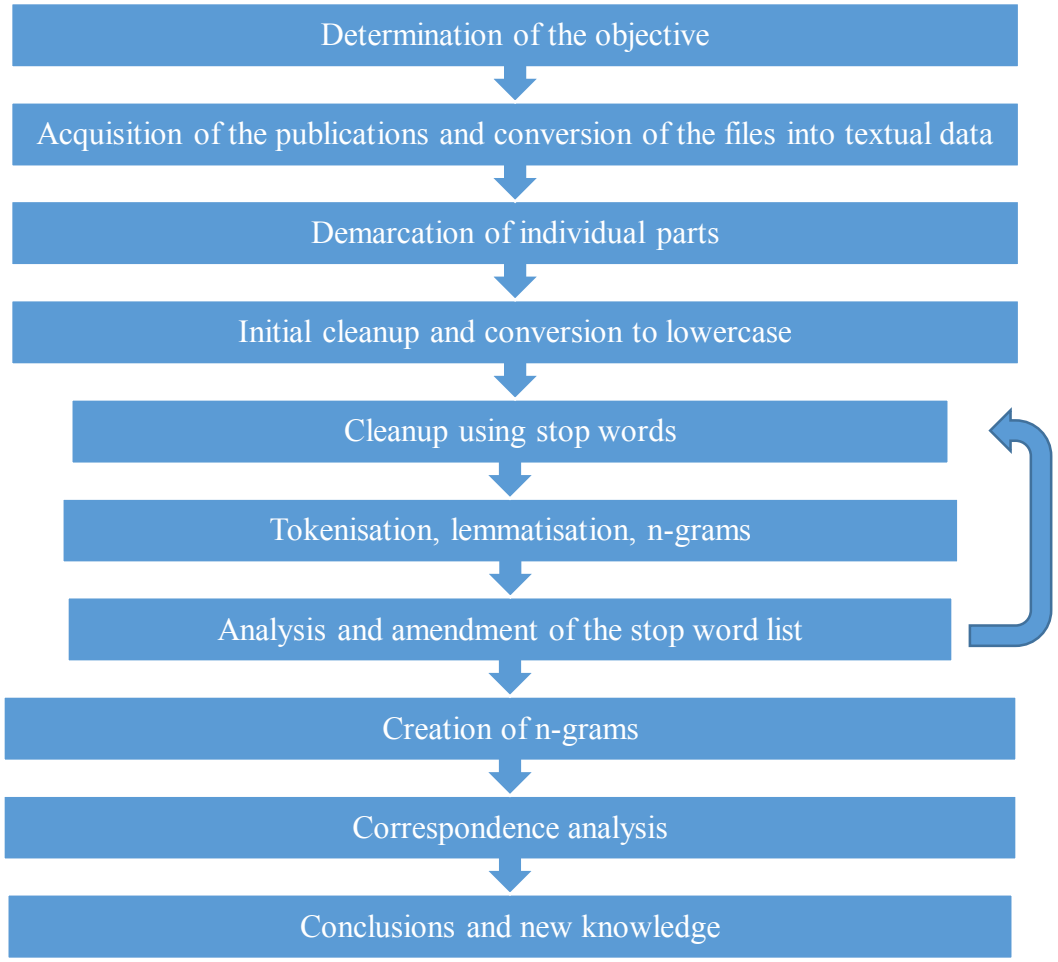

Figure 2. Stages of the research. Source: own work.

The first two stages involved the determination of the study objective and collection of articles to be analysed following conversion into textual data. The articles were found using a website at *https://apps.webofknowledge.com,* where articles indexed by the WoS can be searched. The PDF files were converted to .txt with an online tool available at *https://pdftotext.com/pl/*. The third stage saw all the data in text files loaded into an analytical tool and divided into characteristic parts. A script in R searched for the words 'introduction' and 'references' and divided each text into three separate parts. It the fourth stage, words and phrases characteristic of specific publishers were removed (IEEE Access, sustainable cities and similar), all characters were converted to lowercase. The fifth stage involved cleaning the data using a list of stop words. The sixth stage was tokenisation, lemmatisation and n-grams. Tokenisation meant parsing the data into individual words. Lemmatisation changed plurals into singulars and converted all words into their respective lemmas. Stop words were used to remove words that were insignificant for analytical purposes (such as: the, a, an, in). For each process, selected packages and functions of libraries tidytext, textstem and tm for R were used. Word frequency statistics were built at this stage using n-grams. Approx. 10% of the most common

words in each document were analysed. If necessary, new words or phrases that should be excluded from text mining were added to the stop word list. Following such amendments, the author returned to step five; hence the iterative loop in the method. This was repeated each time the stop words list was amended. The author considered using POS tags (Nakagawa et al., 2007), which flags words with the correct part of speech (such as noun, adjective or verb), but deemed it not necessary for this particular study. The eighth step was to build n-grams, followed by a correspondence analysis in the ninth step. The last stage was to interpret the results to create new knowledge about the investigated documents. The method was applied to whole texts excluding references (TXT), and separately to abstracts only (ABS).

3.2. Results

Using the assumptions and steps described in the previous section, the author investigated 50 documents presented in Table 2. After many iterations of data cleaning at stages four and five, the method yielded a list of the most common words related to the smart city. The study of whole articles (TXT) involved 12,500 words, 40% of which were used once in all the documents. The most common words were data, smart, city, system, network, base, energy, time, node, application, service, device, propose, model, sensor and so on (Fig. 3).

No.	Word	Freq.
1	data	3114
2	smart	2528
3	city	2398
4	system	1689
5	network	1477
6	base	1272
7	energy	1158
8	time	1145
9	node	977
10	application	957
11	service	952
12	device	933
13	propose	920
14	model	909
15	sensor	878

No.	Word	Freq.
16	information	821
17	user	766
18	provide	732
19	process	728
20	set	661
21	figure	648
22	numb	640
23	task	639
24	algorithm	636
25	result	632
26	communication	569
27	technology	551
28	performance	534
29	resource	523
30	cost	515

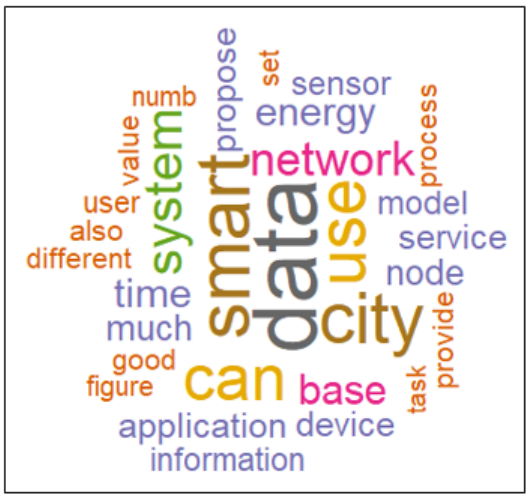

Figure 3. Single words, whole articles without references (TXT). Source: own work using R and RStudio.

For abstracts of the articles (ABS), the analysis yielded approx. 1,500 words, 47% of which were used once in all the documents. The most common words were smart, city, data, system, propose, base, network, energy, paper, service, technology and so on (Fig. 4).

No.	Word	Freq.
1	smart	155
2	city	154
3	data	100
4	system	71
5	propose	68
6	base	52
7	network	50
8	energy	46
9	paper	46
10	service	40
11	technology	37
12	information	36
13	provide	35
14	application	32
15	time	32

No.	Word	Freq.
16	framework	31
17	improve	31
18	resource	31
19	urban	31
20	sensor	30
21	communication	28
22	iot	27
23	performance	26
24	result	26
25	device	24
26	quality	23
27	algorithm	22
28	life	22
29	monitor	22
30	support	22

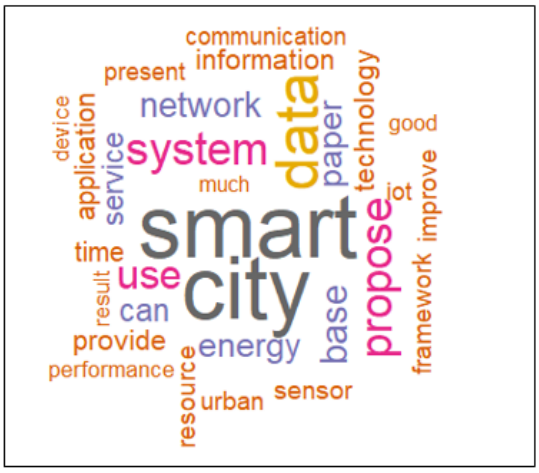

Figure 4. Single words, abstracts (ABS). Source: own work using R and RStudio.

A total of 21 (70%) items out of the 30 most frequent words are the same for TXT and ABS. 71%of the first 100 words were the same for TXT and ABS. The results were ranked, and a list of 30 keywords from the analysed articles related to the smart city was built.

No.	Word	TXT	ABS
1	smart	1	1
2	data	1	1
3	city	1	1
4	system	1	1
5	network	1	1
6	base	1	1
7	energy	1	1
8	propose	1	1
9	service	1	1
10	time	1	1
11	application	1	1
12	information	1	1
13	provide	1	1
14	sensor	1	1
15	device	1	1
		15	15

No.	Word	TXT	ABS
16	technology	1	1
17	node	1	0
18	model	1	0
19	communication	1	1
20	resource	1	1
21	result	1	1
22	algorithm	1	1
23	performance	1	1
24	process	1	0
25	iot	0	1
26	monitor	0	1
27	urban	0	1
28	analysis	0	0
29	method	0	0
30	framework	0	1
		9	10

Figure 5. Ranking of single words. Source: own work using R, Rstudio and MS Excel.

The same method was applied for 2-grams (bigrams). As the number of bigrams was large, only the 200 most common ones (in TXT and ABS) were considered. The grams were listed and ranked, and the most common phrases can be seen (Fig. 6, Fig. 7, Fig. 8).

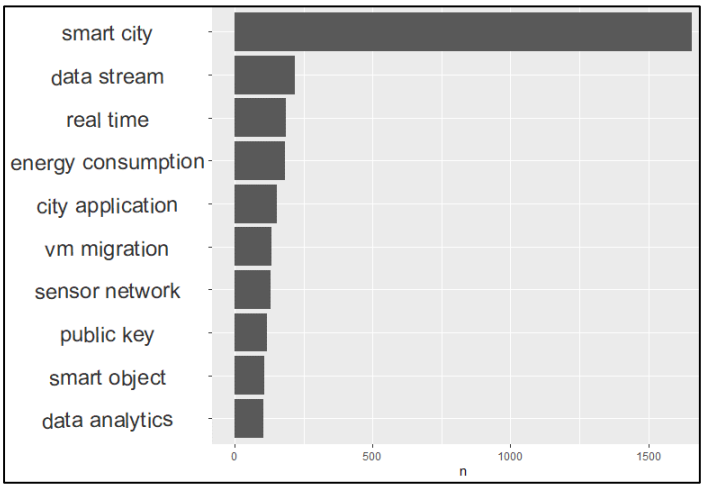

Figure 6. Bigrams (2-grams), whole articles without references (TXT). Source: own work.

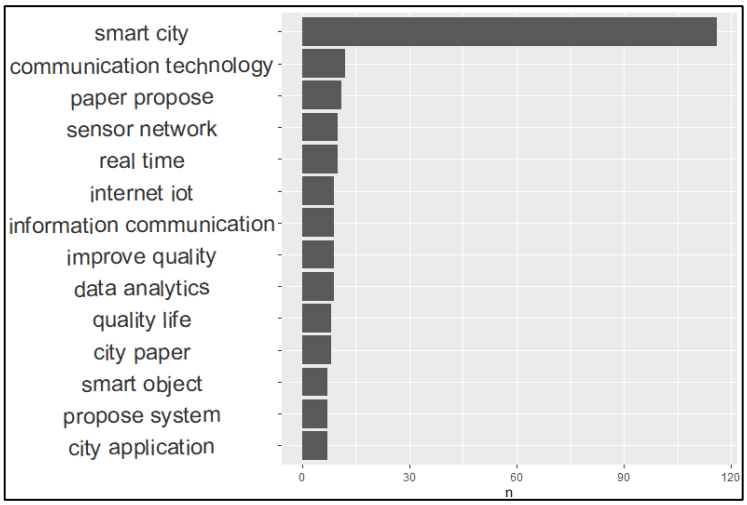

Figure 7. Bigrams (2-grams), abstracts (ABS). Source: own work.

No.	Word	TXT	ABS
1	smart city	1	1
2	real time	1	1
3	sensor network	1	1
4	data analytics	1	1
5	city application	1	1
6	energy consumption	1	1
7	smart object	1	1
8	cloud compute	1	1
9	propose system	1	1
10	communication technology	0	1
11	data stream	1	0
12	decision support	1	1
13	sink node	1	1
14	step size	1	1
15	data process	1	0
16	base station	1	1
17	direction estimation	1	1
18	system smart	0	1
19	wireless sensor	0	0
20	city smart	0	1
		16	17

No.	Word	TXT	ABS
21	admm algorithm	1	0
22	urban fabric	1	0
23	collect data	1	0
24	city service	0	1
25	low cost	0	0
26	fog compute	0	0
27	quality life	0	1
28	service smart	0	0
29	propose framework	0	0
30	multimedia content	0	0
31	healthcare system	1	0
32	mobile cloud	1	0
33	data set	0	0
34	propose method	1	0
35	completion time	0	0
36	machine learn	0	0
37	development smart	0	1
38	energy requirement	0	0
39	smart healthcare	0	1
40	internet iot	0	1
		6	5

Figure 8. Ranking of bigrams (2-grams). Source: own work using R, Rstudio and MS Excel.

Approx. 171,000 trigrams were identified in the TXT set, about 93% of which occurred only once. Approx. 5,614 trigrams were identified in the ABS set, about 98% of which occurred only once. The ranking includes only those phrases that occur more than once in the ABS and TXT sets. The trigrams were ranked by frequency. The author removed very common phrases found only in single articles, which resulted in an interesting list of phrases related to the smart city.

Table 2.
Phrases (3-grams)

No.	Words
1	smart city application
2	wireless sensor network
3	smart city service
4	information communication technology
5	base smart city
6	system smart city
7	energy requirement sink
8	development smart city
9	support vector machine
10	decision support system

No.	Words
11	system delivery capacity
12	smart city provide
13	deploy smart city
14	improve quality life
15	smart city emerge
16	smart healthcare system
17	real time data
18	automatic voice disorder
19	mobile cloud compute
20	intelligent transportation system

Source: own work.

The last analytical stage of the proposed method was to apply word statistics and analysis of correspondence to group documents and relevant technologies. It was a challenge to select keywords for correspondence tables. Different words (variables) yielded different plots. Moreover, the large number of documents and keywords hindered clarity and conclusions. After several dozen trials, the author selected seven keywords related to technology that were the most common in the articles for abstracts and full texts, singular words (Fig. 5) and trigrams (Table 2). The selected words were application, network, service, information, base, system and energy.

Table 3.
Correspondence table for whole articles, TXT from 1 to 20

	Article																			
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
system	37	56	21	13	34	18	47	40	1	44	9	13	11	93	18	33	78	46	61	30
network	8	64	7	32	67	10	20	25	28	3	77	26	23	8	8	44	283	9	7	3
energy	231	101	5	0	78	3	11	147	57	1	64	0	24	8	30	3	18	4	0	3
application	4	44	136	1	7	18	127	0	9	15	11	18	15	6	11	16	122	12	32	10
service	3	55	37	4	6	13	338	1	1	15	0	19	19	12	14	26	31	35	24	1
base	9	23	52	41	23	74	14	2	40	36	7	23	19	26	48	37	59	9	34	25
information	10	26	42	4	39	5	26	37	9	9	17	4	47	15	7	6	9	38	10	19

Source: own work.

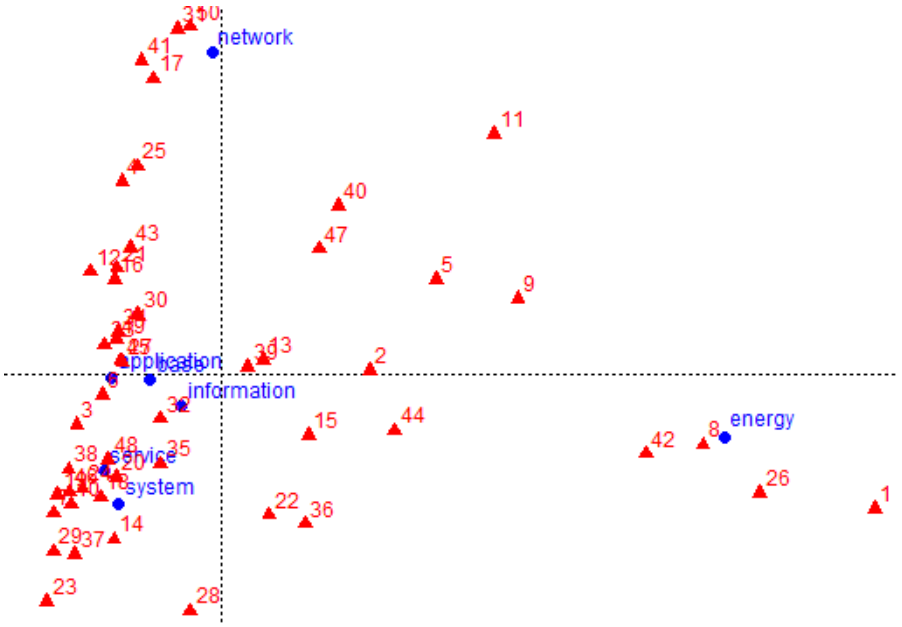

Figure 9. A correspondence analysis, complete texts (TXT). Source: own work.

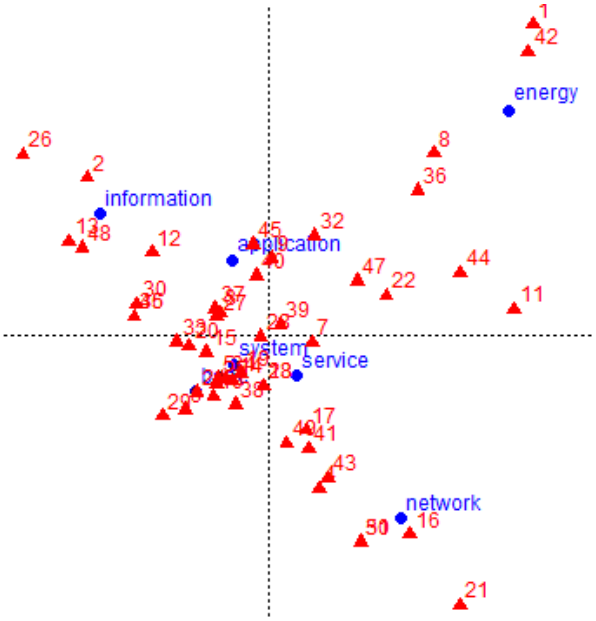

Figure 10. A correspondence analysis, abstracts (ABS). Source: own work.

The graphs show that the analysis of abstracts only partially corresponds to the analysis of whole documents. Some specific articles, however, can be clustered around certain issues and areas. Articles 1, 8 and 42, for example, focus on energy. Analysis of complete texts (TXT) shows that article 26 can join this group as well. Graphs for abstracts (ABS) do not correspond fully to the analysis of whole documents (TXT). Hence, analysis of whole documents together with abstracts yields better and more precise results. An excessive number of words or word variables distorts plot analysis.

In the central part of the plot for abstracts (ABS), there is ‘system’. For complete texts (TXT), these are ‘application’, ‘system’, ‘base’, ‘service’ and ‘information’. These notions can be associated with technology.

Articles can be grouped by areas of application of systems and technologies to improve the functioning of the city. Four examples of areas were identified: security, environment, energy and healthcare (Fig. 11).

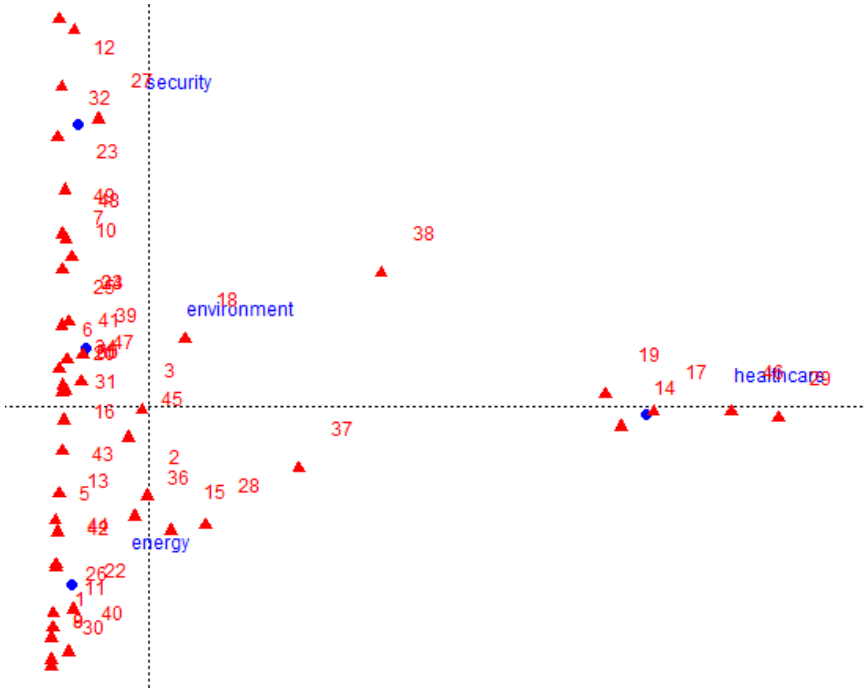

Figure 11. A correspondence analysis, complete texts (TXT) 2. Source: own work.

Further analyses of correspondence involved various combinations of keywords related to technologies or their respective areas of application. Preliminary results confirmed that it was reasonable to conduct correspondence analyses of abstracts and complete texts simultaneously. In this way, similar articles related to the smart city could be preliminarily grouped.

4. Discussion and conclusion

In the current era of global digitalisation and Internet prevalence, the amount of textual data is prolific. IDC predicts that the collective sum of the world's data will grow from 33 zettabytes in 2018 to 175 zettabytes by 2025, for a compound annual growth rate of 61 per cent. The 175 zettabytes figure represents a 9 per cent increase over last year's prediction of data growth by 2025 (Networkworld 2019, Reinsel et al. 2018). Domo's reports Data Never Sleeps 5.0 (2017) and Data Never Sleeps 6.0 (2018) demonstrate the colossal amount of data created in one minute of a year. According to the reports, users published on average 49 thousand photographs on Instagram and performed 3.8 million Google searches in a minute at the end of 2018. The same applies to the continually growing number of research publications.

There is, therefore, a reason to create and use simple methods of data exploration that can be used by managers and researchers in various fields to aid their decision-making. Among them are combinations of text mining and correspondence analysis.

When used for the first time, analysis of research articles on the smart city using text mining is very time-consuming. The primary reason is that scripts need to be written and then adjusted when analysing unstructured texts. Additionally, excessive variability of templates used in PDF research articles hinders analysis and significantly increases the time needed for data clean-up. A mere fifty articles from just two sources were enough to support conclusions regarding research domains appearing in the context of the smart city. They are primarily domains related to technologies and systems employed to improve the functioning of the city. The articles can be grouped by application areas for the systems and technologies. It is especially useful when searching for articles on a specific aspect of the smart city (such as energy). This is not an easy challenge, however. Articles have to be properly selected and thoroughly cleaned. Single-word analysis is not enough as well; two-word and three-word phrases should be considered for whole articles and abstracts. The selection of the right number of articles and words for correspondence analysis is just as important.

Text mining offers ample analytical possibilities, and it all depends on the needs and skills of the user. When applied to research texts, it can be used to pose new research questions as needed based on previous results. This is why it is so important to create simple, useful tools and methods for analysing unstructured text documents, which can be used not only by data science researchers, but also managers or other researchers to make decisions based on reliable, unstructured textual data.

The present paper proposes the foundations of a text mining method for analysing documents concerning the smart city. Anyone attempting to analyse research articles with the proposed method of text mining and procedural steps should account for its limitations. These are related mostly to diagrams, tables and figures in PDF files, which are omitted although they might contain valuable information. Text mining can be employed by various experts focusing on the smart city and constitute an interesting complement for other research methods, such as questionnaire surveys, interviews or observations.

Future work should include further refinement of the assumptions for the method, analyses of a more significant number of research texts and a narrowing down of the domain of the smart city. The method can also be used to monitor the dynamics of changes in the popularity of various new technologies and systems in different functional areas of the smart city over time.

References

1. Albino, V., Berardi, U., & Dangelico, R.M. (2015). Smart cities: Definitions, dimensions, performance, and initiatives. *Journal of urban technology, 22(1),* 3-21.
2. Internet source: Wikipedia – Smart city, Available online https://en.wikipedia.org/wiki/Smart_city, 29.10.2019.

3. Deakin, M. (Ed.) (2013). *Smart cities: governing, modelling and analysing the transition*. Routledge.
4. Deakin, M., & Al Waer, H. (2011). From intelligent to smart cities. *Intelligent Buildings International, 3(3),* 140-152.
5. Internet source: *Web of Science*, Available online https://apps.webofknowledge.com, 28.11.2019.
6. Internet source: *Google Trends*, Available online https://trends.google.com/trends/explore?date=2008-01-01%202020-01-02&q=%22smart%20city%22, 28.11.2019.
7. Internet source: *Networkworld article*, Available online https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html, 28.11.2019.
8. Reinsel, D., Gantz, J., & Rydning, J. (2018). *The digitisation of the world: from edge to core. Framingham: International Data Corporation*. Available online: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf, 11.12.2019.
9. Internet source: *Report: Data Never Sleeps*, Available online https://www.domo.com/learn/data-never-sleeps-5, https://www.domo.com/learn/data-never-sleeps-6, 28.11.2019.
10. Suhaib Peerzada. *What is Text Mining? – The Complete Beginner's Guide* (5.07.2018). Available online https://www.digitalvidya.com/blog/what-is-text-mining-guide, 14.11.2019.
11. Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM, 49(9),* 76-82.
12. Vijayarani S., Ilamathi J., Nithya, Phil M., (2015), Preprocessing Techniques for Text Mining – An Overview. *International Journal of Computer Science & Communication Networks, Vol 5(1)*, 7-16.
13. Silge, J., Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media, Inc. Available online https://www.tidytextmining.com/index.html, 10.11.2019.
14. Internet source: EASE *Guidelines for Authors and Translators of Scientific Articles to be Published in English*. Available online https://ease.org.uk/publications/author-guidelines-authors-and-translators, 11.12.2019.
15. Mogull, S.A. (2017). *Scientific and medical communication: a guide for effective practice*. Routledge.
16. Sollaci, L.B., & Pereira, M.G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the medical library association, 92(3),* 364.
17. Nakagawa, T., & Uchimoto, K. (2007, June). *A hybrid approach to word segmentation and pos tagging*. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 217-220.