

Research/Technical Note

Stratification the Text with Table of Contents

Youri Arzumanyan , **Mikhail Wolfson** , **Alexander Sotnikov** ,
Arian Zakharov* 

Department of Data Economics, The Bonch-Bruевич St Petersburg State University of Telecommunications (SPbSUT), St. Petersburg, Russia

Abstract

The formation of an ensemble of key concepts in relation to texts with highlighted semantic parts (text with a table of contents) is reduced to the stratification process from three procedures - the procedure of text preparation, the procedure of extracting key concepts from semantic parts and dividing the entire text into fragments related to the found key concepts. Quantitative characteristics of the ensemble (the number of words in related fragments) make it possible to solve a number of problems, including determining the predominant content of the text, calculating the parameters of text proximity, identifying concepts of interest to the reader, forming fragments for training neural networks while preserving the author's style, etc. The article briefly describes the formation procedures and provides three examples of using the ensembles of key concepts obtained by stratification. In the first example, the most fully (by the number of words in related fragments) disclosed key concepts in textbooks on the subject of "Project Management" in English, German, French and Russian are determined. The results obtained make it possible, for example, to justify the choice of a specific textbook. In the second example, for ten Russian-language educational and methodological publications on project management, proximity parameters were calculated, including the normalized length of the difference vector, the angle between the ensemble vectors, and the normalized integral characteristic. The results obtained can be used in selecting materials for educational programs and individual courses. In the textbooks participating in the first example, the longest continuous fragments of texts by the number of words, suitable for LLM training, were found.

Keywords

TOC Text, Statistical Text Processing, Ensemble of Key Concepts

1. Introduction

In recent years, the demand for objective, quantitative analysis of textual materials has become increasingly critical in a variety of contexts, ranging from the formation of individualized educational trajectories to the comparative evaluation of textbooks, scientific reports, and technical documentation. These tasks require robust methodologies that can

effectively characterize the informational content of documents, allowing researchers and practitioners to assess, compare, and utilize such materials with greater precision. A significant challenge in this domain is the development of computational techniques capable of extracting meaningful quantitative characteristics from structured texts, which are

*Corresponding author: za54ar@gmail.com (Arian Zakharov)

Received: 13 January 2025; **Accepted:** 17 March 2025; **Published:** 14 April 2025



Copyright: © The Author(s), 2025. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

often distinguished by the presence of a table of contents (TOC) [1] and a high degree of internal organization.

TOC-based documents—including academic textbooks, scientific publications, technical reports, and project proposals—share a common structural feature: they are divided into semantic units, each representing a cohesive and logically distinct component of the document. These semantic parts are typically associated with specific key concepts, which are not only central to the content of the corresponding section but also serve as foundational elements for the material that follows. This structural regularity provides a unique opportunity to analyze the document through the lens of its key concepts, enabling the formation of a conceptual ensemble that can be quantified and systematically evaluated [2-8].

Prior research has established the utility of key concept ensembles in characterizing the content of structured texts. By identifying and analyzing such ensembles, researchers can capture essential thematic elements, evaluate the prominence of specific topics, and assess the textual proximity between different documents [6, 8]. While several methodologies for text analysis exist, many are not optimized for TOC-structured documents, where the semantic segmentation offers distinct advantages that can be leveraged for more accurate and meaningful analysis.

The purpose of this research is to develop and validate a stratification-based approach tailored specifically for TOC-structured texts. This approach consists of three sequential procedures: (1) text preparation, including segmentation and normalization; (2) extraction of key concepts based on frequency and contextual relevance within semantic parts; and (3) division of the entire document into fragments associated with the identified key concepts. The result is a quantifiable ensemble that reflects the internal structure and thematic focus of the document. Importantly, this stratified representation allows for a variety of analytical tasks, such as determining the predominant content areas, computing text similarity metrics, and identifying relevant fragments for further processing, including machine learning applications.

The significance of this research lies in its ability to bridge the gap between qualitative content interpretation and quantitative text analysis. By providing a rigorous, systematic framework for analyzing structured texts, the proposed methodology supports diverse applications. These include the selection of educational resources based on content coverage, the assessment of text similarity for academic or technical comparison, and the preparation of training data for large language models while preserving authorial intent and style. Moreover, the ability to isolate and evaluate continuous text fragments associated with specific key concepts enhances the granularity and applicability of the analysis.

In summary, this study introduces a novel method of text stratification that exploits the inherent structure of TOC-based documents to produce meaningful quantitative insights. Through the development and application of this method, the research contributes a valuable tool to the fields of educa-

tional technology, information retrieval, and computational linguistics, enabling more informed decision-making and deeper understanding of textual materials.

2. Stratification Algorithm

The considered stratification algorithm consists of three stages. The first stage uses a standard text preparation procedure, which includes preserving characters acceptable for the text language, selecting stop words, and stemming/lemnization. In addition to these operations, the text is preliminarily divided into individual sentences and semantic parts, as indicated in the table of contents.

At the second stage, in each semantic part, candidates for key concepts of one, two, three, etc. words are determined by the maximum frequency of occurrence. A mandatory condition for candidates is the absence of stop words at the beginning and end. Upon completion of the search in all semantic parts, the found candidates that are entirely contained in others are removed from the set of candidates. This operation allows us to make overly general concepts more specific. For example, from the pair of concepts “process” and “calculation process” only “calculation process” is retained as a key concept as a more specific one.

At the third stage, the procedure of associating each sentence with one of the candidates for key concepts is performed. The procedure is implemented in two cycles. At the first cycle, sentences that include candidates for key concepts are noted throughout the entire sequence of sentences in the text. If several candidates are involved in a sentence, the candidate with the highest frequency of occurrence in the entire text is selected. If a sentence does not contain candidates, it is associated with the candidate of the previous sentence. At the second cycle, all alternations of associations through one sentence are eliminated. Such alternations indicate that the key concept is only mentioned and is not developed in the immediate future. The sentence with the mentioned key concept takes the association of the previous sentence.

At the end of the third stage, all sentences of the text can be divided and regrouped according to their association with one of the key concept, which allows solving various problems of analysis and synthesis.

3. Examples of Use

To test the functionality of the considered algorithm, we wrote programs in Python and solved the following three problems.

3.1. Determination of the Focus of the Text

Determination of the focus of the text content using the example of textbooks on the subject “Project Management” in English, German, French and Russian [9-12].

The focus of content is understood as the search for the best

fully disclosed key concepts, i.e. concepts that include the maximum number of words in associated sentences. The search results are presented in Table 1.

Table 1. Dominant key concepts.

Text	Dominant key concepts	(23-25) %
[9] English 100% \approx 72,1 thousand words	project schedule	6%
	project team members	4%
	managing a project	4%
	project cost	4%
	project completion	3%
	needed for a project	3%
	kosten	9%
[10] German 100% \approx 80,7 thousand words	et cetera	5%
	soziale kompetenz	4%
	dass projektmanagement	3%
	werden müssen	3%
	livrables du projet	7%
[11] French 100% \approx 29,7 thousand words	méthodologie PM	6%
	doit être	5%
	peuvent être	5%

Text	Dominant key concepts	(23-25) %
[12] Russian 100% \approx 45,5 thousand words	принятие решений	6%
	критический путь	5%
	сетевая модель	3%
	пакеты работ	3%
	цели должны	3%
	текущая стоимость	3%

The obtained results allow, for example, to determine the textbook that better covers the issues that interest the student. In addition, the objectivity of the computational nature of the search for key concepts in case of the appearance of such concepts as “et cetera” in [10] among the dominant concepts allows to assess the author’s style.

3.2. Characteristics of Closeness

Calculation of quantitative characteristics of the similarity of texts using the example of textbooks [12-21] on the discipline indicated in the previous example.

Quantitative estimates of resemblance were calculated using the method described in [6, 8]. The results of the comparison of texts are presented in Table 2.

Table 2. Values of extreme characteristics of resemblance.

Characteristics of closeness	Minimal differences		Maximum differences	
	Meaning	Texts	Meaning	Texts
Normalized length of the difference vector (%)	40.6	[13] \leftrightarrow [17]	72.7	[17] \leftrightarrow [20]
Angle between ensemble vectors in degrees	29	[20] \leftrightarrow [21]	71	[15] \leftrightarrow [19]
Normalized integral characteristic (%)	54.6	[18] \leftrightarrow [21]	79.3	[17] \leftrightarrow [20]

The obtained results can be used for selection of teaching and methodological materials for educational programs and individual courses.

3.3. Formation of Training Models for Tuning LLM Neural Networks

For these purposes, relevant fragments extracted from the approved texts can be used. It is advised to take as relevant fragments the longest (in number of words) continuous fragments of text associated with the key concept of interest.

Table 3 presents examples of key concepts associated with the longest continuous fragments of texts [9-12].

Table 3. Key concepts with the longest continuous fragments.

Text	Key concept	No. of words
[9]	requirements describe the characteristics of the final deliverable	1656
	skills that the project management	1525

Text	Key concept	No. of words
[10]	module planning and project management I bis II	1472
	soziale kompetenz	687
[11]	compétences en gestion de projet	662
	méthodologie PM	569
[12]	принятие решений	1294
	критический путь	1165

4. Conclusions

Figuratively speaking, the presented stratification procedure implies dissolving the entire "fabric" of the text into separate "threads", which are then grouped according to key concepts. The fact that the "threads" contain only whole sentences allows for unambiguous comparison of the "threads" with sections of the original text.

Abbreviations

LLM	Large Language Model
TOC	Table of Contents

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] The Chicago Manual of Style (17th ed.). University of Chicago Press. 2017. ISBN 9780226287058. LCCN 2017020712. OCLC 1055308068.
- [2] Youri Arzumanyan, Mikhail Wolfson, Alexander Sotnikov, Arian Zakharov, Using quantitative methods to analyze the educational program, 9TH INTERNATIONAL CONFERENCE ON ADVANCED INFOTELECOMMUNICATIONS ICAIT, 2020, Conference Proceedings, vol. 2, pp. 601-605, <http://www.sut.ru/doci/nauka/1AEA/APINO/9-APINO-2020,%20%D0%A2.2.pdf>
- [3] Youri Arzumanyan, Arian Zakharov, Yana Sokolova, Comparative analysis of information characteristics of academic disciplines, 9TH INTERNATIONAL CONFERENCE ON ADVANCED INFOTELECOMMUNICATIONS ICAIT, 2020, Conference Proceedings, vol. 2, pp. 606-609 <http://www.sut.ru/doci/nauka/1AEA/APINO/9-APINO-2020,%20%D0%A2.2.pdf>
- [4] Youri Arzumanyan, Mikhail Wolfson, Alexander Sotnikov, Galia Katasonova, Arian Zakharov, Features of modeling educational programs in the development of educational trajectories for training IT specialists, XIX conference "Teaching information technologies in the Russian Federation, Moscow, 19-20 May 2021, Conference Proceedings, pp. 294-295, <https://it-education.ru/conf2021/thesis/4648/>
- [5] Youri Arzumanyan, Mikhail Wolfson, Galia Katasonova, Alexander Sotnikov, Arian Zakharov, Models of educational programs for optimization problems in the design of individual educational trajectories, X INTERNATIONAL CONFERENCE ON ADVANCED INFOTELECOMMUNICATIONS ICAIT, 2021, Conference Proceedings, vol. 3, pp. 330-335, <https://www.sut.ru/doci/nauka/1AEA/APINO/10-APINO-2021.%20T.3.pdf>
- [6] Youri Arzumanyan, Mikhail Wolfson, Galia Katasonova, Arian Zakharov, Alexander Sotnikov, Vector representation of educational programs XI INTERNATIONAL CONFERENCE ON ADVANCED INFOTELECOMMUNICATIONS ICAIT, 2022, Conference Proceedings, vol. 3, pp. 557-561, <https://www.sut.ru/doci/nauka/1AEA/APINO/11-APINO-2022.%20%D0%A2.3.pdf> (accessed 19 January 2025)
- [7] Youri Arzumanyan, Mikhail Wolfson, Arian Zakharov, Alexander Sotnikov, Comparative Analysis of SPbSUT Educational Programs in 2022, XII INTERNATIONAL CONFERENCE ON ADVANCED INFOTELECOMMUNICATIONS ICAIT, 2023, Conference Proceedings, vol. 4, pp. 15-21, <https://apino.sut.ru/@/file/Ua83P4CoSaFyGo14>
- [8] Youri Arzumanyan, Mikhail Wolfson, Arian Zakharov, Alexander Sotnikov, Methods of analysis and design of educational programs using the tools of the Ensemble of Key Concepts, INFORMATION PROCESSES: CONCEPTUAL BASIS OF DIGITAL TRANSFORMATION OF THE ECONOMY, St.-Petersburg, 2024. pp. 44-62.
- [9] Watts, A. Project Management - 2nd Edition. Victoria, B.C.: BCcampus. (Websites) Available from: <https://opentextbc.ca/projectmanagement/> (accessed 19 January 2025)
- [10] Kluge, F. Projektmanagement in Praxis und Lehre der (Landschafts) Architektur ... ein wenig Chaos gehört dazu. [Project management in practice and teaching of (landscape) architecture ... a little chaos is part of it] Münster (Westfalen) 2008, P. 286 (Book) Available from: https://publications.rwth-aachen.de/record/51297/files/Kluge_Florian.pdf (accessed 19 January 2025)
- [11] Le Guide de la Méthodologie de Gestion de Projet PM² 3.0.1. [The PM² Project Management Methodology Guide.] Commission européenne Centre d'Excellence en Gestion de Projets (CoEPM) Bruxelles, Luxembourg. Mars 2021 P. 148 (Book) Available from: https://www.pm2alliance.eu/wp-content/uploads/2023/10/Metho-dologie-de-gestion-de-projet-pm%C2%B2-NO0921037FRN_c.pdf (accessed 19 January 2025)
- [12] Abramov N. V., Motovilov N. V., Naumov N. D. Upravlenie proektami [Project Management]: Textbook – Nizhnevartovsk, 2008. — 197 p. (Book) Available from: <https://files.student-it.ru/download/275982> (accessed 19 January 2025)

- [13] Aleshin A. V., An'shin V. M., Bagrationi K. A. et al. Upravlenie proektami: fundamental'nyj kurs [Project Management: Fundamental Course]: Textbook, ed. by V. M. Anshin, O. N. Ilyina, National Research University "Higher School of Economics" - Moscow: Publishing House of the Higher School of Economics, 2013. 620 p. (Book) Available from: <https://publications.hse.ru/mirror/pubs/share/folder/nvs1ctzplo/direct/148559151.pdf> (accessed 19 January 2025)
- [14] Boronina L. N., Senuk Z. V. Osnovy upravleniya proektami [Fundamentals of Project Management]: Textbook, Ministry of Education and Science of the Russian Federation Ural Federal University. - Yekaterinburg: Ural Federal University Press. 2015. — 112 p. (Book) Available from: <https://elar.urfu.ru/bitstream/10995/30881/1/978-5-7996-1416-4.pdf> (accessed 19 January 2025)
- [15] Denisenko V. I. Upravlenie proektami [Project Management]: Textbook, ed. by Dr. of Technical Sciences, Prof. V. I. Denisenko, Dr. of Economics, Prof. N. M. Filimonova, A. G. and N. G. Stoletov Vladimir State University - Vladimir: VISU Publishing House, 2015. — 108 p. (Book) Available from: <https://dspace.vvsu.ru/bitstream/123456789/4337/1/01451.pdf> (accessed 19 January 2025)
- [16] Ivasenko A. G., Nikonova YA. I., Sizova A. O. Upravlenie proektami [Project Management]: Textbook – Novosibirsk: SSGA, 2007. – 202 p. (Book)
- [17] Mazur I. I., Shapiro V. D., Ol'derogge N. G., Polkovnikov A. V. Upravlenie proektami [Project Management]: textbook for students studying on specialty "Management of organization" ed. by I. I. Mazur and V. D. Shapiro 6th ed. - Moscow: Omega-L Publishing House, 2010. — 960 p. (Book) Available from: <https://topuch.com/download/i-i-mazur-v-d-shapiron-g-ol'derogge-a-v-polkovnikovupravleniep.pdf> (accessed 19 January 2025)
- [18] Osipov D. V. Upravlenie proektami [Project Management]: Textbook for Masters in Management - Moscow: RUT (MIIT), 2017.– 170 c. (Book)
- [19] Strelina E. N. Upravlenie proektami [Project Management]: Textbook for the enlarged group of training directions and specialties 38.00.00 Economics and management – Donetsk: DONNU, 2022. – 310 p. (Book) Available from: <http://repo.donnu.ru:8080/jspui/bitstream/123456789/4962/1/4371.pdf> (accessed 19 January 2025)
- [20] Testina YA. S., CHumakov V. N. Upravlenie proektami [Project Management]: Textbook for Universities – Gatchina: GIEFPT Publishing House, 2023. – 69 p. (Book) Available from: https://sovman.ru/wp-content/uploads/2023/09/ss125_compressed.pdf (accessed 19 January 2025)
- [21] Cycarova N. M. Upravlenie proektami [Project Management]: Textbook - Ulyanovsk State Technical University. - Ulyanovsk: UIGTU, 2021. – 105 p. (Book) Available from: <https://lib.ulstu.ru/venec/disk/2021/21.pdf> (accessed 19 January 2025)

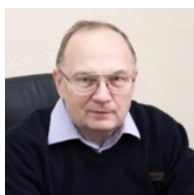
Biography



Youri Arzumanyan graduated from Prof. M. A. Bonch-Bruевич Leningrad Electrotechnical Institute of Communications. He completed his PhD in Radio Engineering in 1979 from the same institution. From 1983 to 2024 he was an associate professor at The Bonch-Bruевич St. Petersburg State University of Telecommunications. From 1993 to 2017 he was Dean of the Faculty of Economics and Management and from 2000 to 2019 he was Head of the Department of Business Informatics. As of January 2025 - Consultant of the Department for new directions of Smart World. Author of 19 scientific articles and 12 inventions. Abstracts of five articles were translated and published in the USA. Author of the textbook 'Fundamentals of Digital Transformation'. Main interests today are the use of entropy approach to solve the problems of sustainability and development of Smart World; author of three articles on this topic. Acted as a leader and executor of research and development projects in the field of information technology and telecommunications.



Mikhail Volfson is an associate professor at The Bonch-Bruевич St. Petersburg State University of Telecommunications, Department of Data Economics. He completed his PhD in E-Business in 2004 from the same institution. In 2007, the scientific title of associate professor was awarded. As of 2009, he is the head of the Business Informatics education program at the university. He is a member of the editorial board of the journal Sociohumanitarian Communications. He is co-author of the monographs 'Information Society. Infocommunications and Business', 'Models and Architectures of Electronic Enterprise'. He is the author of more than 50 scientific articles on e-business and digital transformation. In 2021 he was awarded the honorary title "Honorary Worker of Education of the Russian Federation".



Alexander Sotnikov is a professor at St.-Petersburg State University of telecommunications, Department of Business informatics. He completed his Doctor of Science degree in System Analysis in 2007, and his PhD degree in Telecommunication Networks and Systems in 1983 from the same institution. Recognized for his exceptional contributions, member of International Telecommunication Academy Dr. Sotnikov has been honored with "Master of Communications" Award of the Ministry of Communications of the Russian Federation and Award "Honorary Worker of Education of the Russian Federation" of Higher Education Ministry. He has participated in multiple international research collaboration projects in recent years. Being certified as Siemens OEM, SAP and European Labour Organization specialist, he currently serves on the Editorial Boards of number publications and has been invited as a Keynote Speaker, Technical Committee Member, Session Chair at international conferences.



Arian Zakharov graduated from Prof. M. A. Bonch-Bruевич Leningrad Electrotechnical Institute of Communications. He completed his PhD in Radio Engineering in 1982 from the same institution. From 1987 to 2023 he was an associate professor at The Bonch-Bruевич St. Petersburg State University of Telecommunications. From 1996 to 2013 - general director of the Non-State Private Educational Institution "Educational and Scientific Center for Management, Informatics and Communications "ELITA". Author of more than 40 scientific papers and 7 author's certificates in the field of statistical radio engineering, information technology and programming. He is co-author of the monographs 'Information Society. Infocommunications and Business'. Acted as a leader and executor of re-

search and development projects in the field of telecommunications, information technology and programming. Since 2023 - Consultant of the Department of Data Economics

Research Field

Youri Arzumanyan: Digital Transformation, Smart World, Sustainable Development, Information Theory, Sociophysics.

Mikhail Volfson: E-business organization, Business informatics, Data mining, E-enterprise design, Digital business transformation, Digital marketing.

Aleksandr Sotnikov: Info-communications systems analysis and design, Information processes investigation, Telemedicine networks, systems and services, Information systems and processes simulation.

Arian Zakharov: Computer modeling, Text processing, Programming, Sociophysics.