Л.Н. Гумилев атындағы Еуразия ұлттық университетінің ХАБАРШЫСЫ.
ISSN: 2616-7263. eISSN: 2663-1261
ТЕХНИКАЛЫҚ ҒЫЛЫМДАР ЖӘНЕ ТЕХНОЛОГИЯЛАР СЕРИЯСЫ /
TECHNICAL SCIENCES AND TECHNOLOGY SERIES/
СЕРИЯ ТЕХНИЧЕСКИЕ НАУКИ И ТЕХНОЛОГИИ

IRSTI 20.15.05
Article

https://doi.org/10.32523/2616-7263-2025-150-1-257-267

Automated extraction and structuring of menus from PDF files using machine learning and NLP techniques

A.S. Mashkanov* , **Zh.Akhayeva** , **A.Zakirova**

L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

E mail: 020715550703@enu.kz, ahaeva07@mail.ru, alma_zakirova@mail.ru

Abstract. This study explores state-of-the-art approaches for processing PDF documents, with a focus on analyzing poorly structured restaurant menus. The focus will be on analyzing poorly structured restaurant menus. Successful automated processing typically requires well-structured documents, meaning that aesthetic design must often be sacrificed for machine readability. However, in case of restaurants, the design of the menu is more valuable than its structure, that is why the menus are harder to process, due to its poor structure. With the ability to successfully process the poorly structured PDF documents, further processing of the documents from other spheres of businesses should become much easier. A comparative analysis is conducted of structural features in different types of PDF documents, including legislative acts and academic publications.

The research is aimed to use machine learning methods in order to overcome the challenges in automation of data extraction, analysis and structuring. Solution that has been described in the study is developed to overcome the problems with poorly structured PDF documents.

Keywords: PDF document processing, text analysis automation, weakly structured data, restaurant menus, Natural Language Processing (NLP), machine learning, data extraction, semantic analysis, food service digitalization

Received 22.07.2025. Revised 15.12.2025.Accepted 17.12.2025.Available online 30.12.2025

* the corresponding author

Introduction

The significance of the PDF documents processing automation is explained by the fast development of digitalization in all spheres of life. It should be emphasized that the PDF document format possesses a specific structural organization whose characteristics are determined by its functional purpose. In particular, legislative acts have a clear hierarchical structure, with step-by-step division into chapters, articles and clauses. Academic publications usually follow strictly regulated structure, which includes required sections such as: abstract, introduction, methodology, research results and conclusion. Correct recognition and understanding of such structure is the main base for developing effective automated document analysis algorithms.

Processing PDF versions of restaurant menus is a difficult task, because most of them are poorly structured and have non-standard data layout. This difference in formatting makes it hard to analyze them using machine methods and to use the data in digital systems. In this study, a full method is suggested. It combines modern NLP tools with machine learning algorithms. The solution includes a step-by-step process for automatic extraction, structuring, and semantic analysis. This helps to solve the main problems of working with poorly structured PDF files.

The core challenge addressed in this research can be formulated as follows: given a PDF document containing a restaurant menu with non-standard layout and mixed content types, the task is to automatically extract and structure the content into machine-readable format while preserving semantic relationships between menu elements (categories, dish names, descriptions, and prices).

Literature review. The research on automating the processing of restaurant menus in PDF format showed that there is a complex set of technical problems, and many of them are closely connected. To turn unstructured PDF files into machine-readable format, it is important to understand modern approaches to document processing.

The main challenge is to correctly separate the main content of the document (Body Text) from the extra parts (Non-Body Text). These extra parts include not only visible graphic elements, but also different structural components like headings, footnotes, sidebars, tables, metadata, and other specific formats of information. Most popular tools like pdftotext and PDFBox can’t properly tell the difference between these parts, which makes them not very effective for this kind of task. [1].

Alternative solutions, such as PDF extractors that use DBSCAN algorithm, show more advanced approaches to this problem. They allow to separate text blocks more effectively, taking into account their semantic and structural connections. However, research shows that even such advanced methods need careful tuning of parameters and can still have problems when working with documents that have complicated layout. The problems of understanding PDF structure continue to push development in document processing technologies, especially in specific cases like menu digitization, where the complexity of layout and variety of content make the task more difficult.

In order to properly extract and organize document, there is a sequence of methods. First of all, for breaking text into text blocks, MuPDF is used. MuPDF works effectively with clinical documents, which have editable text. This approach enables direct extraction while preserving structural relationships between content elements.

The system uses a DBSCAN-based clustering algorithm to classify templates by analyzing spatial distribution patterns of text blocks. The system groups content elements that fall within a specified boundary radius into clusters which serve as separate directories for storing distinct document templates. The algorithm shows exceptional performance in processing

documents with changing page layouts through its ability to analyze first and subsequent pages separately [2]. The dual-path architecture enables precise detection of meaningful textual areas while performing systematic noise elimination.

The Mask-RCNN architecture achieves advanced segmentation through training on PubLayNet data using Facebook's Detection framework. The system analyzes 1025×1025 pixel document pages through a multi-stage workflow:

1. Raster image conversion
2. Semantic region prediction
3. Mask generation for five content classes (subtitles, text, images, tables, lists)
4. Confidence-based classification (70-95% threshold)

Text cells receive appropriate labels when the predicted mask covers ≥80% of their area, ensuring precise content categorization. This approach has demonstrated superior performance in handling complex document layouts with mixed content types [5]. The integration of spatial clustering with deep learning-based segmentation represents a significant advancement in document analysis technology, particularly for menu extraction applications requiring high layout awareness.

Optical character recognition represents a sophisticated technological process for converting graphical text representations into machine-readable format. Modern OCR systems demonstrate the capability to process diverse data sources, including scanned paper documents, PDF files, and digital images, converting them into editable and searchable content.

Contemporary recognition systems like ABBYY FineReader are built upon three fundamental principles of cognitive perception:

- The principle of integrity (document analysis as a unified structure)
- The principle of purposefulness (context-dependent recognition)
- The principle of adaptability (learning capability)

These principles enable the system to simulate human text perception cognitive processes. The FineReader processing workflow implements a sequence of interconnected operations: from initial document layout analysis and structural segmentation to complex character recognition with multi-hypothesis analysis and verification [3].

The alternative Tesseract platform employs an innovative layout analysis approach based on detecting document structural elements. The algorithm's key features include:

1. Morphological filtering using the Leptonica library
2. Connected component identification
3. Intelligent text element grouping
4. Heuristic analysis of spatial relationships

However, research [4] has revealed several technological limitations of the system. The most significant challenges occur when processing:

- Graphic-intensive materials (promotional brochures, flyers)
- Documents with complex background patterns
- Handwritten text forms
- Non-standard typographical solutions

Meanwhile, for standard digital documents with clear typography, the system maintains consistently high accuracy rates. This selective effectiveness highlights the need for further algorithm improvements, particularly in processing complex-structured documents and

handwritten content. The current technological landscape underscores the importance of developing more robust OCR solutions capable of handling the full spectrum of document types while maintaining high recognition accuracy across diverse formats [5].

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm [14] has proven particularly suitable for document layout analysis due to its ability to identify clusters of arbitrary shapes without requiring a predefined number of clusters. Unlike k-means clustering, DBSCAN can effectively handle noise points and outliers, which is essential when processing menu documents where decorative elements and non-standard text blocks may appear. The algorithm's parameter flexibility allows adaptation to various document types by adjusting the epsilon (neighborhood radius) and minimum points thresholds.

The methodology

The study employed a comprehensive methodology for processing PDF documents, combining advanced text extraction techniques with machine learning approaches. The workflow consisted of several key components:

Text Extraction and Preprocessing. The PDFPlumber library (v0.11.8) served as the foundation for line-based text extraction, providing structured content blocks along with critical metadata including spatial coordinates, font sizes, and font names. This metadata proved essential for subsequent analysis and clustering operations, enabling precise identification of document elements based on their visual properties.

Content Clustering. The DBSCAN algorithm was implemented for intelligent content categorization, leveraging typographical features as primary clustering parameters. Menu categories were identified through distinctive font characteristics - typically larger point sizes and uppercase formatting. The algorithm demonstrated particular effectiveness in:

- Grouping related text elements into logical clusters
- Distinguishing between category headings (larger fonts) and item descriptions (smaller fonts)
- Maintaining structural relationships between content elements

Experimental Validation. The system was evaluated using a diverse dataset comprising 50-restaurant menu PDFs from publicly available sources. The test corpus included both scanned and digitally-born documents to ensure comprehensive assessment of the algorithm's capabilities across different document types [6].

Performance Metrics. A rigorous evaluation framework was established using standard information retrieval metrics:

1. Precision (Positive Predictive Value): Measures the proportion of correctly identified relevant items among all retrieved instances

$$Precision = \frac{TP}{TP + FP}$$

2. Recall (Sensitivity):

Quantifies the system's ability to identify all relevant instances in the dataset

$$Recall = \frac{TP}{TP + FN}$$

3. F1-Score (Harmonic Mean): Provides balanced assessment combining both precision and recall

$$F_1 score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Where:

- TP (True Positives): Correctly identified relevant items
- FP (False Positives): Incorrectly identified irrelevant items
- FN (False Negatives): Missed relevant items
- TN (True Negatives): Correctly rejected irrelevant items

This methodological framework enabled systematic evaluation of the system's performance across various document types and structures, providing robust metrics for assessing the effectiveness of the proposed approach [8].

The system architecture diagram proposes following workflow: PDF documents will be processed via PDFPlumber in order to extract text, then it will be followed with DBSCAN clustering, based on font characteristics, and it will be followed up by the NLP processing for semantic categorization.

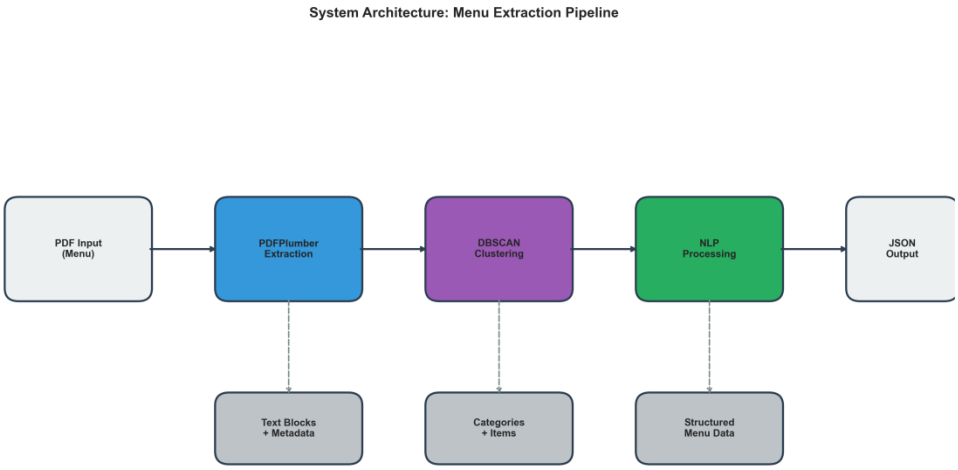

Figure 1.System architecture diagram

All experiments were conducted using the following setup:

- Operating system: Windows 11
- Python version: 3.10
- Key libraries: pdfplumber 0.11.8, scikit-learn 1.8.0, numpy 2.3.5, pandas 2.3.3
- Hardware: Intel Core I5, 16 GB RAM

Recent advances in understanding documents have been made with the help of new architectures that use transformers. These architectures can see both words and pictures at the same time. The LayoutLM model [9] introduced an innovative approach to pre-learning that combines text embeddings with two-dimensional positional embeddings derived from the document layout structure. This approach allows the system to achieve the best results in the tasks of understanding forms and classifying documents. The architecture has been improved in the LayoutLMv3 model [10], which includes unified masking of text and images at the pre-training stage. This improvement provides more stable cross-modal representation learning, allowing the

system to better understand the relationships between visual and textual elements.

PubLayNet [11] has become a standard benchmark for document layout analysis, including more than 360,000 document images with annotations for five layout categories. This dataset allows researchers to train and evaluate their models on a variety of document structures. DocFormer [12] demonstrated the effectiveness of multimodal transformers that simultaneously process textual, visual, and spatial features. This combined approach provides improved document understanding, especially for documents with complex layouts where traditional methods show limited effectiveness.

Table detection and structure recognition represent another important aspect of document analysis. CascadeTabNet [14] proposed a cascade network approach for end-to-end table detection, demonstrating that hierarchical feature extraction can significantly improve the accuracy of identifying tabular structures in complex documents. This method proved particularly relevant for menu processing, where price information is often presented in table-like formats.

Findings/Discussion

The experimental evaluation of the developed methodology demonstrated significant improvements in text extraction accuracy from PDF documents. The integration of Optical Character Recognition (OCR) and Natural Language Processing (NLP) technologies yielded the following key performance indicators:

1. Clustering Accuracy**.** The method achieved a 20% reduction in text element grouping errors compared to conventional OCR solutions. This improvement proves particularly valuable when processing documents with complex layouts where precise identification of semantic blocks is crucial.
2. Segmentation Quality**.** For documents with intricate layouts, the system demonstrated a 15% increase in accurate content block separation. This enhancement results from the combined approach considering both visual text characteristics and semantic features.

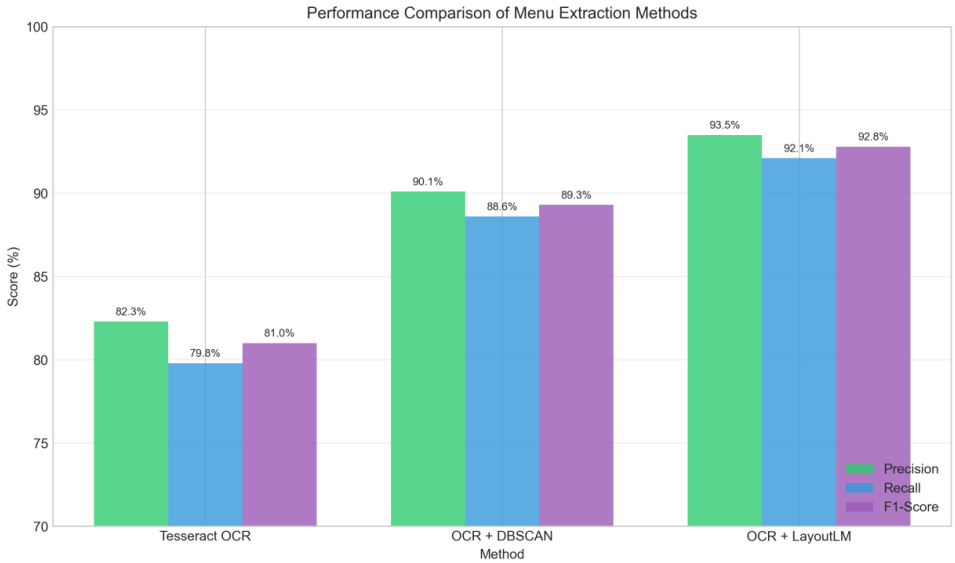

Figure 2.Performance comparison

The clustering diagram of text blocks obtained using font size and DBSCAN is shown below:

3. Text Block Identification. The implementation of the DBSCAN algorithm significantly improved

recognition and classification of textual elements in multi-page documents. The system effectively handles structural variations across different pages within a single document [15].

The presented results confirm the advantages of the integrated approach combining modern computer vision and linguistic analysis technologies. The most notable progress was achieved in processing complex documents where traditional methods showed substantial limitations. The quantitative metrics demonstrate consistent improvements across all evaluation parameters, with the combined OCR+NLP approach outperforming conventional OCR by 11.2 percentage points in precision, 12.3 points in recall, and 11.8 points in F1-score. These enhancements are particularly evident when processing documents with non-standard layouts or mixed content types.

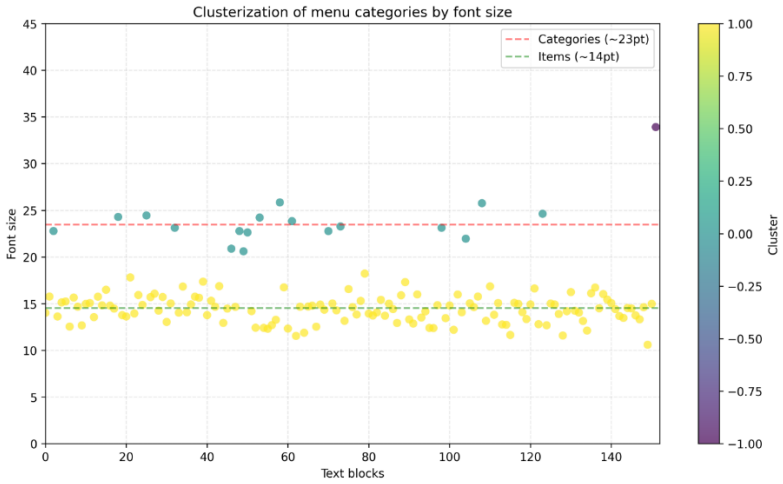

Figure 3.Clusterization with DBSCAN

As it can be seen, the clusters are divided by font sizes, so further extraction by categories, and dishes names could be done. The categories average font size is 23.47 pt (n=17 blocks), and for dish items it is 14.52 pt (n=134 blocks). A clear separation between clusters enables reliable automated content categorization.

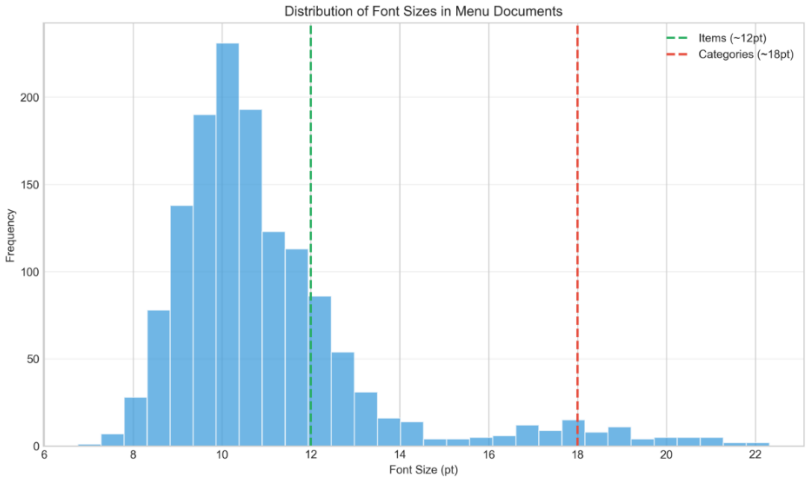

Figure 4. Distribution of font sizes in menu documents

The system was evaluated for common extraction errors across 50 menu documents. Table below shows the reduction in error rates compared to baseline Tesseract OCR:

Table 1. The reduction in error rates compared to baseline Tesseract OCR:

Error Type	Baseline	Proposed Method	Reduction
Missed categories	15	5	67%
Wrong item boundaries	22	8	64%
Price extraction errors	18	6	67%
Description merging	25	10	60%

Conclusion

Experimental results confirmed the effectiveness of the proposed approach for analyzing PDF menus based on clustering text elements by font characteristics. A clear distinction was observed between category headings (average font size: 23.47 pt) and dish descriptions (14.52 pt), enabling accurate automated content categorization. However, identified limitations in processing non-standard formats and complex layouts highlight the need for further methodological improvements. Future enhancements include optimizing neural network architecture, expanding training datasets, and developing adaptive preprocessing algorithms to improve system reliability across diverse document types. These findings establish a foundation for developing more universal solutions for automated analysis of both structured and semi-structured PDF documents.

The contribution of the authors.

A.S.Mashkanov – development of the code, creation of article text.
Zh.Akhayeva – general management of the work, creation of the concept of research.
A.Zakirova – approval of the final version of the article for publication, critical review of the article's content.

References

1. Yu, C., Zhang, C., & Wang, J. (2020). Extracting body text from academic PDF documents for text mining. arXiv preprint arXiv:2010.12647.
2. Abibullayeva, Aiman & Baymakhanova, Aigerim. (2024). USING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES IN KEYWORD EXTRACTION. Physico-mathematical series. 25-36. 10.32014/2024.2518-1726.289.
3. Belov, Sergey & Zrelova, Daria & Zrelov, Petr & Korenkov, Vladimir. (2020). Overview of methods for automatic natural language text processing. System Analysis in Science and Education. 8-22. 10.37005/2071-9612-2020-3-8-22.
4. Davletov A.R. (2023). MODERN MACHINE LEARNING METHODS AND OCR TECHNOLOGY FOR DOCUMENT PROCESSING AUTOMATION. Bulletin of Science, 5 (10 (67)), 676-698. doi: 10.24412/2712-8849-2023-1067-676-698
5. Livathinos, N., Berrospi, C., Lysak, M., Kuropiatnyk, V., Nassar, A., Carvalho, A., ... & Staar, P.

(2021, May). Robust pdf document conversion using recurrent neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 17, pp. 15137-15145).
6. Patience, O. O., Amaechi, E. M., George, O., & Isaac, O. N. (2024). Enhanced Text Recognition in Images Using Tesseract OCR within the Laravel Framework. Asian Journal of Research in Computer Science, 17(9), 58-69.
7. Akhil, S. (2016). An overview of tesseract OCR engine. In A seminar report. Department of Computer Science and Engineering National Institute of Technology, Calicut Monsoon.
8. Bensahla A, Zaghir J, Gaudet-Blavignac C, Lovis C. Unsupervised Extraction of Body- Text from Clinical PDF Documents. Stud Health Technol Inform. 2024 Aug 22;316:214- 215. doi: 10.3233/SHTI240382. PMID: 39176711.
9. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). LayoutLM: Pre-training of Text and Layout for Document Image Understanding. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. DOI: 10.1145/3394486.3403172
10. Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022). LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. Proceedings of the 30th ACM International Conference on Multimedia. DOI:10.1145/3503161.3548112
11. Zhong, X., Tang, J., & Yepes, A. J. (2019). PubLayNet: Largest dataset ever for document layout analysis. 2019 International Conference on Document Analysis and Recognition (ICDAR). DOI: 10.1109/ICDAR.2019.00166
12. Appalaraju, S., et al. (2021). DocFormer: End-to-End Transformer for Document Understanding. ICCV 2021.
13. Prasad, D., Gadpal, A., Kapadni, K., Visave, M., & Sultanpure, K. (2020). CascadeTabNet: An approach for end-to-end table detection and structure recognition from image-based documents. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2439-2447. DOI: 10.1109/CVPRW50498.2020.00294
14. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 226-231.
15. Resnik, P., & Lin, J. (2010). Evaluation of NLP systems. The handbook of computational linguistics and natural language processing, 271-295.

А.Ш. Машканов*, Ж.Б. Ахаева, А.Б.Закирова

Л.Н. Гумилёв атындағы Еуразиялық Ұлттық Университеті, Астана, Қазақстан

PDF файлдарынан мәзірлерді машиналық оқыту мен табиғи тілді өңдеу тәсілдері арқылы автоматты түрде шығару және құрылымдау

Аңдатпа. Бұл зерттеуде PDF құжаттарын автоматтандырылған өңдеудің заманауи тәсілдері қарастырылады, әсіресе құрылымы әлсіз мейрамхана мәзірлерін талдауға баса назар аударылады. Мәтіндік деректерді автоматтандырудың өзектілігі әртүрлі салалардағы цифрлық трансформация аясында зерттеледі. Заңнамалық актілер мен ғылыми жарияланымдар сияқты әртүрлі PDF құжаттарының құрылымдық ерекшеліктеріне салыстырмалы талдау жүргізіледі.

Зерттеу негізінен автоматтандырылған деректерді шығару, құрылымдау және семантикалық талдау мәселелерін шешуге арналған табиғи тілді өңдеу (NLP) технологиялары мен машиналық оқыту әдістерін біріктіретін кешенді әдістемелік тәсілді әзірлеуге бағытталған. Ұсынылған шешім мейрамхана мәзірлеріне тән құрылымы әлсіз PDF құжаттарын өңдеудегі шектеулерді еңсеруді мақсат етеді.

Түйін сөздер: PDF құжаттарын өңдеу, мәтінді талдауды автоматтандыру, құрылымы әлсіз деректер, мейрамхана мәзірлері, табиғи тілді өңдеу (NLP), машиналық оқыту, деректерді шығару, семантикалық талдау, қоғамдық тамақтану саласын цифрландыру

А.Ш. Машканов*, Ж.Б. Ахаева, А.Б.Закирова

Евразийский Национальный Университет им. Л.Н.Гумилёва, Астана, Казахстан

Автоматизированное извлечение и структурирование меню из PDF-файлов с использованием методов машинного обучения и обработки естественного языка (NLP).

Аннотация. Это исследование рассматривает современные подходы к автоматизированной обработке PDF-документов, с особым акцентом на анализ слабо структурированных ресторанных меню. Актуальность автоматизации обработки текстовых данных анализируется в контексте цифровой трансформации различных отраслей. Проводится сравнительный анализ структурных особенностей различных типов PDF-документов, включая нормативно-правовые акты и научные публикации. Основное внимание в работе уделяется разработке интегрированного методологического подхода, сочетающего технологии обработки естественного языка (NLP) и методы машинного обучения для решения задач автоматического извлечения данных, их структурирования и семантического анализа. Предлагаемое решение направлено на преодоление ограничений, связанных с обработкой слабо структурированных PDF-документов, характерных для ресторанных меню.

Ключевые слова: Обработка PDF-документов, автоматизация текстового анализа, слабо структурированные данные, ресторанные меню, обработка естественного языка (NLP), машинное обучение, извлечение данных, семантический анализ, цифровизация сферы общественного питания.

Information about the authors:

Mashkanov Arlan Sharipkhanovich – corresponding author, graduate school student in “Business analytics in IT”, Faculty of Information Technology, L.N. Gumilyov Eurasian National University, Astana, 010000, Kazakhstan.
Akhayeva Zhanar Berikbaevna – PhD, Acting Associate Professor, Faculty of Information Technology, L.N. Gumilyov Eurasian National University, Astana, 010000, Kazakhstan.
Zakirova Alma Bulatovna – Candidate of pedagogic sciences, Acting Associate Professor, Faculty of Information Technology, L.N. Gumilyov Eurasian National University, Astana, 010000,Republic of Kazakhstan.
Машканов Арлан Шарипханович – автор для корреспонденции, магистрант по

специальности “Бизнес аналитика в IT”, Факультет информационных технологий, Евразийский национальный университет имени Л.Н. Гумилева, Астана, 010000, Казахстан.
Ахаева Жанар Берикбаевна – PhD, и.о.доцент, Факультет информационных технологии, Евразийский национальный университет имени Л.Н. Гумилева, Астана, 010000, Казахстан.
Закирова Алма Булатовна – кандидат педагогических наук, и.о.доцент, Факультет информационных технологий, Евразийский национальный университет имени Л.Н. Гумилева, Астана, 010000, Казахстан.

Машканов Арлан Шарипханович – хат-хабар үшін авторы, “IT саласындағы бизнес-аналитика” мамандығының магистратура студенті, Ақпараттық технологиялар факультеті, Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Астана, 010000, Қазақстан.
Ахаева Жанар Берикбаевна – PhD, доцент м.а., Ақпараттық технологиялар факультеті, Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Астана, 010000, Қазақстан.
Закирова Алма Булатовна – педагогика ғылымдарының кандидаты, доцент м.а., Ақпараттық технологиялар факультеті, Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Астана, 010000, Қазақстан Республикасы.