UDC 004.9:004.021

V. Palii, O. Zhurakovska

DATA RECOGNITION IN DOCUMENTS AND CLASSIFICATION ALGORITHM

Abstract: The article considers the actual problem of data recognition in documents and their classification using the "Core Vocabulary", which corresponds to the common data model for describing public service. For solving the issue an algorithm is developed that allows to recognize data in documents and classify it, which is very important while transferring from a document-oriented public services model to a data-oriented model. The algorithm is the basis of the algorithmic software of “Information system for data recognition and classification”. An illustrative example is present.

Keywords: text mining techniques, document analysis, data recognition, data classification, information system for data recognition and classification.

Introduction

In modern world citizens are regularly in need of acquiring administrative services from state authorities or local self-government. Nowadays, in Ukraine it is a common occurrence that first try of acquisition of administrative services fails due to blurry requirements for getting the said services. As a result, people will be forced to go through queuing and dealing with bureaucracy several times before succeeding in getting the needed services. Having stated that, the necessity of simplification, as well as clarification, is becoming more and more obvious. A possible solution to the issue could be a creation of some kind of electronic system that will provide citizens with necessary information on what documents are required or even allowing to receive the providing with administrative services via the Internet.

In addition, such system is currently being developed, however, the current way of processing documents which is used in the system has much handwork required. Having taken things in consideration, it was decided that some way of digitalization is needed, and so Ministry of Digital Transformation of Ukraine started working towards this direction and a law about public electronic registrations was passed.

These processes are related to the necessity to represent public service not as a set of documents, but as a set of data. In this regard, the task of data recognition in documents and its classification using basic dictionaries, which are the basics of common data model for describing public services offered in administration, is relevant [1].

© **V. Palii, O. Zhurakovska**

Research analysis of recognition and classification of data

The problem of data recognition in text is solved using methods of linguistic analysis and machine learning [2].

The way author suggests resolving the task of developing classification part of an algorithm for recognizing of data in document is with the usage of text mining.

Text mining is the process of extracting information from text [2].

Below are listed the text mining steps.

1. Gathering information. During this step the sources which information will be taken from are identified. For example, the source could be DOC or DOXC files, text files, websites and others. The source used in the article is PDF document as it is the most common document type.
2. Text preprocessing. This step preprocesses the source picked in step 1 by filtering and stemming. As a result, some of the unnecessary words will be removed and other words will be reduced to their stems.
3. Information extraction. This step is responsible for finding key features in text which would be then processed by text mining techniques.
4. Text analysis. This step is the key one in text mining because during it the data after being prepared in previous steps is being processed using selected techniques of text mining.
5. Interpretation. This is the last step of text mining at which all the data after being prepared and processed is being interpreted in such shape and form that the user could understand.

As for the main text mining techniques they are the following:

1. Information retrieval. This technique is used to search for information in text documents, performing the search based on a word or phrase. The best examples are the search engines such as Google.
2. Information extraction. Information extraction technique extracts information from textual data. It analyzes unstructured, semi-structured or structured, machine-readable text data sources. This technique focuses on identifying the extraction of entities, attributes, and their relationships.
3. Categorization. In the process of categorizing, a certain number of categories from a predetermined list are matched with a document or its part.
4. Clustering. Clustering is a process of grouping sources the way that each item of each group is closer to other items in the group, according to clustering algorithm, than to items in other groups.
5. Summarization. Summarization is a technique that analyzes a specific text data with the purpose to compress it to brief summary. It helps user to grasp the general content of the document thus reducing time needed to decide if the information in a document is worth spending efforts to read.

Considering the fact that the object of interest of the article is, among other things, the use of text mining in matching the gaps of a document with the core-vocabulary-defined data types the categorization technique is the one that will be used [3].

Classification is of three types: rule-based, machine learning-based and mixed [4, 5].

Rule-based classification classifies text based on the predefined set of rules. These rules consist of pattern and the category that the pattern matches with.

Rule-based classification requires a development of the said rules with use of which it preforms classification, so in case of failure it is easy to find and correct mistakes. However, in case any changes to system are needed it may take considerable time to change or add new rules.

Compared to rule-based classification the use of machine learning classification requires a test dataset, so the model could be trained. In addition, when a change of algorithm is required, it is only a model that will be changed.

In case of developing an algorithm for recognizing data in document based on the core vocabulary the author is inclined to prefer rule-based classification over machine learning one. Having stated that, the reasons are:

– there are no available datasets to use in machine learning classification;

– when implementing the usage of core vocabulary, it is easier to go with rule-based classification [3].

Problem statement

In the purpose of simplification current way of processing the document, which is all about filling the gaps in document with information, is to be evolved into a new one. Current approach is that for getting administrative services one needs to bring a certain number of documents, so the idea is following – citizens bring the necessary documents and as a result of a service they get the document they desire. In the Fig.1 there is “As Is” diagram portraying the way administrative service is provided nowadays. The approach has its drawbacks, like previously stated - some documents exist in multiple forms and the need to bring physical copies stalls the process of service being carried out, so the Ministry came up with a new approach.

In the Fig.2 the new approach of modeling of administrative services, which is based on common data model for describing public services offered in administration [1], is shown. The idea of data-driven approach is that the information needed to be filled in gaps in the documents is now considered as one containing data meaning that information in each gap can be multiple data. For example, field that requires provision of a person first name and last name in terms of data-driven approach is determined to consist of two data field – First name, Last name. This approach allows to store information of each customer and thus refrain from the dire situation where exist multiple documents that are seemingly different but are actually a copy of one another. This way of dealing with documentation is considered to be much easier and more reliable because it is not the documentation itself but the data that will be needed. In order to transfer from document level to data level a lot of work to transform documents into data is required. And so was born an idea to create an algorithm for recognizing the data in document for the sole purpose of having things simplified and digitalized.

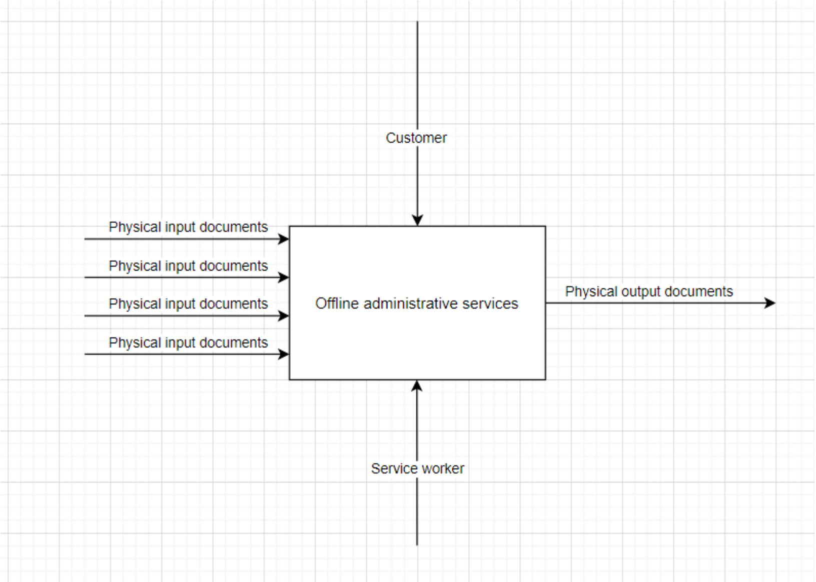


Figure 1. “As Is” diagram of current administrative service approach

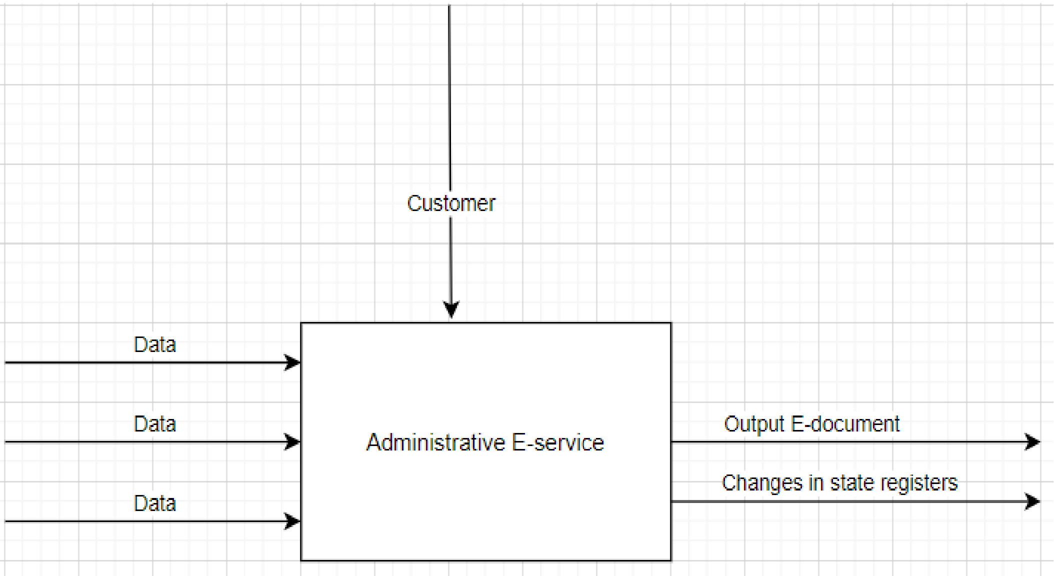


Figure 2. “To Be” diagram of data-driven approach to administrative service

In general, the task is as following:

1. Input is *W*, and the output is $D = \{d_1, d_2, ..., d_n\}$, where *W* is the document that is being analyzed and *D* is the set of data recognized in the document *W*.

2. The input is $D = \{d_1, d_2, ..., d_n\}$ and the output is a map $\{d_i \to P_i\}$, $i = 1...n$, where D is the set of data recognized in the document W and P_i is a property of class Person defined in European Union core vocabulary.

Solving the task

Considering all the stated matters, there is a necessity in development of an algorithm for recognizing data in document. The algorithm consists of two parts – recognizing the parts of the document that are related to that gaps which are to be filled in; and extracting data of the found parts based on core vocabulary of data recommended by European Union. The paper explains usage of the algorithm for recognizing data based on core vocabulary, specifically for the class Person that can be seen in the Fig. 3.

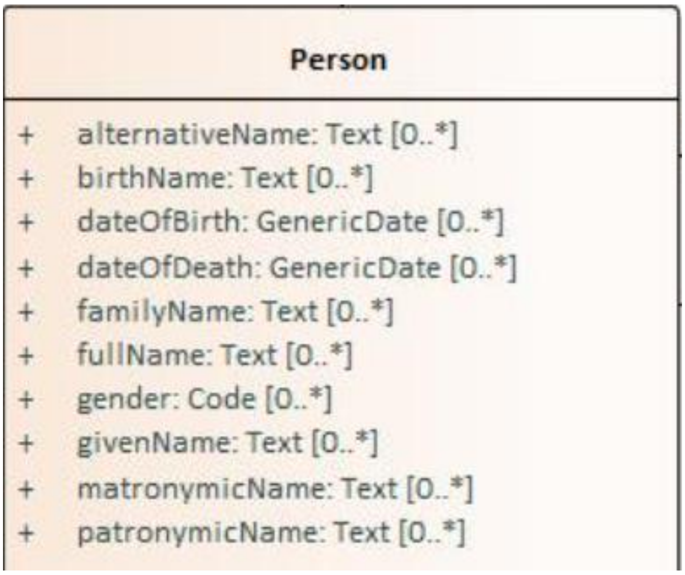

Person
+ alternativeName: Text [0..*]
+ birthName: Text [0..*]
+ dateOfBirth: GenericDate [0..*]
+ dateOfDeath: GenericDate [0..*]
+ familyName: Text [0..*]
+ fullName: Text [0..*]
+ gender: Code [0..*]
+ givenName: Text [0..*]
+ matronymicName: Text [0..*]
+ patronymicName: Text [0..*]

Figure 3. Class Person from European Union recommended core vocabulary

In terms of the algorithm for recognizing of data in document steps: “Gathering information”, “Text preprocessing”, “Information extraction” corresponds to the gap recognition part of the algorithm. In broader way those steps could be called the preprocessing stage. Whereas, steps “Text analysis” and “Interpretation” are responsible for data recognition and data display part of the algorithm.

To find parts of document that corresponds the gap which are to be filled and is to contain data a development of heuristic algorithm was being carried out as there are no organized samples of test data to be used in machine learning.

Now that the classification method for the second part of algorithm is chosen, it is high time to describe the implementation of the first part which is recognizing the parts of the document that are related to that gaps that are to be filled in.

The general idea of the recognition of text related to the gaps can be described by following pseudocode fragments:

```
1. document <–READ(Document.pdf)
2. preprocessedDocument <– preprocessDocument(document)
3. gapPositionsArray <– findGapPositions()
4. dataRelatedArray <– findTextRelatedToGaps()
```

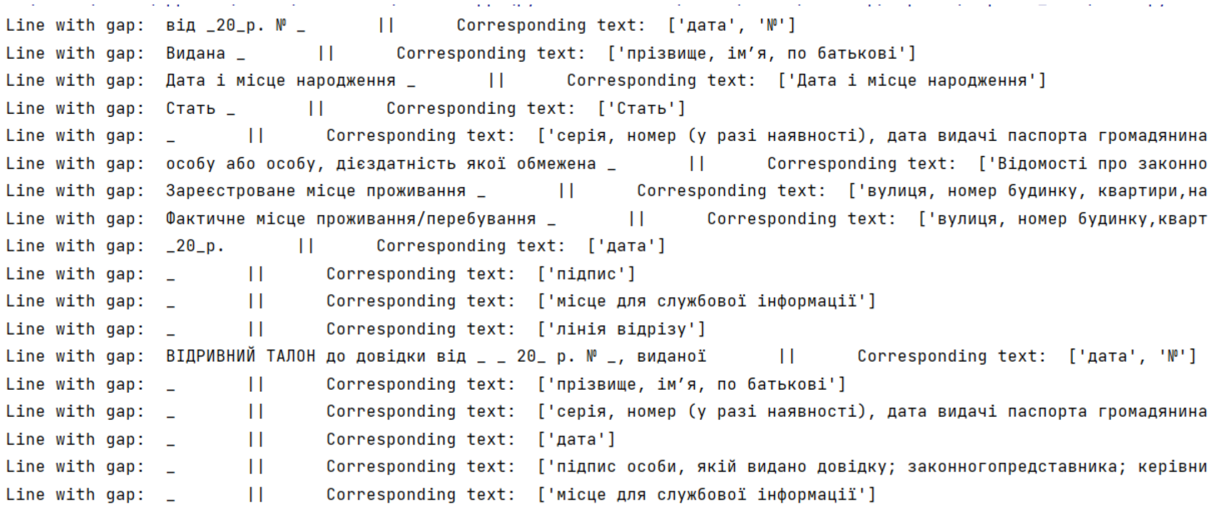
In the first step document in PDF format is read by the algorithm into a string.

Second step is responsible for breaking the string into array of lines according to the document structures and removing unnecessary symbols and lines.

The third step refers to identifying the positions of gaps that are meant to be filled.

The fourth step finds text corresponding to each gap.

The result of the first part of an algorithm for recognizing data in document, which can be seen in the Fig.4, is to be passed to the second part as the input data, this way the classification will define which data from class Person is being required to fill the gap.



```
Line with gap:  від _20_р. № _       ||      Corresponding text:  ['дата', '№']
Line with gap:  Видана _       ||      Corresponding text:  ['прізвище, ім’я, по батькові']
Line with gap:  Дата і місце народження _       ||      Corresponding text:  ['Дата і місце народження']
Line with gap:  Стать _       ||      Corresponding text:  ['Стать']
Line with gap:  _       ||      Corresponding text:  ['серія, номер (у разі наявності), дата видачі паспорта громадянина
Line with gap:  особу або особу, дієздатність якої обмежена _       ||      Corresponding text:  ['Відомості про законно
Line with gap:  Зареєстроване місце проживання _       ||      Corresponding text:  ['вулиця, номер будинку, квартири,на
Line with gap:  Фактичне місце проживання/перебування _       ||      Corresponding text:  ['вулиця, номер будинку,кварт
Line with gap:  _20_р.       ||      Corresponding text:  ['дата']
Line with gap:  _       ||      Corresponding text:  ['підпис']
Line with gap:  _       ||      Corresponding text:  ['місце для службової інформації']
Line with gap:  _       ||      Corresponding text:  ['лінія відрізу']
Line with gap:  ВІДРИВНИЙ ТАЛОН до довідки від _ _ 20_ р. № _, виданої       ||      Corresponding text:  ['дата', '№']
Line with gap:  _       ||      Corresponding text:  ['прізвище, ім’я, по батькові']
Line with gap:  _       ||      Corresponding text:  ['серія, номер (у разі наявності), дата видачі паспорта громадянина
Line with gap:  _       ||      Corresponding text:  ['дата']
Line with gap:  _       ||      Corresponding text:  ['підпис особи, якій видано довідку; законногопредставника; керівни
Line with gap:  _       ||      Corresponding text:  ['місце для службової інформації']
```

Figure 4. The result of the first part of the algorithm

In the Table 1 are shown the supposed results of the algorithm – successfully recognized data in the document that belongs to class Person.

Table 1

Result of classification

Property from class Person	Data recognized in document
fullName	прізвище, ім’я, по батькові
gender	Стать
dateOfBirth	Дата народження

These results allow to classify the recognized data using the “Core Person Vocabulary”. Further development of the research involves expanding the list of dictionaries for classifying other types of data.

Conclusions

The article considers the actual problem of data recognition in documents and their classification using the "Core Vocabulary", which corresponds to the common data model for describing public service. An algorithm has been developed that recognizes data and forms a set of data that is classified at the second stage using the Core Vocabulary.

All in all, the implementation of an algorithm for recognizing data in document is believed to help developers of the Ukrainian state electronic system to quicker evolve from level of information in documents to data level, thus having greatly simplified current way state institution and public services are organized. Practical significance of the developed algorithm is that with its implementation in state “Information system for recognition and classification of data in documents”.

REFERENCES

1. Core Public Service Vocabulary. [Online]. Available: https://ec.europa.eu/isa2/solutions/core-public-service-vocabulary-application-profile-cpsv-ap_en. Accessed on: March 22, 2023.

2. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E., Gutiérrez, J., Kochut, Krys. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques, 2017. [Online]. Available: https://www.researchgate.net/publication/ 318336890_A_Brief_Survey_of_Text_Mining_Classification_Clustering_and_Extraction_Techniques.

3. Core Person Vocabulary. [Online]. Available: https://semiceu.github.io/Core-Person-Vocabulary/releases/2.00/. Accessed on: March 22, 2023.

4. Ranjan, N., Chakkaravarthy, M. A brief survey of machine learning algorithms for text document classification on incremental database. Test Engineering & Management, 25246-25251, 2021. [Online]. Available: https://www.researchgate.net/publication/350451142_A_Brief_Survey_of_Machine_Learning_Algorithms_for_Text_Document_Classification_on_Incremental_Database.

5. Nihar, R., Abhishek, G., Ishwari, D., Payal, G. A survey on text analytics and classification techniques for text documents / International Journal of Development Research, 2021. [Online]. Available: https://www.researchgate.net/publication/354522993_A_SURVEY_ON_TEXT_ANALYTICS_AND_CLASSIFICATION_TECHNIQUES_FOR_TEXT_DOCUMENTS.