Using Part-of-Speech Tagging for Building Networks of Terms in Legal Sphere

Dmytro Lande[a,b,c] and Oleh Dmytrenko[a]

[a] *IRR of NAS of Ukraine, 2, Mykoly Shpaka Street, Kyiv, 03113, Ukraine*
[b] *NTU «Igor Sikorsky KPI», 37, Prosp. Peremohy, Kyiv, 03056, Ukraine*
[c] *SRIIL of NALS of Ukraine, 110-v, Saksaganskogo Street, Kyiv, 01032, Ukraine*

Abstract

This paper considers an important formalization problem and building the terminological ontology of problem subject domains based on content-related text data. As an ontological model, we propose to use a linguistic network model of text representation, the so-called network of key terms. In this network, the nodes are keywords and phrases that appear in the text corpus, and the links between them are semantic-syntactic links between these terms in the text. Using systems of aggregation of thematic information flows from freely available information resources distributed in global computer networks, input sets of text data were prepared. In particular, this paper solves the important and urgent problem of computerized processing of legal information. The task of computerized processing of natural language texts lies at the intersection between linguistic theory and mathematical sciences. Therefore, a wider natural language processing based on Part-of-Speech tagging was used for extraction of the key terms. After the extraction, a statistical weighing of the formed words and phrases was performed. The horizontal visibility graph algorithm was used to build undirected links between key terms. This paper also considers a new method that allows determining the direction of links between terms and weighting these links in the undirected network of words and phrases. This method takes into account the parts of speech tagging and also obeys the principle of inclusion of a word or phrase in their corresponding extended phrases with more words. The approbation of the proposed method was carried out on the example of a freely available legal document «Universal Declaration of Human Rights». After extracting the key terms from this legal document and determining the direction and weight of links between words or phrases using the proposed methods the directed weighted network of terms was built. The considered in this work method for building the terminological networks can be used, in particular, in systems for automatic text structuring and summarizing of legal information, or systems for detecting the duplicates and contradictions in normative legal documents. It will promote the formation and improvement of conceptual and terminological apparatus in the legal sphere and harmonize national and international law.

Keywords

Information space, unstructured data, ontological model, problem subject domain, legal information, text corpus, computerized text processing, Part-of-Speech tagging, network of terms, automatic summarization

1. Introduction

Modern information and communication technologies and the information space, in general, are developing faster than ever before. This process is characterized by a correspondingly rapid increase in data volumes [1]. These large data volumes are produced by elements of the information space, in particular, documents and a variety of data sources such as files, emails, web pages and other sources,

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: dwlande@gmail.com (D. Lande); dmytrenko.o@gmail.com (O. Dmytrenko)
ORCID: 0000-0003-3945-1178 (D. Lande); 0000-0001-8501-5313 (O. Dmytrenko)

regardless of the formats of their presentation. Data is created, recorded, stored, processed and reproduced increasingly often in electronic form. It is important to note the fact that the described above data doubles approximately every 18 months [2]. As a result, over the past five years, humanity has produced more information than during all previous history [3]. For example, the International Data Corporation (IDC) predicts 175 zettabytes (in other words, 175 trillion gigabytes) of new data will be created around the world in 2025 [4]. But such an information surge, or so-called an information explosion, is accompanied not only by an influx of new valuable knowledge. The majority of such data, however, are unstructured data including unnecessary and noisy data, which constitute 95% of big data [5], and only a very small part (about 5%) of all data is a piece of valuable information that can be used in decision-making.

So now the information society is facing a number of problems that no one has faced before. The main problem is the critical discrepancy between the development of modern information systems and the increase of dynamic information flows in global computer networks [6]. Namely, the problem is the lack of appropriate technological solutions and the inability of existing systems to process huge amounts of unstructured data, including text data, and extract knowledge from them at the same rate at which the corresponded data is produced and accumulated. The mentioned above problems lead to the accumulation of unstructured data [7]. In turn, the huge volumes of such messy data make it difficult to find the necessary and relevant information that the Internet user tries to get in response to his request.

Therefore, the huge amount of information flows and dynamic text data that accumulated in global computer networks determine the relevance and importance of the conceptualization process of this data and their further formalization in the form of a certain ontological model.

This leads to the necessity to develop and improve existing technological solutions and create new ones to ensure a sufficiently high speed of processing and analysis of unstructured data.

This process of the global information space formation is important from the point of view of the transformation of the unstructured data accumulated on information resources into the knowledge. In turn, the obtained knowledge can be valuable recommendations in the process of rapid decision-making in various spheres of activity, in particular, telecommunications, cyber, financial, trade, military, political, diplomatic and other spheres.

2. Computerized processing of legal information

Obtaining brief and at the same time the most important and relevant information or informative statements from one or more text documents, so-called summary, abstract or annotation, is an important task of computerized text processing [8]. Generating concise information-rich reports based on short annotations or digests simplifies access to the main content of the text without the need to process a large text document or text corpus.

In the middle of the last century, the works related to automatic text summarization were mentioned [9]. However, due to the globalization of the information space and the continual increase in the number of information flows, the task of an automatic text summarization is more important than ever before. Also, the automatic text summarization rejects information noise, reduces information consumed by humans and promotes rapid access to the main content of the document. As a result, it promotes important management decisions.

Since scientific and technological progress has also affected the legal sphere, the problem of computerized processing of legal information is relevant [10]. The number of normative legal documents submitted in electronic form, and hence the amount of information that an expert in this field has to deal with, is also constantly growing. Although currently there are different systems of automatic summarization [11], improving existing or developing new systems that could process large volumes of legal documents with acceptable performance and quality is still an important task.

The defining feature of legal information is that the related texts are not fully freely accessible and unstructured. This is important to consider the above-mentioned fact when choosing the appropriate method or approach to solve the problem of automatic text summarization in the legal sphere. In general, there are statistical, positional and indicative methods of automatic summarization. In this work, a statistical method was used to calculate the weight values of individual words and phrases.

Based on the conjunction of the statistical method with the linguistic network model, where key terms are nodes and the links between them are semantic-semantic links between terms in a sentence, a new method has been proposed. This method can be used in automatic legal information summarization systems or systems for detecting duplicates and contradictions in legal documents.

3. Text data formalization

An important stage in the complex research of some problem subject domain thematically related to the flow of text data is the presentation of its knowledge in a form that becomes suitable for further automated processing or in other words the formalization of this knowledge. The building terminological ontology of the studied subject domain is one type of its knowledge formalization.

In this work, it is proposed to use a linguistic network model as an ontological model of text data. This choice of model is because, as it turned out, many of the problems that arise when working with information flows lie at the intersection between the mathematical sciences and linguistics theory. The linguistic theory as a branch of general linguistics, in turn, makes it possible to work with natural language texts, knowing their properties, functions and, most importantly, structure. The theory of graphs and complex networks is considered a powerful mathematical theory, within which the problem of formalization of the subject domain can be solved.

Let's consider the mathematical component of conceptualization and further formalization of a certain problem subject domain with which text corpora are meaningfully connected. This paper uses a network model for presenting text data. In other words, texts of a certain thematic orientation can be presented in the form of a network of words and phrases connected by a formal semantic connection. A partial case of such a network model may be a network built based on key terms. In this network, the nodes correspond to the individual key concepts of the subject domain, and the edges are the links between concepts.

From the point of view of linguistics, natural language arises in a number of its problems, which are connected first of all with ambiguity, non-compositionality and self-application of language units. Therefore, when applying the basic techniques of natural language processing, it should be bear in mind that it contains different forms of a word (word forms that have a common basis), derived from another word, and linguistic phrases used to express different meanings. This leads to the fact that the meaning of a single word or phrase in a particular case will depend on the context in which it is [12]. So there is a problem, which is also called inflected language [13]. Since some phrases can be interpreted in two ways, without knowing the context, although knowing the meaning of all other words included in the statement, there is a problem in determining the exact meaning of a complex statement. The above linguistic phenomena significantly complicate the task of establishing the correct reflection of the semantic-syntactic structure of the text into its formal logical representation.

While building the terminological ontologies of the subject domains on the basis of thematic text documents [14] it is important that the terms (words and phrases) used as the names of the concepts that accompany the chosen subject domain are obeyed the principle of unambiguity. It means, the word used as the name should be the name of only one object, if it is a single name. If it is a common name, then this phrase should be a common name for all objects in the same class. Therefore, the linguistic component of natural language text processing is one of the central problems of information technology intellectualization.

4. Basic techniques of natural language processing

In recent years, the tasks of computer processing of dynamic information flows have become increasingly important. In this work, for computerized natural language pre-processing some of the most common techniques are used. In particular, these techniques include text tokenization and removal of stop words.

Tokenization or lexical analysis is the segmentation of a sequence of characters into a sequence of so-called tokens using a scanner or tokenizer that performs the function of lexical analysis. The term "token" should be understood as a certain form of a word. The token is an independent semantic unit, which is considered in aggregate of all its possible forms and meanings. As the initial stage of

computerized text processing, the tokenization allows working with the word as an individual entity, while knowing the context in which this word is used.

To clear the text of words that are a source of noise and are informationally-unimportant, it is recommended to delete co-called stop words [15]. For example, stop words include determiner, prepositions, particles, exclamations, conjunctions, adverbs, pronouns, introductory word, numbers from 0 to 9 (unambiguous). Also, stop words include sequences of characters often used on the Internet (for example, www, HTTP, com, etc.), and others frequently used official, independent parts of speech, symbols and punctuation marks. These words don`t have any additional semantic load. That is why the stop words must be ignored while building terminological ontologies. It is also recommended using a stop dictionary or stop word list that expert in the considered subject domain has formed.

There are various software tools and, in particular, NLTK (Natural Language Toolkit open-source library) modules of the Python NLP (Natural Language Processing) library, which help to easily apply the above methods of pre-processing to different types of texts [16].

After tokenization, a technique such as Part-of-Speech tagging (PoS tagging), or in other words just tagging, is usually used [17]. This natural language processing step is one of the main and basic components of almost any NLP task and helps to extrapolate the language syntax and text structure. The Parts-of-Speech tagging is based not only on the definition of the word but also on the context in which the word is used. That is, the tagging takes into account the connection of the tagged word with neighbouring and related words in a phrase, sentence or paragraph. The main idea of text tagging is relating a word in a text or body to a certain part of speech. Figure 1 shows the main idea of Part-of-Speech tagging in a simple example. For each word in the sentence «One day her mother said» the certain tag (label) that marks a certain part of speech was assigned. For example, the word «one» is referred to as CD (where CD is a tag that marks cardinal number), the word «day» is referred to as NN (where NN marks noun) and so on (where PRP$ marks Possessive Pronoun and VBD marks Verb, past tense).

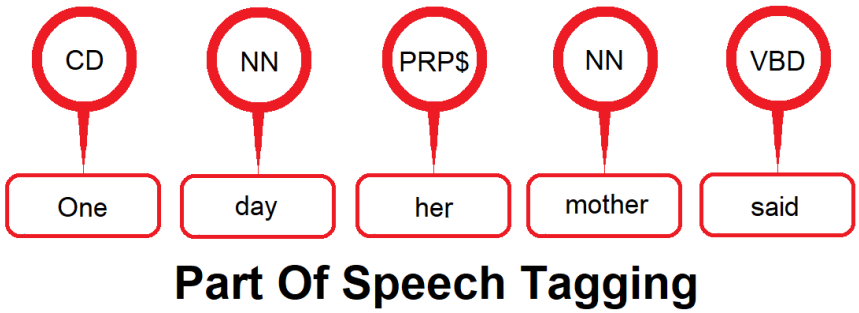

Figure 1: Example of Parts-of-Speech tagging [18]

To mark parts of speech a collection of predefined tags that are assigned to each word in the sentence is used. Figure 2 presents the Penn Treebank list of tags that used for Part-of-Speech tagging task [19].

PoS tagging can be used in searches engines and text corpus analysis tools and algorithms for indexing words and has many other uses as well. Especially PoS tagging can be very useful in case there are words or tokens that can have multiple tags. The tagging helps to distinguish between the occurrences of the word when it used as one part of speech or another. And most importantly, tagging simplifies the context related to a specific subject domain.

The particular parts of speech are represented as word classes or lexical categories. These categories based on the syntactic context of a word or phrase. Therefore, using the Parts-of-Speech as the method for classifying words by parts of speech helps to mark up each word it a text (or corpus) according to its lexical category.

The E. Brill's PoS tagger [21] is one of the first and most widely used English tagger. The stochastic algorithms are also used in addition to a group of rules-based algorithms.

Tag	Part of speach	Tag	Part of speach
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	To
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WPS	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Figure 2: The Penn Treebank PoS Tagset (excluding punctuations) [20]

5. Statistical weighting and key terms extracting using Part-of-Speech tagging

The initial stage of formalization of knowledge about a certain subject domain is the conceptualization or, in other words, the definition of basic objects (individuals, attributes, processes, etc.) and the relationship between them. If we talk about building a terminological ontology as a network based on text corpora, then an important task is to define key terms (key words and phrases). In their symbolic form, these key term actually denote objects, processes or phenomena of the real world or environment.

To define these basic concepts (key terms), it is proposed to perform statistical weighing of words and phrases that the text corpus contain, taking into account Part-of-Speech tagging.

To extract key words and phrases from the text it is necessary to assign them a certain numerical weight. A statistical indicator can be used as one of the weights for representing important words. As a statistical weight of terms, the Term Frequency - Inverse Document Frequency (short is TF-IDF) [22] is commonly used. Although this is not the only approach possible to solve the problem of identifying key terms. But in [23] it was shown that hat the use of the GTF (Global Term Frequency) is more effective when working with thematically related text documents that contained in a text corpus. This statistical indicator shows how the term is important in the global context and determined by the ratio of the total number of this term in all documents to the total number of all terms that the documents contain. It was shown that in contrast to the common statistical indicator TF-IDF, the proposed indicator of the importance of terms make it possible more effectively to find information-important elements of the text when working with a thematically predefined text corpus when the information-important term occurs in almost every document.

6. Method

First of all, it should be noted that the building of the networks of terms is carried out within each separate sentence of the text corpus.

In this work, the NLTK (Natural Language Toolkit) module that developed in the Python programming language were used. For example, "word_tokenize" and "pos_tag" are used to automatically split tokens and assign part of speech tags to each word, respectively.

For stop words removal the sets of stop words freely available by references [24, 25] were applied. In addition to the standard sets of stop word it is also proposed to use the list of stop words a formed by experts.

The proposed method for determining keywords and phrases and the direction of links between them is based on the use of the results obtained through the process of classifying words by parts of speech (Part-of-Speech tagging). Practical research [16] shows that the most commonly used part of speech in English text are determiners (their tag is DT), singular or mass noun (NN), plural noun (NNS), personal pronoun (PRP), verbs and all their forms (VB, VBD, VBG, VBN, VBP, VBZ), adjectives (JJ), including comparative adjectives (JJR) and superlative adjectives (JJS), and adverbs (RB) in particular comparative adverbs (RBR) and superlative adverbs (RBS). In general, individual nouns «NN*» that usually related with people, places, things or concepts, and nouns coupled with adjectives (phrases like «JJ* NN*») are considered as key terms. Also in this work phrases that have the form «$NN*_1$ $NN*_2$», «$JJ*_1$ $JJ*_2$», «$JJ*_1$ $JJ*_2$ NN*», «$JJ*_1$ $JJ*_2$ $NN*_1$ $NN*_2$» are considered to be important and key. As was noted above, determiners, prepositions (IN), coordinating conjunction (CC), individual verbs and their form, adverbs and pronouns are stop words. But in this work we consider the phrases which patterns look like «$V*_1$ to $V*_2$», «$NN*_1$ IN/CC $NN*_2$», «$JJ*_1$ IN/CC $JJ*_2$», «JJ* $NN*_1$ IN/CC $NN*_2$», «$JJ*_1$ IN/CC $JJ*_2$ NN*», «JJ_1 JJ_2 NN_1 IN/CC NN_2», «JJ_1 IN/CC JJ_2 NN_1 IN/CC NN_2» as key. After forming the phrases according to described above patterns and arranging them in a certain order (a sequence is formed where phrases with more words are placed before phrases and words that are part of them), the individual stop words are removed.

The next step is the statistical weighing of words and phrases included in the sequence formed at the previous stage. In this work, GTF (Global Term Frequency) the idea of which is described above is used.

The so-called tuple is formed for each formed phrase in the order of its occurrence in the text. Each tuple consists of three elements: the first element is the term (a word or formed phrase); the next is a tag or combination of tags (for formed phrases) that are assigned to a word depending on to which part of speech this word or phrase belong; the last element of this set is the numeric value of GTF. The defining feature of the proposed technique is that the GTF is calculated taking into account the two first elements of the tuple (the word or phrase and the part of speech to which it belongs). The number of such identical pairs that normalized to the total number of formed terms in the whole text determines the value of the third element of the formed tuple.

The next step is to determine the undirected relationships between the terms in the text. The Horizontal Visibility Graph algorithm (HVG) is used to transforms time series that formed with the consequence of numerical values of GTF into the undirected graph [26]. The idea of the algorithm is that the two nodes t_i and t_j, (in our case, two phrases t_i and t_j), which correspond to the x_i and x_j in the formed time series, are in horizontal visibility if and only if $x_k < \min(x_i; x_j)$ for all t_k where $t_i < t_k < t_j$. In our case, the sequence t_i, $i = 1,..,n$ is the sequence of words and phrases formed within the sentence after the above-described pre-processing (where n is the number of all formed terms). HVG allows building the network structures in which numerical weight assigned to individual words or phrases.

If there is determined using the above HVG algorithm on undirected link between the nodes t_i and t_j of the time series, then it is suggested to establish the direction of this link for pair of node t_i and t_j (where t_i is the source node and t_j is the target node) if only:

- if in the sentence the word (not a phrase) that corresponds to the source node t_j occurs earlier than the term (word or phrase) that corresponds to the target node t_j;
- if in the sentence the phrase (not the word) that corresponds to the target node t_j occurs earlier than the term that corresponds to the source node t_j (Figure 3).

According to the principle of forming a sequence of terms, which described above, and also the proposed rules used to determine the links, the network of key terms (key words and phrases) consists of the words and phrases that included in their corresponding extended phrases with more words. In the built directed network of terms, the major part of the terms is so-called an extension of the corresponding phrases and words. The algorithm proposed in the work [27] uses a similar principle of

determining the direction in the terminological network (the directions build on the principle of entering the term into its corresponding phrase with more number of words).

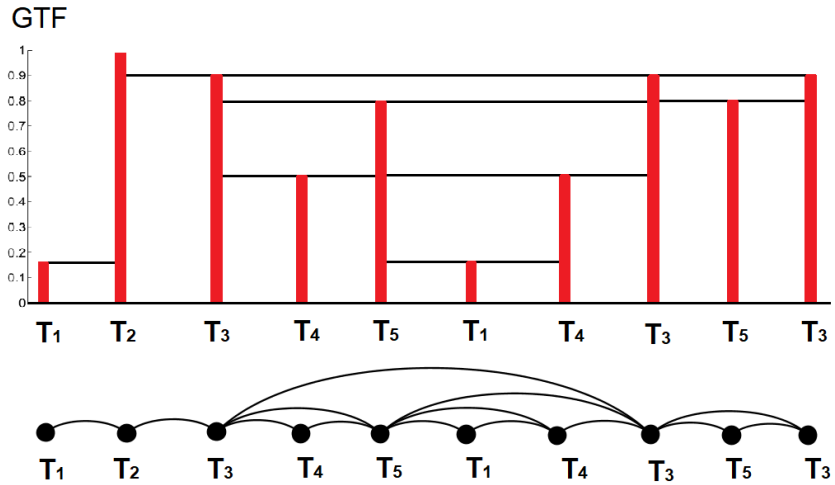

Figure 3: Example of building the directed terminological network

Using the algorithm proposed in [14], the weight values of the links between pair of nodes are determined. After combining ("merging") the nodes that correspond to the same terms (phrases) in the previously built directed network, the number of the same-directed links merged into a single one determines the weight of the merged link between the corresponding nodes.

7. Text data aggregation and corpora forming

During complex research and development of new information systems that are able to process large-scale data, thematic information flows from social media and databases of scientific publications, which are a source of textual data, play an important role. That is, today on the Internet there is a dynamic database available for experiments of such a volume that it was even difficult to imagine before. It is also important to note that the data are practically publicly available on social networks and freely accessible web search engines. In addition, there are various technological possibilities for aggregating this data and forming text corpora, which can then be used as input data sets. Therefore, having the freely available dynamic text arrays and systems of data aggregation from global computer networks opens wide opportunities for improving existing and developing new methods for these data analysis.

In order to carry out objective research and develop new methods and approaches that can be further implemented in information systems, it is important that the input data sets contain the most objective and reliable information. Also, the thematic reflected in these text data sets must be characterized by a sufficient degree of completeness.

The freely available English text «Universal Declaration of Human Rights» published by the United Nations on its website [28] was used to test the methodology.

8. Results of research

Using NLTK (Natural Language Toolkit open-source library) and software modules of the Python NLP (Natural Language Processing) library the initial stages of word processing including tokenization and Part-of-Speech tagging were performed for the text document «Universal Declaration of Human Rights». According to described above methodology, the patterns of phrases were formed. Also, the stop words removing was carried out. At the next stage, for the obtained terms using the global indicator of the importance of the term (GTF) statistical weighing was performed. As a result, for the considered text, the so-called tuple «Term (word or phrase); Tag (part of speech); The

numerical value of GTF» was formed in the order of terms occurrence. The tuples with the largest numerical value of GTF are presented in Table 1 (all tuple is sorted in descending order of their GTF).

In this work, the Gephi software [29] was used to visualize the obtained directed weighted network of words and phrases (Figure 4).

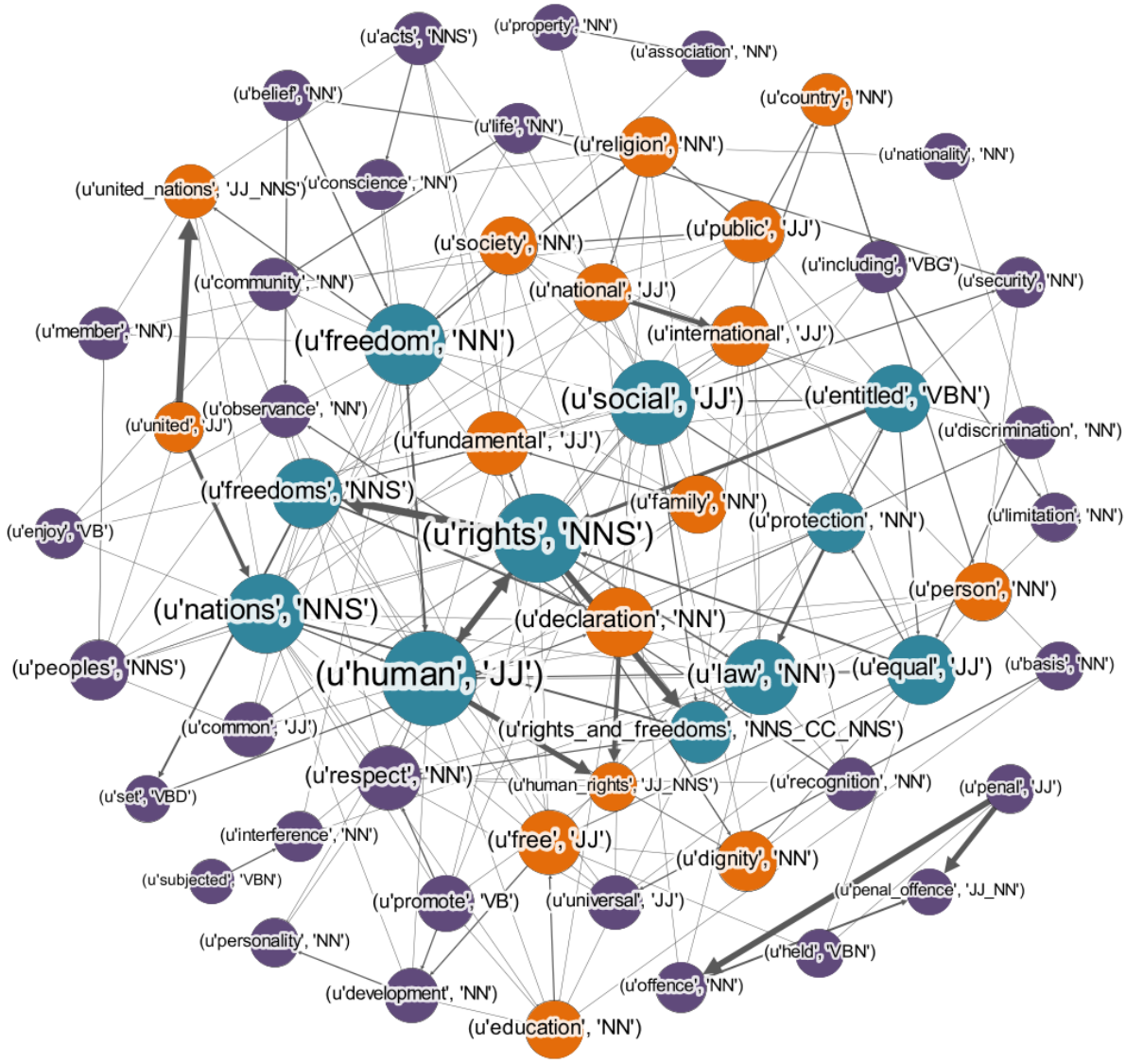

Figure 4: Network of terms built for the legal document «Universal Declaration of Human Rights» (node labels contain the term and its corresponding tag)

Table 1

Top 58 key terms extracted from the «Universal Declaration of Human Rights» and sorted in descending order of their GTF

Terms	Tag	GTF
rights	NNS	0.022105
human	JJ	0.012632
equal	JJ	0.010526
freedoms	NNS	0.010526
freedom	NN	0.010526
law	NN	0.009474
protection	NN	0.009474
entitled	VBN	0.009474

social	JJ	0.008421
rights_and_freedoms	NNS_CC_NNS	0.008421
nations	NNS	0.008421
education	NN	0.007368
free	JJ	0.007368
family	NN	0.006316
human_rights	JJ_NNS	0.006316
fundamental	JJ	0.006316
declaration	NN	0.006316
person	NN	0.005263
society	NN	0.005263
country	NN	0.005263
religion	NN	0.005263
united	JJ	0.005263
national	JJ	0.005263
public	JJ	0.005263
united_nations	JJ_NNS	0.005263
dignity	NN	0.005263
international	JJ	0.005263
discrimination	NN	0.004211
enjoy	VB	0.004211
peoples	NNS	0.004211
promote	VB	0.004211
offence	NN	0.004211
nationality	NN	0.004211
development	NN	0.004211
respect	NN	0.004211
penal_offence	JJ_NN	0.004211
penal	JJ	0.004211
acts	NNS	0.003158
interference	NN	0.003158
set	VBD	0.003158
conscience	NN	0.003158
life	NN	0.003158
community	NN	0.003158
including	VBG	0.003158
common	JJ	0.003158
belief	NN	0.003158
basis	NN	0.003158
property	NN	0.003158
personality	NN	0.003158
limitation	NN	0.003158
recognition	NN	0.003158
member	NN	0.003158
association	NN	0.003158
held	VBN	0.003158
security	NN	0.003158
universal	JJ	0.003158
subjected	VBN	0.003158
observance	NN	0.003158

9. Conclusion

In this paper, a new method for building the networks of terms was considered. The key terms of the built network were extracted using a wider natural language processing based on Part-of-Speech tagging. The analysis of the ontological models obtained by the method for building a directed weighted networks of terms based on text corpus allows making constructive conclusions regarding the subject domain with which the texts are thematically related and can be the basis for decision-making in this domain. The method was tested on the example of enough well-structured legal document «Universal Declaration of Human Rights» that freely available on the Internet. And as a result the network of terms was built. The considered methodology can be used, in particular, in systems of automatic text structuring and summarization of legal information, or systems of detection of duplicates and contradictions in normative legal documents. In general, it will promote the formation and improvement of conceptual and terminological apparatus in the legal sphere and harmonize of national and international law.

10. References

[1] V. Mayer-Schönberger, K. Cukier, Big data: A revolution that will transform how we live, work, and think, Houghton Mifflin Harcourt, 2013.

[2] Humanity Doubles Its Data Creation Every 18 Months, And It Has Powerful Implications, URL: https://www.fluxmagazine.com/data-creation-powerful-implications/.

[3] S. Sagiroglu, D. Sinanc, Big data: A review, in: 2013 international conference on collaboration technologies and systems (CTS), IEEE, 2013, pp. 42-47.

[4] 6 Predictions About Data In 2020 And The Coming Decade, 2020. URL: https://www.forbes.com/sites/gilpress/2020/01/06/6-predictions-about-data-in-2020-and-the-coming-decade/?sh=5d1634224fc3.

[5] A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International journal of information management 35(2) (2015) 137-144.

[6] D. Lande, Analysis of information flows in global computer networks (based on the scientific report at the meeting of the Presidium of the NAS of Ukraine on January 25, 2017), Bulletin of the National Academy of Sciences of Ukraine 3 (2017) 45-53. (in Ukrainian)

[7] R. Feldman, J. Sanger, The text mining handbook: advanced approaches in analyzing unstructured data, Cambridge university press, 2007.

[8] M. Maybury, Advances in automatic text summarization, MIT press, 1999.

[9] H. P. Luhn, The automatic creation of literature abstracts, IBM Journal of research and development 2(2) (1958) 159-165.

[10] D. Lande, O. Dmytrenko, O. Radziievska, Subject Domain Models of Jurisprudence According to Google Scholar Scientometrics Data, in: Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020). Volume I: Main Conference. Lviv, Ukraine, April 23-24, 2020. CEUR Workshop Proceedings (ceur-ws.org), volume 2604, 2020, pp 32-43. ISSN 1613-0073.

[11] Best Text Summarizing Tool for Academic Writing [For Free], 2014. URL: https://ivypanda.com/online-text-summarizer.

[12] A.O. Nikonenko, Review of computer-linguistic methods of natural language texts processing. Artificial Intelligence 3 (2011) 174-181. (in Russian)

[13] K. Ziarek, Inflected Language: Toward a Hermeneutics of Nearness: Heidegger, Levinas, Stevens, Celan, SUNY Press, 1994.

[14] D.V. Lande, O.O. Dmytrenko, Creating the Directed Weighted Network of Terms Based on Analysis of Text Corpora, in: 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC), Kyiv, Ukraine, IEEE, 2020, pp. 1-4. doi: doi.org/10.1109/SAIC51296.2020.9239182

[15] W. J. Wilbur, K. Sirotkin, The automatic identification of stop words, Journal of information science 18(1) (1992) 45-55.

[16] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, O'Reilly Media Inc., 2009.
[17] E. Brill, Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging, Computational linguistics 21(4) (1995) 543-565.
[18] Extract Custom Keywords using NLTK POS tagger in python, 2020. URL: https://thinkinfi.com/extract-custom-keywords-using-nltk-pos-tagger-in-python/.
[19] M. Marcus, B. Santorini, M. A. Marcinkiewicz, Building a large annotated corpus of English: The Penn Treebank, Computational Linguistics 19(2) (1993) 313-330.
[20] B. Santorini, Part-of-speech tagging guidelines for the Penn Treebank Project, Department of Computer and Information Science School of Engineering and Applied Science University of Pennsylvania Philadelphia, PA 19104, 1990.
[21] E. Brill, A simple rule-based part of speech tagger, in: Proceedings of the third conference on Applied natural language processing (ANLC '92). Association for Computational Linguistics, Stroudsburg, PA, USA, 1992, pp. 152-155. doi:10.3115/974499.974526.
[22] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information processing & management 24(5) (1988) 513-523. doi:10.1016/0306-4573(88)90021-0.
[23] D. Lande, Dmytrenko, O. Radziievska, Determining the Directions of Links in Undirected Networks of Terms, in: CEUR Workshop Proceedings (ceur-ws.org). Vol-2577 urn:nbn:de:0074-2318-4. Selected Papers of the XIX International Scientific and Practical Conference «Information Technologies and Security» (ITS 2019), volume 2577, 2019, pp. 132–145. ISSN 1613-0073.
[24] XPO6: Download Stop Word List, 2015. URL: http://xpo6.com/download-stop-word-list/.
[25] Text Fixer: Common English Words List, 2011. URL: http://www.textfixer.com/tutorials/commonenglishwords.php.
[26] G. Gutin, T. Mansour, S. Severini, A characterization of horizontal visibility graphs and combinatorics on words. Physica A: Statistical Mechanics and its Applications 390(12) (2011) 2421-2428.
[27] D. Lande, Building of networks of natural hierarchies of terms based on analysis of texts corpora. *arXiv preprint arXiv:1405.6068*.
[28] Universal Declaration of Human Rights, 2007. URL: https://www.un.org/en/universal-declaration-human-rights/
[29] Gephi, 2017. URL: https://gephi.org.