Elaborative Trademark Similarity Evaluation Using Goods and Services Automated Comparison

Daniil Shmatkov[1,2], Oleksii Gorokhovatskyi[3] and Nataliya Vnukova[1,3]

[1] *Scientific and Research Institute of Providing Legal Framework for the Innovative Development of the National Academy of Law Sciences of Ukraine, Chernyshevska st., 80, 61000, Kharkiv, Ukraine*
[2] *Senior Intellectual Property Counsel, 4B Corporation, Parkovo-Syretska st., 4B, 04112, Kyiv, Ukraine*
[3] *Simon Kuznets Kharkiv National University of Economics, Nauky av. 9a, Kharkiv, 61166, Ukraine*

Abstract

Trademark description comparison is an important problem during registering a new trademark. Manual search for the most similar trademark often faces the plenty of similar trademarks already registered and filed, so subjective evaluation of the trademark similarity imposes some deviations in the comparison. This paper discusses the methods that aim the automatic evaluation of the trademark's similarity/identity based on the text description of goods and services. The analysis included trademarks covering legal services. Various models (count vectorizer, TF-IDF, word2vec, doc2vec) pretrained and trained by ourselves, and sets of vectors comparison approaches (cosine similarity, Tanimoto similarity) have been evaluated. Own simple greedy matching algorithm that can optionally include difference between size of sets which is useful for trademarks comparison has been proposed. The article shows that some combinations of model/similarity measure are successful in searching for correct hierarchy of trademarks, while some other are better in terms of human relevance. The quality and performance analysis of the discussed methods allowed us to choose the most effective of them according to the available data and technical requirements and implement them in practical applications to perform search for the similar or identical trademarks by example. We contribute to the study of methods for automating legal processes and offer a mean to reduce the risk of intellectual property disputes.

Keywords

Trademark, goods and services, legal services, text similarity, word2vec, doc2vec, search

1. Introduction

Trademarks are by far the most popular intellectual property in terms of the number of applications. According to data released by the World Intellectual Property Organization [1] in 2022, there are about five trademark applications per patent application, 12 trademark applications per industrial design application; more specifically, 18.1 million trademark applications were filed globally in 2021 [1]. The competition is so intense that scholars are noticing the emergence of a new "nonsense" type of marks that have flooded registers due to the rise of e-commerce [2] and "the supply of competitively effective trademarks is, in fact, exhaustible" [3].

It can be said that one of the reasons for such a significant difference among intellectual property objects is that trademarks are not subject to such thorough novelty and industrial applicability requirements as inventions and designs [4]. Trademarks are hardly a marker of an enterprise's innovativeness (or demonstrate it in an indirect way [5]). But all this does not mean that the process of brand registration is simple.

COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine

EMAIL: d.shmatkov@gmail.com (D. Shmatkov); oleksii.gorokhovatskyi@gmail.com (O. Gorokhovatskyi); vnn@hneu.net (N. Vnukova)
ORCID: 0000-0003-2952-4070 (D. Shmatkov); 0000-0003-3477-2132 (O. Gorokhovatskyi); 0000-0002-1354-4838 (N. Vnukova)

2. Literature review

Under international treaties and national laws, in order not to confuse the consumer, a word and/or figurative component of one mark should not repeat a word and/or figurative part of another mark already registered in a particular country, if both marks cover identical or similar groups of goods or services. For example, the European Intellectual Property Office points out that "whether a likelihood of confusion exists depends on an assessment of several interdependent factors, including (i) similarity of the goods and services, (ii) the relevant public, (iii) similarity of the signs, taking into account their distinctive and dominant elements and (iv) the distinctiveness of the earlier mark" [6]. There are other circumstantial arguments that may reinforce the above. At the same time, we see that the assessment of similarity/identity of the goods and services and similarity/identity of the signs is integral in that process.

"Comparison of the goods and services must be based on the wording indicated in the respective lists of goods/services" [6]. While this comparison is based on the Nice Classification (International Classification of Goods and Services) in most jurisdictions, "the classification of the goods or services is not conclusive, because similar goods/services may be classified in different classes, whereas dissimilar goods/services may fall within the same class" [6]. At the same time, at the initial stage after registration, the likelihood of confusion must be assessed in the projection of the goods and services for which the mark is registered, and not the goods and services for which it is actually used [7]. Thus, since the purpose of a registered brand is to distinguish a particular goods and services from those of others, the comparison of these parameters is extremely important in this context.

A refusal to register a trademark may be obtained as a result of an examination conducted by an office of a particular country or as a result of a decision made by a special department of such an office based on opposition received from a third party. For example, in 2022, more than 18 thousand decisions on various oppositions were made in the EU [8]. In addition, in various jurisdictions, there are possibilities for the cancellation of a trademark after its registration. If we add here the previously mentioned frequency of registrations in the world (trademarks which may be found to be similar to high degree or identical to a trademark being registered), it becomes quite obvious that the process of registering a trademark is not so simple.

Applicants, and even more so their representatives, are aware of the risks of refusal to register a trademark, therefore, they are trying to reduce these risks by conducting appropriate preliminary searches using various databases and technical means. Scientists suggest, for example, similarity analysis of trademarks spelling, pronunciation, and images using machine learning [9], trademark image search using convolutional neural networks [10–13], artificial intelligence based content, image/pixel, and text similarities search [14], or a neural network model to exploit the semantic, phonetic and visual similarities between two textual trademarks [15]. Of course, such solutions allow applicants to avoid infringement disputes [9], save money for their small and medium enterprises [10], automate the process and increase the accuracy of search and comparison [9, 12, 14]. At the same time, the technical implementation of the mentioned methods probably requires a certain level of user expertise.

Despite the comparison of trademark is the integral process of similarity evaluation for different trademark components, one of the most significant approaches is the partial comparison. It means to quickly narrow the search results for similar trademark just by some of its component (component comparison to form general conclusions is used in disputes in most jurisdictions) and apply the further comparison for the smaller dataset remaining. Additionally, technical approaches to compare graphical, textual and numerical information, that is applicable for trademark comparison are rather different [14].

As justified earlier, one of the most important components to be compared is goods and services description of a trademark that is written according to Nice Classification in most cases. But often, applicants, taking into account the specifics of their business, offer their own formulations of goods and services, which are somewhat different from the formulations presented in the Nice Classification (in fact, therefore, the editions of the Nice Classification are constantly updated). The analysis of that component significantly complements the analysis of word and/or figurative part of a trademark, company history, logistics channels, target audience, etc.

The comparison of two texts or phrases is a problem Natural Language Processing (NLP) deals with. Semantic analysis is mentioned to be important in trademark comparison [14, 16]. Various well-known text comparison and word embedding approaches like count vectorizers, word2vec/doc2vec models, Siamese neural networks [17, 18] were proposed.

The drawback of the majority existing powerful models based on artificial neural networks is the necessity to have dataset to train models on. This is not easy to fulfill this requirement for trademarks subject area as saving/gathering data in search systems is severely limited. For example, the terms of use of the WIPO search tool set limits on the number of requests from a single IP address, automated queries, bulk acquisition, bulk downloading, bulk storing of data, bulk copying, bulk reformatting, bulk sharing and bulk redistributing of data, web scraping etc. [19]; another well-known search tool, TMview [20] prohibits performing any activity that could harm or violate the EUIPO's network performance and/or security, as well as prohibits the extraction of substantial parts of the EUIPO's databases or of the content therein etc.

Our goals in this research are:

- perform comparison of trademarks by only goods and services text fields for the particular class according to Nice Classification (including that it is not clear whether there is a necessity to use semantical similarity or lexical is just enough);
- verify what straightforward approaches are useful to compare trademarks;
- evaluate trademark comparison results numerically that is of interest for applicants/representatives/ stakeholders, and lawyers in order to reduce the likelihood of registration refusal, reduce litigation, arbitration, and negotiation costs, as well as provide researchers and developers with our contribution in studying the problem of automating the process under consideration.

Our contribution includes:

- new intuitive greedy matching numerical vectors algorithm for the comparison of sets of word embeddings which can optionally include difference between size of sets for trademarks comparison application;
- the results of the effectiveness of different approaches to compare sets of word embedding vectors for sentences, in experiments we tried to find such a method that avoids the usage of traditional aggregation of set of word vectors into single one;
- an automated approach that includes a sophisticated assessment of trademark similarity/identity for the next stage of comparison based on goods and services.

3. Text processing workflow

The typical pipeline for NLP text comparison problems starts with data cleaning. This may include removing unnecessary symbols (like HTML tags, emoticons, etc.) punctuation symbols, stop words, digits, and converting to lowercase letters. Sometimes the stemming/lemmatization processes follow that which can normalize the structure of the word, combine the term in different forms into a single, and analyze it properly.

The next important step is tokenization which allows users to split text into smaller parts, e.g., sentences, combinations of words, or separate words. This depends on the quantity of data available and the problem domain.

Feature detection that includes vectorization is the most important step. It is required to convert text data into numerical forms which computers and corresponding models can handle. There are different approaches to convert text to vectors such as one-hot encoding, count vectorizer, N-gram analysis, Term Frequency-Inverse Document Frequency [21, 22] (using just information about the presence of words disregarding their positions and relations is often referred to as “bag of words” approach) and more sophisticated word embedding models which were designed to create similar vectors for the words having close meaning.

Finally, the obtained vectors are compared directly using some distance (a cosine similarity measure is very popular for word vector comparison [23]), resulting in some similarity coefficient. It is worth noting that the similarity value is higher for similar words, sentences, or documents while distance is lower for such cases.

3.1. Data acquisition and cleaning

The gathering of data is often a problem for practical-driven problems and applications. Additionally, the quality of data being used strongly determines both scientific and practical results. Unfortunately, the majority of known online trademark processing utilities [19, 20] don't allow people to export/filter the required data to a sufficient extent to process them offline. For instance, trademark search tool TMview [20] has the option to download information but without the description of goods and services. So, we did some job manually to retrieve that data.

Data cleaning for trademark preprocessing included the removal of entities that do not contain the Nice Classification description for goods and services (empty fields or domestic classes). Also, some trademarks with contradictory data were removed, e.g., information about goods and services for the trademark includes 45 and 42 Nice classes, but actual text descriptions for 23 and 36 classes are provided instead. Additionally, about half of the trademark descriptions were automatically translated into English.

The other important idea to keep in mind is that goods and services descriptions for trademarks may vary significantly, e.g., it is possible that one particular trademark may have 500 words to describe goods and the other may have only 2 words (for example, legal services vs. legal advice, legal consultancy services, legal support services etc.), making the comparison tricky.

3.2. Count vectors (CV)

The first vectorization approach we used to compare the description of goods and services for trademarks was count vectors.

This type of word embedding is very similar to one-hot encoding with the difference that the output vector may contain not only zeros or ones. The value of 1 means the single occurrence of the specific word from dictionary in the text being vectorized, and the position of 1 corresponds to the placement of this word in the vocabulary for the entire corpus. The next occurrence of this word in this sentence increments current count number. All other values for missing words from the vocabulary in this sentence are zeros, so the entire length of this vector is the same as the length of vocabulary. Count vector is very sparse following proper effective storage and comparison implementations.

The main advantage of this approach is its simplicity. The main drawback of this method relates to the fact that the most frequent word is considered to be the most important, so the removal of frequently appearing stop/useless words is the required action here.

3.3. TF-IDF vectors

Term Frequency – Inverse Document Frequency (TF-IDF) is the other statistical method that calculates not only the frequency of the word (term frequency - TF) in the sentence being analyzed but also a sort of importance of the word for the entire corpus (inverse document frequency - IDF) [21, 22, 24]:

$$w_{i,j} = tf_{i,j}\, log_2\left(\frac{N}{df_i}\right), \tag{1}$$

where $w_{i,j}$ is the vector item for word i in document j, $tf_{i,j}$ is the frequency of the term i in document j, N is the quantity of documents if the corpus, df_i is the quantity of documents that contain word i.

Both count vectors and the TF-IDF are statistical measures that operate with unigrams (separate words) and do not take into account the relation between terms and thus cannot catch the semantic meaning of the sentences.

3.4. Word2vec and GloVe

Word2vec and GloVeset of models has been proposed as a way to build high-quality semantic representations of words using huge datasets [25, 26]. This allows us to compare the semantic similarity of the vectors using cosine distance. It was also stressed that linear relationship between words is preserved and multiple degrees of similarity are possible to be present between words. Word2vec utilizes artificial neural networks that typically outperform other word embedding methods last decade (probably, except for more powerful models that appeared recent years).

There are two types of word2vec models, namely Continuous Bag-of-Words (CBOW) and Skip-gram [25]. The CBOW architecture uses the neighborhood around words (defined by windows parameter) as the input to predict the output word. Skip-gram is the opposite architecture that operates with only a single word as input and predicts some words around it, learning relations between words in such a way. The CBOW models typically converge faster and have more information about syntactic similarity, while Ski-gram models require more time and data to train but preserve semantic similarity better.

There are other hyper-parameters to think about during training except of architecture: quantity of iterations over all training sentences, the minimal quantity of word occurrences in the dictionary to be identified as valuable terms, size of the vector representation, train method.

It is worth noting, that word2vec was designed to work effectively on huge datasets, while we are limited both in data quantity and text diversity here. Training the word2vec model requires setting up some hyperparameters which is not a trivial task for the approaches based on neural networks [27].

GloVe method [28] extends and continues the idea presented by the word2vec model by counting contextual information not only from the local neighborhood of words but from the entire (global) corpus. It was mentioned that GloVe word embedding outperforms other models in word similarity and word analogy tasks, which are the problems we are interested in here.

3.5. Distances and comparison

Comparison of word embedding vectors is often performed with the cosine similarity that is normalized dot product between vectors X and Y:

$$d_{cos} = \frac{XY}{\|X\|\|Y\|}. \tag{2}$$

Cosine similarity ranges from -1 to 1 but when vectors contain only positive elements (typical for text representation vectors which are frequencies) it varies from 0 (vectors are different) to 1 (vectors are similar).

Comparison of text in the form of sentences requires the construction of some joint vector for the entire sentence having an embedding vector for each separate word in it. This is not straightforward in common, but often just the averaging of all word vectors is used [29, 30] and works pretty well. More complex approaches include building doc2Vec [31] models instead of word2vec and use information not only about words but also about sentences.

In this work we investigated other simple similarity measures to avoid averaging of embeddings for words in sentence as well as usage of more complicated methods like doc2vec.

We tried a sort of greedy matching vectors (Fig. 1). It utilizes cosine similarity to compare pairs of word vectors. A feature of the algorithm is the requirement to find the closest vector. Despite the drawback that the closest vector may have very low similarity, e.g., two very different words still could be matched, we assign this matching as successful anyway.

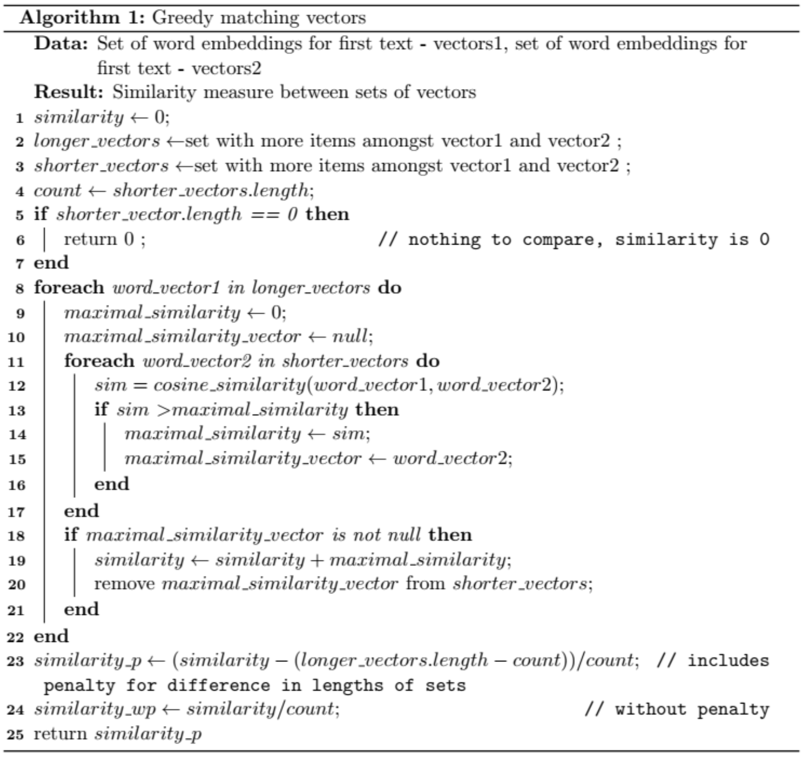
The calculation of two similarity measures is possible here. Line 24 in Fig. 1 calculates the normalized value between 0 and 1 but equals 1 if one set of vectors is a subset of the other (including the extreme case when just one word is compared to the sentence containing this word). The penalized similarity calculated in Line 23 addresses the issue with different lengths of lines but is normalized from one side only, reaching 1 when two sets fit perfectly. During experiments, we looked at both these options.

The other metric we used was Tanimoto similarity between sets of vectors [32, 33] according to:

$$S_T = \frac{|X \cup Y|}{|only\, X| + |only\, Y| + |X \cup Y|}, \tag{3}$$

where X is the first set of vectors (word embedding for the first text), Y is word embeddings set for the second text, $|X \cup Y|$ is the number of word embeddings in both sets, $|only\, X|$, $|only\, Y|$ are the number of vectors only in X and Y respectively. Tanimoto similarity value varies between 0 and 1, for this case

10% cosine distance similarity deviation between separate word embeddings is permitted to consider vectors to be matched (in other words, vectors are same when cosine similarity measure is greater than 0.9). This differs from the algorithm described above when vectors with minimal cosine distances are matched without any conditions.



```
Algorithm 1: Greedy matching vectors
  Data: Set of word embeddings for first text - vectors1, set of word embeddings for
        first text - vectors2
  Result: Similarity measure between sets of vectors
  $similarity \leftarrow 0$;
  $longer\_vectors$ ←set with more items amongst vector1 and vector2 ;
  $shorter\_vectors$ ←set with more items amongst vector1 and vector2 ;
  $count \leftarrow shorter\_vectors.length$;
  if $shorter\_vector.length == 0$ then
      return 0 ;                              // nothing to compare, similarity is 0
  end
  foreach $word\_vector1$ in $longer\_vectors$ do
      $maximal\_similarity \leftarrow 0$;
      $maximal\_similarity\_vector \leftarrow null$;
      foreach $word\_vector2$ in $shorter\_vectors$ do
          $sim = cosine\_similarity(word\_vector1, word\_vector2)$;
          if $sim > maximal\_similarity$ then
              $maximal\_similarity \leftarrow sim$;
              $maximal\_similarity\_vector \leftarrow word\_vector2$;
          end
      end
      if $maximal\_similarity\_vector$ is not null then
          $similarity \leftarrow similarity + maximal\_similarity$;
          remove $maximal\_similarity\_vector$ from $shorter\_vectors$;
      end
  end
  $similarity\_p \leftarrow (similarity - (longer\_vectors.length - count))/count$;  // includes
   penalty for difference in lengths of sets
  $similarity\_wp \leftarrow similarity/count$;                    // without penalty
  return $similarity\_p$
```

Figure 1: Greedy matching vectors algorithm

We also tested Hausdorff distance between sets of word embedding as it is one of the most popular distances to compare numerical sets of vectors. The main drawback is that outliers in vector sets can significantly influence the result.

4. Experimental modeling

Our dataset included information (id, goods and services in the form of class numbers and text description) about 183 filed, registered, and expired in the EU and Ukraine trademarks of 45 Nice class (common description is: "Legal services; security services for the physical protection of tangible property and individuals; dating services, online social networking services; funerary services; babysitting"). The choice of class 45 is determined by the opportunity of using relevant expert opinions within the methodology of this study. The gathered corpus included 975 sentences and 686 unique words (without stop ones).

We used some pretrained word2vec, GloVe and doc2vec models and word2vec and doc2vec models trained on our corpus. We compared trademark descriptions using different approaches and metrics described above to understand which one fits our goals better.

Gensim [29, 30] software was used for experimental modeling as well as pretrained models and examples delivered by it.

The entire list of models, methods and measures includes the following:

1. Cosine similarity for word embedding obtained after count vectorizer.
2. Cosine similarity for words embedding obtained after TF-IDF vectorizer.
3. Cosine similarity between average vectors for word2vec pretrained on "text8" dataset gathered from Wikipedia [34].

4. Hausdorff distance, Tanimoto similarity, and our greedy matching similarity (Fig. 1) between sets of separate word embeddings for trademark goods descriptions sentences obtained by word2vec model pretrained on “text8”.
5. Cosine similarity between average vectors for GloVe “glove-wiki-gigaword-50” model pretrained on Wikipedia texts [34].
6. Tanimoto similarity and our greedy matching similarity (Fig. 1) between sets of separate word embeddings for sentences obtained by GloVe “glove-wiki-gigaword-50” pretrained model.
7. Cosine similarity between average vectors, Tanimoto similarity and our greedy matching similarity (Fig. 1) between sets of separate word embeddings for sentences obtained by word2vec model trained on our corpus (using default CBOW method).
8. Cosine similarity between average vectors, cosine similarity between inferenced by doc2vec model vectors for the texts, Tanimoto similarity and our greedy matching similarity (Fig. 1) between sets of separate word embeddings for sentences obtained by doc2vec model trained on our corpus.

We tested the initial quality for pretrained models on the WS-353 dataset [35, 36]. It contains two sets of English word pairs with corresponding similarity values provided by humans. The typical evaluation measures are Pearson and Spearman rank correlation coefficients between these similarities and ones provided by model.

Pearson and Spearman rank correlation coefficients for “text8” word2vec model are 0.61 and 0.62 respectively, “glove-wiki-gigaword-50” GloVe model got 0.51 and 0.50 scores, so first “text8” model performs better, but there are no guarantees that “text8” will work better for our problem though.

We chose a random trademark that has the following description for 45 Nice class to investigate how different similarity measures perform for the particular trademark search problem:

“Legal services; technical and legal research services; information services relating to legal matters; issuing of legal information; legal advice; legal consultancy services; legal enquiry services; legal services relating to business; legal support services; investigation services; preparation of reports; patent and trade mark agency services; company formation and registration services; provision of information, consultancy and advisory services relating to the aforesaid services.”

We decided to compare the top 5 search results for the trademark listed above and grouped the results for comparable methods. The entire list of trademarks that appeared for different similarity measures is shown in Table 4. The table contains the ID of a trademark (we refer to these ID in tables) and its text description of 45 Nice class to make presentation of results more reliable.

Table 1 contains results of trademarks search by example using the simplest count vectorizer and the TF-IDF vectorizer and the results of average word embedding vectors comparison for different models. For this experiment one can see that results for CV and the TF-IDF are the same expect of amplitudes of the similarity value, which are lower for the TF-IDF. For example, in the top 10 results first nine results for CV are greater than 0.75, while first nine for the TF-IDF are greater than 0.5, and only trademark at the tenth position is different for both these methods.

Results of comparison of trademarks by cosine similarity between average word embeddings are close in first positions. All four models placed trademark #1 in first position, and three models placed trademark #2 in second place. Descriptions of all models #21 – #25 are very short and non-informative, so first model based on “text8” was able to find one long text more than others. This is also very representative for first two rows in Table 1 where CV and the TF-IDF returned only one trademark with long description and four with description containing two words.

Both word2vec and doc2vec results depend on the quality of the dataset and other parameters [37, 38] like the number of training epochs, learning rate, etc. We tested training during 5 and 15 epochs for the word2vec model and found less quantity to be more successful as it produces more suitable results. Increasing the quantity epochs leads to high similarity values for the first trademarks found making it possible to distinguish them by the fifth-sixth number after the decimal point or even not possible at all.

We trained doc2vec during 50 epochs because we were not satisfied with the default 10 epochs in terms of quality for the comparison of inferenced vectors.

The next metric we used to compare sets of word embedding vectors for sentences was Tanimoto similarity, which is often applied to compare sets of numbers (Table 2). There were no trademarks with short description of 45 Nice class here. Results are similar for first two “text8” and “glove-wiki-gigaword-50” pretrained word2vec models and model trained on our corpus. All these three models found trademarks #1, #3, #5, #10 in top 5 of the most similar and all of them placed trademark #1 in first position. The last doc2vec model trained on our dataset showed different results, and some trademarks it found do not appear in any list for other similarity methods.

Table 1

Top5 trademarks found and their similarity scores for CV, TF-IDF vectorizers and for cosine similarity average word embedding vectors for all tested models

Method or model	Top 5 found positions (first is the most similar)				
	1	2	3	4	5
Count vectorizer	1 (0.9475)	21 (0.9071)	22 (0.9071)	23 (0.9071)	24 (0.9071)
TF-IDF vectorizer	1 (0.7750)	21 (0.6577)	22 (0.6577)	23 (0.6577)	24 (0.6577)
Cosine similarity between average word embedding vectors					
pretrained “text8” word2vec	1 (0.9930)	2 (0.9787)	3 (0.9724)	21 (0.9489)	22 (0.9489)
pretrained “glove-wiki-gigaword-50” word2vec	1 (0.9943)	3 (0.9853)	21 (0.9828)	22 (0.9828)	23 (0.9828)
word2vec (our corpus)	1 (0.9999)	2 (0.9999)	21 (0.9999)	23 (0.9999)	24 (0.9999)
our doc2vec (our corpus)	1 (0.9999)	2 (0.9997)	26 (0.9988)	3 (0.9987)	21 (0.9996)

The results obtained by our greedy matching algorithm (Fig. 1) with penalized difference results are shown in the middle of Table 2. Its non-penalized version allows users to find only short occurrences of text trademark descriptions in the initial one and returns trademarks #21 – #25 in all cases. As one can see from Table 2, penalized similarity finds trademarks somewhat close to the results of other methods, but the results for last doc2vec are close to the ones found by this model with Tanimoto similarity above.

The last two experimental similarity values based on Hausdorff distance and inferenced with doc2vec model vectors are again differ compared to other results.

As a result of our qualitative analysis, the hierarchy of trademarks (according to their relevance to the chosen trademark) was correctly arranged using the following methods and models:

- comparison of word embedding sets with greedy matching algorithm for doc2vec and “text8” models;
- Tanimoto similarity between sets of word embeddings using “text8” model;
- cosine similarity between average word embedding vectors using word2vec pretrained on our corpus and count vectorizer.

It is important to note here that the correct identification of the hierarchy does not mean that the most relevant trademarks were selected using the methods. Some methods showed a debatable hierarchy, but in the aggregate, they revealed the top 5 trademarks more consistent (competitive) with the chosen trademark.

Imposing on the analysis an understanding of the specifics of the service, its possible linguistic interpretations for a specific application, and thus identifying the niche (and possibly the target audience) accordingly, we concluded that the most relevant trademarks were identified using the following methods:

- comparison of word embedding sets with greedy matching algorithm based on “text8”;
- cosine similarity between inferenced vectors for our doc2vec model.

The following methods and models showed the least accuracy in the considered context:

- Tanimoto similarity between sets of word embeddings for doc2vec models;
- greedy matching comparison based on word2vec trained on our corpus;
- greedy matching algorithm for doc2vec trained on our corpus.

Table 2
Top 5 trademarks found and their similarity scores for Hausdorff, Tanimoto distances and custom greedy matching algorithm

Method or model	Top 5 found positions (first is the most similar)				
	1	2	3	4	5
Tanimoto similarity between sets of word embeddings					
pretrained “text8” word2vec	1 (0.5846)	9 (0.4493)	3 (0.4429)	5 (0.3704)	10 (0.3690)
pretrained “glove-wiki-gigaword-50” word2vec	1 (0.5846)	3 (0.4638)	9 (0.4493)	5 (0.3659)	10 (0.3647)
word2vec (our corpus)	1 (0.7451)	10 (0.6078)	11 (0.5179)	5 (0.4918)	3 (0.4545)
doc2vec (our corpus)	10 (1)	16 (1)	17 (1)	18 (1)	19 (1)
Comparison of word embedding sets with greedy matching algorithm (Fig. 1)					
pretrained “text8” word2vec	1 (0.6937)	3 (0.6051)	9 (0.5786)	6 (0.5224)	13 (0.4813)
pretrained “glove-wiki-gigaword-50” word2vec	1 (0.7841)	3 (0.7183)	9 (0.6751)	6 (0.6273)	30 (0.6081)
word2vec (our corpus)	1 (0.7276)	15 (0.6965)	10 (0.6863)	11 (0.6443)	16 (0.6273)
doc2vec (our corpus)	10 (0.9780)	16 (0.9423)	17 (0.9423)	18 (0.9423)	19 (0.9423)
Hausdorff distance between sets of word embeddings					
pretrained “text8” word2vec	4 (14.2426)	5 (14.2827)	6 (14.3935)	7 (14.3935)	8 (14.3935)
Cosine similarity between inferenced vectors					
doc2vec (our corpus)	1 (0.9752)	2 (0.8395)	27 (0.8110)	28 (0.7965)	29 (0.7847)

The trademark chosen covered legal services in general, business legal services, information services and research, and intellectual property services. Therefore, trademarks that, for example, covered political advice, non-business legal services, and other very specific, irrelevant legal services, do not pose such a threat of the trademark dilution (and vice versa).

We analyzed also the performance of these metrics to understand their use cases better. We performed the entire set of experiments three times and provide the average time values in seconds, no outliers during modeling were observed. We averaged also time measurements inside each method in the scope of the experiment. The results are presented in Table 3.

Table 3
Performance for all metrics and methods.

Method and description	Average time, sec.
Count vectorizer (creation of CountVectorizer with default Sklearn parameters, calculated on our corpus, cosine similarity calculation time for vectors is included, gathering of corpus is not)	0.0066
TF-IDF (creation of TfidfVectorizer with default Sklearn parameters calculated on our corpus, cosine similarity calculation time for vectors is included, gathering of corpus is not)	0.0072
“Text8” loading and word2vec building on it (default Gensim parameters were used, downloading of “text8” corpus is not included)	44.5

GloVe “glove-wiki-gigaword-50” loading and word2vec building on it (default Gensim parameters were used, downloading of “glove-wiki-gigaword-50” corpus is not included)	13
Training of our word2vec (tokenization of sentences and training of word2vec with default settings)	0.036
Training of our doc2vec model (tokenization of sentences and training of doc2vec with default settings)	1.1
Cosine similarity between average vectors	
Pretrained word2vec “text8”	0.00033
Pretrained word2vec “glove-wiki-gigaword-50” word2vec	0.00033
Word2vec (our corpus)	0.00026
Doc2vec (our corpus)	0.00029
Cosine similarity between inferenced embeddings for our doc2vec model	0.0038
Tanimoto similarity between sets of word embeddings	
Pretrained word2vec “text8”	0.11
Pretrained word2vec “glove-wiki-gigaword-50” word2vec	0.115
Word2vec (our corpus)	0.066
Doc2vec (our corpus)	0.0075
Greedy matching algorithm (Fig. 1)	
Pretrained word2vec “text8”	0.05
Pretrained word2vec “glove-wiki-gigaword-50” word2vec	0.61
Word2vec (our corpus)	0.035
Doc2vec (our corpus)	0.035
Hausdorff similarity between sets of word embeddings	
Pretrained word2vec “text8”	0.0021

As one can see, cosine similarity based on average vectors is the fastest evaluation measure amongst all we tested. But pretrained models that typically allow users to get better results require time to build them, approximately 45 sec. to “text8” and 13 sec. for “glove-wiki-gigaword-50” respectively. Training of own word2vec and doc2vec is much faster but not always feasible in terms of practical applications. Calculation of Tanimoto similarity and greedy matching distance is typically done faster for models training on own data, probably because of short vocabulary compared to pretrained models.

5. Conclusions

Comparison and distinguishing of similar or identical trademarks based on their goods and services description is a vital task at various stages of a brands' life. This problem is even more acute in practical applications when plenty of trademarks should be compared. Additionally, comparison of text items is initially subjective and depends on a person who makes the decision.

The paper describes the research of models and word embeddings matching methods suitable for the comparison of trademarks by goods and services, such models can also be applied similar trademarks search having the initial goods and services text value.

Different well-known models (word2vec, doc2vec) and word vectorization/embeddings comparison methods (count vectorizer, the TF-IDF, cosine similarities, Tanimoto similarity, Hausdorff similarity) have been implemented and tested. We propose also our own simple algorithm based on greedy matching of sets of word embedding vectors that proves its effectiveness during experimental modeling.

It has been shown that a lot of combinations of models and vectors matching can provide correct hierarchy in trademark search problem but two approaches showed good accuracy in terms of relevance: comparison of word embedding sets with the proposed greedy matching algorithm based on pretrained woc2vec “text8” model and cosine similarity between inferenced vectors for doc2vec model, trained on trademark services corpus formed by us. Performance experiments showed that training word2vec and doc2vec models as well as using matching of vectors with the proposed greedy algorithm and known cosine similarity could be implemented effectively enough to make the decision regarding similar trademarks within seconds. Solving this problem with person expertise typically takes hours.

Further work may be related to the usage and research into more sophisticated methods like FastText. BERT for measuring text similarity, increasing of dataset. It seems interesting to extend the proposed ideas to more Nice classes and compare trademarks deeper as well as to expand understanding of how the compliance and non-compliance of the description of goods and services with the Nice Classification of a particular trademark can affect the comparison. In addition, based on the legal practice of trademark disputes, we see a clear need to develop a scale for such a comparison.

By conducting a trademark search, applicants and their representatives using well-known services mentioned in this study can select relevant trademarks using a filter of goods and services, but they cannot further analyze the sample for risks associated with identified marks that may cause consumer confusion. The present study makes a certain contribution to filling this gap and adds new knowledge in the development of the legal tech field.

6. Acknowledgments

We dedicate this work and thank the Armed Forces of Ukraine, Air Forces and anti-aircraft missile operators, all Ukrainian Forces, and divisions, doctors, volunteers, everyone who stands with Ukraine. Our special thanks go to brave energy personnel who works hard to repair and maintain electrical stations and nets after barbarian shelling of Ukraine.

A deep bow to all defenders of Ukraine.

7. List of trademarks and their 45 Nice class description

Table 4
List of encoded trademarks and their 45 Nice class description appeared in experiments.

ID	Nice class 45
1	Legal services ; legal research services; provision of information relating to legal matters; delivery of legal information; provision of legal advice; legal consultancy services; legal information services; business legal services; legal assistance services; legal reporting; agency services in the field of patents and trademarks (legal services); company incorporation and registration services (legal services); as well as providing information, consultancy services and the provision of advice relating to the aforesaid services.
2	Legal assistance for the drafting of contracts; software licensing [legal services]; security consultancy; provision of legal expertise; provision of information on legal matters; licensing of franchise concepts [legal services]; licensing of computer programs [legal services]; provision of legal advice; provision of legal advice relating to franchising; provision of legal advice, information and consultancy services; legal assistance services; consultancy services relating to occupational safety rules; work safety consultancy services; legal consultancy services relating to franchising; services of professional legal advisers relating to franchising; legal information services; legal services ; provision of distinctive signs; services provided by a franchisor or a company offering a partnership, namely transfer (provision) of legal know-how in the field of temporary work, placement and recruitment.
3	Advisory services relating to regulatory affairs; information services relating to regulatory affairs; compilation of regulatory information; safety evaluation; preparation of Regulations; personal background investigations; advisory and information services relating to standards; disciplinary services for members of a professional organisation; advisory and consulting services relating to all the aforesaid services; dispute resolution services; legal services, namely the provision of expert evidence in legal proceedings; arbitration services; mediation services.

4 Legal advice service; information and advice on regulations in the real estate field; legal information and advice in the field of finance and taxation; security consulting services; building security information.

5 Legal services ; notary services; legal advice; legal advice in the field of real estate; legal document and contract drafting services; certification of legal documents; property transfer legal services; drafting of deeds and contracts in the real estate field; consultation and assistance in real estate litigation; testamentary execution; legal services relating to wills; legal advice in the field of taxation; mediation services in legal proceedings relating to real estate; provision of information, expertise and legal and tax advice in the field of real estate; legal research.

6 Legal services; security services for the protection of property and individuals; personal and social services rendered by others to meet the needs of individuals; legal research; legal advice; security consultancy; monitoring of intruder alarms; close protection (escort); alternative dispute resolution services (legal services); litigation services; mediation; legal mediation services; search for missing persons; background investigations of individuals; safety inspection of factories, houses and apartments; rental of safes.

7 Security services for the protection of goods and individuals (with the exception of their transport), advice on anti-theft devices, consultation in the anti-theft field for the sale of retail products at points of sale.

8 Legal services relating to the creation and establishment of companies, legal advice services relating to the creation of companies, formation of companies, management of companies, restructuring, transmission, merger of companies, all legal and judicial services rendered to individuals and businesses.

9 Legal services ; arbitration services; intellectual property advice; registration of domain names; forensic research. Software licensing [legal services]; licensing of intellectual property rights; mediation; legal monitoring services relating to intellectual property. Personal assistance services, namely assistance in carrying out administrative procedures other than for the conduct of business. Supply (rental) of interactive databases allowing access to administrative documents. Certification (quality and origin control) information.

10 Legal services; security services for the protection of property and individuals; accompaniment in society (companions); accompaniment in society (companions), mediation; advisory services relating to national security; provision of information on political matters; promoting the interests of international companies, real estate companies or non-profit companies in the fields of politics, law and regulation (lobbying services); registration of documents containing publicly available administrative data; inspection of factories relating to safety; review of standards and practices to comply with anti-corruption laws and regulations; review of standards and practices to verify their compliance with laws and regulations; legal assistance services.

11 Legal services; consultancy services relating to intellectual property rights, including patents, trade marks and designs; copyright management services; intellectual property licensing services; professional consultancy services relating to legal services; legal services with regard to intellectual property services; management of intellectual property rights as well as searches with regard thereto, also taking into account issues with regard to the establishment, maintenance, enforcement and exploitation of patents, trade marks and other such rights; legal information services; lobbying services other than for commercial purposes, with regard to political issues.

12 Marriage agencies; detective agencies; night guard services; adoption agency services; arbitration services; rental of safes; embalming services; funerary undertaking; lost property return; genealogical research; legal research; security consultancy; intellectual

property consultancy; monitoring intellectual property rights for legal advisory purposes; monitoring of burglar and security alarms; crematorium services; licensing of intellectual property; licensing of computer software [legal services]; organization of religious meetings; opening of security locks; guard services; planning and arranging of wedding ceremonies; missing person investigations; litigation services; babysitting; pet sitting; baggage inspection for security purposes; inspection of factories for safety purposes; evening dress rental; rental of fire extinguishers; clothing rental; rental of fire alarms; registration of domain names [legal services]; personal background investigations; fire-fighting; escorting in society [chaperoning]; horoscope casting; copyright management; dating services; alternative dispute resolution services; funerals; house sitting; mediation; personal body guarding.

13 Legal services ; mediation; security services for the protection of property and individuals; marriage agency services; celebration of religious ceremonies; establishment of horoscopes; funeral services; cremation services; night watch agency services; monitoring of intruder alarms; physical security consultancy services; opening of locks; clothing rental; detective agency services; legal research; intellectual property advice; rental of domain names on the Internet; online social networking services; home childcare.

14 Security services for the protection of property and individuals; operation of an alarm and intervention center (security service); monitoring to protect against hazards such as fire, water, burglary, overheating and power failure; civil protection; babysitting services; security escort for persons, goods and valuables; surveillance and guarding services for persons, goods, valuables, buildings and installations; control and security of apartments and houses in case of absence; stewardship at demonstrations; execution of public tasks in place of the public community, namely local police services; traffic control services; distress response services; processing of alarms and alerts; rapid response services in connection with a rapid response center to avoid risks and repair damage; opening of locks; consultancy in the field of security, in particular in connection with installations and equipment for security technology, warning and fire-fighting; rental of security installations and apparatus (except computers); lost property collection services identifiable by means of identification or control tokens; execution of warrants by or on telephone orders in connection with security services.

15 Social networking services, networking and dating services (dating clubs; provision of social services namely social networking services in the field of personal development, namely self-improvement, self-fulfilment, charitable activities , philanthropic, voluntary, public and community service activities, and humanitarian activities, (online social networking) Information about social networking services in the field of personal development, namely self-improvement, self-fulfillment , charitable, philanthropic, voluntary, public and community service activities, and humanitarian activities.

16 Legal services ; mediation; security service for the protection of property and individuals; marriage agencies; establishment of horoscopes; undertakers; cremation services; night surveillance agencies; monitoring of intruder alarms; security consultancy; opening of locks; clothing rental; detective agencies; legal research; intellectual property advice; online social networking services; home childcare.

17 Legal services ; mediation; security service for the protection of property and individuals; marriage agencies; establishment of horoscopes; undertakers; cremation services; night surveillance agencies; monitoring of intruder alarms; security consultancy; opening of locks; clothing rental; detective agencies; legal research; intellectual property advice; online social networking services; home childcare.

18 Legal services ; mediation; security service for the protection of property and individuals; marriage agencies; establishment of horoscopes; undertakers; cremation services; night

	surveillance agencies; monitoring of intruder alarms; security consultancy; opening of locks; clothing rental; detective agencies; legal research; intellectual property advice; online social networking services; home childcare.
19	Legal services ; mediation; security service for the protection of property and individuals; marriage agencies; establishment of horoscopes; undertakers; cremation services; night surveillance agencies; monitoring of intruder alarms; security consultancy; opening of locks; clothing rental; detective agencies; legal research; intellectual property advice; online social networking services; home childcare.
20	Legal services ; mediation; security service for the protection of property and individuals; marriage agencies; establishment of horoscopes; undertakers; cremation services; night surveillance agencies; monitoring of intruder alarms; security consultancy; opening of locks; clothing rental; detective agencies; legal research; intellectual property advice; online social networking services; home childcare.
21	Legal services.
22	Legal services .
23	Legal services.
24	Legal services.
25	Legal services.
26	Legal services ; legal research.
27	Intellectual property consultancy; intellectual property watching services; licensing of intellectual property; copyright management; legal services.
28	Legal services (advice and litigation) in matters of public law and private law; Litigation Services; Extrajudicial dispute resolution services; Legal research; Forensic research; Software licensing (legal services); Intellectual Property Licensing.
29	Legal services (advice and litigation) in matters of public law and private law; Litigation Services; Extrajudicial dispute resolution services; Legal research; Forensic research; Software licensing (legal services); Intellectual Property Licensing.
30	Coordination of fire safety systems, namely security services for the protection of property and individuals. Audits in the field of fire safety, namely audits in the field of security for the protection of property and individuals (except their transport); Consultations in the field of occupational safety; Advice on occupational safety regulations; Information on health and safety at work; Worker health and safety coordination service for building or civil engineering sites where several self-employed workers or companies are called upon to intervene.

8. References

[1] WIPO, WIPO IP Facts and Figures, 2022. URL: https://www.wipo.int/edocs/pubdocs/en/wipo-pub-943-2022-en-wipo-ip-facts-and-figures-2022.pdf.

[2] G. G. McLaughlin, Fanciful Failures: Keeping Nonsense Marks off the Trademark Register, HARV. L. REV, volume 134, 2021, 1804–1825.

[3] B. Beebe, and J. C. Fromer, Are We Running Out of Trademarks: An Empirical Study of Trademark Depletion and Congestion, Harv. L. Rev., volume 131, number 4, 2017, 945–1045.

[4] D. Shmatkov, Intellectual Property Management of Industrial Software Products: The Case of Triol Corp, in 2021 IEEE 8th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T), IEEE, 2021, pp. 108-112. doi: 10.1109/PICST54195.2021.9772237.

[5] M. J. Flikkema, A. P. De Man, and C. Castaldi, Are trademark counts a valid indicator of innovation? Results of an in-depth study of new benelux trademarks filed by SMEs, in: Carolina Castaldi, Jörn Block, Meindert J. Flikkema (Eds.) Trademarks and Their Role in Innovation, Entrepreneurship and Industrial Organization, in

Trademarks and Their Role in Innovation, Entrepreneurship and Industrial Organization, Routledge, London, 2021, pp. 39–60.

[6] EUIPO, Trade mark and Design guidelines, 2022. URL: https://guidelines.euipo.europa.eu/1803468/1785524/trade-mark-guidelines/chapter-1-general-principles.

[7] K. Cederlund and N. Malovic, Likelihood of confusion must be assessed in light of the goods and services for which a mark is registered, Journal of Intellectual Property Law & Practice, volume 12, issue 4, 2017, pp. 269–270. doi: 10.1093/jiplp/jpx033.

[8] EUIPO, eSearch Case Law, 2022. URL: https://euipo.europa.eu/eSearchCLW/#basic/*/01%2F01%2F2022/31%2F12%2F2022/number/.

[9] C. V. Trappey, A. J. Trappey, and S. C. C. Lin, Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies, Advanced Engineering Informatics, volume 45, 2020, 101120. doi: 10.1016/j.aei.2020.101120.

[10] J. W. Yoon, S. J. Lee, C. Y. Song, Y. S. Kim, M. Y. Jung, and S. I. Jeong, A Study on Similar Trademark Search Model Using Convolutional Neural Networks, Management & Information Systems Review, volume 38, number 3, 2019, 55–80. doi: 10.29214/damis.2019.38.3.004.

[11] T. Lan, X. Feng, Z. Xia, S. Pan, and J. Peng, Similar trademark image retrieval integrating LBP and convolutional neural network, in Image and Graphics: 9th International Conference, ICIG 2017, Springer International Publishing, Shanghai, China, Revised Selected Papers, Part III 9, 2017, pp. 231–242. doi: 10.1007/978-3-319-71598-8_21.

[12] C. A. Perez, P. A. Estévez, F. J. Galdames, D. A. Schulz, J. P. Perez, D. Bastías, and D. R. Vilar, Trademark image retrieval using a combination of deep convolutional neural networks, in 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–7, doi: 10.1109/IJCNN.2018.8489045.

[13] G. Showkatramani, S. Nareddi, C. Doninger, G. Gabel, and A. Krishna, Trademark image similarity search, in HCI International 2018–Posters' Extended Abstracts: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, Proceedings, Part I 20, Springer International Publishing, 2018, pp. 199–205. doi: 10.1007/978-3-319-92270-6_27.

[14] I. Mosseri, M. Rusanovsky, G. Oren, TradeMarker - Artificial Intelligence Based Trademarks Similarity Search Engine, 2019, in: C. Stephanidis (Eds.), HCI International 2019 – Posters, volume 1034 of Communications in Computer and Information Science, Springer, Cham. doi:10.1007/978-3-030-23525-3_13.

[15] Y. Liu, Q. Li, C. Sun, and L. Si, Similar Trademark Detection via Semantic, Phonetic and Visual Similarity Information, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2025–2030. doi: 10.1145/3404835.3463038.

[16] F. Anuar, R. Setchi, Y. Lai, A Conceptual Model of Trademark Retrieval based on Conceptual Similarity, in: Proceedings of the 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013, 2013, pp. 450–459.

[17] N. Poksappaiboon, N. Sundarabhogin, N. Tungruethaipak, S. Prom-on, Detecting Text Semantic Similarity by Siamese Neural Networks with MaLSTM in Thai Language, in: 2nd International Conference on Big Data Analytics and Practices (IBDAP), 2021, pp. 7–11. doi: 10.1109/IBDAP52511.2021.9552077.

[18] T. Ranasinghe, C. Orasan, R. Mitkov, Semantic Textual Similarity with Siamese Neural Networks, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), 2019, pp. 1004–1011. doi: 10.26615/978-954-452-056-4_116.

[19] Global Brand Database, 2022. URL: https://branddb.wipo.int/en.

[20] Tmview, 2022. URL: https://www.tmdn.org/tmview/welcome#/tmview.

[21] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology behind Search, Addison Wesley, 2011.

[22] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK, 2008. doi:10.1017/CBO9780511809071.

[23] T.Brück, M. Pouly, Text Similarity Estimation Based on Word Embeddings and Matrix Norms for Targeted Marketing, in: Proceedings of the 2019 Conference of the North American Chapter of

the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, Vol. 1, Minneapolis, USA, pp. 1827–1836. doi: 10.18653/v1/n19-118.

[24] J. Leskovec, A. Rajaraman, J. Ullman, Mining of Massive Datasets (3rd ed.). Cambridge University Press, Cambridge, UK, 2020. doi:10.1017/9781108684163.

[25] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, 2013, in: J. Bengio, Y. LeCun (Eds.), International Conference on Learning Representations (ICLR). doi: 10.48550/ARXIV.1301.3781.

[26] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed Representations of Words and Phrases and their Compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems, volume 2 (NIPS'13), Curran Associates Inc., Red Hook, NY, USA, pp. 3111–3119.

[27] T. Adewumi, F. Liwicki, M. Liwicki, Word2Vec: Optimal hyperparameters and their impact on natural language processing downstream tasks, Open Computer Science 12(1) (2022) 134–141. doi: 10.1515/comp-2022-0236.

[28] J. Pennington, R. Socher, C. Manning, GloVe: Global Vectors for Word Representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[29] R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks, 2010, pp. 46–50.

[30] Gensim – Topic Modelling in Python, 2022. URL: https://github.com/RaRe-Technologies/gensim

[31] Q. Le, T. Mikolov, Distributed Representations of Sentences and Documents, in: Proceedings of the 31st International Conference on Machine Learning, 2014, pp. 1188–1196.

[32] T. Tanimoto, An elementary mathematical theory of classification and prediction, IBM Internal Report, 1958.

[33] Similarity Measures, 2022. URL: https://docs.eyesopen.com/toolkits/cpp/graphsimtk/measure.html.

[34] What is Gensim-data for? 2018, URL: https://github.com/RaRe-Technologies/gensim-data.

[35] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. “Placing Search in Context: The Concept Revisited”, ACM Transactions on Information Systems, 20.1 (2002): 116-131.

[36] The WordSimilarity-353 Test Collection, 2002. URL: https://gabrilovich.com/resources/data/wordsim353/wordsim353.html.

[37] B. Chiu, G.K. Chrichton, A. Korhonen, S. Pyysalo, How to Train good Word Embeddings for Biomedical NLP, in: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, 2016, pp. 166–174.

[38] L. Savytska, N. Vnukova, I. Bezugla, V. Pyvovarov, M. Sübay, Using Word2vec Technique to Determine Semantic and Morphologic Similarity in Embedded Words of the Ukrainian Language, in: Proceefings of the International Conference on Computational Linguistics and Intelligent Systems, 2021, pp. 235–248.