UDC 004.855.5
DOI https://doi.org/10.32838/2663-5941/2021.4/26

Shaptala R.V.
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Kyselov G.D.
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

VECTOR SPACE MODELS OF KYIV CITY PETITIONS

In this study, we explore and compare two ways of vector space model for Kyiv city petitions creation. In order to automatically analyze freeform texts such as petitions, they need to be converted to a numeric space. By leveraging word vectors based on the distributional hypothesis, namely Word2Vec and FastText, we construct vector models of Kyiv city petitions.

The overall pipeline that we contribute is training word vectors on the dataset of Kyiv city petitions, preprocessing the documents, and applying averaging to create petition vectors. Moreover, this pipeline does not require big data and is applicable to training in a low-resource setting such as the Ukrainian language for which we have only used 4623 unlabeled petitions. No pretrained models and fine-tuning was done for the sake of this research and we provide hyperparameters that were optimal for the experiments.

The advantages and disadvantages of both models are analyzed. Word2Vec-based model gets a higher Silhouette Coefficient score and produces more dense clusters than FastText-based one. This makes it more appropriate for real world applications such as petitions sentiment analysis or clustering. Error analysis confirms this result as FastText pays more attention to the syntactic structure of petitions and words while Word2Vec focuses more on the contexts. To support this claim, we show examples of such behavior for the same textual queries on different urban topics.

Visualizations of the vector spaces after dimensionality reduction via UMAP are demonstrated in an attempt to show their overall structure. They reinforce the resulting Silhouette Coefficient scores by exhibiting denser clusters for the Word2Vec based approach. The resulting models can be used to effectively query semantically related petitions as well as search for clusters of related petitions.

Key words: *vector space model, FastText, Word2Vec, petitions analysis, UMAP.*

Introduction. By now, e-petitions have already matured and are incorporated in a lot of countries' governments. They are no longer experimental and citizens use them actively to make suggestions for public institutions. This is why the analysis of e-petitioning is vital to better understand the relationship between governmental systems and the public [1]. Automatic petition processing can help institutions immensely not only by filtering out noisy petitions, spam, and simply angry threats but also by aggregating people sentiments toward certain changes, events, or orders in an objective manner. Political implications of online petitions are well described in [2].

Unfortunately, a lot of effort is going into manual analysis of petitions which may lead to biased conclusions, is prone to errors, and inefficient. An example of the effort that went into the analysis is [3] where authors were searching for insights in the 'Save the Cretan landscape: Stop golf development at Cavo Sidero' online petition.

Ukraine is no stranger to the e-petition applications. Kyiv city – the capital and the largest city of Ukraine with a population of around 3 million people [4] – has a platform for submitting online petitions to the Kyiv City Council – petition.kievcity.gov.ua. An e-petition makes it possible for citizens to suggest actions to the Kyiv City Council. Our research is aimed at building a model of petitions posted on the above mentioned platform in order to be able to search for relevant petitions given a natural language query.

Similar research has been done by Hagen, L. et al. [5] where the authors have used We The People website data to uncover latent patterns in online petitions. They analyzed linguistic and semantic features of texts and built an LDA [6] model of the provided petitions. While very powerful, the LDA model is more of an exploratory tool that extracts main topics from petitions and lacks the contextual knowledge that models on top of the distributional hypothesis provide [7].

Vector space models. Vector space model is an algebraic model for encoding entities as vectors for the purpose of being able to find similarity of these entities as the degree between vectors. Every vector in such a model encapsulates the semantic structure of an object so that similar objects end up having small degrees between their vectors. The degree is also called similarity. It shows how similar are objects in the vector space instead of the distance between them. The most commonly used similarity function for vector space models is cosine similarity (1). Other noteworthy mentions are Euclidean similarity and Jaccard similarity which may describe similarity between terms better in some cases or applications, for example, in hierarchical vector space models [8]. For our experiments, for petition vector space model we used cosine similarity because it is the similarity measure that is used for underlying word vectors, which are described next. Note that identical vectors are going to have cosine similarity equal to 1.0, so the more two vectors are semantically similar, the more their cosine similarity value has to be closer to 1.0.

$$S(A,B)=\frac{AB}{AB}=\frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\sqrt{\sum_{i=1}^{n}B_i^2}}, \qquad (1)$$

where *S(A, B)* – cosine similarity between vectors *A* and *B*, *n* – dimensionality of vectors.

To embed textual documents into vectors, models based on word embeddings, Term Frequency-Inverse Document Frequency weights, document indexing, and Latent Semantic Analysis are usually used. This paper focuses on models based on word embeddings, out of which the most popular are Google's Word2Vec[9], Stanford's Glove[10], and Facebook's FastText[11]. To better understand how to convert documents into vectors we should first step back and examine word-vector models listed previously.

The Word2Vec method takes leverage of a neural network to find word relationships from text. After the training process is finished, Word2Vec algorithm can be used to find synonyms or semantically similar words and complete sentences with missing parts. Usually, the text that is used to train such a model is huge, for example, the original paper on Word2Vec trained it on Google News Corpus with 100 billion words. Typical size of the underlying vectors is 300. Since our experiments work on a much smaller scale, we choose the dimensionality of vectors to be 100. Word2Vec has two distinct architectures, however, both of them are two-layer neural networks that make use of the distributional hypothesis. The hypothesis claims that words that occur in the same contexts tend to have similar meanings [7]. The first architecture of Word2Vec is called continuous bag-of-words (CBOW) and its goal of learning is to predict the word inside a sentence from its context – a number of neighboring words. The second architecture, called skip-gram, has the opposite learning goal – given a word inside a document predict its neighbors in a certain span. It also puts more weight on closer surrounding words than more distant ones. CBOW takes less time to train than skip-gram but models words that occur in the corpus less frequently worse.

The FastText method is an extension of Word2Vec that takes into account subword information. The authors of the algorithm model morphology by considering subword units, and representing words by a sum of its character n-grams[11], [12]. They extract all of the n-grams in the length range from 3 to 6 from words which lets the model learn prefixes and suffixes as well as other morphological information present in most of the words. This makes FastText model for word vectors expressively more powerful than Word2Vec because words that were not present in the dataset can still be embedded or queried in the vector space. On the other hand, given that FastText learns representations for subword information, for small datasets, like the one that we use, the amount of learned weights and the complexity of the model grows which may lead to less quality with limited data.

One way of training both models is to use negative sampling [13] which minimizes the log-likelihood of sampled negative instances as opposed to training on positive examples only. This method is fast, simple, and widely used for similar tasks.

This paper explores a vector space model for Kyiv city petitions, that is based on the actual texts of petitions. The main question that we addressed was the choice of an approach to create vectors of petitions given vectors of words. We first train word vectors using Word2Vec and FastText which gives us 100-dimensional embeddings for words present in petitions. For every word in the text of a particular petition, the corresponding embedding was taken and averaged across every axis to generate a 100-dimensional vector for the petition as depicted in Fig. 1.

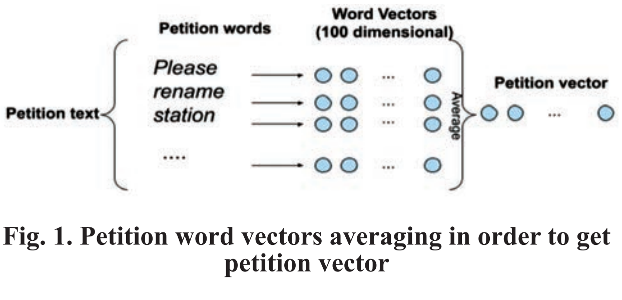


Fig. 1. Petition word vectors averaging in order to get petition vector

As a result, the final vector space model is encoding the high-level meaning of the petition and can be queried for similar petitions, creating a simple exploratory data analysis tool relying purely on the semantic content of petitions. This leads to one of the drawbacks of the model: if petition description is too abstract and poorly constructed, the averaging process creates poor vector representation thus making the space less representative.

Experiments. Word vectors. Based on the abovementioned framework, we trained two sets of word vectors: Word2Vec and FastText. The dataset that we trained on consists of 4623 petitions written in the Ukrainian language that were scraped from the petitions website. To capture semantic relationships between key entities and objects in petitions better we did several preprocessing steps:

1. stop word removal via the stop-words [14] library which provides stop words for 22 languages including Ukrainian.
2. whitespace normalization (strip any irrelevant whitespace before and after the core petition text, as well as any additional spaces in between words).
3. invisible and non-unicode characters removal.
4. infrequent tokens removal (every word that was not present in our dataset more than 20 times was removed with an intuition of being irrelevant or a spelling mistake). This made the training process more stable.

Table 1

Top 10 closest words in Word2Vec space with their cosine distance to queried words

Queried word	**Кличко (Klitschko)**			**КМДА(Kyiv City State Administration)**		
#	***Closest word(Ukrainian)***	***Closest word(English translation)***	***Cosine distance***	***Closest word(Ukrainian)***	***Closest word(English translation)***	***Cosine distance***
1	Віталій	Vitaly	0.747	Київради	Kyiv City Council	0.614
2	мер	mayor	0.586	депутатів	deputies	0.532
3	голова	head	0.458	КМР	Kyiv City Council(abbr.)	0.518
4	голови	head(genitive)	0.417	сайті	site	0.495
5	Київської	Kyiv(adjective)	0.374	рішення	solution	0.460
6	вимогою	requirement	0.368	ради	council	0.452
7	разом	together	0.356	відповідних	according	0.446
8	комісії	commission	0.353	РДА	District state administration	0.445
9	міський	city	0.353	КП	Municipal Enterprise	0.439
10	своїм	their	0.347	розпорядження	order	0.437

Table 2

Top 10 closest words in FastText space with their cosine distance to queried words

Queried word	**Кличко (Klitschko)**			**КМДА(**Kyiv City State Administration**)**		
#	***Closest word(Ukrainian)***	***Closest word(English translation)***	***Cosine distance***	***Closest word(Ukrainian)***	***Closest word(English translation)***	***Cosine distance***
1	Віталій	Vitaly	0.724	Київради	Kyiv City Council	0.577
2	голова	head	0.571	КМР	Kyiv City Council(abbr.)	0.567
3	мер	mayor	0.530	депутатів	deputies	0.501
4	голови	head(genitive)	0.478	сайті	site	0.486
5	Київської	Kyiv(adjective)	0.400	рішення	order	0.468
6	комісії	commission	0.394	РДА	District state administration	0.464
7	разом	together	0.392	ради	council	0.446
8	КМДА	Kyiv City State Administration	0.373	КП	Municipal Enterprise	0.446
9	Шановний	Dear	0.371	питання	question	0.445
10	Віталію	Vitaly(vocative)	0.365	відповідних	according	0.430

We experimented with optimal hyperparameters for both of the models, which are listed here:

- vector dimension – 100.
- training epochs – 1000.
- context window – 5.
- learning rate – 0.025.

Both models use skipgram and negative sampling as part of the internal algorithm and were trained for around 30 minutes with 4 workers on a 2.8 GHz Intel Core i5 processor. No GPU was needed because the size of the dataset is small. We used Gensim [15] as the framework for experiment implementation which allows us to build vector models with an easy-to-use, yet powerful API.

Tables 1 and 2 showcase some of the queries that became possible with the built word models.

Please, note that since FastText is a modification of Word2Vec, the results of queries are similar, both capture semantic relationships, like name and job, or similar institutions of the query. However, as mentioned before, FastText allows us to make queries for words not present in the dataset, which is a huge bonus for word vector models.

Petition vectors. In this section, we detail the results obtained by averaging word vectors for petition contents. In order to give some qualitative measures of newly constructed petition vectors, we are going to show their visualization. The visualization is built by reducing the dimensionality of petition vectors from 100-dimensional to 3-dimensional and plotting this reduced space. The algorithm for dimensionality reduction that we use is UMAP [16] which has increased speed and better preservation of the data's global structure than other dimensionality reduction algorithms. The idea behind UMAP is to first create a high dimensional graph representation of the data and then optimize another low-dimensional graph to be as structurally similar as possible to the constructed graph. We use Tensorboard [17] to make these visualizations and explore the space manually. Tensorboard allows projecting embeddings to a lower-dimensional space via UMAP, t-SNE, or PCA. We chose UMAP because it is faster than t-SNE and more expressive than PCA[18].

For the quantitative measurement of the differences in the researched models we use the Silhouette Coefficient [19] which is defined for a single sample as:

$$S = \frac{a - b}{max(a, b)},$$

where a – average distance between the point and all other points in the same cluster, b – average distance between the point and all other points in the nearest cluster. The Silhouette Coefficient score is defined as the mean of Coefficients of every point. The model with better defined clusters is expected to show higher Silhouette Coefficient score. The Coefficient has a range of –1 to 1 where spaces with highly dense clusters have scores closer to 1 and with highly overlapping clusters closer to 0. Table 2 shows Silhouette Coefficient scores for Word2Vec– and FastText-based petition vectors clustered via DBSCAN [20].

Table 3

Silhouette Coefficient scores

Model	Silhouette Coefficient score
Word2Vec-based	0.468
FastText-based	0.004

As mentioned before, to build vector space models for Kyiv city petitions we used a simple averaging of word vectors of words present in the petition. In fig. 2 you can see Word2Vec- and FastText-based vector spaces after dimensionality reduction visualized.

Both spaces exhibit clustered structure and have petitions of different semantics in different parts of the space. Please, note that Word2Vec-based petition model has visibly more separated clusters than FastText-based one. This is confirmed by the Silhouette Coefficient scores listed previously, where Word2Vec based petition model had a much higher score which means the clusters that are present in its vector space are more dense, while in the FastText-based model they overlap a lot.

A closer look shows that these clusters are indeed semantically divided and several well-defined groups of points exhibit similarity in the topics that they discuss. In the next section, we are going to talk in which way the two built models differ.

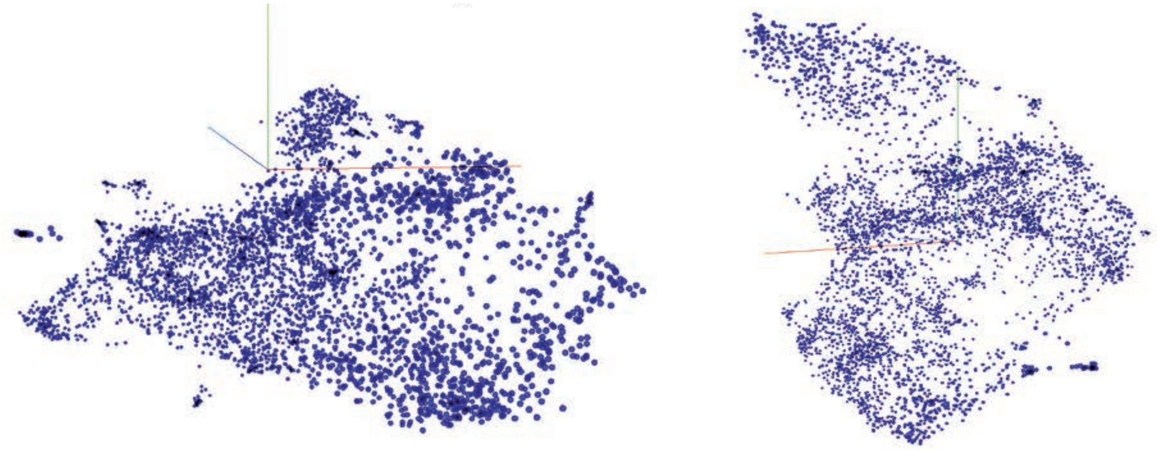
Word2Vec-based. Upon closer inspection, Word2Vec-based model has clearly visible clusters that share some semantic meaning among them. You can see two examples of that on fig. 3.

For more clarity on the insides of the model we provide a few queries and their similarities with the closest petitions in the dataset in the Table 4. As you can see the query that concerns about the ecological situation in Kyiv yields several petitions about poor waste processing and polluted lakes, while a query about hot water supply mostly returns complaints about it to the Kyiv city council. Overall, the quality of Word2Vec-based embeddings is satisfactory for their future use for transfer learning as features to sentiment analysis classifier or any other natural language problem.

FastText-based. FastText-based model shows similar semantic properties to the Word2Vec based model, their underlying core ideas are close after all. However, the overall number of visible clusters is reduced and some of them are clearly clustered on syntactic level instead of desirable semantic level. You can see this in Fig. 4, where (a) is a good semantic cluster, while (b) is similar only by having the same boilerplate start.

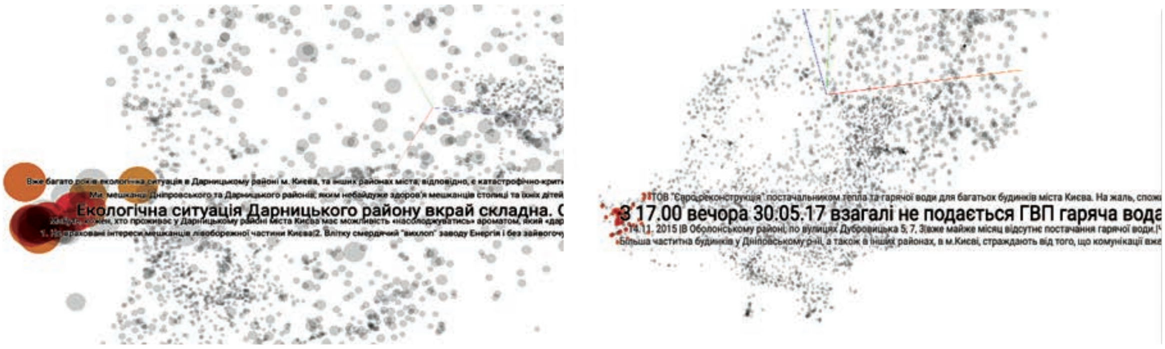
In Table 5 we provide two examples of FastText-based model querying. The first query is just a name of an avenue and the model could find semantically similar petitions that are mostly about its renaming process or about the process of renaming other avenues. The second example shows that the model can also find syntactically similar petitions. Overall, the quality of this model is less-suitable to be used in further semantically significant tasks than Word2Vec-based one.

Conclusions. In this paper, we proposed two methods of construction of vector space models of



a) b)

Fig. 2. Word2Vec (a) and FastText (b) petition vectors U-MAP visualizations



Вже багато років екологічна ситуація в Дарницькому районі м. Києва, та інших районах міста; відповідно, є катастрофічно-крити

Ми, мешканці Дніпровського та Дарницького районів, яким небайдуже здоров'я мешканців столиці та їхніх дітей

Екологічна ситуація Дарницького району вкрай складна. (

Жахливо, кожен, хто проживає у Дарницькому районі міста Києва має можливість «насолоджуватись» ароматом, який «дар

1. Не враховані інтереси мешканців лівобережної частини Києва|2. Влітку смердючий "вихлоп" заводу Енергія і без зайвого

ТОВ "Єврореконструкція" постачальником тепла та гарячої води для багатьох будинків міста Києва. На жаль, спожи

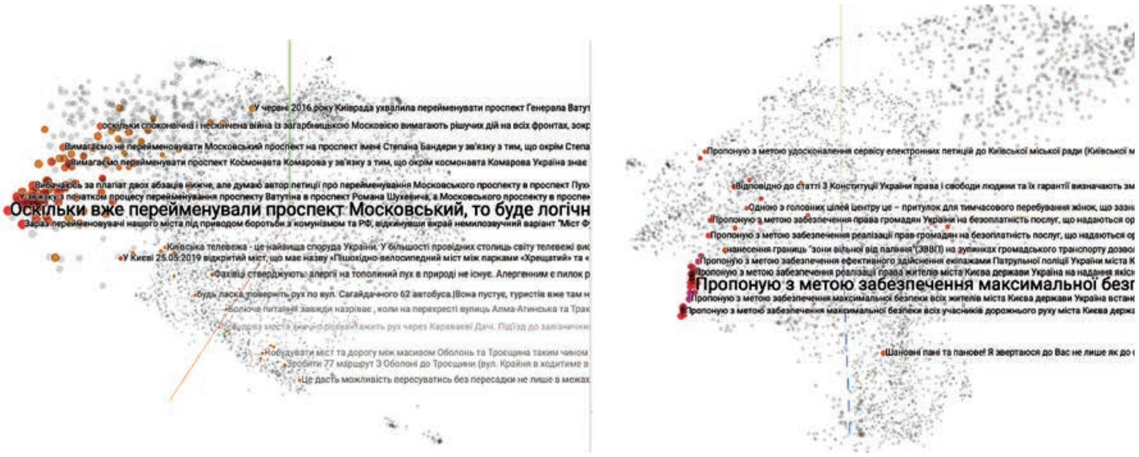
З 17.00 вечора 30.05.17 взагалі не подається ГВП гаряча вода

14.11. 2015 |В Оболонському районі, по вулицях Дубровицька 5, 7, 3|вже майже місяць відсутнє постачання гарячої води.|Ч

Більша частитна будинків у Дніпровському рні, а також в інших районах, в м.Києві, страждають від того, що комунікації вже

a) b)

Fig. 3. Cluster of petitions about ecological situation (a) and water management (b) in Kyiv in the Word2Vec petitions space



У червні 2016 року Київрада ухвалила перейменувати проспект Генерала Вату

оскільки споконвічна і нескінчена війна із загарбницькою Московією вимагають рішучих дій на всіх фронтах, зокр

Вимагаємо не перейменовувати Московський проспект на проспект імені Степана Бандери у зв'язку з тим, що окрім Степа

Вимагаємо перейменувати проспект Космонавта Комарова у зв'язку з тим, що окрім космонавта Комарова Україна знає

Вибачаюсь за плагіат двох абзаців нижче, але думаю автор петиції про перейменування Московського проспекту в проспект Пух

У зв'язку з початком процесу перейменування проспекту Ватутіна в проспект Романа Шухевича, а Московського проспекту в просп

Оскільки вже перейменували проспект Московський, то буде логічн

Зараз перейменовувачі нашого міста під приводом боротьби з комунізмом та РФ, відкинувши вкрай немилозвучний варіант "Міст Ф

Київська телевежа - це найвища споруда України. У більшості провідних столиць світу телевежі вик

У Києві 25.05.2019 відкритий міст, що має назву «Пішохідно-велосипедний міст між парками «Хрещатий» та «

Фахівці стверджують: алергії на тополиний пух в природі не існує. Алергенним є пилок р

будь ласка поверніть рух по вул. Сагайдачного 62 автобуса.|Вона пустує, туристів вже там н

Болюче питання завжди назріває , коли на перехресті вулиць Алма-Атинська та Трак

Будова моста вічно розвантажить рух через Караваєві Дачі. Під'їзд до залізничн

Побудувати міст та дорогу між масивом Оболонь та Троєщина таким чином

Зробити 77 маршрут З Оболоні до Троєщини (вул. Крайня в ходитиме в

Це дасть можливість пересуватись без пересадки не лише в межах

Пропоную з метою удосконалення сервісу електронних петицій до Київської міської ради (Київської м

Відповідно до статті 3 Конституції України права і свободи людини та їх гарантії визначають зм

Одною з головних цілей центру це – притулок для тимчасового перебування жінок, що зазн

Пропоную з метою забезпечення права громадян України на безоплатність послуг, що надаються ор

Пропоную з метою забезпечення реалізації прав громадян на безоплатність послуг, що надаються ор

нанесення границь "зони вільної від паління"(ЗВВП) на зупинках громадського транспорту дозвол

Пропоную з метою забезпечення ефективного здійснення екіпажами Патрульної поліції України міста К

Пропоную з метою забезпечення реалізації права жителів міста Києва держави Україна на надання якіс

Пропоную з метою забезпечення максимальної безп

Пропоную з метою забезпечення максимальної безпеки всіх жителів міста Києва держави Україна встано

Пропоную з метою забезпечення максимальної безпеки всіх учасників дорожнього руху міста Києва держа

Шановні пані та панове! Я звертаюся до Вас не лише як до

a) b)

Fig. 4. Cluster of petitions about the renaming of different city objects (a) in Kyiv in the FastText petitions space and example of suboptimal separation in the FastText petitions space (b)

Table 4

Top 3 closest petitions in Word2Vec space with their cosine distance to queried phrases

	Екологічний стан Києва у Дарницькому районі (Ecological condition of Kyiv in Darnytskyi district)			**відсутнє постачання гарячої води (there is no hot water supply)**		
#	***Closest petition(Ukrainian)***	***Closest petition(English translation)***	***Cosine distance***	***Closest petition(Ukrainian)***	***Closest petition(English translation)***	***Cosine distance***
1	Вже багато років екологічна ситуація в Дарницькому районі м. Києва, та інших районах міста, відповідно, є катастрофічно-критичною: нестерпний сморід в нічний та ранковий час доби як результат недолугої діяльності сміттепереробного заводу «Енергія» ...	For many years, the ecological situation in the Darnytskyi district of Kyiv and other districts of the city, respectively, is catastrophically critical: an unbearable stench at night and in the morning as a result of poor activities of the waste processing plant "Energy" ...	0.69	В Оболонському районі, по вулицях Дубровицька 5, 7, 3 вже майже місяць відсутнє постачання гарячої води. Чіткої відповіді на питання мешканців цих будинків «коли буде вода» від Київенерго та районного ЖКХ немає...	In the Obolonsky area, on Dubrovytska streets 5, 7, 3 there has been no hot water supply for almost a month. There is no clear answer to the question of the residents of these houses "when will there be water" from Kyivenerho and the district housing and communal services ...	0.841
2	заборонено купання в 50 озерах і ставках, зокрема Дідоровських і Мишеловських ставках … в Дарницькому районі ... Купання в цих озерах не рекомендується через незадовільні проби води. Комунальне підприємство «Плесо» заборонило пляжний відпочинок через невідповідність санітарним нормам: за результатами санітарно-мікробіологічних досліджень проб води. Закликаю очистити озеро Сонячне на Позняках та дати можливість людям безпечно відпочивати.	it is forbidden to swim in 50 lakes and ponds, in particular Didorovsky and Mishelovsky ponds … in Darnytskyi district, … . Swimming in these lakes is not recommended due to unsatisfactory water samples. The Pleso municipal enterprise has banned beach holidays due to non-compliance with sanitary norms: according to the results of sanitary-microbiological tests of water samples. I urge you to clean Lake Sunny in Pozniaky and allow people to rest safely.	0.651	Не перший рік виникають нарікання на температуру постачання гарячої води. Єдина можливість додати холодну воду до гарячої виникає...	Not the first year there are complaints about the temperature of the hot water supply. The only way to add cold water to hot water arises ...	0.835
3	Мабуть, кожен, хто проживає у Дарницькому районі міста Києва, має можливість «насолоджуватись» ароматом, який «дарує» нам БОРТНИЦЬКА СТАНЦІЯ АЕРАЦІЇ ТА СМІТТЄСПАЛЮВАЛЬНИЙ ЗАВОД «ЕНЕРГІЯ"» Та, крім неприємного запаху, підприємства також несуть загрозу здоров'ю КОЖНОГО З НАС...	Apparently, everyone who lives in the Darnytskyi district of Kyiv has the opportunity to "enjoy" the aroma that "gives" us BORTNYTSKY AERATION STATION AND incinerator "ENERGY". But in addition to the unpleasant smell, companies also pose a threat to the health of EACH OF US ...	0.642	Оскільки послуги з опалення та постачання гарячої води надаються монополістами, потрібен важіль тиску на постачальників для унеможливлення значного завищення тарифів. Таким засобом впливу стане можливість встановлення індивідуальний систем опалення та підігріву води у будь-яких багатоповерхових будинках.	As heating and hot water services are provided by monopolists, a lever of pressure on suppliers is needed to prevent significant overstatement of tariffs. This means of influence will be the ability to install individual heating and water heating systems in any multi-storey building.	0.818

Table 5

Top 3 closest petitions in FastText space with their cosine distance to queried phrases

	Проспект Московський (Moscow Avenue)			**Пропоную з метою забезпечення ... (In order to ensure ...)**		
#	***Closest petition(Ukrainian)***	***Closest petition(English translation)***	***Cosine distance***	***Closest petition(Ukrainian)***	***Closest petition(English translation)***	***Cosine distance***
1	У зв'язку з початком процесу перейменування проспекту Ватутіна в проспект Романа Шухевича, а Московського проспекту в проспект Степана Бандери було б логічно перейменувати Московський міст, що з'єдує ці два проспекти, в Троєщинський міст.	In connection with the beginning of the process of renaming Vatutin Avenue to Roman Shukhevych Avenue, and Moscow Avenue to Stepan Bandera Avenue - it would be logical to rename the Moscow Bridge, which connects the two avenues, to Troieschyna Bridge.	0.69	Пропоную з метою забезпечення ефективного здійснення екіпажами Патрульної поліції України міста Києва держави Україна своїх посадових обов'язків забезпечити кожен з них сертифікованими приладами лазерними радарами вимірювання швидкості TruCam	In order to ensure the effective implementation of the duties of the crews of the Patrol Police of Ukraine in the city of Ukraine, the state of Ukraine to provide each of them with certified devices - laser radars for measuring the speed of TruCam	0.654
2	Назва Московського проспекту в Оболонському і Московському районах Києва не є історичною. Її було надано 2003 року як дружній політичний жест тодішнього міського голови Києва О. Омельченка щодо його московського колеги Ю. Лужкова...	The name of Moscow Avenue in Obolonsky and Moscow districts of Kyiv is not historical. It was given in 2003 as a friendly political gesture of the then mayor of Kyiv O. Omelchenko towards his Moscow colleague Yu. Luzhkov ...	0.651	Пропоную з метою забезпечення ефективного здійснення екіпажами Патрульної поліції України міста Києва держави Україна своїх посадових обов'язків забезпечити кожен з них сертифікованими приладами - алкотестерами Драгер	In order to ensure the effective implementation by the crews of the Patrol Police of Ukraine of the city of Kyiv, the state of Ukraine, I propose to provide each of them with certified devices - Drager breathalyzers.	0.643
3	Назва «Проспект Правди» (Виноградар) підпадає під Закон про декомунізацію. Окрім того, при перейменуванні його на проспект Павла Шеремета буде символічно, що бічним до цього проспекту є проспект Георгія Гонгадзе, також загиблого київського журналіста.	The name "Pravda Avenue" (Vinogradar) falls under the Law on Decommunization. In addition, when renaming it to Pavel Sheremet Avenue, it will be symbolic that the side of this avenue is Georgy Gongadze Avenue, also a deceased Kyiv journalist.	0.74	Просимо заборонити рух великовагового транспорту територією міста в денний час із метою зменшення руйнування асфальтового покриття та з метою поліпшення організації дорожнього руху і його безпеки, поліпшення екологічного стану та підвищення пропускної спроможності вулично-шляхової мережі м. Києва	Please prohibit the movement of heavy vehicles in the city during the day, in order to reduce the destruction of asphalt pavement and to improve the organization of traffic and its safety, improve the environmental condition and increase the capacity of the road network of Kyiv	0.606

Kyiv city petitions, namely Word2Vec and FastText-based word vector averaging. The main insights are that it is possible to build such vector spaces with limited data and that through our experiments Word2Vec-based algorithm was preferred since it captured more semantic representations instead of syntactic ones. This happened because, innately, FastText works on subword level and while it is useful for getting vectors of unknown character sequences, the process of word vectors averaging does eliminate this advantage, making it a liability. Quantitative results show that Word2Vec-based model is better suited for further clustering and produces more dense clusters than FastText-based one.

The suggested models can be used as a stepping stone in petition analysis pipelines. The vector space models give every petition a numeric representation capturing its semantic meaning that, if included in a classification framework, can help identify citizens' attitudes toward certain events, group and deduplicate petitions with the same intent, or predict if a certain petition is going to get enough votes to pass.

Future work might include trying other aggregation functions to build petition-level vectors, like term-frequency weighting. Other possible research directions include sentiment analysis and automatic clustering of Kyiv city petitions based on built models.

References:

1. R. Lindner and U. Riehm, "Electronic petitions and institutional modernization. International parliamentary e-petition systems in comparative perspective," *JeDEM-eJournal eDemocracy Open Gov.*, vol. 1, no. 1, pp. 1–11, 2009.

2. K. Böhle and U. Riehm, "E-petition systems and political participation: About institutional challenges and democratic opportunities," *First Monday*, 2013.

3. H. Briassoulis, "Online petitions: new tools of secondary analysis?," *Qual. Res.*, vol. 10, no. 6, pp. 715–727, 2010.

4. "Population (1995-2019)", *Kiev.ukrstat.gov.ua*, 2020. [Online]. Available: http://www.kiev.ukrstat.gov.ua/p.php3?c=527&lang=1. [Accessed: 07-Sep-2020].

5. L. Hagen, T. Harrison, O. Uzuner, W. May, T. Fake, and S. Katragadda, "E-petition popularity: Do linguistic and semantic factors matter?," *Gov. Inf. Q.*, 2016.

6. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.

7. Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2–3, pp. 146–162, 1954.

8. M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Advances in neural information processing systems*, 2017, pp. 6338–6347.

9. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

10. J. Pennington, R. Socher, and C. D. 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.

11. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *CoRR*, vol. abs/1607.04606, 2016.

12. A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," *CoRR*, vol. abs/1607.01759, 2016.

13. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space." 2013.

14. A. Coenen and A. Pearce, "Understanding UMAP", *Pair-code.github.io*, 2020. [Online]. Available: https://pair-code.github.io/understanding-umap/. [Accessed: 15-Aug-2020].

15. R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.

16. L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv Prepr. arXiv1802.03426*, 2018.

17. Martin Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems." 2015.

18. A. Coenen and A. Pearce, "Understanding UMAP", *Pair-code.github.io*, 2020. [Online]. Available: https://pair-code.github.io/understanding-umap/. [Accessed: 15-Aug-2020].

19. P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

20. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.

Шаптала Р.В., Кисельов Г.Д. ВЕКТОРНІ МОДЕЛІ ПЕТИЦІЙ МІСТА КИЇВ

У цьому дослідженні ми описуємо та порівнюємо два шляхи створення моделі векторного простору петицій міста Києва. Для автоматичного аналізу текстів вільної форми, таких як петиції, їх потрібно перевести в числовий простір. Використовуючи вектори слів на основі розподільчої гіпотези, а саме Word2Vec та FastText, ми будуємо векторні моделі петицій міста Києва.

Загальний підхід, який ми пропонуємо, – це навчання векторів слів на наборах петицій міст Києва, попередня обробка даних документів та застосування усереднення векторів слів для створення векторів петицій. Більше того, цей підхід не вимагає великих даних і може застосовуватись до навчання у низько-ресурсних середовищах, таких як українська мова, для якої ми використовували лише 4623 петиції без розмітки. Жодних попередньо навчених моделей та їх налаштування для цього дослідження не використовувалось, і ми надаємо гіперпараметри, оптимальні для проведених експериментів.

Проаналізовано переваги та недоліки обох моделей. Модель на основі Word2Vec отримує вищу оцінку Коефіцієнту Силуетта і створює щільніші кластери, ніж модель на основі FastText. Це робить її більш відповідним для реальних застосувань, таких як аналіз настроїв петицій або їх кластеризація. Аналіз помилок підтверджує цей результат, оскільки FastText приділяє більше уваги синтаксичній структурі петицій та слів, тоді як Word2Vec більше зосереджується на контексті. На підтримку цього твердження ми наводимо приклади такої поведінки для однакових текстових запитів на різні міські теми.

Візуалізації векторних просторів після зменшення їх розмірності за допомогою UMAP демонструються, намагаючись показати їх загальну структуру. Вони підкріплюють отримані оцінки Коефіцієнта Силуетта, демонструючи щільніші кластери для підходу на основі Word2Vec. Отримані моделі можна використовувати для ефективного запиту семантично пов'язаних петицій, а також для пошуку груп петицій зі схожими темами або скаргами.

Ключові слова: *векторна модель, FastText, Word2Vec, аналіз петицій, UMAP.*