МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Сумський державний університет

Навчально-науковий інститут бізнесу, економіки та менеджменту

Кафедра економічної кібернетики

«До захисту допущено»
Завідувач кафедри
Віталія КОЙБІЧУК

_____ 2024р.

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня бакалавр

зі спеціальності 051 Економіка,
освітньо-професійної програми «Економічна кібернетика та бізнес-аналітика»
на тему: «Економіко-математичне моделювання державних закупівель через систему Prozorro засобами Data Mining»

Здобувача групи ЕК-01а Штефана Артема Валерійовича

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

Артем ШТЕФАН

Керівник: ст. викл., доктор філософії Сергій МИНЕНКО

Суми – 2024

Ministry of Education and Science of Ukraine

Sumy State University

Educational and Scientific Institute of Business, Economics and Management

Department of Economic Cybernetics

«Admitted to the defense»

Head of Department

Vitaliia KOIBICHUK

_____ 2024y.

QUALIFICATION WORK

to obtain a bachelor's educational degree

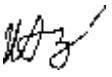
from the specialty 051 Economics,

the educational-professional program «Economic Cybernetics and Business Analytics»

on the topic: «Economic and Mathematical Modelling of Public Procurement Through the Prozorro System Using Data Mining»

Student of the group EC-01a Shtefan Artem

The qualification work contains the results of my own research. The use of ideas, results, and texts of other authors are linked to the corresponding source.



Artem SHTEFAN

Scientific supervisor: senior lecturer, PhD Serhii MYNENKO

Sumy – 2024

ABSTRACT

Shtefan A. V. Economic and Mathematical Modelling of Public Procurement Through the Prozorro System Using Data Mining: the thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science: specialty – 051 Economics (Economic Cybernetics and Business Analytics) / Supervisor S. V. Mynenko. Sumy : Sumy State University, 2024. 69 p.

Public procurement includes a wide range of goods, services, and works, from construction and repair to the supply of medical equipment, computers, transportation, management services, etc. Public procurement authorities must follow procedures established by law, such as publishing tender announcements, ensuring equal access to information for all participants, holding open tenders or competitions, evaluating tender proposals, and concluding contracts with the winners. Tenders are implemented through the electronic public procurement system Prozorro.

However, Russia's full-scale invasion of Ukraine has highlighted the problems associated with the inability to ensure the most efficient use of budget funds. The sources of these problems are both the difficulty of monitoring the thousands of tenders that appear in the system every day and the slow pace of civil society development and, as a result, the lack of proper public control in this area over a long period of time.

Therefore, at this stage of development of Ukraine and its public procurement sector, an important factor is partial automation and consolidation of monitoring of public needs expressed through this sector, reduction of the risk of corruption in the procurement process, etc. For this purpose, it is proposed to use the methods of intellectual analysis of text data (Text Data Mining). This paper proposes a method of applying economic and mathematical modeling of one of them - Topic Modelling.

Manual search for announcements that contradict the principles of fair and transparent bidding according to the Law of Ukraine "About Public Procurement" has a low degree of usefulness, as thousands of tenders are announced by the authorities every day. In addition, the lack of systematic oversight of this process by the Department for Public Procurement and Competition Policy and anti-corruption bodies contributes to the growing influence of the human factor in the form of dishonesty and abuse of office.

In addition, it should be noted that despite the amount of losses incurred by the Ukrainian economy as a result of inappropriate spending of funds through some tender procurements, the interest of the Ukrainian public in this issue shows a disappointing trend. This is a manifestation of the previously mentioned slow development of civil society.

Thus, the need to develop new, at least temporary, approaches to monitoring and administering announced tenders becomes even more apparent. Given that their appearance becomes massive even within one day, the importance of full or partial automation of announcement processing is clear. This is the relevance of tender research today.

It follows from the above that the purpose of this research was to build and demonstrate the feasibility of using topic models for analyzing public procurement activities.

The object of this research is the socio-economic relations that arise between participants in the public procurement process in Ukraine.

The subject of the study was economic and mathematical methods and models of public procurement announced by the relevant authorities through the online platform of the Prozorro system.

As a result of the work, were built and demonstrated the feasibility of using topic models for analyzing public procurement activities: the research objective has been achieved.

It was found that the model based on the BERTopic algorithm is suitable for finding markers that may indicate corruption or the use of public funds with a low

level of utility for society. It was also found that the LDA model can be used to analyze the needs within the sectors of the national economy of Ukraine, as well as the country's socio-economic system.

In the course of the research work, the objectives were achieved:

- The prerequisites and relevance of tender research are described;
- A bibliometric analysis of relevant scientific research in the field of public procurement was conducted;
- Natural Language Processing (NLP) as a method of monitoring public procurement activities is presented;
- A database of tenders was formed using the Prozorro application programming interface;
- The research assumptions were formed;
- Conceptual modeling of topic modeling algorithms, such as Dirichlet latent clustering and BERTopic, was implemented;
- NLP models were built and interpreted the results.

Research methods: synthesis; analysis of relevant publications of representatives of the scientific community, specialists in the field of data mining; bibliometric analysis; topic modeling, cluster analysis.

The source of data for building the models was the database of the Prozorro e-procurement system. The source of knowledge on algorithmization of data collection, data cleaning, and modeling in code was the documentation of the Python programming language and its modules and libraries. The code writing environment was the Spyder integrated development environment.

The scientific and social value of the research lies in the accelerated and consolidated method of monitoring and analyzing public procurement based on machine learning approaches. This will allow for more efficient redistribution of state and local budget funds, more effective anti-corruption measures, etc.

A promising area for further research could be modeling by associative rules to establish links with other variables included in the collected data from Prozorro. Also, to facilitate the work with such models, it is advisable to develop a full-fledged

application based on thematic modeling methods. To explore the possibilities of automating the control of procurement activities, it is advisable to invest in machine learning and artificial intelligence.

The research was carried out within the framework of the research work commissioned by the Ministry of Education and Science of Ukraine «Modeling the mechanisms of de-shadowing and corruption of the economy to ensure national security: the impact of the transformation of financial behavioral patterns», state registration no: 0122U000783. The results of the bachelor's qualification work were published in the article «Financial Fraud Detection on Social Networks Based on a Data Mining Approach» in the professional journal «Financial Markets, Institutions and Risks (FMIR)».

Keywords: economic growth, corruption, public procurement, tender research, monitoring, Prozorro, topic modeling, natural language processing, clustering, LDA, BERTopic.

The content of the qualification work is set out on 69 pages. The list of references of 56 titles is located on 45 – 50 pages. The work contains 1 table, 13 figures, and appendices A, B, C, D, E, F.

The year of completion of the qualification work is 2024.

The year of defense of the qualification work is 2024

Міністерство освіти і науки України
Сумський державний університет
Навчально-науковий інститут бізнесу, економіки та менеджменту
Кафедра економічної кібернетики

ЗАТВЕРДЖУЮ
Завідувачка кафедри
к.е.н., доцентка
_____ Віталія КОЙБІЧУК
“ ” квітня 2024 р.

ЗАВДАННЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ НА
ЗДОБУТТЯ ОСВІТНЬОГО СТУПЕНЯ БАКАЛАВРА
(спеціальність 051 Економіка «Економічна кібернетика та бізнес аналітика»)

студенту 4 курсу, групи ЕК-01а

Штефану Артему Валерійовичу

1. Тема роботи: «Економіко-математичне моделювання державних закупівель через систему Prozorro засобами Data Mining»
затверджена наказом по університету від «__» ____ 20__ року № _____
2. Термін подання студентом закінченої роботи «30» травня 2024 року
3. Мета кваліфікаційної роботи: побудова та демонстрація доцільності використання тематичних моделей для аналізу державної закупівельної діяльності.
4. Об’єкт дослідження: соціально-економічні відносини, що виникають між учасниками процесу публічних закупівель в Україні.
5. Предмет дослідження: економіко-математичні методи та моделі державних закупівель, оголошувані відповідними органами влади через онлайн-платформу системи Prozorro.
6. Кваліфікаційна робота виконується на матеріалах бази даних системи електронних закупівель Prozorro
7. Орієнтовний план кваліфікаційної роботи, терміни подання розділів керівникові та зміст завдань для виконання поставленої мети

Розділ 1 Теоретичні засади моделювання державних закупівель засобами Data Mining (Topic Modelling) – 10 травня

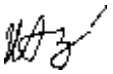
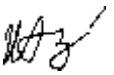
У розділі 1: описати передумови та актуальність тендерних досліджень; провести бібліометричний аналіз релевантних наукових досліджень у сфері державних закупівель; представити обробку природної мови (NLP) як метод моніторингу діяльності у сфері публічних закупівель; сформувати базу тендерів

за допомогою інтерфейсу прикладного програмування Prozorro; сформулювати припущення дослідження.

Розділ 2 Побудова тематичних моделей державних закупівель через систему Prozorro

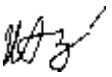
У розділі 2: реалізувати концептуальне моделювання алгоритмів тематичного моделювання, таких як латентне розміщення Діріхле та BERTopic; побудувати NLP-моделі та інтерпретувати отримані результати.

8. Консультації з роботи:

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1	Миненко С. В., старший викладач кафедри		01.04.2024 
2	Миненко С. В., старший викладач кафедри		

9. Дата видачі завдання: «1» квітня 2024 року

Керівник кваліфікаційної роботи С. В. Миненко

Завдання до виконання одержав  А. В. Штефан

ЗМІСТ

ВСТУП

Процес державних закупівель, поняття про який виникло в умовах Перебудови 1988 року в СРСР, внаслідок виникнення проблем із виконанням чергового п’ятирічного народногосподарського плану, пройшов еволюцію від інструменту, за допомогою якого робилися спроби нівелювати нестачу ресурсів у власності держави для досягнення запланованих рівнів економічних показників [1, 2], до процедури, метою якої, на сьогоднішній день, є забезпечення ефективного використання коштів держави та дотримання принципів доступності торгів, чесної конкуренції між усіма надавачами товарів та послуг, відкритості щодо вибору постачальників з боку всіх замовників. Оскільки цей процес є аналогічним стосовно державних органів та інших економічних агентів, станом на середину 2020-х років, термін «державні закупівлі» став еквівалентним поняттю «публічні закупівлі».

У цій роботі пропонується сконцентруватися на публічних закупівлях у частині саме замовників-органів державної влади, тому, здебільшого, надалі використовуватиметься термін «державні закупівлі».

Державні закупівлі включають широке коло товарів, послуг і робіт, від будівництва і ремонту до постачання медичного обладнання, комп'ютерів, транспорту, послуг з управління тощо. Органи, що здійснюють державні закупівлі, повинні дотримуватися процедур, визначених законом, таких як публікація оголошень про торги, забезпечення рівного доступу до інформації для всіх учасників, проведення відкритих торгів або конкурсів, оцінка тендерних пропозицій та укладання договорів з переможцями [3, 4]. Реалізація тендерів відбувається через електронну систему публічних закупівель Prozorro [5].

Однак повномасштабне вторгнення РФ в Україну висвітлило проблеми, пов’язані із нездатністю забезпечити максимально ефективне використання бюджетних коштів. Джерелами цих проблем є як складність моніторингу тисяч тендерів, що з’являються у системі щодня, так і повільний темп розвитку

громадянського суспільства і, як наслідок, відсутність належного громадського контролю у цій сфері протягом великого періоду часу.

Тому, на даному етапі розвитку України та сфери її публічних закупівель, важливим фактором є часткова автоматизація та укрупнення моніторингу суспільних потреб, що виражаються через цю сферу, зниження рівня ризиків корупційних проявів у процесі здійснення закупівельної діяльності тощо. Для цього пропонується використовувати методи інтелектуального аналізу текстових даних (Text Data Mining). У цій роботі запропоновано спосіб застосування економіко-математичного моделювання одного з них – тематичного моделювання (Topic Modelling).

Із зазначеного вище випливає, що метою цього дослідження є побудова та демонстрація доцільності використання тематичних моделей для аналізу державної закупівельної діяльності.

Об’єктом даного дослідження є соціально-економічні відносини, що виникають між учасниками процесу публічних закупівель в Україні.

Предметом дослідження виступають економіко-математичні методи та моделі державних закупівель, оголошувані відповідними органами влади через онлайн-платформу системи Prozorro.

По ходу проведення дослідження вимагається досягти реалізації таких завдань:

- Описати передумови та актуальність тендерних досліджень;
- Провести бібліометричний аналіз релевантних наукових досліджень у сфері державних закупівель;
- Представити обробку природної мови (NLP) як метод моніторингу діяльності у сфері публічних закупівель;
- Сформувати базу тендерів за допомогою інтерфейсу прикладного програмування Prozorro;
- Сформувати припущення дослідження;
- Реалізувати концептуальне моделювання алгоритмів тематичного моделювання, таких як латентне розміщення Діріхле та BERTopic;

– Побудувати NLP-моделі та інтерпретувати отримані результати.

Методи дослідження: синтез; аналіз відповідних публікацій представників наукової спільноти, фахівців у сфері інтелектуального аналізу даних; бібліометричний аналіз; тематичне моделювання, кластерний аналіз.

Джерелом забезпечення даними для побудови моделей є база даних системи електронних закупівель Prozorro. Джерелом знань щодо алгоритмізації збору даних, очистки даних та моделювання в коді є документація мови програмування Python та її модулів, бібліотек. Середовище написання коду – інтегроване середовище розробки Spyder.

Наукова та суспільна цінність дослідження полягає у пришвидшеному та укрупненому способі моніторингу та аналізу сфери державних закупівель, що базується на підходах машинного навчання. Це дозволить ефективніше перерозподіляти кошти державного, місцевого бюджетів, результативніше проводити заходи щодо антикорупційної боротьби тощо.

Наукове дослідження було виконано в межах науково-дослідної роботи за замовленням МОН України «Моделювання механізмів детінізації та декорумпізації економіки для забезпечення національної безпеки: вплив трансформації фінансових поведінкових патернів», № держреєстрації: 0122U000783. Результати кваліфікаційної роботи бакалавра оприлюднені у статті «Financial Fraud Detection on Social Networks Based on a Data Mining Approach» у фаховому журналі «Financial Markets, Institutions and Risks (FMIR)» [6].

РОЗДІЛ 1. ТЕОРЕТИЧНІ ЗАСАДИ МОДЕЛЮВАННЯ ДЕРЖАВНИХ ЗАКУПІВЕЛЬ ЗАСОБАМИ DATA MINING (TOPIC MODELLING)

1.1. Передумови та актуальність тендерних досліджень

У січні 2023 року засобами масової інформації почала поширюватися інформація про те, що Міністерство оборони України закуповує продукти харчування на потреби армії за цінами, вищими від встановлених у точках роздрібної торгівлі у 2 – 3 рази. Згідно опублікованого контракту МО розміром 13 млрд. грн., постачальником виявилося товариство з обмеженою відповідальністю «Актив компані», розмір статутного капіталу якого складав 1 тис. грн. Реалізація шахрайської схеми виявилася можливою завдяки повному закриттю оголошень на Prozorro про закупівлі на потреби армії [7]. Внаслідок цього, співробітниками Державного бюро розслідувань, спільно зі Службою безпеки України, було затримано злочинців, які привласнювали кошти, виділені на харчові потреби Збройних сил України [8].

Через ризики, що виникли після приховування доступу до оборонних закупівель на Prozorro, Міністерством економіки України було оголошено про відкриття інформації про незбройні закупівлі [9], однак протягом першого півріччя 2023 року з’явилася серія журналістських розслідувань від різних інформаційних агенцій про неефективне використання державних коштів з можливими проявами корупції через державні закупівлі у цивільному секторі [10].

Пошук оголошень, які суперечать принципам чесних та прозорих торгів згідно Закону України «Про публічні закупівлі», в ручному режимі має низький ступінь корисності, оскільки щодня владними органами оголошуються тисячі тендерів. Крім того, відсутність систематичного нагляду за цим процесом з боку Департаменту сфери публічних закупівель та конкурентної політики [11], антикорупційних органів, сприяє зростанню впливу людського фактору у вигляді недобросовісності та зловживань службовим становищем.

Крім того, необхідно відмітити, що не дивлячись на об’єми збитків, яких зазнає економіка України внаслідок недоцільних витрат коштів через деякі тендерні закупівлі, інтерес до даної проблеми, з боку української громадськості, показує невтішну тенденцію. За даними, отриманими за допомогою інструменту Google Trends, популярність теми «government procurement» не є стабільно високою (Рисунок 1.1).

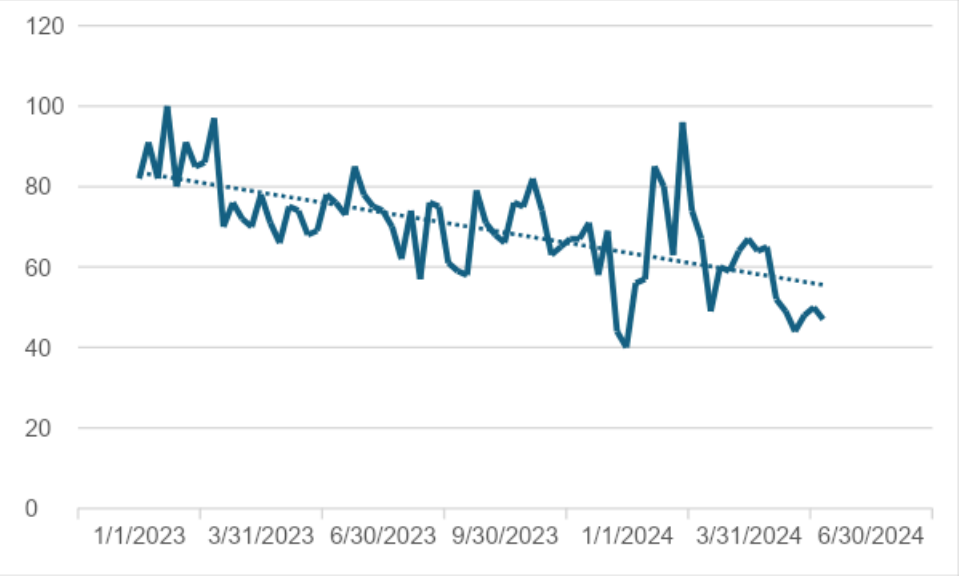


Рисунок 1.1 – Популярність запитів за темою «government procurement» в Україні у період 01.01.2023 – 26.05.2024 (ум. од. популярності)
(Розроблено автором за джерелом [12])

Це є проявом згаданого раніше процесу повільної розбудови громадянського суспільства.

Таким чином, необхідність розробки нових, принаймні, тимчасових підходів до моніторингу та адміністрування оголошуваних тендерів стає ще більш очевидною. З огляду на те, що їх поява набуває масового характеру навіть у межах однієї доби, зрозумілою є важливість повної або часткової автоматизації обробки оголошень. Саме у цьому виражається актуальність тендерних досліджень на сьогодні.

Для вивчення можливостей автоматизації контролю закупівельної діяльності доцільно інвестувати у машинне навчання, штучний інтелект. У даній роботі на практиці реалізується один із типів моделювання, що базується на неконтрольованому машинному навчанні.

У наступному пункті розглянемо науковий контекст в Україні щодо досліджень у сфері державних закупівель, у якому виконано цю роботу.

1.2. Бібліометричний аналіз релевантних наукових досліджень у сфері державних закупівель

Аналіз релевантних публікацій вітчизняних науковців, присвячених темі державних закупівель, індексованих пошуковою системою наукової літератури Google Scholar, протягом останнього року показав, що над дослідженнями відкритої та ефективної реалізації тендерів, конкурентного середовища, проявів корупції та запобігання їй тощо у сфері публічних закупівель працює велика кількість фахівців з галузей економіки та права.

Так, Артеменком О. В., Волковою Л. О. та Світличним О. П. досліджувалися поняття воєнного стану та державних закупівель під час встановлення цього правового режиму. Було розглянуто особливості порядків здійснення торгів, умов реалізації, підтверджень правомірного їх проведення за алгоритмом документально [13].

Длугопольським О. В. та Чапраком Ю. В. було сформульовано основні дилеми у сфері закупівель. Серед них:

- Підзвітність та відповідальність;
- Шахрайство і бюрократія;
- Зловживання делегуванням (дилема принципал-агента);
- Ефективність коротко- та довгострокових витрат;
- Розширення повноважень.

Автори вважають, що у довгостроковій перспективі критичної важливості набудуть реформи, спрямовані на вдосконалення процесів, що виникають у ході

закупівельної діяльності, підвищення стійкості системи публічних закупівель до девіацій [14].

Під час виконання дослідження тенденцій ринку публічних закупівель в умовах дії воєнного стану, Кильницькою Є. В., Глуховою С. В., Колодяжною Т. В. було розраховано кількісні, проаналізовано якісні показники публічних закупівель. Було виявлено значне зростання динаміки сум за тендерними договорами із середини 2022 року, порівняно з попередніми роками [15]. Враховуючи сукупність ризиків, описаних вище за текстом, можна припустити, що згідно даних, отриманих авторками, об’єми коштів, які виводилися у тінь, прямо пропорційно зросли.

На думку Пантелеймоненка А. О., Мільки А. І. та Павленко О. С., функціонування сфери державних закупівель є одним із визначальних факторів, що зумовлюють наповнюваність державного бюджету завдяки забезпеченню потреб країни у благах, необхідних для вирішення соціально-економічних проблем, підтримання обороноздатності країни та безпеки її громадян, забезпечення конституційних прав на освіту та охорону здоров'я, а також відтворення та збільшення валового внутрішнього продукту країни. Тому тендери мають бути чітко визначеними, проведення торгів та вибір переможця має бути контрольованим та прогнозованим, відповідати поточним запитам українського суспільства [16].

Ідея щодо автоматизації обробки підозрілих тендерів співпадає з пропозицією Педченка Н. С., Кудацького О. М. та Педченка М. Г. Однак відмінність полягає у кардинально різних підходах: зазначені автори пропонують автоматизувати цей процес на рівні вимог до подачі тендерних пропозицій [17].

Описані результати деяких з релевантних, на даний момент, досліджень виявляють широкий діапазон проблематики у сфері державних закупівель та підтверджують актуальність проведення дослідження, задане у межах даної наукової роботи.

Для всебічного дослідження переваг та недоліків, перспектив тощо державних закупівель в Україні, слід залучати якомога більший методичний інструментарій. У цій роботі основним методом дослідження є тематичне моделювання – техніка, що використовується в обробці природної мови (Natural Language Processing, NLP).

1.3. Обробка природної мови (NLP) як метод моніторингу діяльності у сфері публічних закупівель

Обробка природної мови – це розділ інформатики, у рамках якого вивчаються особливості існування та функціонування мов, а також можливості їх обробки за допомогою електронно-обчислювальної техніки для багатоцільового моделювання. Базою для їх реалізації є математичні, статистичні алгоритми та, зрештою, машинне навчання.

Спершу NLP використовувалися набори нескладних правил, які дозволяли виконувати поверхневу обробку простих текстових конструкцій, однак з розвитком цифрової індустрії, експоненційно прогресуючим науково-технологічним рухом, виникла проблема неможливості масштабування тих моделей під впливом зростаючих потоків даних, у тому числі – текстових. Тому логічним продовженням розвитку цієї галузі стало використання статистичних методів, машинного навчання. Це стало причиною появи нових великих моделей, які здатні обробляти та генерувати нові масиви тексту, засновані на правилах людської мови та загальної бази знань людства [18]. Відомим прикладом моделі, побудованої на принципах NLP, є ChatGPT (Chat Generative Pre-Trained Transformer).

Загалом, приклади використання NLP зручно систематизувати та представити у вигляді таблиці (табл. 1.1):

Таблиця 1.1 – Приклади використання обробки природної мови
(Розроблено автором на основі [18])

Приклад	Опис
Виявлення спаму	Найбільш застосовувані технології для знаходження спаму в електронних листах використовують обробку природної мови, а саме – класифікацію текстових фрагментів для аналізу наявності таких, що можуть сигналізувати про спам чи фішинг в отриманому листі. Виявлення спаму – одна з небагатьох проблем NLP, вирішення якої експерти вважають найбільш успішним.
Машинний переклад	Якісні онлайн-перекладачі. Тут NLP використовується для врахування міжмовних відмінностей для того, щоб максимально зберігати контекст, який може бути втрачено під час простого почергового перекладу слів.
Чат-боти та голосові помічники	Використовують розпізнавання тексту та голосу для генерації влучних відповідей на запити, що робляться користувачами таких додатків.
Аналіз настроїв у тексті	NLP є важливим бізнес-інструментом для аналізу рівня задоволеності товарами чи послугами. Такий підхід дозволяє масово аналізувати емоційне забарвлення відгуків, коментарів тощо та полегшує прийняття деяких управлінських рішень.
Підсумовування тексту	Підсумовування тексту використовує методи NLP для обробки величезних обсягів цифрового тексту і створення анотацій дослідницьких баз даних або для зайнятих читачів, які не мають часу на читання повного тексту.

Тематичне моделювання – це метод обробки природної мови, що використовується для автоматичного вираження кола тем у наборах текстових даних. Він є важливим для організації, розуміння та вилучення інформації з великих текстових масивів даних. Алгоритми тематичного моделювання виявляють приховані теми, аналізуючи закономірності повторюваності слів у документі [19, 20].

Таким чином, NLP як метод моніторингу діяльності у сфері публічних закупівель виражається через впровадження у цей процес алгоритмів тематичного моделювання. Це дасть можливість значно прискорити процес аналізу тендерів, що оголошуються через електронну систему Prozorro, шляхом автоматизації масового визначення закупівельної тематики.

Для досягнення поставленої мети необхідно зібрати інформацію з Prozorro, що стосується закупівель. Вочевидь, здійснювати ручний збір даних не є доцільним, тому пропонується розробити додаток, задачею якого буде автоматичний збір тендерів, оголошуваних органами державної влади.

1.4. Формування бази тендерів за допомогою інтерфейсу прикладного програмування Prozorro

Для збору тендерів було вирішено використовувати інтерфейс прикладного програмування (Application Programming Interface, API). Причинами для цього є наступне:

- Веб-скрапінг (збір даних за тегами, класами в HTML-коді) є достатньо повільним, що пов’язується з наступною проблемою;
- Динамічна структура сторінок зі списками тендерів на Prozorro. Під час скрапінгу у момент, коли завантажується нове оголошення, скрипт, без додаткових ускладнень, зупинятиме свою роботу.

Інтерфейс прикладного програмування (API) – це своєрідний дозвіл певної системи звертатися зовнішньому клієнту (додатку) до даних, що містяться на сервері, напряму за допомогою набору, встановлених її розробниками, правил [21]. Такий метод отримання даних характеризується відносною швидкістю, порівняно зі скрапінгом, веб-сторінок. Prozorro надає відкритий доступ до свого API, а також документацію до нього [22]. Тому переходимо до розробки програми для збору закупівельної інформації.

Для написання додатку зручно використовувати мову програмування Python. Крім базових методів цієї мови необхідно використати такі бібліотеки:

- Requests. Ця бібліотека потрібна для виконання запитів до API Prozorro [23];
- Pandas для обробки та збереження даних у форматі CSV [24];
- Time для додавання затримки у разі перевищення кількості запитів на сервер [25];
- Concurrent.futures. Цей модуль потрібен для забезпечення багатопотокового збору даних з метою підвищення продуктивності роботи скрипту [26].

Елементи коду програми для збору даних розміщено у додатках до кваліфікаційної роботи (Рисунок В.1 – В.4)

Розглянемо функції, що реалізуються під час викачування даних про закупівлі, а також елементи тендерів, що безпосередньо зберігаються у датасет.

Розроблений скрипт складається із трьох функцій, що постійно взаємодіють між собою. Серед них:

- save_to_csv. Ця функція виконує збереження даних, що збираються, після кожної вдалої ітерації та наприкінці виконання коду;
- get_tender_data_inside здійснює запит до Prozorro API для отримання детальної інформації про тендер;
- rec_tenders. Ця функція є основною та найбільшою функцією програми. Вона викликає попередньо описану функцію, завдяки багатопотоковості (9), водночас для декількох тендерів зі списку, відсортованого від найновіших (descending: true). Крім того, дана функція збирає інформацію з отриманих елементів, фільтрує тендери, оголошені органами влади (if procuring_entity and procuring_entity.get('kind') == 'authority') та зберігає дані про них у список, який потім передається у функцію save_to csv.

Серед даних про закупівлі, які було вирішено збирати, є такі:

- Заголовок;
- Сума, грн;
- Орган;
- Регіон;
- Опис;
- Посилання на сторінку тендера.

Для побудови тематичних моделей доцільно використовувати змінну «Заголовок» або «Опис», оскільки саме ці змінні містять текстову інформацію про закупівлю. Усі інші змінні мають допоміжну функцію: наприклад, якщо у певному тематичному кластері буде виявлено підозрілі ключові слова, то в наборі даних можна знайти тендери, що містять такий ключ, за допомогою функції швидкого пошуку. А змінні «Сума, грн», «Орган» та «Регіон» дадуть більш повне розуміння пріоритету перевірки конкретного тендера. Відповідно, змінна «Link» дозволить швидко перейти на сторінку цього оголошення.

У результаті збору даних про закупівлі, було сформовано набір даних (Рисунок С.1), обсяг якого становить 58451 спостереження. Часовий інтервал публікацій першого та останнього зібраного тендеру: 07.05.2024 – 13.03.2024. Цього має вистачити для досягнення поставленої мети роботи.

Однак, перед розглядом моделей, їх побудовою та аналізом результатів, сформуємо основні припущення дослідження.

1.5. Формування припущень дослідження

Як було визначено, сфера державних закупівель відіграє важливу роль у грошових надходженнях до державного, місцевого бюджетів та зумовлює стійкість і можливості зростання економіки України. Крім того, в умовах агресивної війни РФ проти України, через цю сферу активно реалізується збройне та незбройне забезпечення сил оборони України.

Проте, в умовах, з одного боку, тимчасово недостатнього рівня громадського контролю у закупівельних процесах, а з іншого – складності моніторингу тендерів на предмет їх відповідності Закону України «Про публічні закупівлі», особливо, внаслідок збільшення частоти публікації оголошень з 2022 року, дотримання принципів прозорості та чесності не вдається забезпечити повноцінно, не дивлячись на відкритість електронної системи публічних закупівель Prozorro.

Необхідно підкреслити, що ідея використання тематичного моделювання для моніторингу активності державних закупівель не включає подальшого обмеження доступу до платформи пересічним користувачам, ускладнення умов подачі тендерів чи цензурування тощо, а має на меті лише сприяти забезпеченню принципів чесності та прозорості у сфері для підвищення рівня економічного розвитку та обороноздатності України.

Виходячи з попередньо вказаного, визначено такі припущення щодо економіко-математичного моделювання державних закупівель через систему Prozorro за допомогою моделей тематичної кластеризації:

1. Ключові слова, згруповані у кластери, здатні доносити тематичне навантаження аналізованих тендерів до дослідника та бути вдало інтерпретованими на рівні людського сприйняття.

2. Тематичні моделі є ефективним інструментом для аналізу державної закупівельної діяльності.

3. Потенціал тематичних моделей не обмежується використанням лише для виявлення аномальної активності.

Сформульовані припущення мають бути перевірені через безпосереднє використання алгоритмів тематичного моделювання на практиці.

РОЗДІЛ 2. ПОБУДОВА ТЕМАТИЧНИХ МОДЕЛЕЙ ДЕРЖАВНИХ ЗАКУПІВЕЛЬ ЧЕРЕЗ СИСТЕМУ PROZORRO

2.1. Концептуальне моделювання алгоритмів латентного розміщення Діріхле та BERTopic

Алгоритми тематичного моделювання, які пропонується застосувати у даній роботі, нині є одними з найбільш широко використовуваних [27]. Причому, запропонована Дж. К. Прічардом, М. Стіфенсом і П. Доннелі у 2000 році, у контексті дослідження генотипів [28], Latent Dirichlet Allocation (LDA) – це вже класична модель, можливості якої ґрунтовно вивчені та описані науковою спільнотою. Сьогодні використовується фахівцями у багатьох сферах людської діяльності, здебільшого як техніка машинного навчання для задач екстракції ключових слів, що характеризують тематичний спектр сукупності текстів [29].

Алгоритм BERTopic, розроблений М. Гроотендорстом у 2020 році [30, 31], вважається найбільш прогресивним і надійним з точки зору кількісної оцінки [32]. Він використовує вже натреновані трансформери для виокремлення тем та формування тематичних кластерів [33]. Крім того, як показує застосування на практиці, він вимагає меншої кількості зусиль для налаштування моделей, порівняно з LDA.

Нижче пропонується розглянути узагальнену модель обробки тексту та виведення результатів за допомогою LDA та BERTopic (Рисунок 2.1):

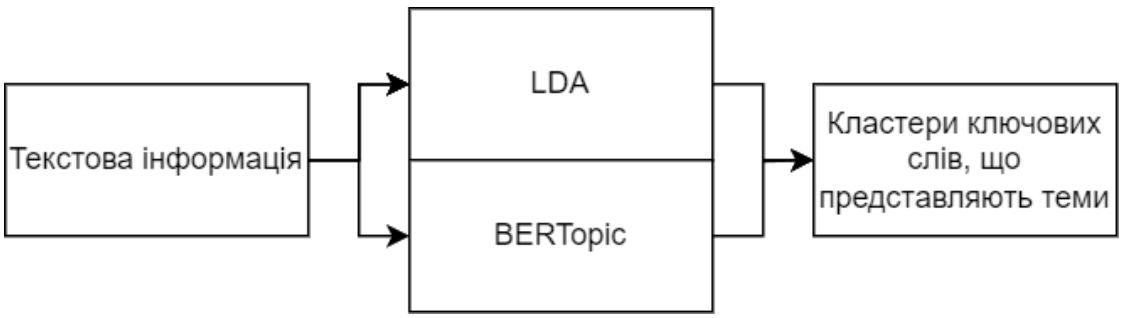


Рисунок 2.1 – Узагальнене представлення тематичного моделювання за допомогою LDA та BERTopic

(Розроблено автором)

Для кращого розуміння принципу роботи розглядуваних алгоритмів, пропонується реалізувати їх базові концептуальні моделі.

2.1.1. Концептуальна модель LDA

Латентне розміщення Діріхле – це стохастична модель, розроблена за парадигмою некерованого машинного навчання. Вона призначена для роботи з великими колекціями текстів (англ. – corpus) [34]. Текстові документи (спостереження), за такого підходу, сприймаються як випадкові суміші мовних одиниць за прихованими темами, кожна з яких характеризується відповідною частотою слів, які трапляються у масиві тексту.

Даний алгоритм складається з таких етапів генерування для кожного документа w у корпусі D:

1. Вибір кількості слів N у документах за розподілом Пуассона з параметром ξ. На цьому етапі визначаються довжини документів у колекції.
2. Тематичний розподіл θ у межах спостереження встановлюється згідно розподілу Діріхле з параметром α. Цей крок дозволяє встановити ймовірності наявності певної теми у конкретно взятому документі.
3. Для кожного з N слів w_n:
   - Вибір теми z_n з мультиноміального розподілу (більше, ніж двох) θ тем.
   - Вибір w_n слова, що репрезентуватиме z_n тему з імовірністями β.

Варто зауважити, що в описаній моделі зроблено кілька припущень, що дозволять спростити сприйняття концепції досліджуваного алгоритму.

Перш за все, розмірність k-розподілу Діріхле (відповідно, розмірність змінної z, що характеризує теми) вважається фіксованою. Це означає, що до створення безпосередньо моделі LDA, кількість тем, на яку буде поділено

корпус може бути навіть емпірично визначена дослідником. Однак у даному проекті використано метрику когерентності слів. Даний показник описано у підрозділі 2.2.

Другий аспект, вартий уваги, полягає у тому, що ймовірності слів визначаються матрицею β розміром $k \times V$ де $\beta_{ij} = p(w_j = 1|z_i = 1)$, яка до навчання тематичної моделі вважається фіксованою. Тобто β – це матриця ймовірностей, що характеризують вірогідність потрапляння того чи іншого слова зі словника унікальних слів V аналізованої колекції у певну тему.

По-третє, припущенням про Пуассонівський розподіл, здебільшого, нехтують, надаючи перевагу реальній кількості слів, що містяться у спостереженнях (документах).

Випадковий розподіл тем θ (вектор) набуває значень у (k-1)-розмірному просторі, якщо $\theta_i \geq 0$ і $\sum_{i=1}^{k} \theta_i = 1$. Тоді функція густини ймовірності має вигляд (2.1):

$$p(\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1}, \quad (2.1)$$

де α – це вектор, що складається з $\alpha_i > 0$ і $\Gamma(x)$ – це гамма-функція.

Параметр α та β спричиняють таку спільну ймовірність розподілу тем θ, сукупності тем z та N слів w (2.2):

$$p(\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(\theta) p(w_n|z_n, \beta), \quad (2.2)$$

де $p(\alpha)$ – визначення ймовірності для θ;

$\prod_{n=1}^{N} p(\theta)$ – для кожного слова n спостереження визначається ймовірність потрапляння у конкретну тему z_n з розподілу θ;

$\prod_{n=1}^{N} p(z_n, \beta)$ – визначення ймовірності слова w_n за теми z_n зі словника матриці β.

Інтегруючи по θ і підсумовуючи по z, отримаємо граничний розподіл документа (2.3):

$$p(\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(\theta)p(w_n|z_n, \beta)\right)d\theta. \quad (2.3)$$

На даному етапі стає можливим обчислення ймовірності корпусу документів (спостережень у вибірці даних, що містять текстову інформацію) (2.4):

$$p(\alpha, \beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(\theta_d)p(w_{dn}|z_{dn}, \beta)\right)d\theta_d. \quad (2.4)$$

Таким чином, відбувається оцінка того, наскільки добре модель LDA пояснює оброблювану колекцію текстів, на основі чого слова збираються у тематичні кластери і кінцевий користувач може ознайомитися зі зведенням тем у досліджуваному наборі текстів, водночас маючи розуміння, наскільки вони близькі між собою за змістом.

Графічно модель LDA представлена на рисунку нижче (Рисунок 2.2):

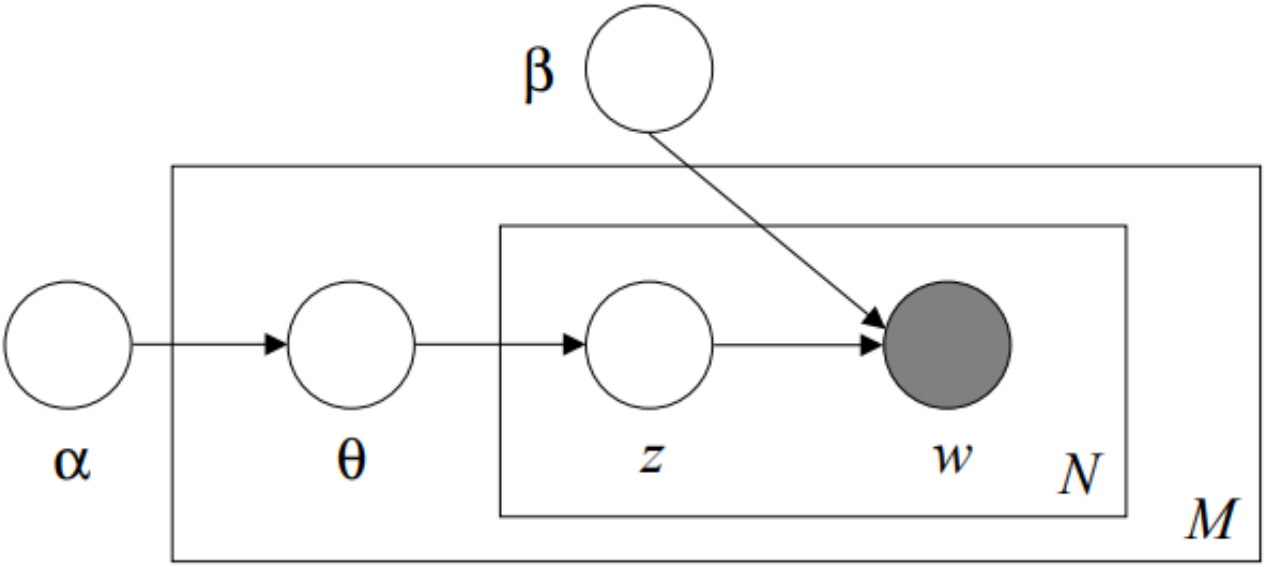


Рисунок 2.2 – Принцип функціонування LDA [35]

Вищенаведена схема показує, що змінні моделі розподілені на три рівні:

1. Рівень корпусу. На даному рівні знаходяться параметри α і β, де α визначає розподіл Діріхле у документах, а β є матрицею, яка характеризує розподіл слів для кожної теми. Зазначені параметри винесені на цей рівень, оскільки залишаються постійними величинами для усієї модельованої колекції текстів.

2. Рівень документа (одного тексту з колекції). Він представлений змінною θ, яка характеризує розподіл тем в окремо взятому спостереженні d. Він визначається один раз у межах одного тексту.

3. Рівень слова. На внутрішньому рівні змінна z_{dn} вказує на тему, до якої належить слово w_{dn} у документі d на місці n.

Наостанок необхідно зробити уточнення про те, що модель виконує імовірнісну оцінку належності до певної теми не одного тексту з вибірки, а кожного слова в конкретному тексті. Таким чином, одне спостереження модель може розподіляти на декілька тем, при цьому зберігаючи контекст усієї колекції текстів та надаючи розуміння тематичного спектру дослідника вибірки [35].

Надалі пропонується розглянути принцип роботи другого вказаного алгоритму – BERTopic.

2.1.2. Концептуальна модель BERTopic

Ключовою відмінністю алгоритму BERTopic від попереднього є те, що він використовує вже навчені мовні моделі різного об’єму, тоді як LDA працює з наявною вибіркою текстів. Однак BERTopic підтримує також напівкероване, динамічне моделювання тем, а також моделювання з учителем [36].

Найкращим способом зображення алгоритму BERTopic є послідовність, що складається із п’яти етапів (Рисунок 2.3).

Згідно порядку роботи алгоритму, першим етапом є конвертація тексту в числове вираження. Тут використовуються трансформатори речень, які вбудовують текст у векторний простір так, що подібні фрагменти визначаються завдяки косинусній подібності [37]. Це сприяє якісному виконанню завдання кластеризації. BERTopic влаштований таким чином, що у ході побудови моделі може використовуватися будь-який серед існуючих трансформерів речень [38]. У випадку моделювання державних закупівель через систему Prozorro, було вирішено використовувати «paraphrase-multilingual-mpnet-base-v2» [39], оскільки вона є найбільшою попередньо натренованою мовною моделлю, яка навчалася на великій кількості кириличних текстів і, у тому числі, текстах українською мовою. Однак, на практиці, у такого підходу є вагомий недолік, у разі, якщо корпус має великий об’єм: модель будуватиметься відчутно довше, оскільки пропускна здатність цього трансформатора становить лише 2500 речень на секунду в ідеальних умовах (з використанням графічного процесора NVIDIA GPU V100).

Рисунок 2.3 – Принцип функціонування BERTopic
(Розроблено автором на основі [33])

Другим кроком у процесі моделювання є зменшення розмірності отриманих числових представлень, оскільки, за замовчуванням, під час кластеризації даних з високою розмірністю може виникнути явище прокляття розмірності [40]. BERTopic підтримує різні підходи до вирішення даної задачі, однак у даній роботі використано метод UMAP (Uniform Manifold Approximation and Projection) [41]. Це техніка, використання якої допомагає зберегти велику частину структури набору даних при зменшенні його розмірності. Цю структуру важливо зберегти, оскільки вона містить інформацію, необхідну для ефективного формування кластерів семантично схожих документів.

Попередньо уникнувши проблеми прокляття розмірності, модель виконує задачу кластеризації за допомогою HDBSCAN – методу кластеризації на основі щільності отриманих числових даних. Перевага використовуваного на даному етапі підходу полягає у тому, що HDBSCAN здатен формувати кластери дуже різної форми (Рисунок 2.4), що дозволяє уникати білого шуму в отримуваних кластерах.

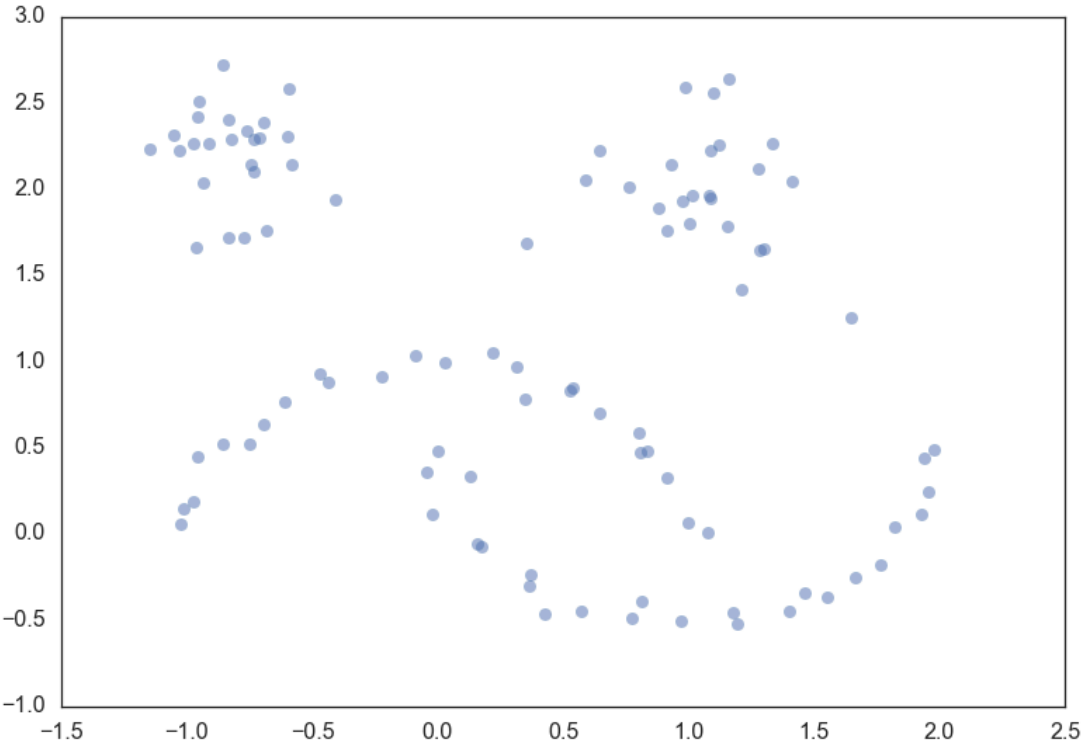


Рисунок 2.4 – Приклад результату кластеризації методом HDBSCAN [42]

У контексті тематичного моделювання це означає, що зрозумілість теми, висвітлюваної тією чи іншою групою слів, у результаті побудови моделі, буде вищою, порівняно з потенційними результатами, отриманими, наприклад, під час кластеризації методом k-середніх. Варто зазначити, що на даному етапі кластеризація проводиться на рівні цілих спостережень (документів), а не окремих слів. Цим зумовлений наступний крок – токенізація.

На цьому кроці відбувається частотний аналіз слів у кластеризованих документах, на основі чого формується «мішок слів», з якого пізніше

визначається те, які слова репрезентуватимуть певну тему. Однак для проведення частотного аналізу необхідно реалізувати токенізацію документів. Токенізація – це процес розбиття суцільного тексту на окремі одиниці (слова). Фактично одне слово – це токен. Це важливо, оскільки саме окремі слова включаються у тематичні кластери, які подаються кінцевому користувачеві.

З отриманого представлення «мішка слів» на попередньому кроці необхідно встановити, що відрізняє один кластер від іншого. Визначити те, які слова є типовими для одного з кластерів, але не описують всі інші кластери. Для вирішення цієї задачі використовується метрика c-TF-IDF. З її допомогою виділяються найважливіші слова з кожного кластера, які, у свою чергу, описують теми з досліджуваного корпусу (2.5):

$$W_{x,c} = \|tf_{x,c}\| \log log\left(1 + \frac{A}{f_x}\right), \quad (2.5)$$

де $W_{x,c}$ – значущість слова x у кластері c;

$\|tf_{x,c}\|$ – нормалізовані частоти слів x у кластері c;

A – середня кількість слів, що включалися у кластери;

f_x – частота слова x у всіх кластерах

Тут виділяється частота слова x у класі c, де c відноситься до кластера, сформованого раніше. В результаті отримуються частоти tf на основі класів. Вони є нормалізованими, щоб врахувати різницю в розмірах тем. Потім береться логарифм одиниці плюс середня кількість слів у класі A, поділений на частоту слова x у всіх класах. Одиниця додається до логарифму, щоб значення були додатними. Надалі виконується візуалізація, щоб дослідник міг ознайомитися із результатами моделювання [33].

Як зазначалося раніше, BERTopic, як правило, має вищі оцінки моделі за зв’язністю (когерентністю) слів. Однак, коли мова йде про майнінг тексту, то

кількісне оцінювання якості NLP-моделей відходить на другий план. Натомість, найбільш важливою є легкість інтерпретації тематичних складових людиною.

Згідно дослідження Г. Аксельборна та Дж. Бергрена про використання технік тематичного моделювання для розуміння клієнтів, результати для BERTopic та LDA відрізнялися, значною мірою, щодо оптимальної кількості тем та людської інтерпретації якості отриманих моделей. LDA дав значно кращі результати у сенсі людської інтерпретації якості тем. З іншого боку, за допомогою BERTopic було отримано модель з кращим показником узгодженості слів [43].

Враховуючи вищезазначене, робити попередній висновок щодо доцільності використання тієї чи іншої моделі для тематичного аналізу державних закупівель не є доречним. Тому на даному етапі пропонується реалізувати фізичне проектування обох економіко-математичних моделей державних закупівель через систему Prozorro, проаналізувати отримані результати та дати їм економічну інтерпретацію.

2.2. Побудова NLP-моделей та інтерпретація отриманих результатів

Виконання задачі тематичного моделювання, як і збору даних з електронної системи публічних закупівель Prozorro, пропонується проводити за допомогою мови програмування Python. Моделювання виконуватиметься у межах таких етапів:

1. Розробка моделі LDA:
   - Попередня обробка тексту;
   - Визначення кількості тем за допомогою коефіцієнту узгодженості (когерентності) слів;
   - Побудова моделі.
2. Розробка моделі BERTopic.
3. Візуалізація та аналіз результатів.

Всі ілюстративні матеріали, що демонструють усі блоки використаного програмного коду, розміщено у додатках до кваліфікаційної роботи (Рисунок D.1 – D.7).

Попередня підготовка корпусу необхідна для побудови моделі на основі алгоритму латентного розміщення Діріхле, оскільки вона дозволить отримати кращі результати тематичного моделювання [44]. Крім того, ті бібліотеки та їх методи, що використовуватимуться у даній роботі, у процесі побудови моделі на основі алгоритму LDA, вимагають попередньо обробленого тексту. Цей етап включає у себе видалення зайвих пробілів, приведення тексту до нижнього регістру, видалення несуттєвих фрагментів тексту, токенізацію, видалення стоп-слів, стемінг або лематизацію [45].

Перед виконанням цього етапу, вочевидь, необхідно використати модуль os [46] для зміни робочої директорії на ту, де зберігаються матеріали дипломної роботи, включно із зібраним датасетом. Для зчитування набору даних використаємо бібліотеку Pandas [24]. Зразок структури даних, завантажених у проект, зображено нижче (Рисунок 2.5):

Заголовок	Сума, грн	Орган	Регіон	Опис	Link
Колесо для тачки	570	Управління поліції охорони у Львівській області	Львівська область	nan	https://prozorro.gov.ua/tender/f79afa798f12412aabaaa3c7cde27be9
Послуги з організації перевезення відправлень	49998	Бахмутська міська рада	Донецька область	nan	https://prozorro.gov.ua/tender/dabafb4e7a354b8b8705e62190d00347
Послуги з благоустрою населених пунктів-утримання зелених насаджень	9.49469e+06	Департамент благоустрою та інфраструктури Дніпровської міської ради	Дніпропетровська область	nan	https://prozorro.gov.ua/tender/1807d654a37d4d76a9a58c440a635edc
Послуги з поточного ремонту покрівлі шляхом…	16175	Головне управління Держгеокадастру в Івано-Франківській області	Івано-Франківська область	nan	https://prozorro.gov.ua/tender/d1f6cc21fd77452dbd01932b0d17136c
Крісла	14670	Головне управління Національної поліції в Рівненській області	Рівненська область	nan	https://prozorro.gov.ua/tender/0a480424ec9e49ccaf2310c22eda8570

Рисунок 2.5 – Зразок завантажених даних

Як видно з рисунка вище, перші п’ять спостережень датафрейму у стовпці «Опис» не містять даних. Проаналізуємо склад набору даних, використавши метод info() (Рисунок 2.6). У зібраному наборі даних (tenders_df) міститься 58451 спостереження, однак у трьох із шести стовпців присутні пропущені значення. Найбільше їх саме у стовпці «Опис»: він містить лише 3924 ненульових значення. Враховуючи цю обставину, використовувати дану змінну в якості колекції текстів не є доречним рішенням. Оскільки змінна «Заголовок» є найбільш повно представленою і такою, що дає уявлення про предмет

оголошуваної закупівлі, було вирішено використовувати саме її у якості корпусу.

```
In [3]: tenders_df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 58451 entries, 0 to 58450
Data columns (total 6 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Заголовок  58451 non-null  object
 1   Сума, грн  58441 non-null  float64
 2   Орган      58451 non-null  object
 3   Регіон     56557 non-null  object
 4   Опис       3924 non-null   object
 5   Link       58451 non-null  object
dtypes: float64(1), object(5)
memory usage: 2.7+ MB
```

Рисунок 2.6 – Зведення про досліджуваний датасет

Створивши колекцію документів (tenders_headlines), переходимо до попередньої обробки тексту з метою побудови моделі LDA. Якщо приведення тексту до нижнього регістру, видалення зайвих пробілів на початку, наприкінці рядка та між словами здійснюється за допомогою базових функцій lower(), strip() та split() відповідно, то для видалення шуму з тексту у вигляді посилань, знаків пунктуації, символів, кодів з Єдиного закупівельного словника [47] тощо необхідно використати можливості регулярних виразів (модуль re) [48].

Для токенізації тексту зручно застосовувати токенізатор слів з інструментарію для обробки природної мови (nltk) [49]. На перший погляд, набори токенів уже підготовлені до безпосереднього моделювання тем (Рисунок 2.7). Однак, зупинивши підготовку тексту на цьому етапі, буде отримано кластери тем з високим рівнем складності розуміння тематичного спектру аналізованої колекції текстів. Причиною для цього виступає наявність токенізованих прийменників, сполучників, певних абревіацій чи, загалом, слів, які не мають основного смислового навантаження у тексті. Фактично такі токени, через високу їх частотність, з більшою вірогідністю репрезентуватимуть кластери, чим знижуватимуть якість інтерпретабельності виокремлених тем.

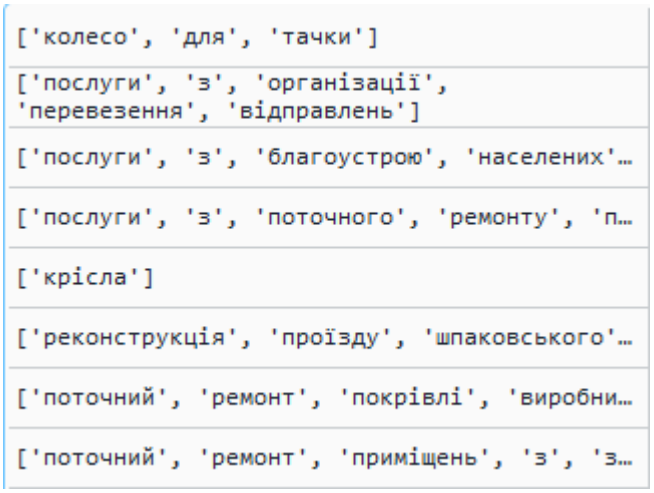


Рисунок 2.7 – Токенізовані документи

З огляду на попередньо викладене, наступним кроком, у межах попередньої обробки документів, буде видалення стоп-слів. Для цього використано готовий список стоп-слів українською мовою [50], емпірично розширений специфічними, для сфери публічних закупівель, зразками.

Завершальним етапом попередньої обробки тексту є стемінг або лематизація. Це два процеси, що мають однакову мету застосування – приведення слів у різних відмінках до словникової форми, однак різні підходи для її досягнення. Стемінг – це процес видалення зі слів суфіксів та закінчень. Лематизація, за умови успішного застосування, гарантовано приводить слово до початкової його форми [51]. Перший з двох підходів є більш жорстким і не завжди після видалення суфікса приводить слово до початкової форми. Однак інший значно рідше успішно застосовується, особливо до мов з кириличною системою письма. У даному випадку, обрано лематизацію (WordNetLemmatizer), оскільки таким чином має зберегтися більший обсяг інформації.

Після повного циклу попередньої обробки корпусу отримано результат, графічно відтворений нижче (Рисунок 2.8). Дані у такому вигляді вже є придатними до використання для побудови та навчання LDA моделі.

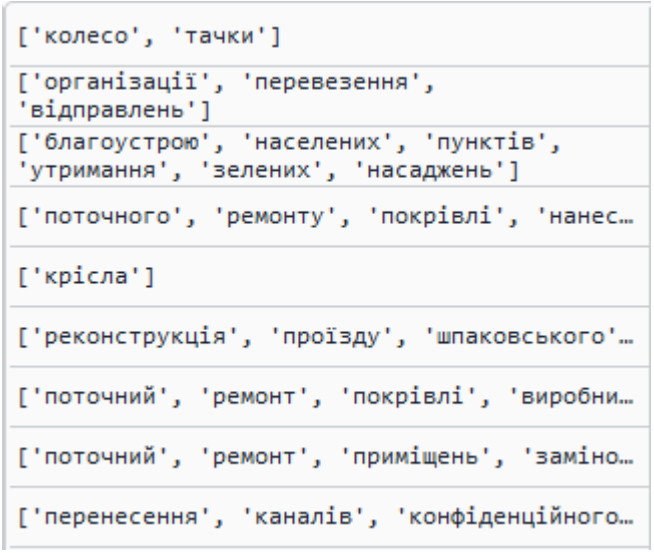


Рисунок 2.8 – Зразок документів, готових до моделювання

Для створення першої тематичної моделі зручно використовувати інструментарій бібліотеки Gensim [52], зокрема LdaModel. За допомогою параметра num_topics можна вказати будь-яке значення кількості кластерів для розбиття [53]. Проте у даній роботі у якості орієнтиру взяті не емпіричні уявлення щодо оптимальної кількості тем, а коефіцієнт зв’язності слів, попередньо згаданий у даній роботі.

Зв’язність (когерентність, узгодженість) – це тест, який можна використовувати для оцінки ефективності тематичної моделі. Когерентність зазвичай використовується для аналізу взаємозв'язку між словами або подібності між ними. У тематичному моделюванні узгодженість вимірює якість даних, порівнюючи семантичну схожість між словами, що часто повторюються в темі. Оцінка зв'язності - це шкала від 0 до 1, в якій хороша когерентність (висока схожість) має оцінку 1, а погана когерентність (низька схожість) має оцінку 0 [54]. Необхідно зазначити, що гіпотетична узгодженість корпусу, рівна одиниці, означатиме те, що це набір однакових слів. Таким чином, зроблено наступне припущення: кількість тем (кластерів) з рівнем зв’язності слів у темах близько 50 – 55% є прийнятною і може бути використана для фінальної моделі.

На цьому кроці складається словник унікальних слів за допомогою Dictionary (модуль з бібліотеки Gensim), виконується частотний аналіз слів.

Після цього розроблено цикл for, для якого було експертним шляхом встановлено діапазон кількостей тем для розрахунку когерентності від 2 до 80 включно. У рамках цього циклу будуються моделі для кожної кількості тем із вказаного діапазону, для кожної з них розраховується коефіцієнт когерентності.

У результаті розрахунків було отримано найкращу узгодженість слів у кластерах на рівні 47,98% для моделі з кількістю тем: 44. Відповідно, фінальну LDA модель було побудовано саме з таким числом тематичних кластерів.

Як зазначалося раніше, BERTopic потребує значно менше налаштувань, не вимагає попередньої обробки тексту тощо завдяки кардинально відмінному підході відносно попередньої (використання вже навчених моделей вбудовування).

Переходимо до розгляду результатів моделювання, їх пояснення та порівняння. Нижче наведено карти відстаней між темами для BERTopic та LDA моделей (Рисунок 2.9, 2.10). Варто зазначити, що для візуалізації міжтемних відстаней LDA використовувався окремо імпортований інструмент pyLDAvis [55].

Одразу відмітимо різницю у підборі кількості тем при реалізації NLP моделей: за допомогою BERTopic в автоматичному режимі було виділено набагато більше тем (1422), порівняно із заданою кількістю для моделі LDA (44). Це пов’язано з тим, що дана модель концентрується на досягненні якнайвищого рівня когерентності слів, а отже – максимальному збігу ознак у структурах лексичних значень або їх еквівалентності [56]. Тому теми тут мають менші обсяги токенів.

Із вищезазначеного можна зробити припущення, що BERTopic доцільно використовувати для більш детального аналізу субтем, Це може бути корисно для контролю корупційних ризиків, стимулювання оптимізації та перерозподілу витрат бюджетних коштів в умовах війни тощо.

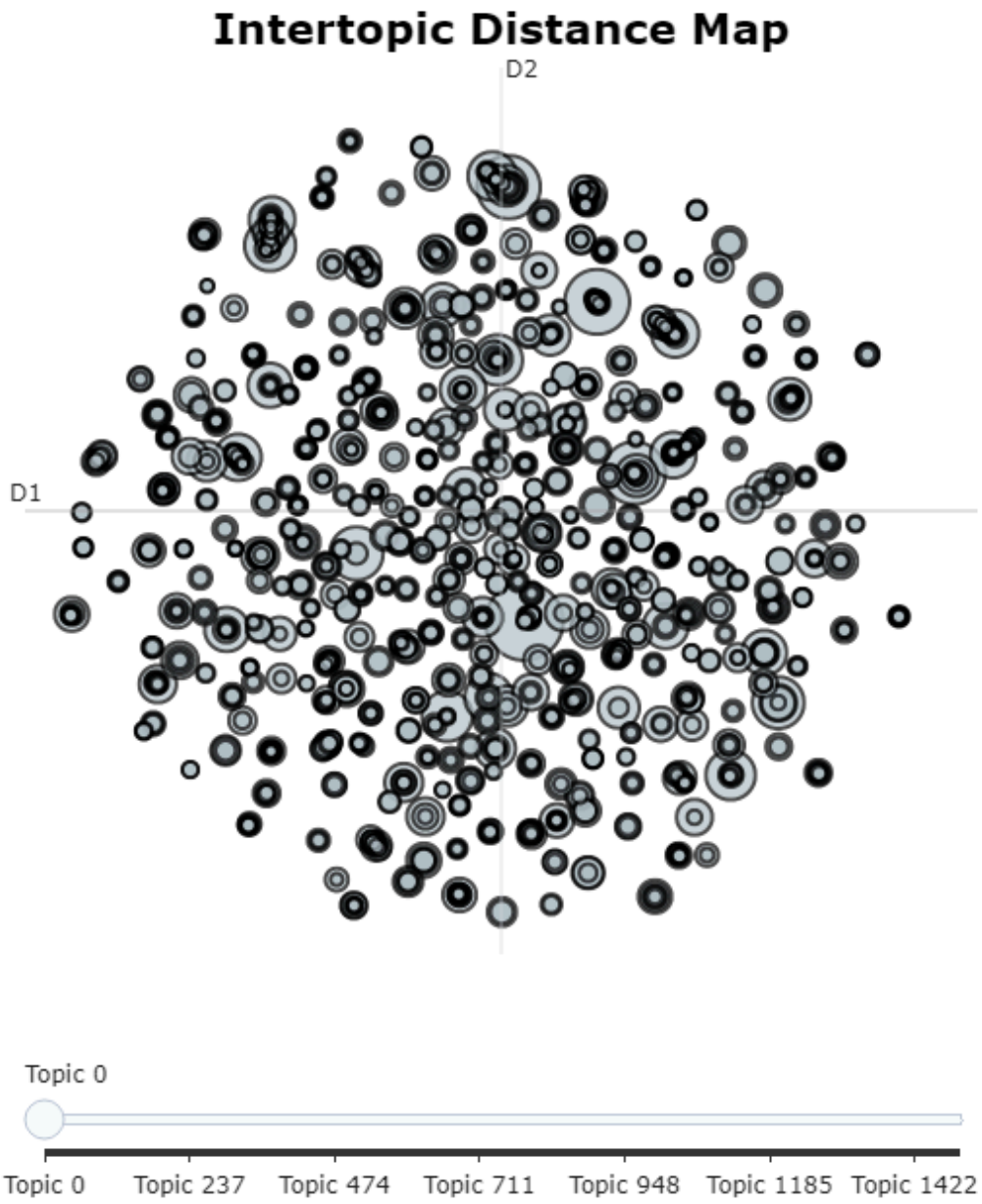


Рисунок 2.9 – Карта відстаней між темами BERTopic моделі

Припущення щодо LDA моделі є таким, що вона дає більш цілісне, але узагальнене уявлення про коло предметів закупівель, оголошення про які потрапили у сформовану вибірку даних з Prozorro. Це може бути корисно для аналізу потреб у межах галузей національної економіки України, соціально-економічної системи країни.

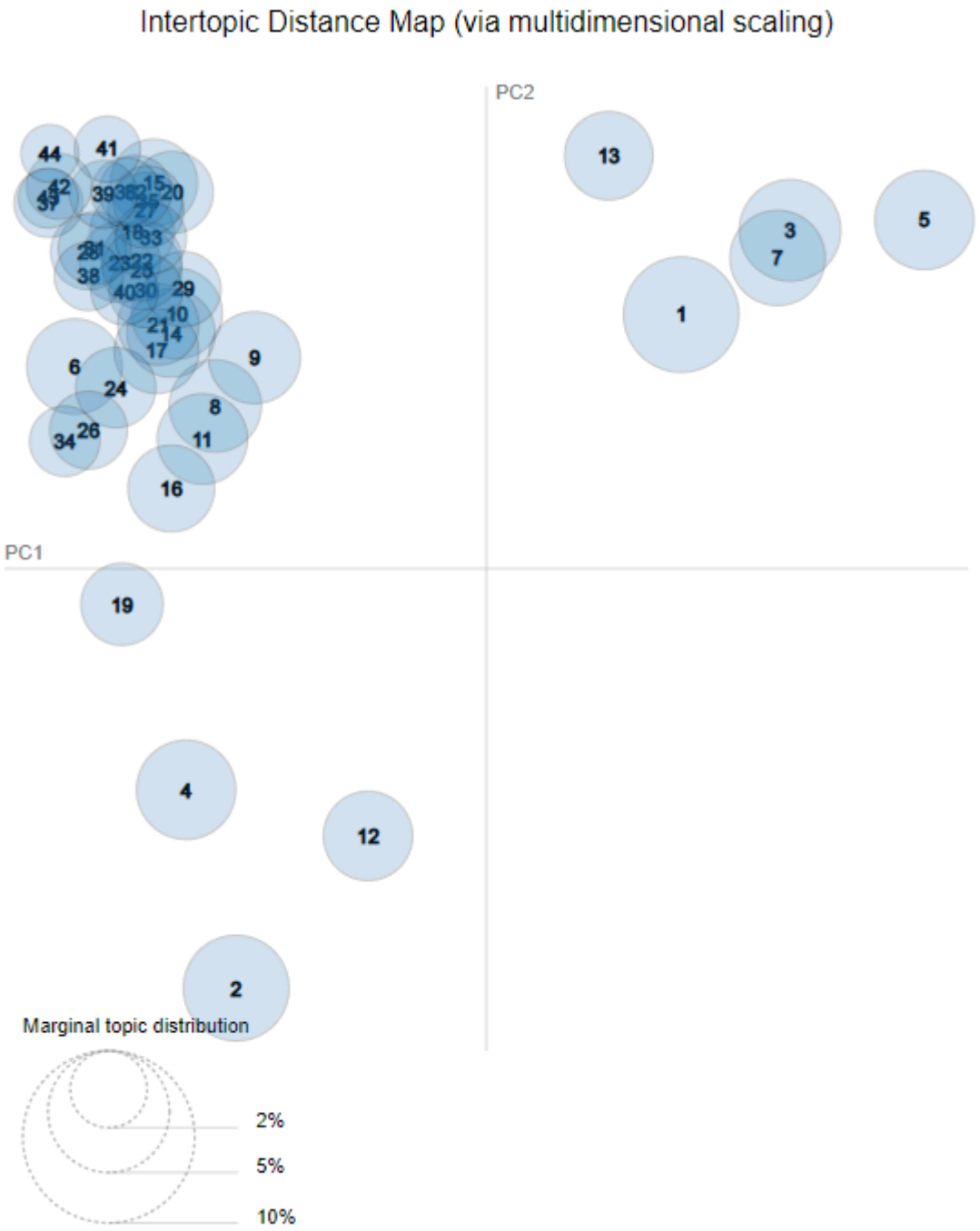


Рисунок 2.10 – Карта відстаней між темами LDA моделі

Інакше кажучи, побудовані моделі обидві можуть бути використані з різною метою. LDA модель підходить для аналізу тенденцій потреб, а BERTopic – для аналізу аномальної активності. Порівняємо приклади діаграм найбільш релевантних термінів у межах конкретних кластерів обох моделей (Рисунок 2.11, 2.12).

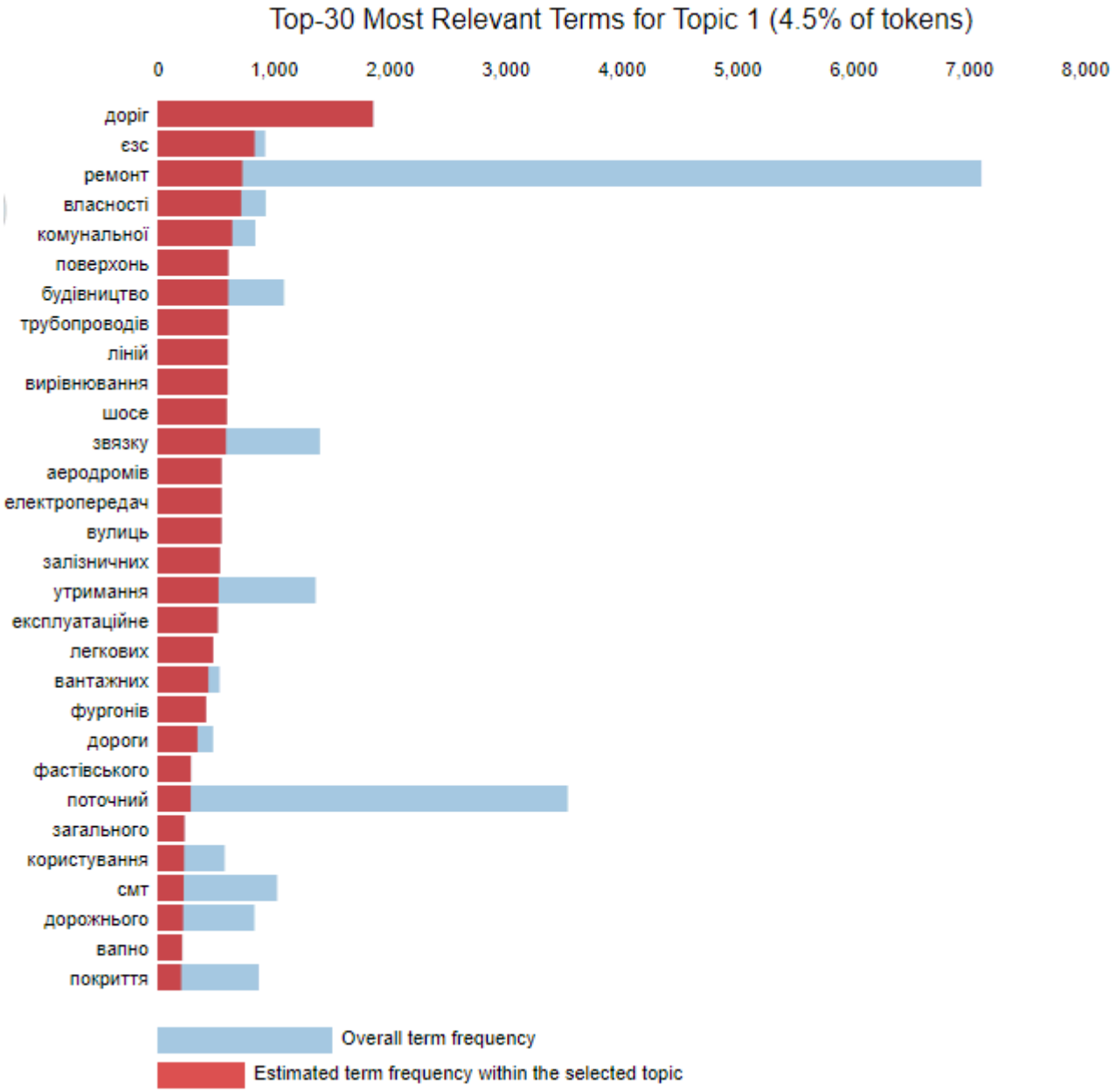


Рисунок 2.11 – Найбільш релевантні терміни для теми 1 (LDA)

На рисунку 2.11 зображено ключові слова, що характеризують найбільший кластер (номер 1, містить 4,5% токенів), утворений моделлю. Найменший кластер, відповідно, знаходиться під номером 44. Як стає зрозуміло зі складу слів у першому кластері, він характеризує тематику, яка стосується комунальних послуг та матеріально-технічної бази для їх здійснення. Вочевидь, житлово-комунальне господарство є невід'ємною частиною нормального функціонування економічної (забезпечення водою, газом, електроенергією підприємств тощо), соціальної (забезпечення комфортних умов для існування

населення) сфер і держави загалом, тому частка державних закупівель тут залишатиметься високою. Крім того, досить близькими до цієї теми є кластери під номерами 7 та 3 (Рисунок 2.10), у яких тематика стосується обслуговування та розбудови житлового фонду України (Рисунок Е.1, Е.2).

Таким чином, збираючи з рівними інтервалами часу сукупності текстів оголошень однакового об’єму, можна аналізувати актуальність конкретної теми за допомогою часток токенів корпусу, охоплюваних нею. Це дозволить формувати часові ряди на їх основі, прогнозувати зростання чи скорочення обсягів потреб у тій чи іншій сфері та, відповідно до цього, коригувати внутрішню політику держави, націлену на розробку програм з поліпшення становища відповідної системи.

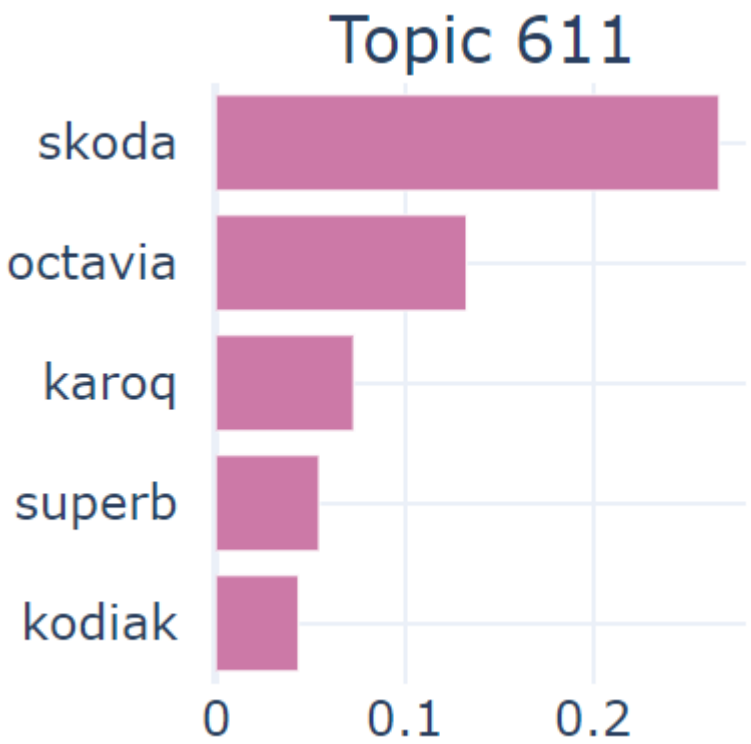


Рисунок 2.12 – Найбільш релевантні терміни для теми 611 (BERTopic)

Проаналізувавши кластери, утворені під час BERTopic моделювання (Рисунок F.1), було виявлено тему, що характеризує оголошення про закупівлі, пов’язані з автомобілями Skoda (Рисунок 2.12). Перш за все, вартує уваги

значно вищий рівень конкретики всередині тематичної групи, порівняно з прикладом LDA моделі (Рисунок 2.11), але це не є основною причиною зацікавленості темою номер 611. Першочерговою причиною її розгляду є можливе порушення відповідними тендерами таких принципів, викладених у статті 5 Закону України «Про публічні закупівлі»:

- Максимальна економія, ефективність та пропорційність;
- Запобігання корупційним діям і зловживанням.

Крім того, згідно частини 4 статті 23 Закону України «Про публічні закупівлі», технічні специфікації товару чи послуги не повинні містити посилання на конкретні марку чи виробника [4]. Зазначене вище пояснюється тим, що подібна конкретизація виключає можливість учасникам тендеру зробити інші, більш вигідні пропозиції, що дозволило би скоротити витрати бюджетних коштів.

У зібраному датасеті було виявлено 60 заголовків, що містять ключ «skoda» і потенційно мають порушення принципів, викладених у ЗУ «Про публічні закупівлі».

На прикладі цієї моделі було продемонстровано потенціал її застосування для ефективнішої редукції ризиків, пов’язаних із корупційними схемами; оптимізації витрат коштів бюджету тощо, оскільки виявлення одного кластера, що включає підозрілі токени, дає розуміння, за яким ключем можна знайти посилання чи ряд посилань, що ведуть на сторінки тендерів, які можуть вимагати додаткових ревізій чи оскарження.

На даному етапі пропонується перейти до висновків з проведеного дослідження кваліфікаційної роботи.

ВИСНОВКИ

У результаті виконання роботи було побудовано та продемонстровано доцільність використання тематичних моделей для аналізу державної закупівельної діяльності: мети дослідження досягнуто.

Було виявлено, що модель на основі алгоритму BERTopic придатна для пошуку маркерів, які можуть вказувати на корупційні прояви чи використання державних коштів з низьким рівнем корисності для суспільства. Також встановлено, що модель LDA може використовуватися для аналізу потреб у межах галузей національної економіки України, соціально-економічної системи країни. Таким чином, висунуті гіпотези підтвердилися.

У ході виконання наукової роботи було досягнуто поставлених завдань:

- Описано передумови та актуальність тендерних досліджень;
- Проведено бібліометричний аналіз релевантних наукових досліджень у сфері державних закупівель;
- Представлено обробку природної мови (NLP) як метод моніторингу діяльності у сфері публічних закупівель;
- Сформовано базу тендерів за допомогою інтерфейсу прикладного програмування Prozorro;
- Сформовано припущення дослідження;
- Реалізовано концептуальне моделювання алгоритмів тематичного моделювання, таких як латентне розміщення Діріхле та BERTopic;
- Побудовано NLP-моделі та інтерпретовано отримані результати.

Стейкхолдерами проведеного дослідження можуть бути фахівці з таких відомств:

- Національне агентство з питань запобігання корупції (НАЗК) – антикорупційний орган, що має превентивну функцію;
- Національне антикорупційне бюро України (НАБУ) – орган, що розслідує корупційні злочини, пов’язані зі значними сумами державних коштів;

– Департамент сфери публічних закупівель Міністерства економіки України.

Крім того, це може бути цікаво науковій спільноті, представники якої працюють над дослідженнями у цій сфері, та навіть активним громадянам України.

Перспективним напрямком подальших досліджень може бути моделювання за асоціативними правилами для встановлення зв’язків з іншими змінними, включених у збирані дані з Prozorro. Також, для полегшення роботи з такими моделями, доцільною є розробка повноцінного додатку, в основі якого лежатимуть методи тематичного моделювання.

На цьому виконання наукової роботи завершено.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Пінькас Г. І. Основні етапи розвитку державних закупівель в Україні. *Проблеми і перспективи розвитку банківської системи України*. 2007. № 21. С. 53–60. URL: https://essuir.sumdu.edu.ua/bitstream-download/123456789/55653/5/Pinkas_Derzhavni_zakupivli.pdf.

2. Зельдіна О. Р., Курепіна О. Ю. Розвиток законодавства України щодо публічних закупівель. *Економіка та право*. 2019. Т. 2, № 53. С. 70–77. URL: https://economiclaw.kiev.ua/index.php/economiclaw/article/view/850/827.

3. ЩО ТАКЕ ДЕРЖАВНІ ЗАКУПІВЛІ - E-TENDER.UA. *E-TENDER*. URL: https://e-tender.ua/news/shcho-take-derzhavni-zakupivli-e-tenderua-1319.

4. Про публічні закупівлі : Закон України від 25.12.2015 р. № 922-VIII : станом на 19 квіт. 2024 р. URL: https://zakon.rada.gov.ua/laws/show/922-19#Text.

5. Головна | Prozorro. *Prozorro*. URL: https://prozorro.gov.ua/en.

6. Bozhenko, V., Mynenko, S. & Shtefan, A. (2022). Financial Fraud Detection on Social Networks Based on a Data Mining Approach. Financial Markets, Institutions and Risks, 6(4), 119-124.

7. Ніколов Ю. Тилові пацюки Міноборони під час війни «пиляють» на харчах для ЗСУ більше, ніж за мирного життя. *Зеркало недели | Дзеркало тижня | Mirror Weekly*. URL: https://zn.ua/ukr/economic-security/tilovi-patsjuki-minoboroni-pid-chas-vijni-piljajut-na-kharchakh-dlja-zsu-bilshe-nizh-za-mirnoho-zhittja.html.

8. ДБР затримало підприємців, які продавали серед іншого військовим яйця по 17 гривень (ВІДЕО) - Державне бюро розслідувань. *Головна - Державне бюро розслідувань*. URL: https://dbr.gov.ua/news/dbr-zatrimalo-pidpriemciv-yaki-prodavali-sered-inshogo-vijskovim-yajcya-po-17-griven.

9. Prozorro відкриває інформацію про незбройні оборонні закупівлі | Міністерство економіки України. *Міністерство економіки України*.

URL: https://www.me.gov.ua/News/Detail?lang=uk-UA&id=a3312b2f-66ec-4aaa-873b-62294333ca0f&title=Zakupivli.

10. Скандальні тендери. *Інформатор UA*. URL: https://informator.ua/uk/tags/skandalni-tenderi.

11. Департамент сфери публічних закупівель та конкурентної політики | Міністерство економіки України. *Міністерство економіки України*. URL: https://www.me.gov.ua/Documents/Detail?lang=uk-UA&id=ff65b101-93c0-405b-a231-616e7692d885&title=Departament.

12. government procurement. *Google Trends*. URL: https://trends.google.com/trends/explore?date=2023-01-01%202024-05-27&geo=UA&q=/g/120m5_ps&hl=en-US.

13. Svitlychnyi O. P., Artemenko O. V., Volkova L. O. Some features of public procurement during matrial law. *Analytical and Comparative Jurisprudence*. 2023. No. 1. P. 309–316. URL: https://doi.org/10.24144/2788-6018.2023.01.50.

14. Dluhopolskyi O., Chaprak Y. Key directions of research and dilemmas in the sphere of public procurement: theoretical framework. *Innovation and Sustainability*. 2023. No. 1. P. 51–63. URL: https://doi.org/10.31649/ins.2023.1.51.63.

15. Кильницька Є. В., Глухова С. В., Колодяжна Т. В. Дослідження тенденцій ринку публічних закупівель під час дії воєнного стану. *Проблеми сучасних трансформацій. Серія: економіка та управління*. 2023. № 8. URL: https://doi.org/10.54929/2786-5738-2023-8-10-01.

16. Пантелеймоненко А., Мілька А., Павленко О. ПУБЛІЧНІ ЗАКУПІВЛІ В УКРАЇНІ В УМОВАХ ВОЄННОГО СТАНУ. *Економіка та суспільство*. 2023. № 51. URL: https://doi.org/10.32782/2524-0072/2023-51-20.

17. Педченко Н.С., Кудацький О.М., Педченко М.Г. ПУБЛІЧНІ ЗАКУПІВЛІ В УКРАЇНІ: ПЕРЕВАГИ НА НЕДОЛІКИ. *Scientific Bulletin of PUET: Economic Sciences 1(107) 2023*. 2023. № 2 (108). URL: https://doi.org/10.37734/2409-6873-2023-2-4.

18. What Is Natural Language Processing? | IBM. *IBM in Deutschland, Österreich und der Schweiz*. URL: https://www.ibm.com/topics/natural-language-processing.

19. What Is Topic Modeling? A Beginner's Guide. *Levity | Streamline Your Freight Email Operations with AI Automation*. URL: https://levity.ai/blog/what-is-topic-modeling#:~:text=Topic%20modeling%20is%20a%20type,predefined%20tags%20or%20training%20data.

20. Topic modeling in NLP: Approaches, implementation and use cases. *LeewayHertz - AI Development Company*. URL: https://www.leewayhertz.com/topic-modeling-in-nlp/#:~:text=Topic%20modeling%20is%20a%20technique,insights%20from%20large%20textual%20datasets.

21. What is an API? - Application Programming Interface Explained - AWS. *Amazon Web Services, Inc.* URL: https://aws.amazon.com/what-is/api/#:~:text=API%20stands%20for%20Application%20Programming,other%20using%20requests%20and%20responses.

22. OpenProcurement API – openprocurement.api 2.5 documentation. *OpenProcurement API – openprocurement.api 2.5 documentation*. URL: https://prozorro-api-docs.readthedocs.io/uk/latest/index.html.

23. Requests: HTTP for Humans™ – Requests 2.32.3 documentation. *Requests: HTTP for Humans™ – Requests 2.32.3 documentation*. URL: https://requests.readthedocs.io/en/latest/.

24. pandas - Python Data Analysis Library. *pandas - Python Data Analysis Library*. URL: https://pandas.pydata.org/.

25. time - Time access and conversions. *Python documentation*. URL: https://docs.python.org/3/library/time.html.

26. concurrent.futures - Launching parallel tasks. *Python documentation*. URL: https://docs.python.org/3/library/concurrent.futures.html.

27. Egger R., Yu J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*. 2022. Vol. 7. URL: https://doi.org/10.3389/fsoc.2022.886498.

28. Pritchard J. K., Stephens M., Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*. 2000. Vol. 155, no. 2. P. 945–959. URL: https://doi.org/10.1093/genetics/155.2.945.

29. Pan X., Xue Y. Advancements of Artificial Intelligence Techniques in the Realm about Library and Information Subject- A Case Survey of Latent Dirichlet Allocation Method. *IEEE Access*. 2023. P. 1. URL: https://doi.org/10.1109/access.2023.3334619.

30. MaartenGr - Overview. *GitHub*. URL: https://github.com/MaartenGr.

31. Keita Z. Meet BERTopic– BERT's Cousin For Advanced Topic Modeling. *Medium*. URL: https://towardsdatascience.com/meet-bertopic-berts-cousin-for-advanced-topic-modeling-ea5bf0b7faa3#:~:text=This%20tool%20was%20developed%20by,words%20in%20the%20topic%20descriptions.

32. 10 Leading Language Models For NLP In 2022. *TOPBOTS*. URL: https://www.topbots.com/leading-nlp-language-models-2020/.

33. BERTopic. *GitHub Pages*. URL: https://maartengr.github.io/BERTopic/index.html.

34. Definition of 'corpus'. *Collins*. URL: https://www.collinsdictionary.com/dictionary/english/corpus#:~:text=A%20corpus%20is%20a%20large,compilation%20More%20Synonyms%20of%20corpus.

35. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 2003. P. 993–1022. URL: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=http://githubhelp.com.

36. Reaves B. What is BERTopic? Metabob's AI team is testing another topic modeling technique. *Medium*. URL: https://medium.com/metabob/what-is-bertopic-metabobs-ai-team-is-testing-another-topic-modeling-technique-e78d242f472b.

37. GitHub - UKPLab/sentence-transformers: Multilingual Sentence & Image Embeddings with BERT. *GitHub*. URL: https://github.com/UKPLab/sentence-transformers.

38. Pretrained Models – Sentence Transformers documentation. *SentenceTransformers Documentation – Sentence Transformers documentation*. URL: https://www.sbert.net/docs/sentence_transformer/pretrained_models.html.

39. sentence-transformers/paraphrase-multilingual-mpnet-base-v2 · Hugging Face. *Hugging Face – The AI community building the future.* URL: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2.

40. Karanam S. Curse of Dimensionality–A "Curse" to Machine Learning. *Medium*. URL: https://towardsdatascience.com/curse-of-dimensionality-a-curse-to-machine-learning-c122ee33bfeb.

41. GitHub - lmcinnes/umap: Uniform Manifold Approximation and Projection. *GitHub*. URL: https://github.com/lmcinnes/umap.

42. How HDBSCAN Works – hdbscan 0.8.1 documentation. *The hdbscan Clustering Library – hdbscan 0.8.1 documentation*. URL: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html.

43. Axelborn H., Berggren J. Topic Modeling for Customer Insights. A Comparative Analysis of LDA and BERTopic in Categorizing Customer Calls. Umeå : Umeå University, 2023. 65 p. URL: https://umu.diva-portal.org/smash/get/diva2:1763637/FULLTEXT01.pdf.

44. Bismi I. Topic Modelling using LDA. *Medium*. URL: https://medium.com/@iqra.bismi/topic-modelling-using-lda-fe81a2a806e0.

45. Świtała M. S., Chlebus M. So close and so far. Finding similar tendencies in econometrics and machine learning papers. Topic models comparison. *Working Papers*. 2020. Vol. 16, no. 322. P. 1–36.

URL: https://www.researchgate.net/figure/Econometrics-and-statistics-chosen-topics-LDA-model_fig2_341914817.

46. os - Miscellaneous operating system interfaces. *Python documentation*. URL: https://docs.python.org/3/library/os.html.

47. Міністерство з питань стратегічних галузей промисловості України - Єдиний закупівельний словник ДК 021:2015. *Головна | Міністерство з питань стратегічних галузей промисловості України*. URL: https://mspu.gov.ua/diyalnist/oboronni-zakupivli/yedinij-zakupivelnij-slovnik-dk-0212015.

48. re - Regular expression operations. *Python documentation*. URL: https://docs.python.org/3/library/re.html.

49. NLTK :: Natural Language Toolkit. *NLTK :: Natural Language Toolkit*. URL: https://www.nltk.org/.

50. GitHub - skupriienko/Ukrainian-Stopwords: the list of ~2000 ukrainian stopwords (with numbers). *GitHub*. URL: https://github.com/skupriienko/Ukrainian-Stopwords.

51. What Are Stemming and Lemmatization? | IBM. *IBM in Deutschland, Österreich und der Schweiz*. URL: https://www.ibm.com/topics/stemming-lemmatization#:~:text=The%20practical%20distinction%20between%20stemming,be%20found%20in%20the%20dictionary.

52. Gensim: topic modelling for humans. *Radim Řehůřek: Machine learning consulting*. URL: https://radimrehurek.com/gensim/.

53. Kapadia S. Topic Modeling in Python: Latent Dirichlet Allocation (LDA). *Medium*.
URL: https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0.

54. Pickett M. Exploring Coherence Metrics for Optimizing Topic Models of Humpback Song. 2020. P. 1–18. URL: https://www.mbari.org/wp-content/uploads/Pickett.pdf.

55. pyLDAvis: Topic Modelling Exploration Tool That Every NLP Data Scientist Should Know. *neptune.ai*. URL: https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know.

56. Клименко І. М. Міжмовні лексико-семантичні відношення (на матеріалі англійської та української фразеології). С. 21–27. URL: https://core.ac.uk/download/pdf/268531765.pdf.

ДОДАТКИ

ДОДАТОК А

ABSTRACT

Shtefan A. V. Economic and Mathematical Modelling of Public Procurement Through the Prozorro System Using Data Mining: the thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science: specialty – 051 Economics (Economic Cybernetics and Business Analytics) / Supervisor S. V. Mynenko. Sumy : Sumy State University, 2024. 69 p.

Public procurement includes a wide range of goods, services, and works, from construction and repair to the supply of medical equipment, computers, transportation, management services, etc. Public procurement authorities must follow procedures established by law, such as publishing tender announcements, ensuring equal access to information for all participants, holding open tenders or competitions, evaluating tender proposals, and concluding contracts with the winners. Tenders are implemented through the electronic public procurement system Prozorro.

However, Russia's full-scale invasion of Ukraine has highlighted the problems associated with the inability to ensure the most efficient use of budget funds. The sources of these problems are both the difficulty of monitoring the thousands of tenders that appear in the system every day and the slow pace of civil society development and, as a result, the lack of proper public control in this area over a long period of time.

Therefore, at this stage of development of Ukraine and its public procurement sector, an important factor is partial automation and consolidation of monitoring of public needs expressed through this sector, reduction of the risk of corruption in the procurement process, etc. For this purpose, it is proposed to use the methods of intellectual analysis of text data (Text Data Mining). This paper proposes a method of applying economic and mathematical modeling of one of them - Topic Modelling.

Manual search for announcements that contradict the principles of fair and transparent bidding according to the Law of Ukraine "About Public Procurement" has a low degree of usefulness, as thousands of tenders are announced by the authorities every day. In addition, the lack of systematic oversight of this process by the Department for Public Procurement and Competition Policy and anti-corruption bodies contributes to the growing influence of the human factor in the form of dishonesty and abuse of office.

In addition, it should be noted that despite the amount of losses incurred by the Ukrainian economy as a result of inappropriate spending of funds through some tender procurements, the interest of the Ukrainian public in this issue shows a disappointing trend. This is a manifestation of the previously mentioned slow development of civil society.

Thus, the need to develop new, at least temporary, approaches to monitoring and administering announced tenders becomes even more apparent. Given that their appearance becomes massive even within one day, the importance of full or partial automation of announcement processing is clear. This is the relevance of tender research today.

It follows from the above that the purpose of this research was to build and demonstrate the feasibility of using topic models for analyzing public procurement activities.

The object of this research is the socio-economic relations that arise between participants in the public procurement process in Ukraine.

The subject of the study was economic and mathematical methods and models of public procurement announced by the relevant authorities through the online platform of the Prozorro system.

As a result of the work, were built and demonstrated the feasibility of using topic models for analyzing public procurement activities: the research objective has been achieved.

It was found that the model based on the BERTopic algorithm is suitable for finding markers that may indicate corruption or the use of public funds with a low

level of utility for society. It was also found that the LDA model can be used to analyze the needs within the sectors of the national economy of Ukraine, as well as the country's socio-economic system.

In the course of the research work, the objectives were achieved:

- The prerequisites and relevance of tender research are described;
- A bibliometric analysis of relevant scientific research in the field of public procurement was conducted;
- Natural Language Processing (NLP) as a method of monitoring public procurement activities is presented;
- A database of tenders was formed using the Prozorro application programming interface;
- The research assumptions were formed;
- Conceptual modeling of topic modeling algorithms, such as Dirichlet latent clustering and BERTopic, was implemented;
- NLP models were built and interpreted the results.

Research methods: synthesis; analysis of relevant publications of representatives of the scientific community, specialists in the field of data mining; bibliometric analysis; topic modeling, cluster analysis.

The source of data for building the models was the database of the Prozorro e-procurement system. The source of knowledge on algorithmization of data collection, data cleaning, and modeling in code was the documentation of the Python programming language and its modules and libraries. The code writing environment was the Spyder integrated development environment.

The scientific and social value of the research lies in the accelerated and consolidated method of monitoring and analyzing public procurement based on machine learning approaches. This will allow for more efficient redistribution of state and local budget funds, more effective anti-corruption measures, etc.

A promising area for further research could be modeling by associative rules to establish links with other variables included in the collected data from Prozorro. Also, to facilitate the work with such models, it is advisable to develop a full-fledged

application based on thematic modeling methods. To explore the possibilities of automating the control of procurement activities, it is advisable to invest in machine learning and artificial intelligence.

The research was carried out within the framework of the research work commissioned by the Ministry of Education and Science of Ukraine «Modeling the mechanisms of de-shadowing and corruption of the economy to ensure national security: the impact of the transformation of financial behavioral patterns», state registration no: 0122U000783. The results of the bachelor's qualification work were published in the article «Financial Fraud Detection on Social Networks Based on a Data Mining Approach» in the professional journal «Financial Markets, Institutions and Risks (FMIR)».

Keywords: economic growth, corruption, public procurement, tender research, monitoring, Prozorro, topic modeling, natural language processing, clustering, LDA, BERTopic.

The content of the qualification work is set out on 69 pages. The list of references of 56 titles is located on 45 – 50 pages. The work contains 1 table, 13 figures, and appendices A, B, C, D, E, F.

The year of completion of the qualification work is 2024.

The year of defense of the qualification work is 2024.

АНОТАЦІЯ

Штефан А. В. Економіко-математичне моделювання державних закупівель через систему Prozorro засобами Data Mining: робота на здобуття кваліфікаційного ступеня бакалавра: спеціальність – 051 Економіка (Економічна кібернетика та бізнес-аналітика) / науковий керівник С. В. Миненко. Суми : Сумський державний університет, 2024. 69 с.

Державні закупівлі включають широке коло товарів, послуг і робіт, від будівництва і ремонту до постачання медичного обладнання, комп'ютерів, транспорту, послуг з управління тощо. Органи, що здійснюють державні закупівлі, повинні дотримуватися процедур, визначених законом, таких як публікація оголошень про торги, забезпечення рівного доступу до інформації для всіх учасників, проведення відкритих торгів або конкурсів, оцінка тендерних пропозицій та укладання договорів з переможцями. Реалізація тендерів відбувається через електронну систему публічних закупівель Prozorro.

Однак повномасштабне вторгнення РФ в Україну висвітлило проблеми, пов’язані із нездатністю забезпечити максимально ефективне використання бюджетних коштів. Джерелами цих проблем є як складність моніторингу тисяч тендерів, що з’являються у системі щодня, так і повільний темп розвитку громадянського суспільства і, як наслідок, відсутність належного громадського контролю у цій сфері протягом великого періоду часу.

Тому, на даному етапі розвитку України та сфери її публічних закупівель, важливим фактором є часткова автоматизація та укрупнення моніторингу суспільних потреб, що виражаються через цю сферу, зниження рівня ризиків корупційних проявів у процесі здійснення закупівельної діяльності тощо. Для цього пропонується використовувати методи інтелектуального аналізу текстових даних (Text Data Mining). У цій роботі запропоновано спосіб застосування економіко-математичного моделювання одного з них – тематичного моделювання (Topic Modelling).

Пошук оголошень, які суперечать принципам чесних та прозорих торгів згідно Закону України «Про публічні закупівлі», в ручному режимі має низький ступінь корисності, оскільки щодня владними органами оголошуються тисячі тендерів. Крім того, відсутність систематичного нагляду за цим процесом з боку Департаменту сфери публічних закупівель та конкурентної політики, антикорупційних органів, сприяє зростанню впливу людського фактору у вигляді недобросовісності та зловживань службовим становищем.

Крім того, необхідно відмітити, що не дивлячись на об’єми збитків, яких зазнає економіка України внаслідок недоцільних витрат коштів через деякі тендерні закупівлі, інтерес до даної проблеми, з боку української громадськості, показує невтішну тенденцію. Це є проявом згаданого раніше процесу повільної розбудови громадянського суспільства.

Таким чином, необхідність розробки нових, принаймні, тимчасових підходів до моніторингу та адміністрування оголошуваних тендерів стає ще більш очевидною. З огляду на те, що їх поява набуває масового характеру навіть у межах однієї доби, зрозумілою є важливість повної або часткової автоматизації обробки оголошень. Саме у цьому виражається актуальність тендерних досліджень на сьогодні.

Із зазначеного вище випливає, що метою цього дослідження була побудова та демонстрація доцільності використання тематичних моделей для аналізу державної закупівельної діяльності.

Об’єктом проведеного дослідження є соціально-економічні відносини, що виникають між учасниками процесу публічних закупівель в Україні.

Предметом дослідження виступали економіко-математичні методи та моделі державних закупівель, оголошувані відповідними органами влади через онлайн-платформу системи Prozorro.

У результаті виконання роботи було побудовано та продемонстровано доцільність використання тематичних моделей для аналізу державної закупівельної діяльності: мети дослідження досягнуто.

Було виявлено, що модель на основі алгоритму BERTopic придатна для пошуку маркерів, які можуть вказувати на корупційні прояви чи використання державних коштів з низьким рівнем корисності для суспільства. Також встановлено, що модель LDA може використовуватися для аналізу потреб у межах галузей національної економіки України, соціально-економічної системи країни.

У ході виконання наукової роботи було досягнуто поставлених завдань:

- Описано передумови та актуальність тендерних досліджень;
- Проведено бібліометричний аналіз релевантних наукових досліджень у сфері державних закупівель;
- Представлено обробку природної мови (NLP) як метод моніторингу діяльності у сфері публічних закупівель;
- Сформовано базу тендерів за допомогою інтерфейсу прикладного програмування Prozorro;
- Сформовано припущення дослідження;
- Реалізовано концептуальне моделювання алгоритмів тематичного моделювання, таких як латентне розміщення Діріхле та BERTopic;
- Побудовано NLP-моделі та інтерпретовано отримані результати.

Методи дослідження: синтез; аналіз відповідних публікацій представників наукової спільноти, фахівців у сфері інтелектуального аналізу даних; бібліометричний аналіз; тематичне моделювання, кластерний аналіз.

Джерелом забезпечення даними для побудови моделей була база даних системи електронних закупівель Prozorro. Джерелом знань щодо алгоритмізації збору даних, очистки даних та моделювання в коді стала документація мови програмування Python та її модулів, бібліотек. Середовище написання коду – інтегроване середовище розробки Spyder.

Наукова та суспільна цінність дослідження полягає у пришвидшеному та укрупненому способі моніторингу та аналізу сфери державних закупівель, що базується на підходах машинного навчання. Це дозволить ефективніше

перерозподіляти кошти державного, місцевого бюджетів, результативніше проводити заходи щодо антикорупційної боротьби тощо.

Перспективним напрямком подальших досліджень може бути моделювання за асоціативними правилами для встановлення зв’язків з іншими змінними, включених у збирані дані з Prozorro. Також, для полегшення роботи з такими моделями, доцільною є розробка повноцінного додатку, в основі якого лежатимуть методи тематичного моделювання. Для вивчення можливостей автоматизації контролю закупівельної діяльності доцільно інвестувати у машинне навчання, штучний інтелект.

Наукове дослідження було виконано в межах науково-дослідної роботи за замовленням МОН України «Моделювання механізмів детінізації та декорумпізації економіки для забезпечення національної безпеки: вплив трансформації фінансових поведінкових патернів», № держреєстрації: 0122U000783. Результати кваліфікаційної роботи бакалавра оприлюднені у статті «Financial Fraud Detection on Social Networks Based on a Data Mining Approach» у фаховому журналі «Financial Markets, Institutions and Risks (FMIR)».

Ключові слова: економічне зростання, корупція, державні закупівлі, тендерні дослідження, моніторинг, Prozorro, тематичне моделювання, обробка природної мови, кластеризація, LDA, BERTopic.

Зміст кваліфікаційної роботи викладено на 69 сторінках. Список використаних джерел із 56 найменувань розміщено на 45 – 50 сторінках. Робота містить 1 таблицю, 13 рисунків та додатки A, B, C, D, E, F.

Рік виконання кваліфікаційної роботи – 2024 рік.

Рік захисту кваліфікаційної роботи – 2024 рік.

ДОДАТОК В

Програма для збору даних з Prozorro

```
import requests
import pandas as pd
import time
from concurrent.futures import ThreadPoolExecutor
```

Рис. В.1 – Імпорт бібліотек для скрипту збору даних

```
def save_to_csv(tender_data, filename='tender_urls.csv'):
    if tender_data:  # Перевірка, чи є дані для збереження
        df = pd.DataFrame(tender_data)
        df.to_csv(filename, index=False, encoding='utf-8-sig')
        print(f"Дані збережено: {len(tender_data)} записів")
    else:
        print("Немає даних для збереження")
```

Рис. В.2 – Код функції для збереження даних у форматі CSV

```
def get_tender_data_inside(session, tender_id):
    url_for_parse = f"https://public.api.openprocurement.org/api/2.5/tenders/{tender_id}"
    try:
        response = session.get(url_for_parse)
        response.raise_for_status()  # Запускає виняток, якщо HTTP статус не є успішним
        return response.json()
    except requests.exceptions.HTTPError as e:
        if e.response.status_code == 429:
            print("Отримано 429 Too Many Requests, чекаємо перед повторним запитом...")
            time.sleep(10)  # Затримка на 10 секунд перед повторним запитом
            return get_tender_data_inside(session, tender_id)  # Рекурсивний повторний запит
        raise  # Передача винятку, якщо статус код не 429
```

Рис. В.3 – Код функції для виконання HTTP-запиту до API Prozorro

```python
def rec_tenders():
    url = "https://public.api.openprocurement.org/api/2.5/tenders"
    params = {"offset": "",
              "descending": "true"}
    tenders_data = []
    page_count = 0

    with requests.Session() as session:
        try:
            while True:
                response = session.get(url, params=params)
                response.raise_for_status()
                data = response.json()
                tenders = data.get("data", [])

                with ThreadPoolExecutor(max_workers=9) as executor:
                    futures = [executor.submit(get_tender_data_inside, session, tender['id']) for tender in tenders]
                    for future in futures:
                        tender_details = future.result()
                        if tender_details:
                            procuring_entity = tender_details['data'].get('procuringEntity')
                            if procuring_entity and procuring_entity.get('kind') == 'authority':
                                tenders_data.append({
                                    'Заголовок': tender_details['data']['title'],
                                    'Сума, грн': tender_details['data'].get('value', {}).get('amount', ''),
                                    'Орган': tender_details['data']['procuringEntity']['identifier']['legalName'],
                                    'Регіон': tender_details['data']['procuringEntity']['address'].get('region', ''),
                                    'Опис': tender_details['data'].get('description', ''),
                                    'Link': f"https://prozorro.gov.ua/tender/{tender_details['data']['id']}"
                                })
                save_to_csv(tenders_data)  # Збереження даних після кожної успішної ітерації

                n_page = data.get('next_page')
                if n_page and n_page.get('offset'):
                    params['offset'] = n_page['offset']
                    page_count += 1
                    print(f"Оброблено сторінку {page_count}")
                else:
                    break
        except Exception as e:
            print(f"Помилка під час збору даних: {e}")
        finally:
            save_to_csv(tenders_data)  # Фінальне збереження даних
            print("Виконання програми зупинено. Останній стан даних збережено.")

rec_tenders()
```

Рис. В.4 – Код функції, що реалізує основний процес збору даних

ДОДАТОК С

Результат збору даних через API Prozorro

	A	B	C	D	E	F	G	H	I	J	K	L
1	Заголовок	Сума, грн	Орган	Регіон	Опис	Link						
2	Колесо для тачки	570	Управління поліції с	Львівська область		https://prozorro.gov.ua/tender/f79afa798f12412aabaaa3c7cde27be9						
3	Послуги з організац	49998	Бахмутська міська р	Донецька область		https://prozorro.gov.ua/tender/dabafb4e7a354b8b8705e62190d00347						
4	Послуги з благоустр	9494687	Департамент благо	Дніпропетровська область		https://prozorro.gov.ua/tender/1807d654a37d4d76a9a58c440a635edc						
5	Послуги з поточного	16175	Головне управління	Івано-Франківська область		https://prozorro.gov.ua/tender/d1f6cc21fd77452dbd01932b0d17136c						
6	Крісла	14670	Головне управління	Рівненська область		https://prozorro.gov.ua/tender/0a480424ec9e49ccaf2310c22eda8570						
7	Реконструкція проїз	11309342	Управління капітал	Житомирська область		https://prozorro.gov.ua/tender/85c9a63b33204e6e8c718f1812b81ff9						
8	Поточний ремонт п	1936057	Банкнотно-монетни	Київська область	Контактна особа за	https://prozorro.gov.ua/tender/7b510c2b4e0d475684e66d3706be5696						

Рис. С.1 – Зразок зібраних даних

ДОДАТОК D

Реалізація тематичного моделювання

```
import os
import pandas as pd
import nltk
import re
nltk.download('punkt')
from nltk.tokenize import word_tokenize
nltk.download('stopwords')
from nltk.stem import WordNetLemmatizer
nltk.download('wordnet')
nltk.download('omw-1.4')
import gensim
from gensim.models import CoherenceModel, LdaModel
from gensim.corpora.dictionary import Dictionary
import pyLDAvis.gensim_models as gensimvis
import pyLDAvis
from bertopic import BERTopic
```

Рис. D.1 – Імпорт бібліотек, необхідних для обробки тексту, побудови моделей та візуалізацій

```
os.chdir('C:/Users/artsh/Desktop/diploma')
tenders_df = pd.read_csv('tender_urls.csv')
tenders_df.head()
```

Рис. D.2 – Код для зміни робочої директорії та завантаження даних у проект

```
tenders_df.info()
tenders_df.isnull().sum()
len(tenders_df)
```

Рис. D.3 – Код для отримання описових статистик про завантажений датафрейм

```
tenders_headlines = tenders_df['Заголовок']

#Приведення усього тексту до нижнього регістру та видалення зайвих пробілів
tenders_headlines = tenders_headlines.map(lambda x: x.lower())
tenders_headlines = tenders_headlines.map(lambda x: x.strip())
tenders_headlines = tenders_headlines.map(lambda x: " ".join(x.split()))

#Видалення несуттєвих фрагментів тексту
tenders_headlines = tenders_headlines.map(lambda x: re.sub(r'http\S+|www\S+|https\S+', '', x, flags=re.MULTILINE))
tenders_headlines = tenders_headlines.map(lambda x: re.sub(r'\@\w+|\#','', x))
tenders_headlines = tenders_headlines.map(lambda x: re.sub(r'[^\w\s\u0400-\u04FF]','', x))
tenders_headlines = tenders_headlines.map(lambda x: re.sub(r'\b\d{7,}\b', '', x))

#Токенізація тексту
tenders_headlines = tenders_headlines.map(lambda x: word_tokenize(x))

#Обробка стоп-слів
with open('stopwords_ua.txt', 'r', encoding='utf-8') as file:
    stops = [line.strip() for line in file.readlines()]

for i in range(0,len(tenders_headlines)):
    tenders_headlines[i] = [word for word in tenders_headlines[i] if not word in list(stops)]

#Лематизація
lemmatizer = WordNetLemmatizer()
for i in range(0,len(tenders_headlines)):
    words = []
    for word in tenders_headlines[i]:
        words.append(lemmatizer.lemmatize(word))
    tenders_headlines[i] = words
```

Рис. D.4 – Код для попередньої обробки корпусу для моделі LDA

```
dictionary = Dictionary(tenders_headlines)
tenders_corpus = [dictionary.doc2bow(tender) for tender in tenders_headlines]
for i in range (2,81):
    lda_model = LdaModel(tenders_corpus, num_topics=i, id2word=dictionary)
    coherence_lda_model = CoherenceModel(model=lda_model, texts=tenders_headlines, dictionary=dictionary)
    coherence_lda = coherence_lda_model.get_coherence()
    print(f'Кількість тем = {i} має коефіцієнт когерентності {coherence_lda}')
```

Рис. D.5 – Код для визначення оптимальної кількості тем для моделі LDA

```
#Побудова моделі
best_lda = LdaModel(corpus=tenders_corpus, id2word=dictionary, num_topics=44, passes=5)

#Візуалізація
pylda_visualisation = gensimvis.prepare(best_lda, tenders_corpus, dictionary)
pyLDAvis.save_html(pylda_visualisation, 'pyLDA visualisation.html')
pyLDAvis.display(pylda_visualisation)
```

Рис. D.6 – Код для побудови моделі LDA, а також візуалізації та збереження результатів

```
tenders_headlines_bert = tenders_df['Заголовок']
bert = BERTopic(top_n_words=10, calculate_probabilities=True, embedding_model='paraphrase-multilingual-mpnet-base-v2')
bertopic, probs = bert.fit_transform(tenders_headlines_bert)
bert.visualize_topics().write_html('bert_top_vis.html')
bert.visualize_barchart(top_n_topics=800).write_html('barchart.html')
```

Рис. D.7 – Код для побудови моделі BERTopic, візуалізації та збереження результатів

ДОДАТОК Е

Додаткові приклади візуалізацій моделі LDA

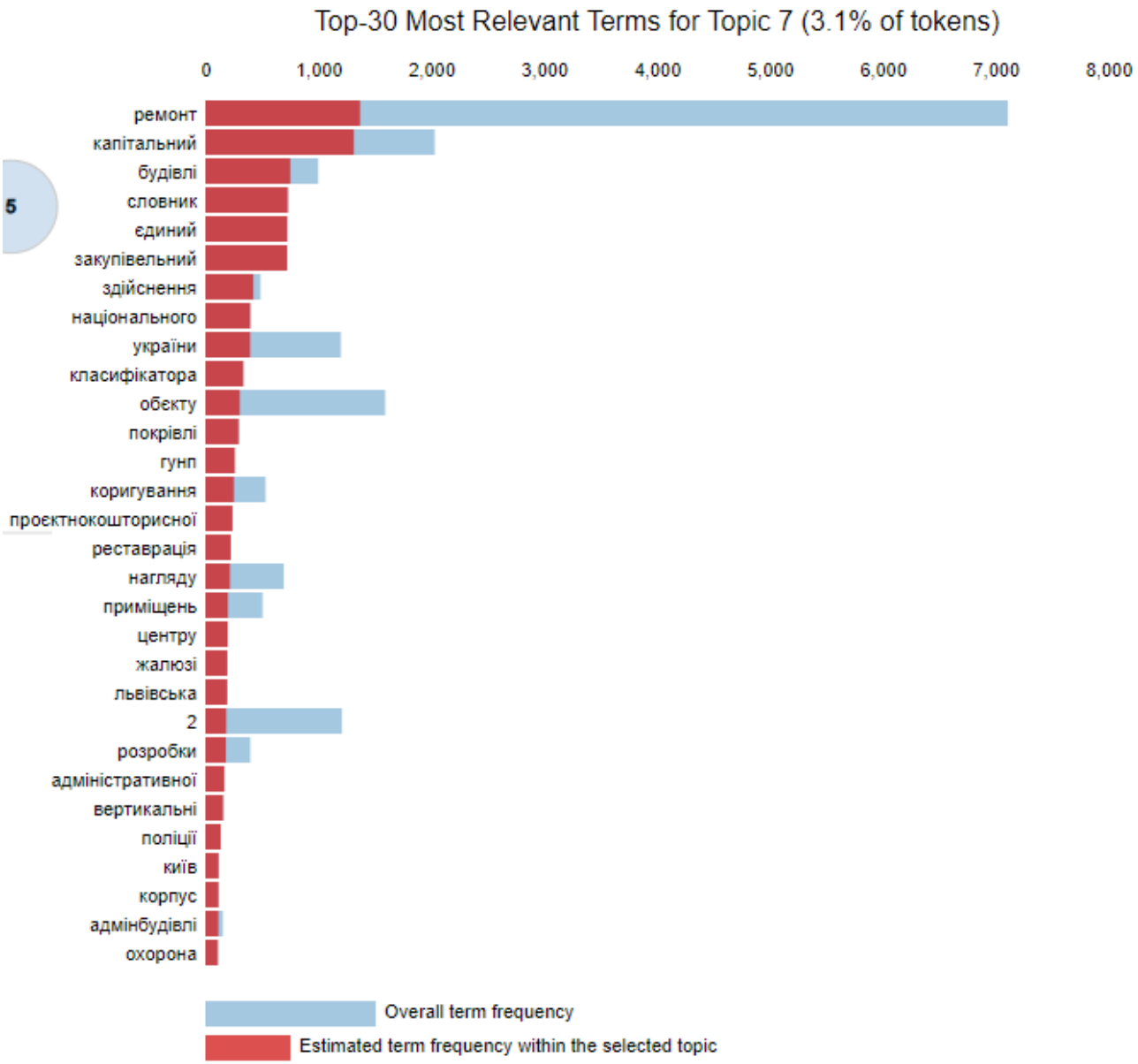


Рис. Е.1 – Найбільш релевантні терміни для теми 7

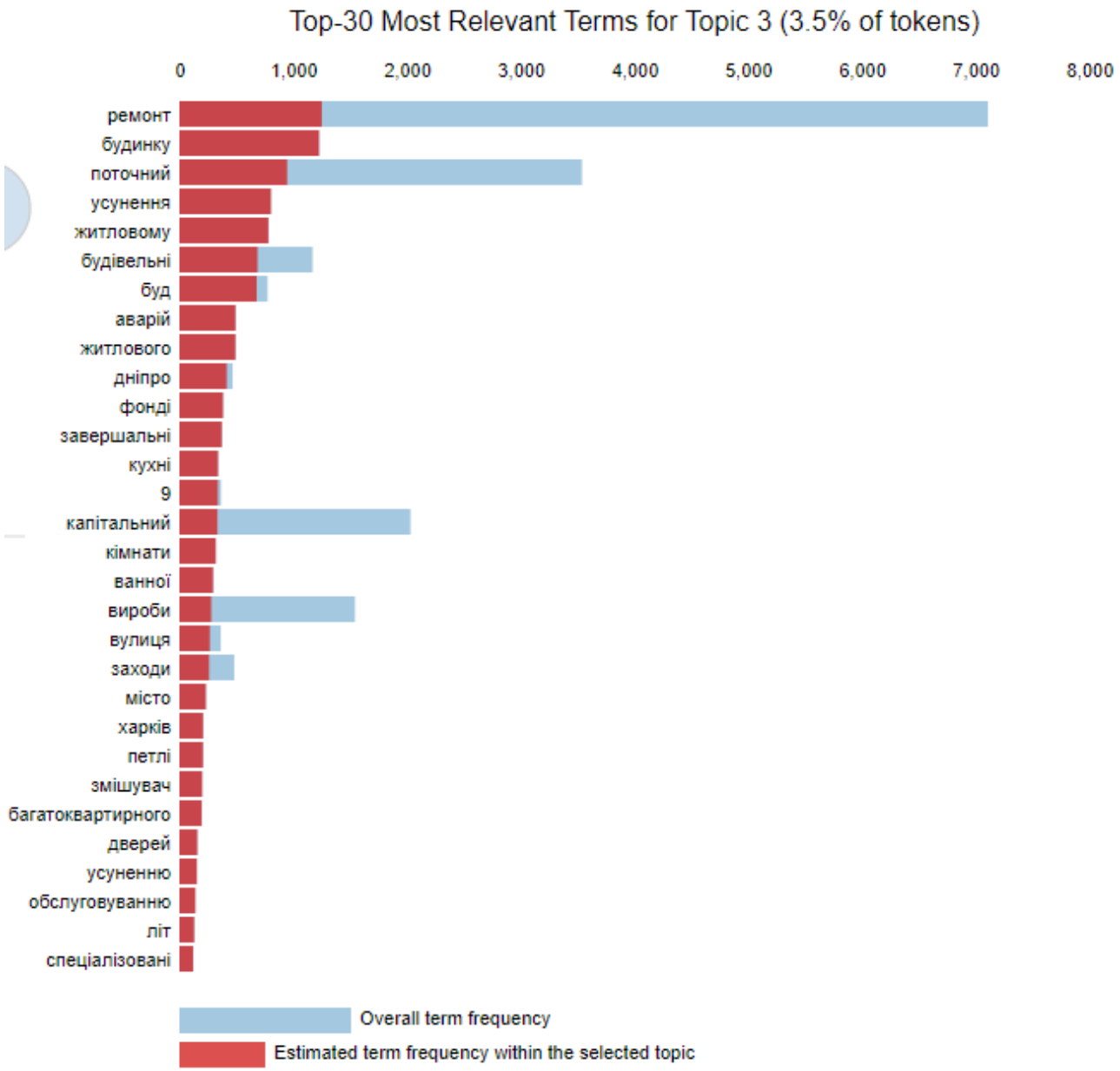


Рис. Е.2 – Найбільш релевантні терміни для теми 3

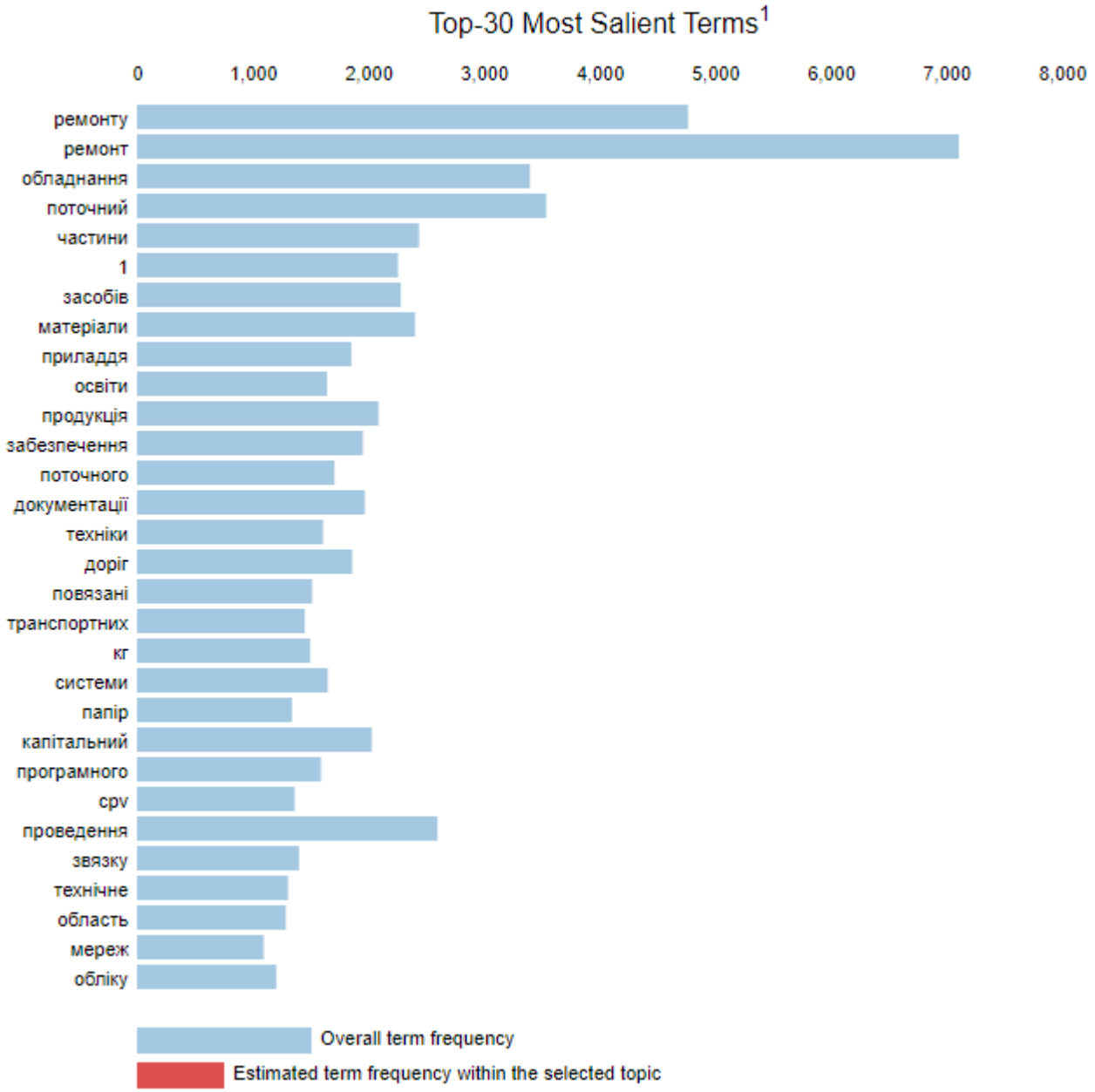


Рис. Е.3 – 30 найпоширеніших термінів корпусу

ДОДАТОК F

Зразок масиву внутрішньокластерних діаграм BERTopic моделі

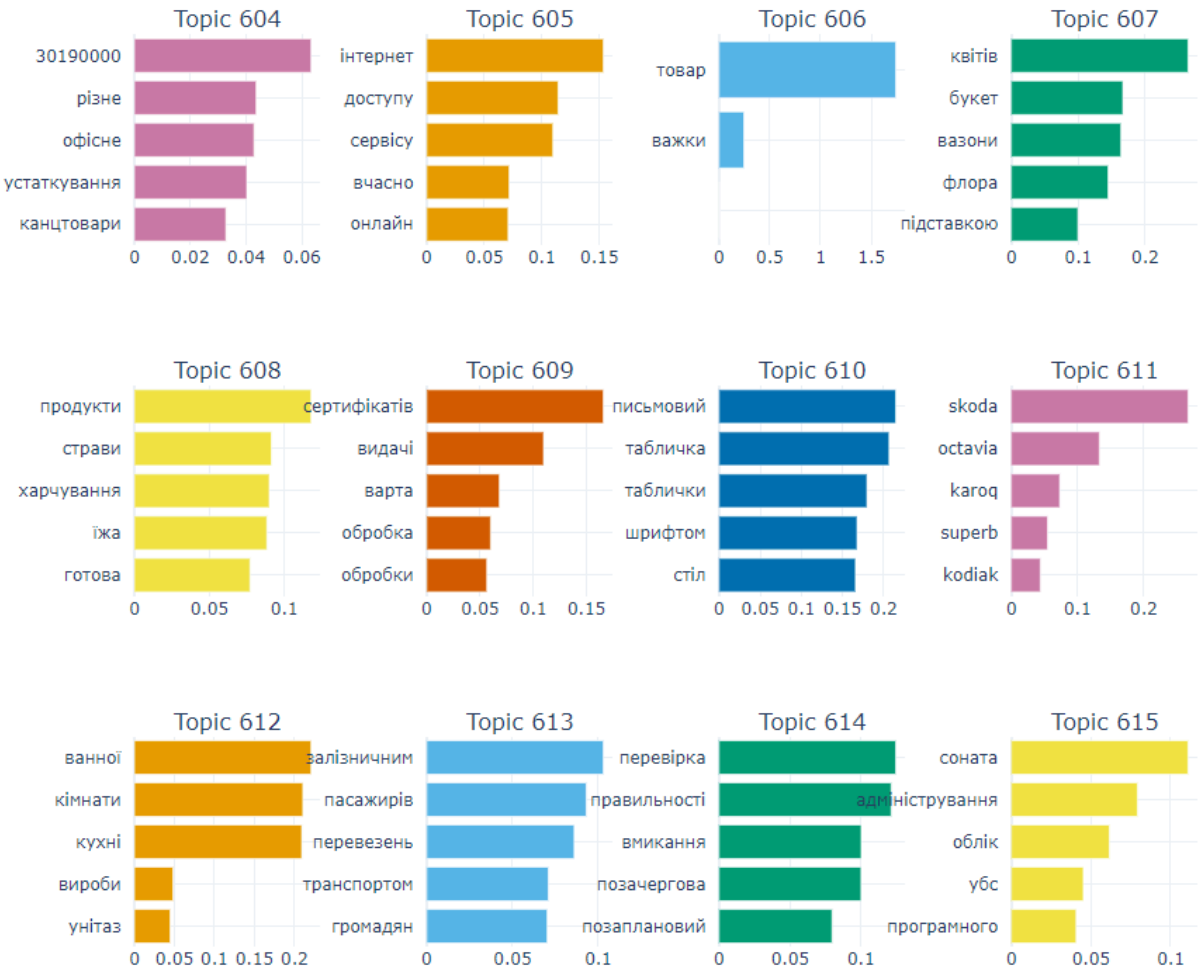


Рис. F.1 – Приклад частот термінів у темах