ДОДАТОК А

Текст для перевірки токенізації

У місті вже декілька днів без масштабних обстрілів, але техніка має бути готовою виїхати у будь-який момент. Влад прийшов на службу у 2022-му. А щоб доборотися, показати, що я не буду сидіти зі зв'язаними руками, поки ви з моєї Батьківщини, з моєї країни робите бозна-що. У школі №11 вчився Данило Дідік, який загинув внаслідок теракту біля Палацу спорту 2015 року. Удар стався близько 22:37, повідомив очільник ОВА Олег Синєгубов. Протягом тижня 15-21 квітня четверо жителів Харківщини загинули внаслідок російських обстрілів. Площа пожежі склала близько 100 кв.м. 37-річна волонтерка зазнала поранень внаслідок удару РФ по Куп'янську-Вузловому, повідомили в облпрокуратурі. Подати пропозиції та зауваження можна до 25 квітня: надіславши листа за адресою: пл. Конституції, 7; електронною поштою: rename@city.kharkiv.ua; за телефоном +380577607968. Як ідеться у повідомленні, із 24 жовтня в ПП "Рава-Руська — Гребенне" замінюватимуть дорожнє покриття. Понад 17 000 доларів з людини: на Львівщині затримали ймовірного організатора виїзду чоловіків за кордон. За шість тижнів у прокаті "Думками навиворіт 2" отримав $601 млн на внутрішньому ринку США та ще $861 млн за кордоном. Так загальні касові збори сягнули $1,461 млрд. Синоптики обіцяють упродовж тижня в Україні похолодання до -21°C та 15 сантиметрів снігу. Місцями стовпчик термометра опуститься до -11 градусів. Облаштували покрівлю: у Запоріжжі триває відновлення багатоповерхівки по вул. Незалежної України. Друга черга — внутрішнє оздоблення квартир та сходової клітини під'їздів №3,4, опорядження приміщень квартир під'їздів №1,2,4 та сходових клітин під'їздів №1-4; "Проєкт закону (про демобілізацію — ред.) практично розроблений і готовий до виходу і передачу до Верховної Ради через Кабінет міністрів", – розповів Гаврилюк. У грудні 2023 року міністр оборони Рустем Умєров в інтерв'ю Суспільному сказав, що міністерство працює над встановленням чітких термінів служби для військовослужбовців. Актуальну інформацію про стан доріг та допомогу можна отримати за телефоном цілодобової гарячої лінії: +38 (095) 568 38 77. На дорогах

— ожеледиця. У суботу можливі пориви північно-західного вітру до 15-20 м/с. За їхньою інформацією, площа укриття становить 234,5 м². “Якщо у червні це було ~100 км²/місяць та 3,4 км²/добу, то у вересні це вже стало ~400 км²/місяць та 13,4 км²/добу! Пік прийшов на листопад — 610 км²/місяць та 20,3 км²/добу! Вдумайтеся, ми втрачаємо 20 кілометрів квадратних кожної доби”, — йдеться у дописі. Богдан Бунчак та кураторки збирають 1 000 000 грн: за найбільший донат можна отримати уламок з тіла художника. Де у Харкові можна здати кров? вул. Клочківська, 366 (пн-пт — з 08:00 до 15:00, кожну суботу — з 08:00 до 13:00); ТРЦ "Нікольський", -1 поверх (вівторок, четвер, субота — з 09:00 до 14:00). Аби створити безпечний простір для громадян, сервісний центр МВС №6341 переїхав до укриття. Адреса не змінилася — м. Харків, вул. Шевченка, буд. 26.

ДОДАТОК Б

Токенізований текст у форматі csv

У,місті,вже,декілька,днів,без,масштабних,обстрілів,",",але,техніка,має,бути,готовою,виїхати,у,будь-який,момент,.,Влад,прийшов,на,службу,у,2022-му,.,А,щоб,доборотися,",",показати,",",що,я,не,буду,сидіти,зі,зв'язаними,руками,",",поки,ви,з,моєї,Батьківщини,",",з,моєї,країни,робите

бозна-

що,.,У,школі,№11,вчився,Данило,Дідік,",",який,загинув,внаслідок,теракту,біля,Палацу,спорту,2015,року,.,Удар,стався,близько,22:37,",",повідомив,очільник,ОВА,Олег,Синєгубов,.,Протягом,тижня,15-

21,квітня,четверо,жителів,Харківщини,загинули,внаслідок,російських,обстрілів,.,Площа,пожежі,склала,близько,100,кв.м,.,37-річна,волонтерка

зазнала,поранень,внаслідок,удару,РФ,по,Куп'янську-Вузловому,",",повідомили,в,облпрокуратурі,.,Подати,пропозиції,та,зауваження,можна,до,25,квітня,:,надіславши,листа,за,адресою,:,пл.,Конституції,",",7,;,електронною,поштою,:,rename@city.kharkiv.ua,;,за,телефоном,+380577607968,.,Як,ідеться,у,повідомленні,",",із,24,жовтня,в,ПП,"""""

Рава-Руська,—,Гребенне,"""",замінюватимуть,дорожнє,покриття,.,Понад,17 000,доларів,з,людини,:,на,Львівщині,затримали,ймовірного,організатора,виїзду,чоловіків,за,кордон,.,За,шість,тижнів,у,прокаті,"""",Думками,навиворіт,2,"""",отримав,$601,млн,на,внутрішньому,ринку,США,та,ще,$861,млн,за,кордоном,.,Так,загальні,касові

збори,сягнули,"$1,461",млрд,.,Синоптики,обіцяють,упродовж,тижня,в,Україні,похолодання,до,-

21°C,та,15,сантиметрів,снігу,.,Місцями,стовпчик,термометра,опуститься,до,-11,градусів,.,Облаштували,покрівлю,:,у,Запоріжжі,триває,відновлення,багатоповерхівки,по,вул.,Незалежної,України,.,Друга,черга,—

,внутрішнє,оздоблення,квартир,та,сходової,клітини,під'їздів,"№3,4"

",",опорядження,приміщень,квартир,під'їздів,"№1,2,4",та,сходових,клітин,під'їздів,№1-4,;,"""",Проєкт,закону,(,про,демобілізацію,—

,ред.,),практично,розроблений,і,готовий,до,виходу,і,передачу,до,Верховної,Ради,через,Кабінет,міністрів,"""",",",–
,розповів,Гаврилюк,.,У,грудні,2023,року,міністр,оборони,Рустем,Умєров,в,інтерв'ю

Суспільному,сказав,",",що,міністерство,працює,над,встановленням,чітких,термінів,служби,для,військовослужбовців,.,Актуальну,інформацію,про,стан,доріг,та,допомогу,можна,отримати,за,телефоном,цілодобової,гарячої,лінії,:,+38 (095) 568 38 77,.,На,дорогах,—,ожеледиця,.,У,суботу,можливі,пориви,північно-західного,вітру,до,15-20,м/с,.,За,їхньою,інформацією,",",площа

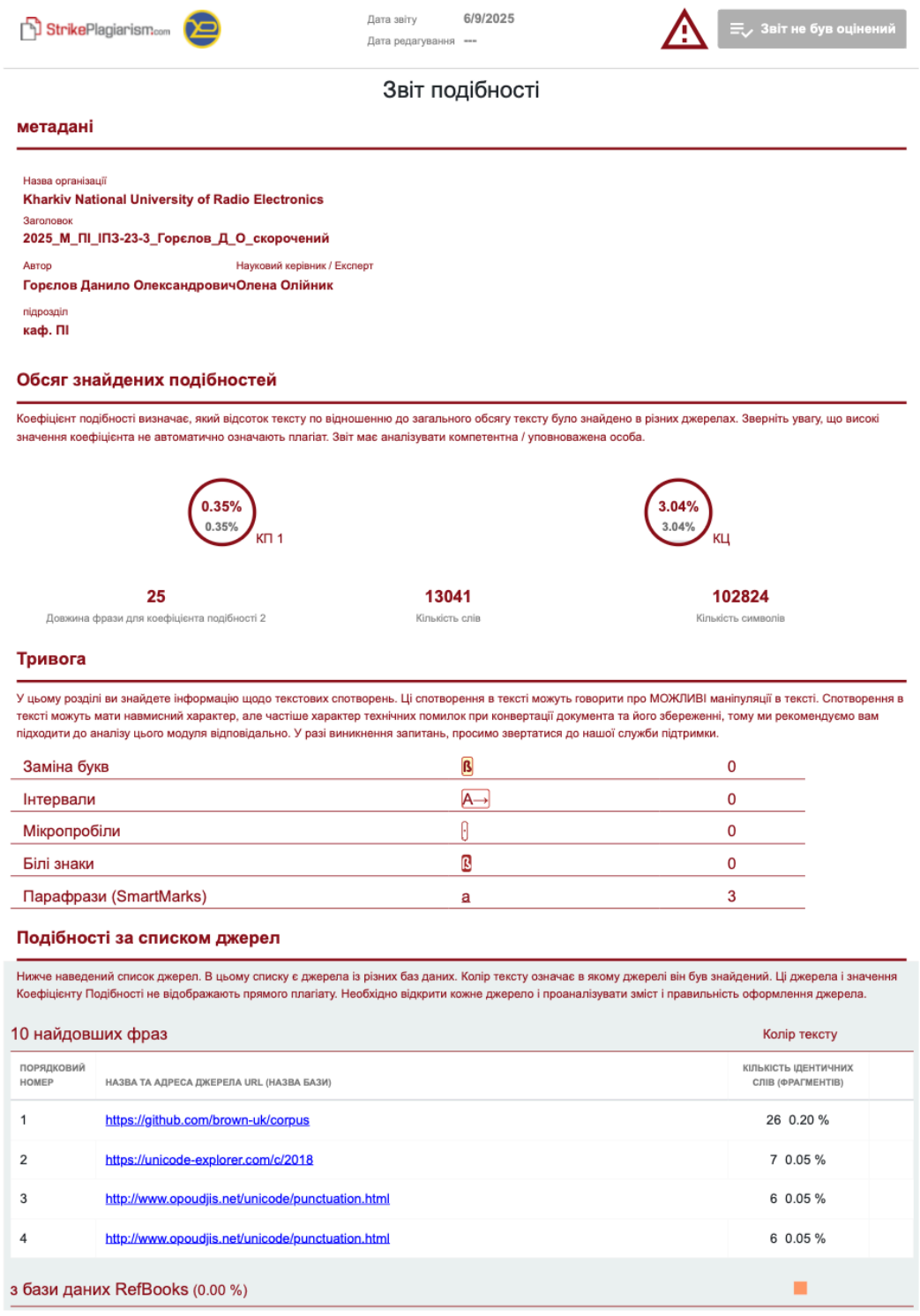
укриття,становить,"234,5",м²,.,“,Якщо,у,червні,це,було,~100,км²/місяць,та,"3,4",км²/добу,",",то,у,вересні,це,вже,стало,~400,км²/місяць,та,"13,4",км²/добу,!,Пік,прийшов,на,листопад,—
,610,км²/місяць,та,"20,3",км²/добу,!,Вдумайтеся,",",ми,втрачаємо,20,кілометрів,квадратних,кожної,доби,”,","

—,йдеться,у,дописі,.,Богдан,Бунчак,та,кураторки,збирають,1 000 000,грн,:,за,найбільший,донат,можна,отримати,уламок,з,тіла,художника,.,Де,у,Харкові,можна,здати,кров,?,вул.,Клочківська,",",366,(,пн-пт,—
,з,08:00,до,15:00,",",кожну,суботу,—,з,08:00,до,13:00,),;

ТРЦ,"""",Нікольський,"""",",",-1,поверх,(,вівторок,",",четвер,",",субота,—
,з,09:00,до,14:00,),.,Аби,створити,безпечний,простір,для,громадян,",",сервісний,центр,МВС,№6341,переїхав,до,укриття,.,Адреса,не,змінилася,—
,м.,Харків,",",вул.,Шевченка,",",буд.,26,.,,,

ДОДАТОК В

Звіт результатів перевірки на унікальність тексту в базі хнуре



StrikePlagiarism.com

Дата звіту 6/9/2025
Дата редагування ---

Звіт не був оцінений

Звіт подібності

метадані

Назва організації
Kharkiv National University of Radio Electronics
Заголовок
2025_М_ПІ_ІП3-23-3_Горєлов_Д_О_скорочений
Автор
Горєлов Данило Олександрович
Науковий керівник / Експерт
Олена Олійник
підрозділ
каф. ПІ

Обсяг знайдених подібностей

Коефіцієнт подібності визначає, який відсоток тексту по відношенню до загального обсягу тексту було знайдено в різних джерелах. Зверніть увагу, що високі значення коефіцієнта не автоматично означають плагіат. Звіт має аналізувати компетентна / уповноважена особа.

0.35% 0.35% КП 1

3.04% 3.04% КЦ

25	13041	102824
Довжина фрази для коефіцієнта подібності 2	Кількість слів	Кількість символів

Тривога

У цьому розділі ви знайдете інформацію щодо текстових спотворень. Ці спотворення в тексті можуть говорити про МОЖЛИВІ маніпуляції в тексті. Спотворення в тексті можуть мати навмисний характер, але частіше характер технічних помилок при конвертації документа та його збереженні, тому ми рекомендуємо вам підходити до аналізу цього модуля відповідально. У разі виникнення запитань, просимо звертатися до нашої служби підтримки.

Заміна букв		0
Інтервали		0
Мікропробіли		0
Білі знаки		0
Парафрази (SmartMarks)	a	3

Подібності за списком джерел

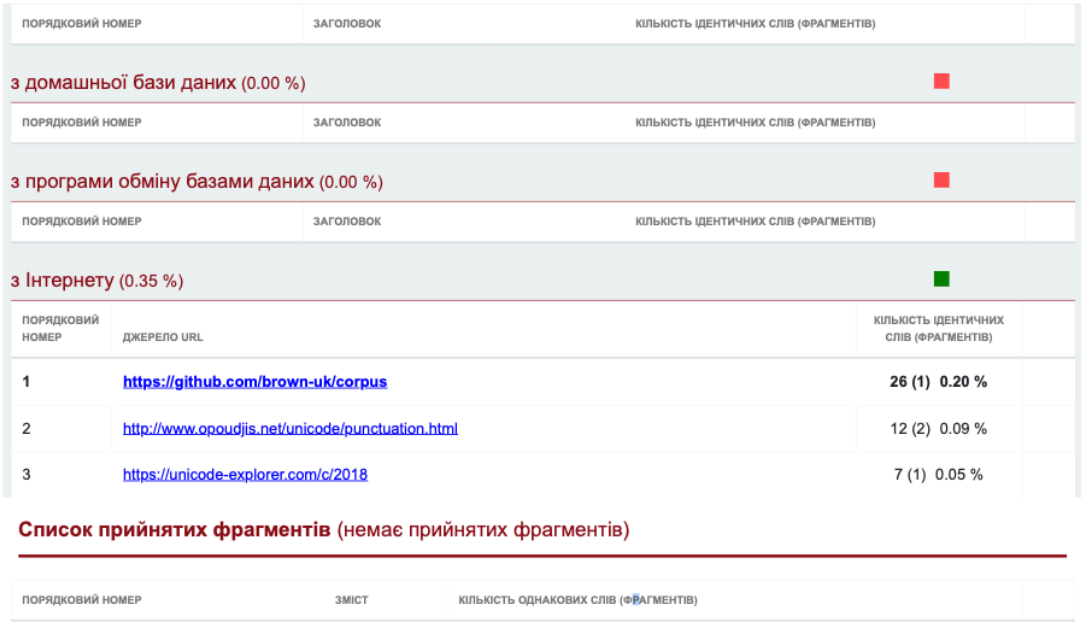
Нижче наведений список джерел. В цьому списку є джерела із різних баз даних. Колір тексту означає в якому джерелі він був знайдений. Ці джерела і значення Коефіцієнту Подібності не відображають прямого плагіату. Необхідно відкрити кожне джерело і проаналізувати зміст і правильність оформлення джерела.

10 найдовших фраз — Колір тексту

ПОРЯДКОВИЙ НОМЕР	НАЗВА ТА АДРЕСА ДЖЕРЕЛА URL (НАЗВА БАЗИ)	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	https://github.com/brown-uk/corpus	26 0.20 %
2	https://unicode-explorer.com/c/2018	7 0.05 %
3	http://www.opoudjis.net/unicode/punctuation.html	6 0.05 %
4	http://www.opoudjis.net/unicode/punctuation.html	6 0.05 %

з бази даних RefBooks (0.00 %)

Рисунок В.1 – Звіт подібності, перша сторінка



ПОРЯДКОВИЙ НОМЕР	ЗАГОЛОВОК	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)

з домашньої бази даних (0.00 %)

ПОРЯДКОВИЙ НОМЕР	ЗАГОЛОВОК	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)

з програми обміну базами даних (0.00 %)

ПОРЯДКОВИЙ НОМЕР	ЗАГОЛОВОК	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)

з Інтернету (0.35 %)

ПОРЯДКОВИЙ НОМЕР	ДЖЕРЕЛО URL	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	https://github.com/brown-uk/corpus	26 (1) 0.20 %
2	http://www.opoudjis.net/unicode/punctuation.html	12 (2) 0.09 %
3	https://unicode-explorer.com/c/2018	7 (1) 0.05 %

Список прийнятих фрагментів (немає прийнятих фрагментів)

ПОРЯДКОВИЙ НОМЕР	ЗМІСТ	КІЛЬКІСТЬ ОДНАКОВИХ СЛІВ (ФРАГМЕНТІВ)

Рисунок В.2 – Звіт подібності, друга сторінка

ДОДАТОК Г

Слайди презентації

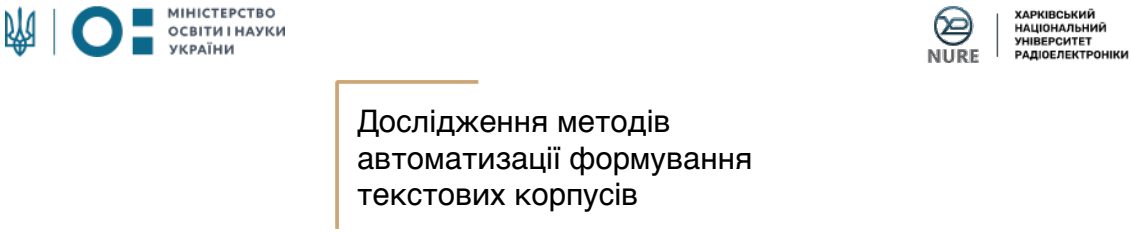


Рисунок Г.1 – Слайд №1

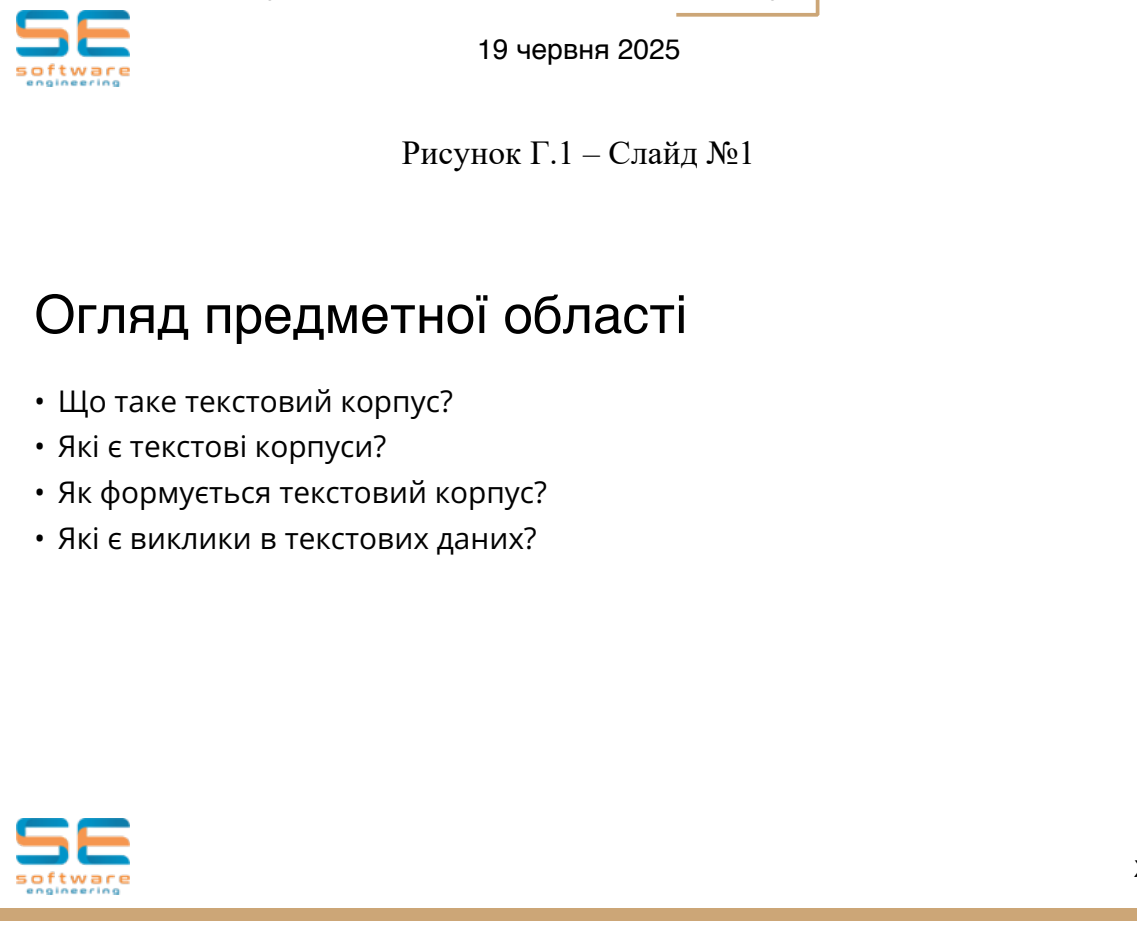


Рисунок Г.2 – Слайд №2

Огляд предметної області

Що таке текстовий корпус?

Текстовий корпус – це структурована збірка текстових даних. Корпуси текстів є основою для розвитку систем обробки природної мови

Які є текстові корпуси?

English Corpora, BNC, БрУк, UberText 1.0, UberText 2.0, Mova.info, ГРАК

x

Рисунок Г.3 – Слайд №3

Огляд предметної області

Як формується текстовий корпус?

1. Обрання джерел. Новинні сайти, книги, соціальні мережі
2. Вилучення тексту. HTML-парсинг, OCR, STT, API
3. Нормалізація. Уніфікація форматування, символів, конструкцій
4. Оцінка якості. Виявлення та виправлення помилок
5. Токенізація. Розбиття на аналітичні одиниці
6. Розмітка та збереження. Структурна анотація

x

Рисунок Г.4 – Слайд №4

Огляд предметної області

Які є виклики в текстових даних?

- Форматування. Різні типи лапок, апострофів, телефонних номерів
- Мовні особливості. Діалекти, суржик, регіональні варіанти
- Помилки написання. Граматичні, орфографічні та інші помилки

Які наслідки?

Неякісні корпуси текстів, шум для NLP-моделей, збільшення розмірності

Рисунок Г.5 – Слайд №5

Дослідження

Метою роботи є підвищення якості україномовних текстових корпусів та систем їх формування

Для досягнення поставленої мети потрібно:

- Дослідити текстові корпуси
- Дослідити методи та інструменти для формування текстових корпусів

Рисунок Г.6 – Слайд №6

Аналіз підходів до формування корпусів

- БрУК. Висока якість, напівручна обробка, низька масштабованість, мінімальна розмітка
- Uber Text 1.0. Низька якість, код недоступний, немає структури та розмітки
- UberText 2.0. Середня якість, повна автоматизація, проблеми нормалізації та токенізації, код частково недоступний, немає розмітки
- ГРАК. Висока якість, напівручна обробка, низька масштабованість, недоступний для завантаження

x

Рисунок Г.7 – Слайд №7

Аналіз літературних джерел по нормалізації

- Aliero та ін. Загальний огляд підходів до нормалізації
- Starko та ін. (ГРАК). Використання LanguageTool та ручна обробка
- Вакуленко. Нормалізація для text-to-speech
- Chaplynskyi (UberText 2.0). Використання LT, додаткові вимоги до тексту
- Cudak та ін. Нормалізація телефонних номерів для порівняння POI
- Van der Goot та Çetinoğlu. Проблеми лексичної нормалізації в текстах зі змішаними мовами.

x

Рисунок Г.8 – Слайд №8

Аналіз інструментів для токенізації

- SpaCy. Використовує правила на основі префіксів, суфіксів та інфіксів
- Stanza. Базується на навчених нейронних мережах та універсальному токенізаторі
- NLTK. Пропонує кілька варіантів токенізації, охоплює регулярні вирази та правила
- Language Tool. Використовує правила та n-грами
- Lang-UK. Використовує спеціально розроблений для української мови алгоритм, заснований на регулярних виразах

X

Рисунок Г.9 – Слайд №9

Підсумки аналізу джерел

Проблеми:

- Немає гнучкого та масштабованого рішення для формування корпусів
- Немає методів нормалізації телефонних номерів, апострофів та лапок в новинних текстів
- Токенізація ігнорує семантичну єдність конструкцій

Зрештою, необхідно розробити вдосконалену систему автоматичного формування текстових корпусів (САФТК)

X

Рисунок Г.10 – Слайд №10

Опис вимог до програмної системи

- Необхідно розробити модульну систему, що забезпечуватиме повний цикл обробки тексту. Система повинна відповідати вимогам гнучкості, розширюваності та забезпечувати високу якість текстів
- Система повинна нормалізовувати телефонні номери, лапки та апострофи згідно з офіційними рекомендаціями
- Система повинна виконувати семантичну токенізацію зі збереженням цілісності семантичних конструкцій та дотримуватись ДСТУ 3582:2013
- Інші вимоги, наприклад, підхід до анотації файлів

X

Рисунок Г.11 – Слайд №11

Архітектура програмної системи

Модульний підхід. Незалежні компоненти з чіткими інтерфейсами

Конвеєрна обробка. Джерело → Запис → Парсинг → Нормалізація → Оцінка та виправлення→ Токенізація → Збереження

Гнучкість. Легка заміна компонентів без зміни загальної логіки

XML-формат. metadata, errors, warnings, document, h, p, ol, ul, li, s, t, q

X

Рисунок Г.12 – Слайд №12

Покращення нормалізації

- Нормалізація пробілів. a б → a б
- Нормалізація апострофів. ‘п'ятниця’ → "п’ятниця"
- Нормалізація лапок. a "b «c» d" → a «b “c” d»
- Нормалізація номерів. +380991234567 → +380 (99) 123-45-67

X

Рисунок Г.13 – Слайд №13

Покращення токенізації

Сформовано тестові набори. будь-який; бозна-що; 2022-му; 18.05.2021; 22:37; $1,461; м/с; м²; км²/місяць; тощо

Обрано токенізатор для покращення. NLTK (69.23%) та SpaCy (53.85%)

Виконано модифікації. Оновлено суфікси, префікси, інфекси, додано ретокенізацію

X

Рисунок Г.14 – Слайд №14

Методологія експерименту

- Оцінка нормалізації на Uber Text 2.0 News Cleaned
- Оцінка токенізації на уривках новин «Суспільне Новини»
- Оцінка системи на обробці архіву «Суспільне Новини»

software engineering

X

Рисунок Г.15 – Слайд №15

Результати нормалізації

Телефоні номери до нормалізації:

№	Формат	Кількість збігів	Відсоток від загальної кількості
1	(0XX) XXX-XX-XX	11827	19.258451
2	0XX-XXX-XX-XX	6407	10.432814
3	0XX XXX XX XX	4964	8.083111
4	0XXXXXXXXX	4232	6.891161
5	+380 (XXX) XX-XX-XX	4047	6.589917

Телефоні номери після нормалізації:

№	Формат	Кількість збігів	Відсоток від загальної кількості
1	+380 (XX) XXX-XX-XX	51489	83.84192
2	+380 (XXX) XX-XX-XX	9923	16.15808

software engineering

X

Рисунок Г.16 – Слайд №16

Результати нормалізації

Розподіл апострофів до та після нормалізації:

Символ	До нормалізації		Після нормалізації	
	Кількість	Відсоток	Кількість	Відсоток
U+0027	7465048	60.5074	5519	0.0448
U+2019	4693904	38.0461	1226	0.0100
U+02BC	144521	**1.1714**	12302968	**99.9403**
U+0060	20895	0.1694	397	0.0032
U+2018	12824	0.1039	209	0.0017
U+02B9	214	0.0017	0	0
U+02BB	10	0.0001	0	0
Загалом	12337416	100.0000	12310319	100.0000

software engineering

X

Рисунок Г.17 – Слайд №17

Результати нормалізації

Розподіл лапок до та після нормалізації:

Символ	До нормалізації		Після нормалізації	
	Кількість	Відсоток	Кількість	Відсоток
U+0022	29002764	51.743	7780	0.014
U+00AB	12495485	**22.293**	26895177	**47.961**
U+00BB	12333176	**22.003**	26840011	**47.863**
U+201D	1065203	**1.900**	1166659	**2.080**
U+201C	1046085	**1.866**	1167062	**2.081**
U+201E	109199	0.195	0	0.000
U+201F	16	0.000	0	0.000
U+275D	4	0.000	0	0.000
U+275E	4	0.000	0	0.000
Загалом	56051936	100.000	56076689	100.000

software engineering

X

Рисунок Г.18 – Слайд №18

Результати токенізації

Порівняльний аналіз результатів роботи розробленого токенізатора з еталонною розміткою показав повний збіг усіх 507 токенів, що підтверджує коректність роботи алгоритму та його відповідність встановленим вимогам.

Рисунок Г.19 – Слайд №19

Результати роботи системи

- Корпус з 413 новин за 31 березня 2025 року
- Час обробки одного документа триває від 3 до 10 секунд

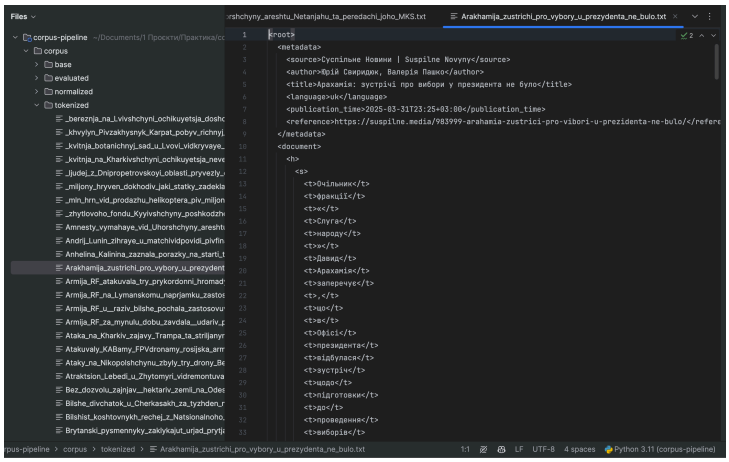


Рисунок Г.20 – Слайд №20

Аналіз отриманих результатів

- Сформовано високоякісний анотований корпус з 413 новин
- Реалізовано гнучку САФТК з модульною архітектурою
- Покращено семантичну токенізацію зі збереженням цілісності конструкцій
- Стандартизовано непослідовні елементи згідно з офіційними рекомендаціями

software engineering

X

Рисунок Г.21 – Слайд №21

Публікація результатів


Text Normalization Challenges in Ukrainian News Corpus: Problem Analysis and Research Framework

software engineering

X

Рисунок Г.22 – Слайд №22

Підсумки

Система готова до практичного використання з відкритим кодом та модульною архітектурою для легкої адаптації під різні завдання корпусної лінгвістики.

Перспективи розвитку:

- Розширення семантичної токенізації
- Розширення нормалізації на міжнародні телефонні номери
- Оптимізація компоненту пошуку помилок
- Інтеграція розумного виправлення помилок
- Реалізація механізму відстеження прогресу
- Розширення на інші типи джерел

software engineering

X

Рисунок Г.23 – Слайд №23

Які є запитання?

software engineering

X

Рисунок Г.24 – Слайд №24

ДОДАТОК Д

Апробація результатів роботи

210

Text Normalization Challenges in Ukrainian News Corpus: Problem Analysis and Research Framework

Danylo O. Horielov[a] and Oleksander V. Vechur[a]

[a] *Kharkiv National University of Radio Electronics, 14 Nauky Ave., Kharkiv, 61166, Ukraine*

Abstract

Ukrainian news texts contain significant formatting inconsistencies that affect their quality for computational linguistics applications. This paper identifies three critical normalization problems in Ukrainian news corpora: inconsistent phone number formats, mixed apostrophe symbols, and improper quotation mark handling. We propose a systematic research framework using rule-based normalization approaches designed for Ukrainian orthographic conventions. Our methodology includes plans for evaluation on the UberText 2.0 corpus and development of specialized tools for Ukrainian text processing. The research aims to establish standardized formats that improve text consistency for language model training and natural language processing applications.

Keywords

Text normalization, ukrainian language, phone numbers, apostrophes, quotation marks, natural language processing

1. Introduction

Ukrainian news texts are valuable resources for natural language processing and language model training. However, these texts contain formatting inconsistencies that reduce their quality and usability for computational analysis. While general text preprocessing handles basic normalization like symbol standardization, Ukrainian news texts need specialized attention for language-specific formatting problems.

We have identified three main normalization challenges through analysis of Ukrainian news sources. First, phone numbers appear in many formats across news articles, making information extraction difficult. Second, apostrophe symbols vary significantly, with multiple Unicode characters used for the same purpose. Third, quotation marks show inconsistent usage with improper nesting and mixed styles within articles.

These issues require systematic solutions that follow Ukrainian orthographic standards and practical formatting guidelines. This research proposes a framework to address these specific normalization challenges in Ukrainian news texts.

2. Related work

Text normalization transforms text into a standardized form for natural language processing applications. Aliero et al. [1] identify four main approaches: rule-based methods using predefined rules, statistical methods learning from data, neural network methods using deep learning, and hybrid methods combining multiple approaches. For specialized tasks like phone number standardization, rule-based approaches remain common.

For Ukrainian texts, Starko et al. [2] described preprocessing approaches for the GRAC corpus, including Latin-Cyrillic character confusion correction and basic normalization using LanguageTool's Ukrainian module. Vakulenko [3] focused on standardizing letters, numerals, and measurements for

MIT@AIS'2025s: 1st International Scientific and Practical Conference "Modern Information Technologies and Artificial Intelligence Systems", May 19–22, 2025, Kharkiv-Yaremche, Ukraine
EMAIL: horielov.danylo@gmail.com (A. 1); alexander.vechur@nure.ua (A. 2)
ORCID: 0009-0005-9393-6231 (A. 1); 0000-0001-9605-1475 (A. 2)

Рисунок Д.1 – Тези опубілковані на 1-й Міжнародній науково-практичній конференції «Сучасні інформаційні технології та системи штучного інтелекту MIT@AIS-2025», перша сторінка

text-to-speech applications. Chaplynskyi [4] introduced UberText 2.0 with comprehensive preprocessing pipelines for Ukrainian text normalization.

Phone number normalization has been addressed by Cudak et al. [5], who focused on removing formatting characters for computational comparison rather than standardizing display formats. Van der Goot and Çetinoğlu [6] explored lexical normalization in code-switched multilingual texts, showing improvements in downstream tasks.

However, existing approaches do not specifically address the combination of phone number standardization, quotation mark nesting, and apostrophe disambiguation that characterizes Ukrainian news texts.

3. Problem analysis

3.1. Phone number format inconsistencies

Ukrainian news sources show different variations in phone number formats. Common patterns include +380XXXXXXXXX, +380 (XX) XXX XX XX, and +38 (0XX) XXX-XX-XX. These differences come from various editorial policies, automated content systems, and different source formatting rules.

The Ukrainian government design system recommends +380 (XX) XXX-XX-XX for standard numbers [7]. We propose adding +380 (XXX) XX-XX-XX for special three-digit service numbers to better distinguish toll-free and paid services from regular phone numbers.

3.2. Apostrophe symbol variations

Ukrainian orthography uses apostrophes to separate certain vowels from preceding consonants. However, news texts use various Unicode symbols (U+0027, U+02BC, U+201E etc.) for the same purpose. This variation makes automated processing difficult and creates confusion when distinguishing apostrophes from quotation marks.

While official Ukrainian orthography does not specify which apostrophe symbol to use, standardization is necessary for consistent processing [8]. The U+02BC symbol offers advantages due to its curved appearance, default use in Ukrainian keyboards, and visual distinction from quotation marks.

3.3. Quotation mark handling issues

Ukrainian news texts show inconsistent quotation mark usage with various symbols (U+0022, U+201D, U+201E etc.). Problems include improper nesting, mixing different styles within articles, and confusion between quotation marks and apostrophes.

Ukrainian orthographic standards require guillemets (U+00AB and U+00BB) for outer quotations and curly quotes (U+201C and U+201D) for inner quotations [8]. Proper implementation needs contextual analysis to determine opening versus closing positions and systematic handling of nested quotation levels.

4. Proposed research framework

4.1. Phone number normalization approach

Our approach will focus exclusively on Ukrainian phone numbers to ensure high accuracy. The methodology involves pattern recognition based on Ukrainian country codes and valid operator prefixes. We plan to implement multiple regular expression patterns to capture format variations while preventing false matches through verification checks.

The normalization process will convert all valid formats into two standardized forms: +380 (XX) XXX-XX-XX for regular numbers and +380 (XXX) XX-XX-XX for special service numbers. This approach maintains semantic distinction while achieving format consistency.

Рисунок Д.2 – Тези опубілковані на 1-й Міжнародній науково-практичній конференції «Сучасні інформаційні технології та системи штучного інтелекту MIT@AIS-2025», друга сторінка

4.2. Apostrophe disambiguation strategy

The proposed algorithm will distinguish between apostrophes and quotation marks through contextual analysis. Key factors include character patterns, position within text, and surrounding punctuation context. The approach will convert symbols between letters to apostrophes while treating symbols at text boundaries or near punctuation as potential quotation marks.

4.3. Quotation mark standardization method

The quotation mark normalization process will operate in multiple phases: initial consolidation of different symbols, contextual determination of opening versus closing positions, and implementation of nested quotation styles. The approach will consider adjacent characters, text boundaries, existing quotation marks, and punctuation context.

The methodology will implement Ukrainian orthographic standards with alternating styles for multiple nesting levels, ensuring proper visual organization and readability.

5. Evaluation plan and expected results

5.1. Evaluation methodology

We plan systematic evaluation using the UberText 2.0 corpus, specifically the News Cleaned subcorpus containing 21.37 GB of Ukrainian news articles. The evaluation framework will include pattern extraction to analyze format distribution before and after normalization, statistical analysis of conversion accuracy, and error categorization for improvement.

Success metrics will include format consolidation rates for phone numbers, apostrophe standardization percentages, and proper quotation mark nesting implementation. The evaluation will also assess processing warnings and errors to identify cases requiring additional attention.

5.2. Expected outcomes

This research aims to provide specialized normalization tools designed for Ukrainian news texts. Expected results include significant reduction in format variations, improved text consistency for processing, and enhanced corpus quality for language model training.

Based on preliminary analysis, we expect to consolidate hundreds of phone number formats into two standard formats, achieve high apostrophe standardization rates, and implement proper quotation mark handling according to Ukrainian conventions. The methodology development will contribute to Ukrainian computational linguistics resources.

6. Future research directions

Following successful implementation and evaluation, future work will focus on extending phone number normalization to international formats, developing adaptive approaches for corpus-specific patterns, and integrating the normalization tools into comprehensive text preprocessing systems.

Additional research directions include optimizing performance for large-scale processing, addressing other normalization aspects such as hyphen standardization, and exploring neural approaches for more complex normalization challenges in Ukrainian text processing.

7. Conclusion

This paper establishes a research framework for addressing critical normalization challenges in Ukrainian news texts. The proposed methodology targets specific problems of phone number

Рисунок Д.3 – Тези опубілковані на 1-й Міжнародній науково-практичній конференції «Сучасні інформаційні технології та системи штучного інтелекту MIT@AIS-2025», третя сторінка

inconsistencies, apostrophe symbol variations, and quotation mark handling through specialized rule-based approaches designed for Ukrainian orthographic conventions.

The research will contribute to improving Ukrainian language resource quality and supporting more effective natural language processing applications for Ukrainian texts. Implementation and evaluation on large-scale corpora will provide valuable insights for Ukrainian computational linguistics and establish best practices for news text preprocessing.

8. Declaration on generative AI

During the preparation of this work, the author(s) used ClaudeAI in order to: Text Translation, Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content

9. References

[1] A.A. Aliero, B.S. Adebayo, H.O. Aliyu, A.G. Tafida, B.U. Kangiwa, N.M. Dankolo, Systematic Review on Text Normalization Techniques and its Approach to Non-Standard Words, International Journal of Computer Applications 185(33) (2023) 44-55.

[2] V. Starko, A. Rysin, M. Shvedova, Ukrainian text preprocessing in GRAC, in: Proceedings of the 16th IEEE International Conference on Computer Sciences and Information Technologies (CSIT 2021), Vol. 2, IEEE, 2021, pp. 101–104.

[3] M. Vakulenko, Normalization of Ukrainian letters, numerals, and measures for natural language processing, Digital scholarship in the humanities 38(3) (2023) 1307-1321.

[4] D. Chaplynskyi, Introducing UberText 2.0: A corpus of modern Ukrainian at scale, in: Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP 2023), ACL, 2023, pp. 1–10.

[5] M. Cudak, M. Piech, R. Marcjan, Sparse data classifier based on the first-past-the-post voting system, Computer Science 23(2) (2022) 277–296.

[6] R. van der Goot, Ö. Çetinoğlu, Lexical normalization for code-switched data and its effect on POS-tagging, arXiv preprint arXiv:2006.01175 (2020).

[7] Orfohrafiia ta syntaksys [Orthography and syntax], Dyzain systema derzhavnykh saitiv Ukrainy [Design system of ukrainian government websites], 2019. URL: https://design.gov.ua/ua/teksty-i-kontent/orfografiya-ta-sintaksis.

[8] Natsionalna akademiia nauk Ukrainy, Ukrainskyi pravopys [Ukrainian orthography], Naukova Dumka, Kyiv, 2019.

Рисунок Д.4 – Тези опубілковані на 1-й Міжнародній науково-практичній конференції «Сучасні інформаційні технології та системи штучного інтелекту MIT@AIS-2025», четверта сторінка

Рисунок Д.5 – Сертифікат учасника 1-ї Міжнародної науково-практичної конференції «Сучасні інформаційні технології та системи штучного інтелекту MIT@AIS-2025»

ДОДАТОК Е

Експертний висновок результатів перевірки кваліфікаційної роботи на відповідність оформлення вимогам ДСТУ 3008: 2015

Експертний висновок результатів перевірки кваліфікаційної роботи

студент (посада) | програмної інженерії (кафедра) | ІПЗм-23-3 (група)

Горєлов Данило Олександрович
(прізвище, ім'я, по батькові)

Зауваження

Пункт ДСТУ 3008-2015	Зміст пункту	Сторінка кваліфікаційної роботи
1	2	3
	7.1 Загальні положення	
	7.3 Нумерація сторінок звіту	
	7.4 Нумерація розділів, підрозділів, пунктів, підпунктів	
	7.5 Рисунки	
	7.6 Таблиці	
	7.7 Переліки	
	7.8 Примітки	
	7.9 Виноски	
	7.10 Формули та рівняння	
	7.11 Посилання	
	7.13 Список авторів	
	7.14 Скорочення та умовні познаки	
	7.15 Додатки	

зауважень немає

Експерт _____ (підпис) Олена ОЛІЙНИК (прізвище, ініціали)

11.06.2025

Рисунок Е.1 – Експертний висновок результатів перевірки кваліфікаційної роботи на відповідність оформлення вимогам