МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Київський національний університет імені Тараса Шевченка

Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

Тональний словник української мови

Кваліфікаційна робота
освітнього ступеня «бакалавр»
за спеціальністю 035 «Філологія»,
спеціалізацією 035.10 «Прикладна лінгвістика»,
галузі знань 03 «гуманітарні науки»
ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова»
Оксани ТОЛОЧКО

Науковий керівник:
Валентина РОБЕЙКО
Науковий консультант:
к.техн.н. Микола КОСТІКОВ

Рецензент:
к.техн.н. Микола САЖОК

«Допущено до захисту»
Протокол № 11 засідання кафедри
української мови та прикладної лінгвістики
ННІФ від 01.06.2023
Завідувач кафедри_____**Сергій Різник**

КИЇВ – 2023

ЗМІСТ

ВСТУП

Останні кілька десятиліть автоматичний аналіз настроїв, або тональний аналіз, сентимент-аналіз (клас методів аналізу в комп'ютерній лінгвістиці, призначений для виявлення в текстах емоційно забарвленої лексики [23]) широко застосовується в різних сферах життя. Наприклад, це може бути аналіз настроїв у новинних текстах чи аналіз відгуків на Інтернет-сайтах із метою фільтрації відгуків на позитивні та негативні. Великий обсяг інформації, який зростає та змінюється щодня, специфіка мови кожної зі сфер спілкування, мови новин зокрема, роблять таку задачу доволі складною.

Одним із найголовніших та водночас найскладніших та найбільш трудомістких етапів процесу розробки системи сентимент-аналізу є створення тонального словника окремої мови.

Нині для української мови є лише один тональний словник української мови в вільному доступі (проект групи lang-uk [40]), обсяг якого становить 6 859 слів. Востаннє цей тональний словник редагувався експертами в вересні 2016 року. Оскільки активний словниковий склад мови постійно змінюється, виникає потреба в створенні нового тонального словника для української мови, який буде містити найбільш вживані та найбільш частотні слова, що й зумовлює ***актуальність*** нашої роботи. Надалі створений словник та детальні інструкції й описи його створення будуть опубліковані в мережі "Інтернет" з можливістю вільного доступу.

Метою роботи є створення тонального словника української мови. Для досягнення поставленої мети необхідно виконати такі ***завдання***:

– проаналізувати сучасні системи сентимент-аналізу;

– проаналізувати наукові праці щодо створення та побудови тональних словників;

– опрацювати, систематизувати та доповнити дані тонального словника;

– провести тональну та лінгвістичну розмітку слів;

– проаналізувати створений словник та описати всі тонкощі його побудови, зокрема статистичні показники;

– створити програму автоматичного добування статей новин із сайтів новин в Інтернет-просторі;

– створити програму автоматичної побудови частотного словника для статей новин, щоб з'ясувати достатність обсягів створеного тонального словника (частка покриття тексту).

Об'єктом роботи є питомі слова української мови та слова іншомовного походження.

Предметом є тональність слів, яка досліджується на матеріалі двох попередньо анотованих, а також власноруч зібраних списків слів.

Матеріал дослідження:

1) два списки слів, які були попередньо частково розмічені експертами-студентами Донецького національного університету імені Василя Стуса, філологічного факультету, кафедри загального та прикладного мовознавства й слов'янської філології. Обсяг першого списку становив 40 567 слововживань, другого – 7 000 лексем;

2) чотири списки слів, які були відібрані й упорядковані нами в ході роботи.

У ході роботи було використано різноманітні ***методи*** емпіричного та теоретичного дослідження (спостереження, порівняння, експеримент, сходження від абстрактного до конкретного, формалізація), а також загальнонаукові ***методи*** (абстрагування, аналіз і синтез, індукція та дедукція, моделювання).

Теоретичне значення роботи полягає в виокремленні критеріїв для аналізу та класифікації тональних словників, визначенні понять "тональний словник" і "сентимент-аналіз".

Практичне значення роботи – створення тонального словника для української мови, який містить питомі українські слова та слова іншомовного походження, а також оцінку їхньої тональності. Створений словник може бути використано як інструмент для автоматичного сентимент-аналізу українськомовних текстів, особливо для текстів медійного стилю.

Новизною цієї роботи є створення тонального словника для української мови з тональною, лінгвістичною й екстралінгвістичною розміткою.

Структура роботи – вступ, чотири розділи, загальні висновки, список використаних джерел та програмних бібліотек і 19 додатків.

РОЗДІЛ 1. ПРИНЦИПИ ПОБУДОВИ ТА КЛАСИФІКАЦІЇ ТОНАЛЬНИХ СЛОВНИКІВ

Кожен, хто прагне вільно володіти тією чи іншою мовою, звертається за допомогою до словників. Зі шкільного курсу учні знають про тлумачний, орфографічний, орфоепічний, етимологічний словники, словники синонімів, антонімів, омонімів тощо.

У наш час стрімкий розвиток технологій та галузі автоматичної обробки природної мови потребує створення ще одного виду словників – тональних. Це саме ті словники, які орієнтовані на опрацювання природної мови. На думку багатьох мовознавців, розробка такого типу словників допоможе в автоматичному визначенні тональності того чи іншого тексту, з якої б сфери діяльності людини він не був узятий. Тоді виникає ряд запитань у кожного, хто щойно прочитав ці рядки: що таке тональний словник? Для чого все-таки потрібен тональний словник? Навіщо з'ясовувати тональність тексту?.. Відповіді на ці та інші запитання буде подано в цій роботі.

1.1. Поняття тонального словника

Насамперед варто чітко окреслити, що таке тональний словник. Словник – це "книга, в якій в алфавітному чи тематичному порядку подано слова якоїсь мови (з тлумаченням, перекладом на іншу мову і т. ін.)", відповідно до одного чи декількох висвітлених аспектів у ній [29, Том 9, с. 367]. Варто зауважити, що кожен словник має своє конкретне призначення [33]. На думку багатьох вчених та письменників, словник – це дещо більше, ніж просто книга з впорядкованими даними. "Словники – це не лише довідники, але й елемент національної культури, адже в слові втілено багато граней народного життя. Все багатство й різноманіття лексичних запасів певної мови зібрано в словниках" [34]. Оскільки словників дуже багато й вони мають різне призначення, всі вони були поділені та класифіковані між собою на типи та види.

Коли ми говоримо про певний тип словника, то майже відразу можна приблизно описати призначення словника та його складові елементи. А ось із тональним словником не все так просто. То чому ж так складно дати визначення тональному словникові?..

На думку А. Романюка та М. Романишин, тональний словник, у найпростішому його вигляді, становить список слів і словосполучень зі значенням тональності для кожного слова [32, с. 63-64].

Тоді виникає нове запитання: а що таке тональність?

У нас є відповідь і на це запитання. Якщо брати термін “тональність”, без посилання на конкретний вид мистецтва чи діяльності, то в тлумачному словнику натрапимо на таке визначення: “характер, сила звучання тону, голосу” [29, Том 10, с. 186]. Але цього визначення недостатньо для характеристики тональності слів. Тепер згадаймо, для чого створюються ці тональні словники? Так, якщо говорити досить просто, то тональні словники створюються з метою визначення тональності того чи іншого тексту. І тому стає зрозуміло, що поняття тональності стосується конкретної сфери – мови. Стає очевидним, що тональність – це ще один із аспектів значення слова. В тлумачному словнику знаходимо визначення тональності в літературі: “основна емоційна настроєність твору” [Там само]. Але й це ще не відповідає поставленій меті цієї роботи, адже робота проводиться не над творами художньої літератури, а над текстами, які подаються у вигляді відгуків чи статей. Тому тональність слід трактувати як “емоційне ставлення автора висловлювання до деякого об’єкту (об’єкту реального світу, події, процесу або їх властивостями / атрибутам), виражене в тексті” [23]. Та емоційна складова, яка виражена на рівні лексеми або комунікативного фрагмента, називається лексичною тональністю (або лексичним сентиментом) [Там само]. Це саме той аспект щодо тональності, який цікавить нас у цій роботі. Тепер розуміння тонального словника стає більш чітким.

Тому, трохи видозмінивши початкове трактування тонального словника, спробуємо дати чітке визначення цьому поняттю.

Отже, **тональний словник – це словник, що становить список слів і словосполучень зі значенням основної емоційної настроєності для кожного слова**.

Також тональний словник, або словник аналізу настроїв (дослівно: sentiment analysis dictionary) – це той словник, який містить інформацію про емоції або полярність, які виражені словами, фразами або поняттями [4]. Це говорить про те, що реєстровою одиницею словника може бути не лише окреме слово, а навіть словосполучення та цілі фрази. На практиці словник зазвичай надає один або кілька балів для кожного слова. Потім ми можемо використовувати ці бали для обчислення загальних настроїв вхідного речення на основі окремих слів [Там само].

Отже, коли поняття тонального словника з'ясовано, можна розглянути його різновиди.

1.2. Різновиди тональних словників

Процес побудови та створення словників – це основне завдання особливої галузі лінгвістичної науки – лексикографії [34]. Тональні словники – це саме ті словники, які створюються з метою сентимент-аналізу тексту, або аналізу тональності тексту.

Аналіз тональності тексту (сентимент-аналіз, тональний аналіз, англ. Sentiment analysis, англ. Opinion mining) – це "клас методів контент-аналізу в комп'ютерній лінгвістиці, призначений для автоматизованого виявлення в текстах емоційно забарвленої лексики й емоційної оцінки авторів (думок) по відношенню до об'єктів, мова про які йде в тексті" [16; 23].

Можна виділити два різновиди тональних словників, які між собою мають незначну, але вагому різницю:

1) тональний словник як публічно доступний лексичний ресурс, який є автоматизованим [16];
2) тональний словник як набір проаналізованих даних (dataset) [28].

Якщо коротко спробувати охарактеризувати кожен різновид, то перший вид тональних словників – це достатньо добре опрацьований обсяг слів із залученням експертів та автоматизацією результатів на досить високому та професійному рівні за допомогою використання елементів машинного навчання, “такі як прихований семантичний аналіз, метод опорних векторів” та “мішок слів” [23]. Такий вид тональних словників можна доповнювати згодом автоматично, без повторного залучення експертів, оскільки базовий ресурс слів створюється з метою навчити комп’ютер, щоб у подальшому він самостійно (із певним контролем людини) міг визначати тональність кожного нового слова.

Наступний вид тональних словників – це набір проаналізованих даних. Такі словники теж можуть залучати експертів для їх створення або ж проводити тестування серед “натовпу” людей (термін, який вживається досить широко на позначення цього процесу – краудсорсинг [28]). Найчастіше тональні словники цього типу створюються на основі сентимент-анотованого корпусу.

Представлені два різновиди словників розглядаються як дві окремі категорії аналізу тональності:

1) автоматизований аналіз тональності;

2) ручний (або аналіз тональності експертами), відповідно до власних виокремлених різновидів [16].

Найбільш помітні відмінності між цими двома підходами лежать в ефективності системи й точності аналізу. У комп’ютерних програмах автоматизованого аналізу тональності застосовують алгоритми машинного

навчання, інструменти статистики і обробки природної мови, що дозволяє обробляти великі масиви тексту, включаючи веб-сторінки, онлайн-новини, тексти дискусійних груп в мережі “Інтернет”, онлайн-огляди, веб-блоги та соціальні медіа [Там само].

Отже, щоб визначити, що конкретний новостворений тональний словник можна віднести до певного виду, нами було виокремлено такі критерії, на яких потрібно акцентувати свою увагу:

1) ким створюється цей словник (залучення експертів чи переважно тестування серед широкого загалу людей);

2) як створюється словник (комп’ютеризація, автоматизація чи вручну);

3) який рівень залучення комп’ютерних технологій (високий чи низький);

4) який рівень обробки даних комп’ютером (вищий чи нижчий відповідно).

1.3. Принципи формування тональних словників, їхня порівняльна характеристика

Оскільки тональні словники почали створюватися для різних мов, в основному для англійської, та розширювати сферу свого застосування, також починають розроблятися різні структуровані та упорядковані принципи їх формування.

Нами було проаналізовано п’ять тональних словників: один із української мови [40], три з англійської [4; 16; 17] та один з російської [28]. Варто зазначити, що SentiWords (тональний словник англійської мови) походить від SentiWordNet (SWN), що зумовлює збіг даних за багатьма критеріями [17]. Кількість проаналізованих словників невелика, оскільки кількість тональних словників у вільному доступі теж доволі незначна.

У ході роботи було вирішено здійснити аналіз та порівняння тональних словників за кількома параметрами (далі П1 – П8): П1 – ким здійснювалася оцінка тональності слів, П2 – яка шкала для визначення тональності застосовувалася, П3

– обсяг словника, П4 – розподіл балів щодо частин мови, П5 – чи застосовувалися у ході створення словника графічні моделі та чи представлено графічний інтерфейс для користувача, П6 – із якою метою створювався словник, П7 – чи вказана на сайті словника інформація щодо подальших планів розробників щодо словника, П8 – на основі поданих даних визначити, до якого різновиду належить словник.

Першим було проаналізовано "***Тональний словник української мови" (tone-dict-uk.tsv)*** [40]. П1 – двома експертами. П2 – оцінка в діапазоні від -2 до +2 (без 0) та є примітка, що досліджено слова, які мають не нейтральну тональність [34]. П3 – 6 859 слів. П4 – відсутній. П5 – присутні (таку інформацію подано в файлі tone-dict-uk-auto.tsv) [40]. П6 – авторами словника на досліджуваному нами ресурсі нічого про це не сказано. Робимо припущення, що основна мета – це створення тонального словника української мови, який може доповнюватися та досліджуватися лінгвістами. П7 – про подальші плани теж нічого не сказано. І знову дозволимо собі припустити, що плани щодо розвитку тонального словника спираються на сформульовану нами мету. П8 – автоматизований.

"Багатомовний тональний словник" (SentiWordNet), який було проаналізовано на основі англійської мови [16; 4]. П1 – п'ятьма експертами. П2 – два вектори: позитивно-негативна полярність та суб'єктивно-об'єктивна; триєдине маркування: тональність може бути позитивна, негативна або об'єктивна. Присутня теза про те, що об'єктивна тональність може розглядатися як нейтральна (не відноситься до позитивної чи нейтральної тональності) [17]. П3 – словник постійно доповнюється, тож складно охарактеризувати його обсяги. П4 – присутній. П5 – присутні. П6 – для науковців, які знаходяться у пошуках необхідної інформації та для користувачів мережі Інтернет. П7 – тестування нових алгоритмів для тегування, продовження подальшого розвитку SentiWordNet за межами вже реалізованої "Версії 1.0". П8 – автоматизований.

“SentiWords” (тональний словник англійської мови) [17; 4]. П1 – оскільки SentiWords походить від SentiWordNet, то оцінка п’ятьма експертами зберігається. П2 – оцінка в діапазоні від -1 до +1. П3 – 155 тисяч слів. П4 – присутній. П5 – присутні з використанням алгоритмів. П6 – використання словника без необхідності спочатку розбирати вхідний текст, як це здійснюється у SentiWordNet. Це у свою чергу означає, що оцінки призначаються безпосередньо словам (полярність слів, незалежних від контексту). П7 – така інформація не подається. П8 – автоматизований.

“VADER” (тональний словник англійської мови) [4]. П1 – десятеро людей-анатотарів. П2 – оцінки варіюються від -4 до +4, замість звичайного діапазону від -1 до +1. П3 – трохи більше 7 000 слів. П4 – відсутній. П5 – відсутні. П6 – лексикон створений та спеціально налаштований для соціальних медіа, а також охоплює смайли та абревіатури. П7 – така інформація не подається. П8 – ручний.

“Тональний словник російської мови” (датасет РуСентіЛекс-2017) [28]. П1 – експертами (експертна розмітка), проте кількість не вказана. П2 – оцінка в діапазоні від -1 (максимально можлива негативна оцінка) до +1 (максимально можлива позитивна оцінка). П3 – 28 197 слів. П4 – відсутній. П5 – відсутні. П6 – потреба класифікувати списки слів у відповідності з їхньою полярністю, у свою чергу списки слів будуть отримувати автоматичним алгоритмом. П7 – зрозуміти причину тональності тих чи інших слів; розмежувати слова, які пов’язані з почуттями, емоціями тощо через те, що такі типи слів з більшою ймовірністю схильні до культурного та соціального впливу. П8 – ручний.

Щодо другого параметру можна зробити такі висновки: тональність слова можна подати двома способами: як категорію або в межах якого завгодно обсягу шкали (тут в роботі було помічено такі діапазони шкал: від -2 до +2; від -1 до +1; від -4 до +4, коли 0 включають або ні).

Останній, восьмий, критерій огляду тональних словників (до якого різновиду належить словник (ручний або автоматизований)) спрямований на те, щоб підсумувати всі попередні критерії й, спираючись на вище подані два різновиди тональних словників, охарактеризувати належність того чи іншого проаналізованого словника до певного різновиду.

1.4. Використання тонального словника: сентимент-аналіз

1.4.1. Поняття сентимент-аналізу

Сентимент-аналіз, або тональний аналіз, або аналіз настроїв – це новий тип аналізу тексту із залученням лінгвістичних знань, який широко використовується для досягнення різного роду цілей. Існує декілька трактувань цього поняття. За одним із них, сентимент-аналіз – це процес визначення, чи є написаний текст позитивним, негативним або нейтральним [15]. В іншому джерелі знаходимо, що сентимент-аналіз – це автоматизований процес визначення, чи текст виражає позитивну, негативну або нейтральну думку про продукт / продукцію або тему [12]. При зверненні до третього джерела отримуємо таке розуміння сентимент-аналізу: форма аналітики тексту, яка оцінює відношення / ставлення покупця / клієнта щодо таких аспектів, як сервіс / обслуговування чи продукт / продукція, які покупець / клієнт описує у тексті. Далі йде пояснення того, що текстом може бути що завгодно: речення, коментар чи увесь документ [18].

При зіставленні двох перших визначень, можна зробити висновок, що вони є майже синонімічними. В обох визначеннях фігурує тональність, яка може бути позитивною, негативною чи нейтральною. У другому визначенні здійснена деталізація того, що може містити у собі текст і хто є автором такого тексту.

Якщо ж порівняти два перших визначення із третім, то можна сказати, що вони відрізняються між собою лише словесно, проте зберігають основну ідею трактування терміна. На основі поданих даних можна зробити висновок, що **сентимент-аналіз – це такий процес аналізу текстових даних, який дає змогу**

визначити ставлення людини-клієнта до продукції, наданих послуг чи якості обслуговування цього клієнта, до того ж ставлення може бути позитивним, нейтральним чи негативним.

1.4.2. Потреба сентимент-аналізу

Сентимент-аналіз у період XXI століття є дуже важливим та необхідним. Це зумовлено тим, що нині існує потреба швидко та ефективно обробити величезні масиви інформації, які є навколо нас [12]. Вручну такі обсяги проаналізувати нереально, а от із застосуванням комп'ютерних технологій цілком можливо. Обробка даних комп'ютером має окрім швидкості ще одну вагому перевагу – об'єктивність [Там само].

Можна дійти висновку, що **сентимент-аналіз** має ряд **переваг** [12]:

- масштабованість (тобто, здатність швидко обробляти великі масиви даних);
- аналіз у реальному часі (сентимент-аналіз здатен помітити кризові ситуації, які можуть трапитися, та вказати на них);
- послідовність критеріїв оцінювання (автоматизований сентимент-аналіз постійно буде класифікувати тексти однаково впродовж тривалого часу, адже на нього не впливають такі людські фактори, як вірування, настрій, втома тощо).

1.4.3. Короткий огляд систем сентимент-аналізу

Існує базовий алгоритм здійснення сентимент-аналізу [15], який є прямолінійним за своїм характером та складається із трьох обов'язкових етапів:

1) розбити текстовий документ на складові частини;
2) ідентифікувати / визначити кожний компонент та частини, що містять сентимент (англ. sentiment-bearing phrase);
3) надати сентимент-оцінку для кожного компонента за визначеною шкалою.

Проте може бути ще й четвертий етап, який є факультативним: поєднати оцінки заради здійснення багаторівневого / багатомовного сентимент-аналізу.

Незважаючи на базовий алгоритм, шляхи реалізації сентимент-аналізу можуть бути різними.

Стандартно виокремлюють такі три системи сентимент-аналізу [12]:

1) **проста система сентимент-аналізу, яка базується на правилах** – це система, яка використовує набір правил, які створені людиною вручну, а також працює на основі тонального словника, який був попередньо створений та розмічений. Такі дані допомагають комп'ютеру здійснити сентимент-оцінку конкретного тексту [15; 12]. Набір правил залучає техніки опрацювання природної мови, такі як токенізація, стемінг та парсинг [12];

2) **автоматична система, або система машинного навчання** – система, яка працює на основі методів машинного навчання. Прикладом техніки машинного навчання може бути класифікація тексту, яка широко застосовується для сентимент-аналізу [Там само]. Хоча початково системи на основі машинного навчання використовувалися, щоб покрити недоліки системи сентимент-аналізу, яка базується на правилах [15]. Впровадження **системи машинного навчання** здійснюється в два етапи: тренування та передбачення [12]. Розрізняють автоматичні системи, які навчалися за участю експерта (supervised) [3] та без його участі [15];

3) **гібридна система** – це така система, яка поєднує в собі машинне навчання та традиційні правила (систему, яка базується на правилах [12]) [15]. Така система має кращі результати за рахунок того, що кожен із двох підходів компенсує недоліки один одного [15; 12].

Окрім трьох вищезазначених систем сентимент-аналізу, існує ще одна система, яка відтворює найбільш прогресивний підхід до обробки даних природної мови. Йдеться про **системи, які працюють на основі нейронних мереж та системи глибокого навчання** (англ. Deep Learning techniques, Artificial Neural Networks відповідно) [18]. Саме таким системам в останні роки віддають

найбільшу перевагу при обробці природних мов. Для сентимент-аналізу найчастіше використовуються такі нейромережеві архітектури як LSTM (Long Short-Term Memory, або довга короткотривала пам’ять) [18] та Transformer [20].

1.4.4. Застосування сентимент-аналізу

Сфери застосування сентимент-аналізу щороку збільшуються, що говорить про потребу аналізу такого роду. Нині сентимент-аналіз використовується для аналізу ситуації у політиці, аналізу обслуговування клієнтів, аналізу відгуків як клієнтів, так і співробітників компаній [15; 12], а також у соціології, маркетингу, медицині, психології тощо [32, с. 63]. Дослідження тональності тексту потрібне для того, щоб зрозуміти, наскільки емоційно забарвленим є той чи інший текст. У найпростішому випадку – це класифікація тексту як позитивного, негативного чи нейтрального. У складнішому випадку – визначення конкретних емоцій або обчислення почуття щодо конкретної сутності [4]. Коли буде створено тональний словник для конкретної мови, який буде містити всі загальновживані та найбільш частотні лексеми, то стане можливим дослідження емоційного забарвлення відгуків про готелі, банки, ресторани, коментарів про переглянуті фільми чи серіали, повідомлень у блогах та соціальних мережах про конкретні політичні події та ін. Оскільки сентимент-аналіз широко використовується у багатьох сферах людської діяльності, виникає велика кількість програмного забезпечення, яке дає змогу здійснити сентимент-аналіз. Станом на початок 2022 року було визначено 13 програм-переможців у цій сфері [19]: Google Cloud Natural Language API [5], Lexalytics Salience [7], MeaningCloud [8], VisualText [22], Microsoft Azure Cognitive Service Text Analytics AP [9], IBM Watson Natural Language Understanding [6], Twinword [21], Rosette Text Analytics [13], NetOwl [10], Angoos KnowledgeREADER [1], Averbis [2], PrediCX [11], SAS Sentiment Analysis [14]. Інформація щодо них була систематизована та подана у таблиці 1.4.4.1., яка подана в додатках до роботи (див. Додаток Ї).

На початку 2023 було представлено топ 18 програм щодо здійснення сентимент-аналізу (які зібрано на основі двох джерел [55, 56]): BytesView [57], MonkeyLearn [58], Google Cloud Natural Language API [5], Microsoft Azure [9], Qualaroo [59], NICE Interaction Analytics [60], Rosette Text Analytics [13], Aylien [61]. Зокрема створено ряд програм, що працюють на основі штучного інтелекту [56]: OpenText [62], Brand24 [63], ParallelDots [64], Lexalytics Salience [7], Hi-Tech BPO [65], Social Mention [66], Social Searcher [67], Sentiment Analyzer [68], MeaningCloud [8] та Tweet Sentiment Visualization [69]. Цікаво, що серед найкращих програм протягом двох років лишаються Google Cloud Natural Language API [5], Microsoft Azure [9], Rosette Text Analytics [13], Lexalytics Salience [7] та MeaningCloud [8].

Висновки до першого розділу

У цьому розділі було розглянуто ті теоретичні засади, що стосуються понять “тональний словник” та “сентимент-аналіз”. На основі аналізу досліджених наукових джерел було сформовано власні трактування таких понять як “тональність” та “тональний словник”. Також було розглянуто, які є різновиди тональних словників, спираючись на власне зіставлення та порівняння тих тональних словників, які є у вільному доступі. Було проаналізовано три визначення “сентимент-аналізу”, на основі яких було сформульовано найбільш всеохопне трактування цього терміна. Було розглянуто основі етапи та типи сентимент-аналізу, які існують на сьогодні. Зокрема було окреслено потреби сентимент-аналізу та його практичне використання. Подані теоретичні матеріали допомагають яскраво показати необхідність створення тонального словника саме для української мови. Практична реалізація словника аналізу настроїв для української мови представлена в наступному розділі цієї роботи.

РОЗДІЛ 2. АВТОМАТИЧНЕ ОПРАЦЮВАННЯ ТЕКСТІВ

У цій роботі неодноразово буде згадано про автоматичну обробку природної мови. Адже ця галузь посідає важливе місце в лінгвістичних дослідженнях. Одним із аспектів автоматичної обробки природної мови є здійснення автоматичного опрацювання текстів (далі – АОТ).

Нині існує безліч способів та інструментів, які дозволяють здійснити АОТ швидко, легко та отримати бажану інформацію.

Цей розділ присвячений теоретичному викладу основної інформації щодо тонкощів вебскрапінгу (вебскрейпінгу) та створення автоматичних частотних словників.

У цій роботі частотні словники будуть необхідні при зіставленні лексичних одиниць у тональному та частотному словнику задля розуміння достатності обсягів тонального словника.

2.1. Вебскрапінг (вебскрейпінг) як один із сучасних способів роботи з текстами

Ми живемо в епоху діджиталізації, а це означає, що все наше життя зазнає істотних змін. Адже всі види інформації інтенсивно трансформуються у цифрову із залученням цифрових технологій.

АОТ не стало винятком. У сфері прикладної лінгвістики вебскрапінг (вебскрейпінг) став одним із найбільш розповсюджених методів АОТ. Що ж означають ці два терміни?

Якщо говорити зрозумілими словами, то **вебскрапінг** (вебскрейпінг; англ. web scraping) – це витягування / видобування необхідних даних з вебсайту з метою подальшого використання отриманої інформації у зручному для дослідника чи користувача форматі [42].

Є чимало теоретичної інформації та прикладних реалізацій з вебскрапінгу (вебскрейпінгу). Спробуємо зосередитися на основній характеристиці представленого процесу.

Вебскрапінг (вебскрейпінг) можна здійснити двома шляхами [Там само]:

- вручну;
- автоматично.

Цілком очевидно, що видобувати дані вручну є досить довгим процесом. До того ж такий підхід навряд чи буде високоефективним, коли необхідно опрацювати великі обсяги даних.

Тому зараз найбільш розповсюдженим способом є залучення автоматичних засобів задля досягнення бажаної мети щодо отримання бажаної інформації.

Однією із проблем при здійсненні вебскрапінгу (вебскрейпінгу) є те, що переважна більшість вебсторінок є вкрай перенасиченими та містять величезні кількості непотрібних даних, які треба вилучати [Там само].

Зокрема, якщо говорити про вебсайти новин, то можна зіштовхнутися з рядом проблем, які часом буває не так легко й вирішити.

Отже, **вебскрапінг (вебскрейпінг) – це один із найефективніших та найбільш корисних способів здобуття даних з вебсайту**, особливо якщо здійснювати цей процес шляхом залучення комп'ютерних технологій.

2.2. Автоматичне укладання частотного словника як одне із завдань прикладної лінгвістики

2.2.1. Поняття частотного словника

Який розділ мовознавства займається укладанням частотного словника (далі – ЧС)? Звісно, укладанням будь-якого типу словників займається лексикографія.

Лексикографія – це "розділ мовознавства, який займається теорією і практикою укладання словників" [43, с. 249]. Загалом "укладання словників вимагає великих теоретичних знань і "доброго чуття мови", тобто розуміння

відтінків значення слова, особливостей його вживання, сполучуваності з іншими словами тощо” [Там само].

З часом укладанням ЧС почала займатися окрема лінгвістична дисципліна – статистична лексикографія [50, с. 88].

Цікаво, що статистична лексикографія все більше тяжіє до створення словників комплексних типів. Мова йде про такі словники, які б в описі тих чи інших мовних одиниць унаочнювали різні ознаки їхніх форми, змісту або використання в тексті. Прикладом частотного словника комплексного типу може бути будь-який частотний словник, який нині автоматично укладається на мовному порталі mova.info [71] для окремого тексту, вибраного користувачем, або для цілого стилю.

В другій половині XX та на початку XXI ст. істотно “зріс інтерес мовознавців до створення ЧС” [45, с. 217].

Що ж таке частотний словник? Що вирізняє ЧС серед інших відомих нам словників? Для чого укладають ЧС?

Насправді можна натрапити на кілька трактувань поняття ЧС, проте всі вони містять у собі спільне ядро.

На думку лінгвіста П.М. Алексєєва, ЧС – це така лексикографічна праця, в якій представлено сукупність слів або інших лінгвістичних одиниць (словоформи, словосполучення), які зареєстровані укладачем цієї праці в досліджуваному тексті (або текстах) з вказівкою на частоту їх вживання в цьому тексті (текстах). Саме чітка вказівка щодо частотності використання слова в тексті (текстах) відрізняє ЧС від інших словників. В ЧС можна побачити тільки ту лексику, яку укладач виявив в конкретному тексті (групі текстів чи навіть корпусі текстів) [44, с.53].

ЧС – це такий вид словника, що складається зі списку одиниць підрахунку (здебільшого слів), які розташовані за спадом їх частоти вживання в сукупності текстів. Саме такий ранговий список дає можливість визначити “найчастотніші

слова та словоформи, обчислити показники покриття тексту, індекси винятковості, концентрації тощо” [45, с. 217].

ЧС – це словник, в якому для кожного слова подано кількість його вживань у тексті, або частоту його появи в такому тексті [49].

Таким чином, після здійснення аналізу поданих визначень, можна сформулювати кінцеве визначення поняттю “частотний словник”.

Частотний словник – це такий вид словника, який укладається на основі досліджуваного тексту чи масиву текстів, де подано частоту появи мовних одиниць у досліджуваному матеріалі.

“ЧС як модель організації лексики допомагає не тільки вирішити проблему невизначеності стосовно словникового запасу слів та його реалізації в мові, а також дозволяє встановити ієрархію, в основі якої є частота використання слова” [44, с. 53].

Довгий час лексичний склад досліджуваного тексту (текстів) аналізувався лише з літературного боку. Проте згодом було помічено, що частотність лексичної одиниці – це насправді “важлива характеристика слова, оскільки вона свідчить про активність його функціонування в тексті, про його вагу у статистичній структурі тексту тощо” [70, с.89].

Стало зрозуміло, що лише шляхом залучення статистичних методів можна розкрити всі можливі закономірності функціонування лексики (чи інших одиниць мови) [51, с. 7].

Укладання ЧС є досить-таки умовно новим напрямом лінгвістичних досліджень. Спершу необхідно було розробити теоретичну базу, яка б дозволяла здійснювати створення ЧС за чітко визначеним алгоритмом.

Проблему статистичної структури тексту, укладання ЧС досліджували зарубіжні й вітчизняні мовознавці: Ґ. Альтманн, І. Попеску, Р. Кьолер, Ґ. Віммер,

Ю. Тулдава, В. Перебийніс, П. Алексєєв, А. Шайкевич, А. Павловський, М. Рушковський та ін. [45, с. 217-218].

Спочатку ЧС словники укладалися вручну, що займало надзвичайно багато часу та необхідно було залучити велику кількість людських ресурсів задля створення такого словника. Було помічено, що такий підхід до укладання, попри певні домовленості між експертами-укладачами, не є універсальним, адже існує великий ризик помилок чи неточностей, що зумовлені людським фактором.

Перспективним є підхід автоматичного укладання ЧС. Безумовно, комп'ютер значно полегшує та пришвидшує опрацювання текстової інформації [Там само, с. 219], дозволяє уніфіковано та неупереджено представити отримані результати.

Якщо говорити про те, до якого типу словників належать ЧС, то не можна не погодитися з ідеєю В. І. Перебийніс. На її думку, ЧС словник належить до текстоорієнтованих словників, тобто є таким, який укладено виключно на матеріалі тексту або корпусу текстів [47, с. 52].

Отже, **частотний словник – це такий вид текстоорієнтованого словника, який укладається на основі досліджуваного тексту чи масиву текстів, де подано частоту появи мовних одиниць у досліджуваному матеріалі.** Укладанням частотних словників займається особливий розділ лексокографії чи лінгвостатистики – статистична лексикографія. Частотні словники можна укладати двома шляхами: вручну або із залученням можливостей комп'ютерних технологій.

2.2.2. Короткий огляд щодо створення частотних словників в українському мовознавстві

В українському мовознавстві питаннями виявлення найчастотніших слів різних стилів займалися, зокрема, Н. П. Дарчук [48], Т. О. Грязнухіна [Там само], С. Бук, М. Муравицька та інші.

Одним із найвідоміших ЧС в українському мовознавстві є **2-томний “Частотний словник сучасної української художньої прози”** (1981) за ред. В. С. Перебийніс [51].

Цей словник був укладений вручну колективом співробітників відділу структурно-математичної лінгвістики Інституту мовознавства ім. О. О. Потебні НАН України. Об’єктом опису в ньому були окремі слова, причому як повноправні реєстрові одиниці до словника були введені також і дієприслівники. Загальна вибірка текстів для укладання цього словника становила 500 тис. слововживань у текстах творів 25 українських письменників, які були опубліковані в період між 1945–1970 рр. [50, с. 90].

А вже у 1998 р. на базі 2-томного “Частотного словника сучасної української художньої прози” з прямим абеткуванням одиниць, або абеткуванням за початком слів було укладено з допомогою комп’ютера й видано у паперовому вигляді однотомний **“Обернений частотний словник сучасної української художньої прози”** [52], який становить один з можливих граматичних словників сучасної української мови з розгорнутою інформацією про особливості вживання тих чи інших словоформ у прямій та авторській мові художніх прозових творів [50, с. 92].

Методика статистичного опрацювання текстів та вироблений формат статті “Частотного словника сучасної української художньої прози” були використані в двох комп’ютерних частотних словниках, укладених співробітниками відділу структурно-математичної лінгвістики Інституту мовознавства ім. О. О. Потебні НАН України разом з працівниками лабораторії комп’ютерної лінгвістики при кафедрі української мови та прикладної лінгвістики Київського національного університету ім. Тараса Шевченка, – **частотних словниках сучасної української публіцистики** та **сучасного українського наукового стилю**. За обсягом опрацьованих текстових вибірок (300 тис. слововживань кожна) це невеликі словники. Їхньою помітною особливістю є “застосування автоматичних процедур

лематизації, або виведення канонічних форм змінюваних слів на основі результатів автоматичного морфологічного аналізу текстів. Автоматично здійснювалося й саме укладання корпусу словників: формування реєстру одиниць та статистичні підрахунки" [Там само].

Оскільки в пріоритеті автоматичне створення ЧС, останні роки співробітники лабораторії комп'ютерної лінгвістики при кафедрі української мови та прикладної лінгвістики Київського національного університету ім. Тараса Шевченка [Там само, с.93] мають змогу надати користувачеві мовного порталу mova.info [71] інтегровані динамічні частотні словники, які автоматично формуються для будь-якого обраного на цьому порталі тексту чи корпусу текстів. Користувач може отримати ЧС лексем та ЧС словоформ, відсортувати їх за спадом частоти чи алфавітом, а для ЧС лексем – ще й за семами, частинами мови та ін.

Отже, в українському мовознавстві активно займаються укладанням частотних словників. Нині особливу увагу для та створення та подальшого вивчення привертають ті частотні словники, що укладаються автоматично.

2.2.3. Класифікації частотних словників

Хоча укладанням ЧС почали займатися лише в середині XX столітті, нині існує розгалужена класифікація ЧС. Давайте спробуємо коротко представити типи ЧС.

Залежно від типу лексичних одиниць ЧС поділяються на [70, с.90]:

- ЧС словоформ, слів (лексем);
- ЧС основ слів (використовуються в інформатиці);
- ЧС слів у певних значеннях (семантичний частотний словник);
- ЧС словосполучень.

За характером вибірки ЧС поділяють на:

- ЧС усієї мови;
- ЧС певного функціонального стилю;

- ЧС письменника;
- ЧС конкретного твору.

В. С. Перебийніс запропонувала свою багатовекторну класифікацію, яка довзоляє поділяти частотні словники за такими критеріями:

1) одиниця опису;

2) обсяг аналізованих текстів;

3) тип текстів;

4) повнота представлення лексики певного корпусу опрацьованих текстів;

5) спосіб упорядкування одиниць опису;

6) типи статистичних (частотних) характеристик і повнота їхнього представлення у словнику [50, с. 88].

Наприклад, за одиницями опису розрізняють:

- ЧС слів;
- ЧС словоформ;
- ЧС словосполучень;
- ЧС морфем;
- ЧС буквосполук [Там само].

За обсягом вибірки ЧС поділяють на:

- великі (з вибіркою в 1 млн. і більше слововживань);
- середні (обсяг вибірки від 400 тис. до 999 тис.);
- невеликі (з вибіркою від 100 тис. до 399 тис.);
- мікрословники з вибіркою, меншою за 100 тис. слововживань [Там само].

Звісно, класифікація ЧС за кількісною характеристикою може бути різною у наукових джерелах.

За способом подання мовних одиниць, для яких встановлено частотні характеристики, виділяють:

- повні словники, що містять всі одиниці, вжиті в аналізованих текстах;

- неповні, що містять лише одиниці з частотою, яка дорівнює або перевищує певний заданий поріг – граничний показник частоти [Там само, с. 89].

Отже, станом на сьогодні існує розгалужена система щодо класифікації частотних словників за різними параметрами.

2.2.4. Використання частотних словників

Чому виникла потреба у створенні ЧС? Для чого їх використовують?

Оскільки ЧС у порівнянні з іншими видами словників є відносно новим типом словника, то використання ним часто ускладнюється й обмежується. На це є кілька причин: з одного боку, це пов’язано з недостатнім знайомством мовознавців зі статистикою, а з іншого – з тим, що мовознавці не бачать можливих сфер його використання і методів роботи з ним [51, с. 18-19].

Насправді нині практичне застосування ЧС є досить широким: “відбір лексичного мінімуму при вивченні іноземних мов, створення ефективних систем стенографії, атрибуція непідписаних рукописів, створення економних алгоритмів кодування текстів для ЕОМ” (комп’ютерних систем), а також для “систем машинного опрацювання текстів як наприклад, машинний переклад, інформаційний пошук, автоматичне реферування й анотування літератури” [Там само, с. 7].

Практичні потреби статистичного обстеження текстів для опрацювання їх на ЕОМ (комп’ютерах) “набувають особливо великої ваги в період науково-технічної революції. Тому якщо перші частотні словники створювалися з метою вдосконалення систем стенографії та для відбору лексичного мінімуму, то в наші дні частотні словники створюються в основному для потреб інформаційного пошуку та машинного перекладу” [Там само].

ЧС насамперед використовують для створення ефективних методик викладання мови як іноземної, для виділення ключових слів (так наприклад, цей напрям є дуже поширеним в інформатиці), для створення раціональних кодів (у

теорії зв’язку) [46, с. 93]. Адже якщо говорити про створення посібників з іноземної мови, то досить-таки часто виникає проблема відбору слів для діалогів, для текстів на побутові теми. А ЧС дає можливість вирішити це питання [51, с. 19].

Цікаво, що ЧС можна використати і для створення різного роду посібників з української мови для шкіл [Там само].

Також ЧС дозволяє здійснювати теоретичні та практичні дослідження у межах обраної мови. Тому що ЧС може надати дослідникові багатий матеріал щодо функціонування лексичної і лексико-граматичної системи мови у досліджуваному часовому зрізі [Там само].

Використання ЧС дозволить дослідникові оперувати не такими приблизними оцінками, як “частовживане”, “малопоширене” і подібні, а точними статистичними оцінками не лише щодо частоти, а й також щодо ступеня поширеності і рівномірності вживання слова чи групи слів [Там само].

Будь-який ЧС дає змогу простежити, як з роками відбувається еволюція мови та сказати, коли з’являються нові слова та вирази і коли зникають старі [45, с. 219].

Оскільки інтенсивно створюються ЧС, то це надає нам можливість здійснювати зіставно-типологічні дослідження. Насамперед мова йде про використання ЧС з метою їх зіставлення задля того, щоб простежити закономірності функціонування лексики досліджуваних мов. Можна зіставляти ЧС різних мов з погляду розподілу в них слів за частотою, співвідношення частин мови, лексико-семантичних груп тощо [51, с. 20].

Отже, у період технологічного розвитку частотні словники є одним із найголовніших інструментаріїв задля проведення різного роду лінгвістичних та статистичних досліджень. Зокрема, якщо говорити про лінгвістичні дослідження,

то ЧС може бути основою для створення таких розділів граматики, як словозмінна, словотвірна та лексична парадигматики.

Висновки до другого розділу

У цьому розділі коротко окреслено теоретичні аспекти щодо вебскрапінгу (вебскрейпінгу) та особливостей створення та укладання частотних словників.

Вебскрапінг (вебскрейпінг; web scraping) – це один із можливих засобів видобування будь-яких даних з вебсайту, щоб у подальшому отриману інформацію можна було використати у зручному для дослідника чи користувача форматі.

Частотний словник – це такий словник, який репрезентує сукупність слів або інших лінгвістичних одиниць, що були присутні у межах досліджуваного тексту чи корпусів текстів. Саме частотні словники "дають кількісну характеристику тексту, допомагають встановити особливості функціонування в ньому лексичних одиниць, а також статистичні його закони" [45, с.217]. Виходить, що текст – єдине джерело для укладання ЧС. Загальними тенденціями щодо розвитку створення та укладання ЧС у сучасній лексикографії є те, що все частіше відбувається застосування комп'ютерних програм опрацювання тексту з метою укладання ЧС.

РОЗДІЛ 3. РОЗРОБКА ТОНАЛЬНОГО СЛОВНИКА ДЛЯ УКРАЇНСЬКОЇ МОВИ

Після аналізу наукових публікацій розпочалася робота над створенням тонального словника для української мови. Варто зауважити, що тональний словник створюється з метою подальшої роботи з текстами зі сфери новин. Тому підбір слів для словника відбувається на основі новинних текстів. Робота зі словниками та списками слів проводилася спершу на основі таблиць Microsoft Excel 2016 [72], а потім – з використанням Google таблиць [73]. Згодом всі дані було перенесено до папки на Google Диск (посилання на папку подано в розділі "Додатки", ознайомитися з готовим тональним словником можна в Додатку Й).

Для роботи було надано:

1) попередньо частково розмічений список слів обсягом 40 567 слововживань (далі для зручності називатимемо його ТС №1) у первісному, неопрацьованому вигляді подано у Додатку А;

2) тональний словник, який було сформовано вже нами на основі 14 окремих, також попередньо розмічених, списків слів. Кожен такий список містив близько 500 слів. У результаті зведення вийшов другий тональний словник обсягом 6966 слів та словосполучень (далі для зручності називатимемо його ТС №2), який подано в Додатку В.

Попередня розмітка (анотування) складалася з оцінок тональності, яка була здійснена студентами-експертами, за що хочеться їм подякувати..

Аналіз обох словників із метою з'ясування значення лексем здійснювався на основі використання 11-томного академічного тлумачного "Словника української мови" (далі СУМ) [29], ресурсу СЛОВНИК.ua [35] та Вікіпедії [24].

Оскільки нині існує безліч способів щодо того, за якою шкалою тональності оцінювати слова та словосполучення [28; 4; 40; 17], було вирішено взяти найбільш оптимальний та зручний вигляд шкали, де є можливими лише наступні оцінки:

Таблиця 3.1. Шкала тональності

Оцінка	Значення оцінки
-2	дуже негативно
-1	негативно
0	нейтрально
1	позитивно
2	дуже позитивно

Власне саме такі оцінки й містили попередньо розмічені ТС №1 та №2, що в свою чергу спрощує нашу роботу, адже не потрібно створювати ніяких додаткових функцій щодо переведення оцінок із однієї шкали в іншу.

У першому списку слова, які повторювалися, відповідно містили кілька оцінок. А от ТС №2 відразу містив три або п’ять експертних оцінок. Вважаємо такий підхід до створення тонального словника одним із найкращих, адже він забезпечує переведення суб’єктивної оцінки в розряд об’єктивної.

3.1. Етапи роботи створення тонального словника української мови

Робота над створенням кінцевого варіанту тонального словника здійснювалася в кілька етапів:

Проміжний етап – формування власних списків слів, які хотілося б включити до кінцевого варіанту тонального словника (далі ТС);

І етап – робота із ТС №1;

ІІ етап – робота із ТС №2;

ІІІ етап – створення кінцевого варіанту тонального словника шляхом зведення опрацьованих ТС №1 та ТС №2 з додаванням власного списку слів та словосполучень.

3.1.1. Проміжний етап

Проміжний етап здійснювався між І та ІІ етапами, коли було вже проаналізовано ТС №1. Метою цього етапу є створення окремих списків слів, які є досить частотними за своїм вживанням у новинах та які було поділено за певними тематичними спрямуваннями.

Ми вважаємо за необхідне включити до складу ТС назви обласних центрів України [27], населених пунктів [30], списки днів тижня, місяців та пір року, список найуживаніших абревіатур тощо. З тематичними списками слів можна ознайомитися у файлі “README. Додаток Г”, який додано до Додатку Г. Також було сформовано список слів (див. Додаток Д), що стосуються країн світу, їх столиць [37; 36; 39; 38], утворені від них прикметники та назви мешканців (“README. Додаток Д”). Потім всі п’ять стовпців, які були в Додатку Д, було поєднано в один, відсортовано в алфавітному порядку й подано в Додатку Е.

У результаті на проміжному етапі роботи було сформовано дев’ять списків слів, які будуть додані до кінцевого варіанту ТС й для яких буде виставлено оцінку їхньої тональності.

Також ще одним завданням на цьому етапі було з’ясування того, чи потребують дієслова додаткового введення в словник своїх пар за таким граматичним значенням як доконаний та недоконаний вид. Щоб це з’ясувати, було створено Google форму (яку подано в Додатку Ж з аналізом отриманих даних), де респондентам пропонувалося надати оцінку тональності для 22 дієслів, що утворюють 11 дієслівних пар.

Форму було заповнено 25 людьми й аналіз отриманих даних показав, що дев’ять пар із 11 мають майже однакову оцінку тональності дієслівної пари: або є відсоток однієї оцінки, яка суттєво переважає над іншими, або дві оцінки мають повний збіг у відсотковому співвідношенні. Незважаючи на отримані результати,

діаграми відсоткового співвідношення оцінок до дієслівної пари відрізнялися, що говорить про те, що все-таки є потреба включати дієслівну пару за її відсутності.

3.1.2. І етап створення словника

На першому етапі проводилася робота зі ТС №1, який початково містив 40 567 слововживань та був поданий у вигляді таблиці, яка була оформлена за допомогою Microsoft Excel [72]: у першому стовпці подано слово, абревіатуру, словосполучення тощо, тобто реєстрову одиницю словника, а в другій – лише одна оцінка.

Робота з даними ТС №1 здійснювалася у кілька кроків: структурування, щоб словник виглядав як єдине ціле; виключення дублів; вилучення окремих одиниць зі словника.

Варто сказати, що вилучалися ті одиниці, які, на нашу думку, не є властивими для медійного стилю [25]. У першу чергу йдеться про вилучення слів ненормативної лексики. Також нами було прийняте рішення вилучати власні імена та прізвища людей, які є маловідомими, а отже, маловживаними й тому нейтральними за тональністю.

Детальніше ознайомитися з тим, як відбувалися ці кроки можна у файлі “README. Додаток А”.

Наступна фаза роботи з ТС №1 – “Формування І частини тонального словника”. На цьому щаблі проводилася робота із ТС №1, що зумовлювала попередній вигляд майбутнього словника. Зараз словник містить 11 стовпців (реєстрові одиниці, примітки, оцінки експертів та середня оцінка). Детальніше ця інформація описана в файлі “README. Додаток А”.

Особлива увага зверталася на слова іншомовного походження, для яких не існує єдиного можливого їхнього написання в українській мові [26].

Варто також сказати про те, що одне слово було змінено дещо інакше, ніж інші, адже проблема була у правильності написанні цього слова. Йдеться про таке слово як “україномовний”, яке було змінено на “українськомовний” [31].

Отже, на першому етапі проводилася робота із упорядкуванням та систематизацією даних ТС №1, з метою подальшої роботи із цим словником. У результаті проведеної нами роботи обсяг словника зменшився до 12 471 слова / словосполучення, які подано в Додатку Б.

3.1.3. ІІ етап створення словника

На другому етапі здійснювалася робота над ТС №2, який не був поданий у вигляді одного файлу, як це було у випадку із ТС №1, а в вигляді 14 різних файлів. Взагалі два опрацьованих нами словники багато в чому відрізнялися.

ТС №2 був більш розміченим та інакше побудований (якщо порівнювати ТС №2 із ТС №1). Файл містив три або п’ять експертних оцінок, а не одну (як у ТС №1). Тому був стовпчик, де проводився розрахунок середнього арифметичного виставлених оцінок. Також були примітки експерта, який робив попередню розмітку словника.

Перш ніж почати роботу з цими словниками, потрібно було об’єднати всі 14 словників в один, привести отриманий словник до одного вигляду, тобто структурувати його. В результаті було отримано той тональний словник, який описувався вище як ТС №2 й обсяг якого становив 7 004 слів.

Вигляд ТС №2 описано у файлі “README. Додаток В”.

Наступним кроком було впорядкування слів за алфавітом. Під час упорядкування було помічено повтори слів, які довелося видалити, зберігаючи оцінки тональності, якщо вони відрізнялися для цих дублів. У результаті роботи було створено ТС №2 обсягом 6 966 слів.

3.1.4. ІІІ етап створення словника

На цьому етапі необхідно було об’єднати ТС №1, ТС №2 та власний список слів (Додаток Є, який містить у собі поєднання Додатку Г, Додатку Е та перелік власних слів), щоб подати кінцевий варіант тонального словника.

У ході роботи із двома тональними словниками було сформовано міні-словник слів (Додаток Є). Цей словник призначений для подальшого включення у кінцевий варіант словника. Вибір слів базувався на аналізі ТС №1 та перегляді новин самостійно. В результаті обсяг міні-словника становить 1 135 слів, але більшість із них можуть повторюватися з основним масивом слів у зведеному словникові. Це в свою чергу означає те, що знову з’являться дублі, які потрібно буде виключити, зберігаючи оцінки тональності.

Вирішено було розбити роботу на кілька кроків:

КРОК 1. Об’єднати ТС №1 та ТС №2, усунути дублі та за потреби об’єднати оцінки.

КРОК 2. Об’єднати ТС №1 та ТС №2 із Додатком Є.

На першому кроці слова із ТС №1 і ТС №2 знову упорядковувалися за алфавітом. Процес усунення дублів детально описано у файлі “README. Усунення дублів. ТС №1 та ТС №2”.

Але навіть при об’єднанні двох словників залишалися слова, в яких було одна, дві, три, чотири, п’ять або шість оцінок. Тому було вирішено поділити всі слова на дві категорії: повна та неповна, або часткова, оцінка тональності. Повна оцінка тональності означає, що слова матимуть максимальну кількість можливих оцінок у нашому словнику, тобто вісім оцінок. Це в свою чергу означає, що всі комірки, які надані для експертних оцінок, будуть заповненими. Неповна, або часткова, оцінка – слова матимуть лише п’ять оцінок, серед яких п’ятою буде наша оцінка. Якщо ж слово містило рівно п’ять оцінок, де нашою була вже шоста

оцінка, тоді таке слово в кінцевому результаті буде мати повну оцінку, адже будуть залучатися експерти задля представлення повної оцінки.

Окрім оцінки тональності слів, здійснювався аналіз і розв'язання проблем для тих слів, з якими ми зіштовхнулися на попередніх етапах роботи. Сірі та темно-сірі комірки могли або ставати зеленого кольору й отримувати лише дві оцінки (наша та оцінка із ТС №1), або взагалі вилучатися зі списку. Відбір здійснювався за принципом логічних міркувань, наскільки часто буде зустрічатися це слово? Якщо слово маловживане й може зустрітися лише кілька разів та в надто специфічних, особливих ситуаціях, тоді це слово вилучалося. Слово, яке додавалося до складу словника, аналізувалося, за потреби йому приписувалися всі ті характеристики, про які вже було сказано на І етапі роботи (дієслівна пара, частиномовна належність слова чи будь-який інший вид примітки). Хочеться додати, що на цьому кроці було розв'язано питання семантики для слова "купа" та встановлено, за яким саме значенням цьому слову було надано оцінку попередніми експертами. Також, окрім попередніх приміток, додавалася примітка, що стосувалася написання окремих слів за новим українським правописом [41]. Таким чином, окремі слова мають подвійний варіант написання (див. "соціальний", до якого додано примітку "соціяльний"). Особлива увага зверталася на слова із подвійним наголосом, які в подальшому потребують контекстного аналізу. Для слів із подвійним наголосом створювалася відповідна примітка з жовтогарячою заливкою комірки для приміток.

На другому кроці, перш ніж об'єднати ТС №1, ТС №2 із Додатком Є, було вирішено порівняти ці два списки слів (один із яких – це зведений проаналізований варіант тонального словника, що складається з ТС №1 та ТС №2) за допомогою власноруч створеної міні-програми (див. Додаток З), яка написана мовою програмування Python 3 [74] шляхом використання середовища розроблення PyCharmEdu [75]. Ця програма перевіряє два вхідних csv-файли на

збіг слів і в кінцевому результаті виводить список слів, які є в другому списку (Додаток Є), але відсутні в першому (зведений словник), тобто це перелік слів, які варто додати до основного масиву словника. За допомогою цієї програми нам не потрібно буде повторно усувати можливі дублі вручну. Після того як ми порівняємо два файли за допомогою створеної програми, ми просто додаємо список виведених програмою слів до кінцевого варіанту словника та проаналізуємо їх так само, як й інші слова. Цей код не використовувався на попередніх етапах, тому що в тих списках, які ми аналізували раніше, окрім слів, були ще й оцінки. А їх потрібно було врахувати.

У результаті роботи створеного коду було виявлено 613 збігів слів. Але ця цифра є неточною, оскільки є слова з різними варіантами написання апострофів, про що було сказано раніше. Тож однакові слова з різним написанням вимагали вилучення одного із варіантів. Також вилучалися словосполучення, такі як, наприклад, “Сполучені Штати Америки”, оскільки в словнику вже була оцінена відповідна абревіатура (США).

У Додатку І подано тестовий файл, у якому до вже створеного списку слів у результаті об’єднання ТС №1 і ТС №2 було додано два списки слів. Ці списки були отримані в результаті запуску створеного нами коду. Комірки слів із першого списку в Додатку І, що містить список збігів, виділено світло-зеленим кольором, а в примітках додано слово “збіг”, яке теж було зафарбовано в світло-зелений колір. Комірки слів із другого списку виділено пурпуровим кольором, а в примітках додано “додати” й також зафарбовано в пурпуровий колір відповідно. Цей наданий Додаток І доводить те, що створений код працює добре та правильно, як і було заплановано. Щоправда, слова з написанням апострофа було проаналізовано кодом некоректно, оскільки два csv-файли містять два різні варіанти написання апострофа: «'» і «’». Але якщо в ході роботи зі словником буде помічено однакові

слова з різним написанням апострофа, тоді один із варіантів буде вилучено нами вже вручну й відповідно вилучене слово не буде оцінено.

На цьому етапі слова отримали всі свої оцінки, що дало змогу нам вилучити ті слова, результати яких дорівнювали нулю. Це здійснено з тією метою, що при запуску майбутньої програми, яка буде здійснювати аналіз слів у тексті за допомогою створеного нами словника, автоматично буде приписуватися оцінка “нуль” словам, які не було знайдено в словнику.

3.2. Статистичні дані про створений тональний словник

У результаті роботи було створено тональний словник української мови обсягом 14 857 слів. Нижче подано таблиці та діаграми, які наочно відображають статистичні дані. У таблиці 3.2.1. подано інформацію про зменшення обсягів списків слів, над якими здійснювалася робота, та про створений словник. Діаграма 3.2.2. демонструє відсоткове співвідношення між повною та частковою відміткою про тональність. А діаграма 3.2.3. демонструє кількісне співвідношення між частинами мови, які отримали лінгвістичну розмітку.

Таблиця 3.2.1. У скільки разів / на скільки зменшився обсяг словника після роботи з ним

Назва словника	***Початковий обсяг словника***	***Стало***	***У скільки / на скільки разів зменшився обсяг словника***
ТС №1	40 567	12 471	у 3,25 рази
ТС №2	7 004	6 966	на 38 слів
Тональний словник української мови	19 437	14 895	у 1,3 рази

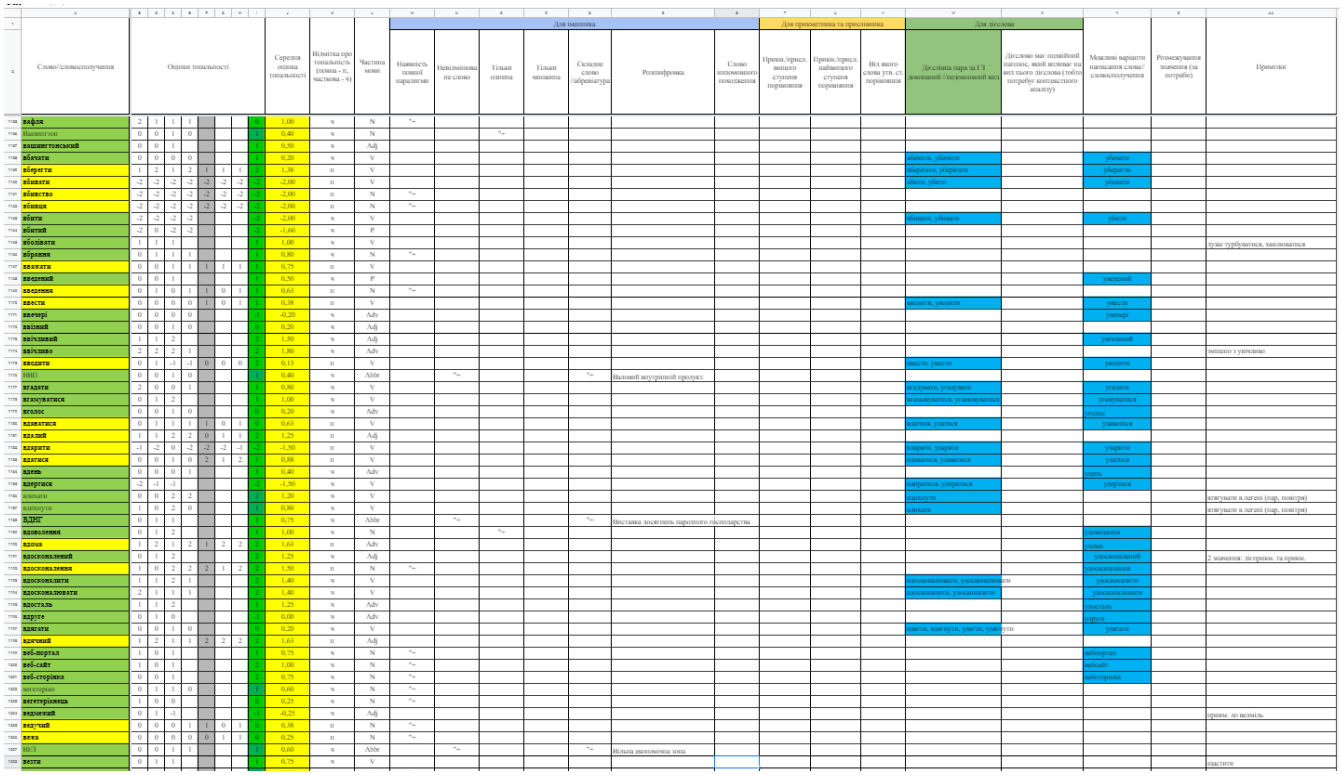
Діаграма 3.2.2. Відсоткове співвідношення слів із повною (п) та частковою (ч) відміткою про тональність (станом на 18.05.23)

Діаграма 3.2.3. Кількісний показник частин мови в тональному словнику (станом на 19.05.23)

3.3. Представлення тонального словника

Коли ТС №1, ТС №2 та Додаток Є було об’єднано, загальний вигляд словника було змінено. Нижче подано вигляд створеного тонального словника (див. Додаток Й).

Малюнок 3.3.1. Фрагмент вигляду тонального словника української мови



Як можна побачити вище, будова словника така: 1-й стовпчик – подано слово або словосполучення, стовпці 2-9 – оцінки тональності, 10-й стовпчик – середня оцінка тональності слова, 11-й стовпчик – відмітка про тональність, яка може бути повною (п) або частковою (ч), 12-й стовпчик – вказівка, якою саме частиною мови є аналізоване слово чи словосполучення. Окрім слів та їхніх оцінок (стовпці 1-10), у словник також додано лінгвістичні характеристики слів та словосполучень (стовпці 13-24), що в подальшому допоможуть покращити ефективність роботи програми, яка буде автоматично “прописувати” оцінку слову чи словосполученню у тексті. 25-й стовпчик містить можливі варіанти написання слова (наприклад, до слова “весь” у цьому стовпці подано такі варіанти як “ввесь,

увесь”). А 26-й стовпчик присвячений розмежуванню значення слів та словосполучень за потреби. Останній, 27-й, стовпчик виокремлений для різного роду приміток.

З інформацією щодо символьних скорочень для частин мови й опису лінгвістичних характеристик для кожної частини мови можна ознайомитися у файлі “README. Вигляд власного тонального словника”, яку подано у Додатку Й.

Таку повну лінгвістичну характеристику станом на 19.05.23 зроблено для всіх слів (14 857). Відбулися зміни щодо зафарбовування комірок. Тепер жовтим позначено ті комірки, які стосуються слів, що мають повну оцінку тональності (тобто всі вісім оцінок). А зеленим – слова з неповною оцінкою (п’ять оцінок відповідно). Слова, які в словнику дублюються, мають жовтогаряче забарвлення, щоб акцентувати на них нашу увагу. Адже в подальшому вони будуть досліджуватися більш детально, щоб з’ясувати, яке з двох чи трьох значень варто врахувати при тональній розмітці цього слова. Слова, які додано в масив словника завдяки роботі коду, виділено пурпуровим кольором.

3.3.1. Дослідження слів, що набули нових значень

У зв’язку з подіями у нашому житті добір тих чи інших лексичних одиниць постійно змінюється. Окрім цього, змінюється чи наповнюється новими смислами і семантичне значення слів.

За останні півтора року у нашому мовленні почали фігурувати слова, які набули нового значення. І це пов’язано з подіями, які розпочалися 24 лютого 2022 року.

Тому нами було вирішено перевірити, чи впливає нове семантичне значення слів на тональну оцінку цих слів. Перевірка була здійснена у вигляді опитування шляхом заповнення респондентами Google форми (яку подано в Додатку Р з аналізом отриманих даних). У цій формі респондентам пропонувалося надати оцінку тональності для 12 слів, що утворюють 6 пар. Кожна з пар – це одне й те

саме слово, яке представлено у двох різних значеннях. Слова було взято з Інтернет-статті [53]. Щоб респондент зміг розпізнати значення слова, їх було подано у контексті (реченнях). Ці речення було сформульовано так, як вони могли б зустрітися у текстах новин.

Форму було заповнено протягом листопада-грудня 2022 року 64 людьми й аналіз отриманих даних показав, що п’ять пар із 6 мають різну оцінку тональності у межах досліджуваної пари. Було помічено такі тенденції:

- слово з “нейтрального” стало або “дуже негативним” в новому значенні (2 пари: йдеться про такі пари слів як “розтяжка” та “приліт”), або “дуже позитивним” (1 пара: “бавовна”);
- слово з “негативного” набуло у другому значенні статусу “дуже негативне” (2 пари: “град” та “тривога”);
- відповідно, одна пара слів не змінила свою оцінку щодо тональності (“дуже негативне”: “смерч”), але відсоткове співвідношення щодо цієї оцінки серед респондентів суттєво відрізняється.

Отже, зважаючи на отримані результати, ми можемо бачити, що діаграми відсоткового співвідношення оцінок у межах пар слів відрізнялися. А це говорить нам про те, що все-таки є потреба включати слово у двох значеннях до складу створюваного тонального словника за відсутності одного або обох досліджених значень.

3.3.2. Автоматичне зіставлення тонального словника зі списком нових слів

Вже було згадано, що в сучасному українському медійному просторі розширюється семантика раніше відомих слів. Також виникає ряд неологізмів та оказіоналізмів, що пов’язано з нашими реаліями сьогодення. Все це впливає на добір лексичного інструментарію задля висвітлення тієї чи іншої новини. Тому

тональний словник потребує постійного доповнення, щоб якомога повніше охопити новинний текст.

У ході роботи це і було зроблено. На основі аналізу Інтернет-джерел [54], які подають перелік слів, що увійшли до нашого постійного вжитку, було створено новий список слів (див. Додаток О). Цей список надає інформацію щодо слів, які можуть зустрітися у текстах новин нині.

Нами було вирішено здійснити автоматичне зіставлення створеного списку слів із тональним словником, щоб з'ясувати, які слова варто додати до тонального словника.

Для цього етапу роботи нами було обрано одну із найбільш використовуваних на сьогодні мов програмування – мову програмування Python версії 3.8 [74]. Розробка програми здійснювалась у середовищі PyCharm Edu 2019.3.1 [75].

Представлення повного коду для цього етапу надано у Додатку П.

У коді програми можна побачити, що:

➔ код поділено на блоки (кожен блок відділено порожнім рядком);

➔ майже кожен рядок коду має пояснення у вигляді закоментованого рядка.

Створення коду, який автоматично порівнює вміст двох файлів у форматі csv, можна представити у вигляді таких кроків:

- імпортування бібліотек, необхідних для подальшої роботи;
- відкриття та прочитання обох файлів;
- представлення обох файлів у форматі списку;
- створення порожнього списку, який буде наповнюватися елементами з потрібного файлу;
- створення циклу, який автоматично прибирає пробіли на початку та наприкінці кожного елемента файлу, а також замінює порожні комірки на дефіси;

- наповнення попередньо створених списків відредагованими даними;
- створення порожніх списків, які у подальшому будуть наповнюватися необхідними даними;
- створення циклу, який наповнює один із порожніх списків (tonsum_1) виключно словами з тонального словника (це здійснено з тією метою, щоб спростити подальші кроки виконання необхідної роботи);
- створення циклу, який порівнює слова в обох списках (які попередньо були надані як два csv-файли) та наповнює один зі списків (difference) словами, які варто додати до тонального словника;
- створення циклу, який порівнює слова в обох списках (які попередньо були надані як два csv-файли) та наповнює один зі списків (coincidence) словами, які присутні в обох списках;
- виведення отриманих результатів щодо збігів та відмінностей у консоль (задля того, щоб переконатися, що все виконано правильно);
- відкриття нового текстового файлу, який буде наповнюватися словами, які варто додати до тонального словника;
- у результаті запуску створеного коду: 1) в консолі ми бачимо списки слів, які збігаються та відрізняються у межах двох вхідних файлів; 2) а також ми отримуємо текстовий файл слів, які варто включити до тонального словника.

Примітки:

- ❖ створений код не враховує можливі варіанти написання слів та словосполучень (що можна зробити в майбутньому);
- ❖ не було створено окремого списку для слів зі списку з нових слів (як це було зроблено для списку слів з тонального словника), оскільки в цьому не було потреби, бо список нових слів містить лише одну колонку – слова та словосполучення;

- універсальність коду полягає в тому, що списком нових слів може бути список, який укладено так само, як і тональний словник, і тоді є можливим внесення відсутніх слів безпосередньо у сам тональний словник (це стане можливим за умови доопрацювання описаного етапу у межах коду);
- недоліком щодо запису відсутніх слів у текстовий файл у режимі “write” (“w”) є те, що при кожному новому запуску цього коду попередні дані будуть стиратися та перезаписуватися новими даними. Хоча в певній мірі цей недолік можна сприймати і за перевагу (у випадку, якщо ми не ставимо перед собою мети накопичувати списки слів, які відсутні у тональному словнику).

3.4. Тональний словник української мови в Інтернет-просторі

3.4.1. Версія 1.0.

У березні 2023 року першу версію створеного тонального словника було опубліковано на одному із найбільших вебсервісів для спільної розробки програмного забезпечення – GitHub. Перша версія словника містить слова та словосполучення з оцінками тональності.

Посилання на створений проект: Oksana504/sentimentdictionary-uk: Тут подано тональний словник української мови. (github.com).

У файлі “README.md” надано всю основну інформацію щодо тонального словника та формати файлів, які подано.

Оскільки тональний словник представлено у форматі csv, який не є дуже зручним для читання користувачами, було вирішено представити тональний словник у вигляді таблиці.

Перетворення формату csv у таблицю здійснювалося автоматично шляхом залучення однієї із найбільш використовуваних на сьогодні мов програмування – мови програмування Python версії 3.8 [74]. Розробка програми здійснювалась у середовищі PyCharm Edu 2019.3.1 [75].

Процес перетворення формату csv у таблицю можна представити у вигляді таких етапів:

- імпортування бібліотек, необхідних для подальшої роботи;
- відкриття тонального словника у форматі csv;
- створення циклу та порожнього списку, який буде наповнюватися інформацією з тонального словника та водночас заміняти порожні комірки на тире;
- створення заголовків для колонок таблиці тонального словника;
- наповнення таблиці даними із тонального словника;
- збереження отриманої таблиці в txt форматі.

Ознайомитися з повним виглядом коду можна у Додатку Н.

Було помічено, що попри те, що тональний словник було взято в кодуванні utf-8, створену таблицю було збережено у кодуванні ANSI. Згодом при додаванні таблиці до проекту на GitHub було помічено, що дані таблиці некоректно відображаються на пристроях користувачів GitHub. Тому створену таблицю на GitHub представлено ще й у кодуванні utf-8.

Нами було вирішено зберегти представлення таблиці у двох кодуваннях, тому що таблиця відображається по-різному на кожному пристрої. Це зумовлено двома факторами: кодуванням локально, на ПК користувача, або кодуванням на самому сервері GitHub.

3.4.2. Версія 2.0.

Навесні 2023 року було завершено роботу з новою (другою) версією тонального словника, яку було доповнено лінгвістичною та екстралінгвістичною інформацією. Кількість даних істотно збільшилася.

Нову версію словника теж було додано у GitHub. Ця версія теж потребувала табличного представлення даних. Тож для неї теж була здійснена автоматична конвертація із csv-формату в таблицю в txt-форматі шляхом залучення Python

версії 3.8 [74]. Розробка програми здійснювалась у середовищі PyCharm Edu 2019.3.1 [75].

Етапи роботи майже повністю збігаються з тими, що було здійснено для першої версії із кількома доповненнями:

- здійснена заміна символу виду “+ на математичний знак плюс (“+”). Така дивна розмітка зумовлена особливостями роботи з Google таблицями, де будь-який математичний знак на початку комірки сприймається за математичний символ;
- кількість назв колонок була доповненою, що зумовлено додаванням нової інформації;
- додано кодування “utf-8” при записі таблиці у текстовий файл. Це було здійснено, щоб уникнути проблем з кодуванням, з якими ми зіштовхнулися при укладанні табличного представлення для першої версії словника.

  Ознайомитися з повним виглядом коду можна у Додатку Н.

Висновки до третього розділу

У цьому розділі подано детальний опис роботи з попередньо наданими даними та описано етапи створення тонального словника української мови. У кінцевому результаті представлено створений тональний словник для української мови як очікуваний результат роботи. Також наведено окремі статистичні показники щодо всіх трьох словників, над якими проводилася робота.

Окрім цього, представлено результати двох опитувань, які довели доцільність внесення дієслівних пар за граматичним значенням “доконаний / недконаний вид” та слів, які за останні півтора року набули нового значення. Зокрема, було здійснено автоматичне зіставлення двох файлів у форматі csv з метою пошуку слів, які вже присутні у кінцевому варіанті тональному словнику, та тих слів, які варто додати до тонального словника.

Одним із здобутків є оприлюднення двох версій створеного тонального словника у різних форматах на одному із найбільших вебсервісів для спільної розробки програмного забезпечення – GitHub.

РОЗДІЛ 4. АВТОМАТИЧНЕ ОПРАЦЮВАННЯ ТЕКСТІВ НОВИН

Оскільки попередньо було створено тональний словник української мови, було вирішено апробувати отриманий словник на текстах новин.

Для цього було обрано два вебсайти новин: “tsn.ua” та “hromadske.ua”. Такий вибір пояснюється тим, щоб продемонструвати можливість застосування створеного тонального словника для різних новинних джерел в мережі “Інтернет”. Зокрема, обрані сайти є одними із найпопулярніших новинних сайтів за останній час.

Було вирішено створення програми поділити на декілька етапів:

1. Автоматичне отримання текстів новин із сайту “tsn.ua”.
2. Автоматичне отримання текстів новин із сайту “hromadske.ua”.
3. Автоматичне опрацювання отриманих текстів з метою укладання частотних словників; зіставлення одиниць частотного та тонального словників, щоб дізнатися, чи достатньо слів та словосполучень у тональному словнику для здійснення подальшого семантичного аналізу.

Розбиття автоматичного отримання текстів для кожного із сайтів на окремі етапи пояснюється особливостями подання html-сторінки на кожному із обраних нами сайтів.

Для практичного створення представлених вище етапів роботи нами було обрано одну із найбільш використовуваних на сьогодні мов програмування – мову програмування Python версії 3.8 [74]. Розробка програми здійснювалась у середовищі PyCharm Edu 2019.3.1 [75].

Для вебскрапінгу (вебскрейпінгу) сайтів новин було використано одну із бібліотек Python – Beautiful Soup [76], яка дозволяє легко отримувати інформацію з вебсторінок.

Кожен з етапів було реалізовано в окремому файлі, які пов’язані між собою. Представлення отриманих програм та їх результатів подано нижче.

4.1. Вебскрапінг (вебскрейпінг) (для сайту “tsn.ua”)

Структура html-сторінки сайту “tsn.ua” є досить добре структурованою та зрозумілою. Як ми побачимо згодом, з неї значно легше видобувати тексти новин, ніж із сайту “hromadske.ua”.

Перевагою html-сторінки сайту новин “tsn.ua” є те, що посилання на всі новини за день має досить зручну структуру. Наприклад, посилання на всі новини за 15 травня 2023 року має такий вигляд: https://tsn.ua/news?day=15&month=05&year=2023 (також див. Screenshot (мал. 4.1.1.) нижче). Тобто, досить-таки легко із самого навіть посилання ми можемо дізнатися, за який день ми беремо новини (у цьому прикладі це 15 (day=15) травня (month=05) 2023 року (year=2023)) і, за потреби, ми можемо з легкістю змінювати день, місяць та рік задля подальшого аналізу.

Малюнок 4.1.1. Вигляд посилання на сторінку новин за день на сайті новин “tsn.ua”

Виходить, що ми можемо автоматично отримати із сайту “tsn.ua” всі новини за один день. Було досліджено, що за день на сайті публікується від 40 до 65 новин.

Створення коду для цього етапу роботи було оформлено у вигляді функції, яка в подальшому буде викликатися на етапі АОТ та автоматичного укладання ЧС. Повний вигляд коду представлено у Додатку К.

У коді програми можна побачити, що:

- ➔ код поділено на блоки (кожен блок відділено порожнім рядком);
- ➔ майже кожен рядок коду має пояснення у вигляді закоментованого рядка.

Створення функції, яка автоматично видобуває текстовий вміст статей новин із сайту “tsn.ua”, можна представити у вигляді таких кроків:

- імпортування бібліотек, необхідних для подальшої роботи;
- видобування із сайту “tsn.ua” посилання на кожну новину окремо у вигляді циклу;
- створення порожнього рядка та списків, які у подальшому будуть наповнюватися необхідними текстовими даними;
- створення циклу, який автоматично поетапно видобуває текст новин із отриманих посилань;
- створення внутрішнього циклу для об’єднання заголовків новин в один список;
- створення ще одного циклу, який збирає докупи елементи статей та упорядковує їх в єдиний список текстів новин;
- у результаті запуску створеної функції ми отримуємо єдиний список текстів новин.

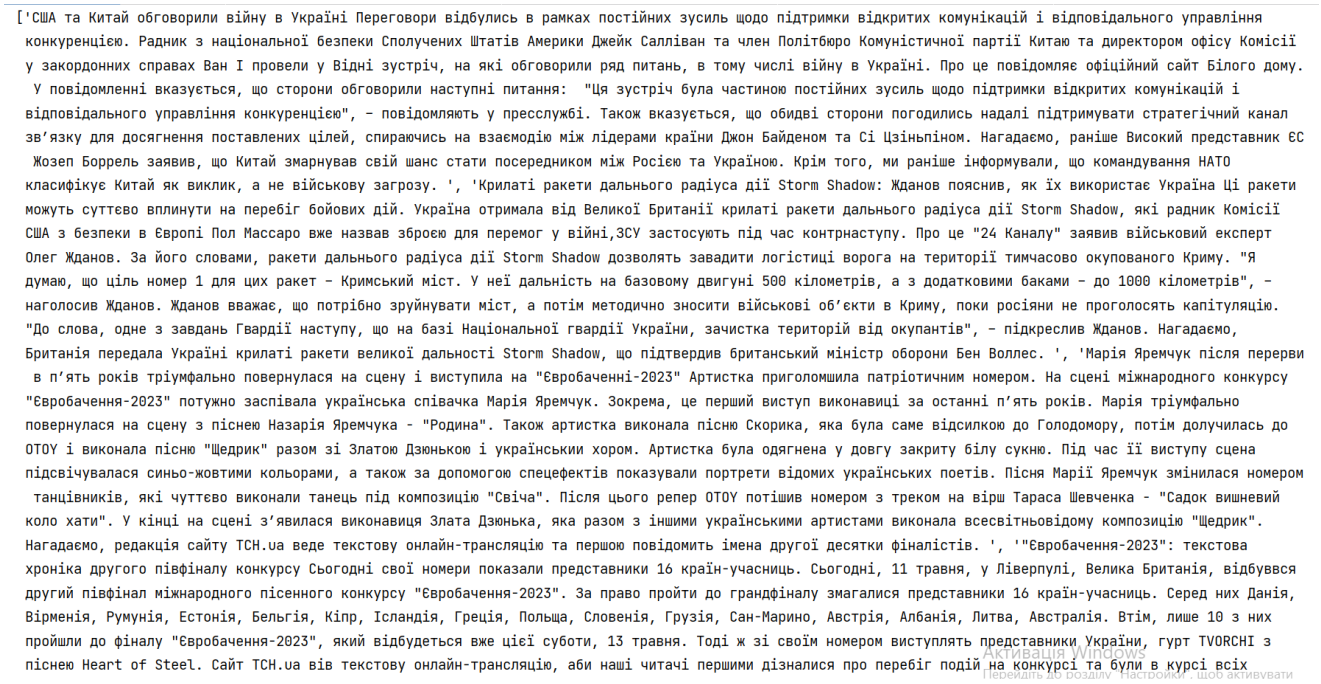
Варто зазначити, що:

- ❖ код функції написаний таким чином, що ми можемо видобувати не тільки всі статті за день, а й створювати обмеження щодо кількості статей, які ми хочемо взяти для подальшого аналізу (див. рядок 31 у коді-додатку К);
- ❖ при видобуванні даних необхідно було провести декілька замін символів Юнікоду на порожні елементи (див. рядок 37 у коді-додатку К).

Хочеться зауважити, що проблеми з окремими символами виникали і при подальших етапах роботи. Вони усуваються шляхом ручного прописування тих символів, з якими ми зіштовхнулися у ході виконання цієї роботи. Проте такий підхід не є універсальним і потребує подальшого вдосконалення коду задля вирішення цієї проблеми.

Один із можливих результатів роботи функції представлено нижче.

Малюнок 4.1.2. Фрагмент вигляду списку всіх текстів новин за 11 травня 2023 року



['США та Китай обговорили війну в Україні Переговори відбулись в рамках постійних зусиль щодо підтримки відкритих комунікацій і відповідального управління конкуренцією. Радник з національної безпеки Сполучених Штатів Америки Джейк Салліван та член Політбюро Комуністичної партії Китаю та директором офісу Комісії у закордонних справах Ван І провели у Відні зустріч, на які обговорили ряд питань, в тому числі війну в Україні. Про це повідомляє офіційний сайт Білого дому. У повідомленні вказується, що сторони обговорили наступні питання: "Ця зустріч була частиною постійних зусиль щодо підтримки відкритих комунікацій і відповідального управління конкуренцією", – повідомляють у пресслужбі. Також вказується, що обидві сторони погодились надалі підтримувати стратегічний канал зв'язку для досягнення поставлених цілей, спираючись на взаємодію між лідерами країни Джон Байденом та Сі Цзіньпіном. Нагадаємо, раніше Високий представник ЄС Жозеп Боррель заявив, що Китай змарнував свій шанс стати посередником між Росією та Україною. Крім того, ми раніше інформували, що командування НАТО класифікує Китай як виклик, а не військову загрозу. ', 'Крилаті ракети дальнього радіуса дії Storm Shadow: Жданов пояснив, як їх використає Україна Ці ракети можуть суттєво вплинути на перебіг бойових дій. Україна отримала від Великої Британії крилаті ракети дальнього радіуса дії Storm Shadow, які радник Комісії США з безпеки в Європі Пол Массаро вже назвав зброєю для перемог у війні,ЗСУ застосують під час контрнаступу. Про це "24 Каналу" заявив військовий експерт Олег Жданов. За його словами, ракети дальнього радіуса дії Storm Shadow дозволять завадити логістиці ворога на території тимчасово окупованого Криму. "Я думаю, що ціль номер 1 для цих ракет – Кримський міст. У неї дальність на базовому двигуні 500 кілометрів, а з додатковими баками – до 1000 кілометрів", – наголосив Жданов. Жданов вважає, що потрібно зруйнувати міст, а потім методично зносити військові об'єкти в Криму, поки росіяни не проголосять капітуляцію. "До слова, одне з завдань Гвардії наступу, що на базі Національної гвардії України, зачистка територій від окупантів", – підкреслив Жданов. Нагадаємо, Британія передала Україні крилаті ракети великої дальності Storm Shadow, що підтвердив британський міністр оборони Бен Воллес. ', 'Марія Яремчук після перерви в п'ять років тріумфально повернулася на сцену і виступила на "Євробаченні-2023" Артистка приголомшила патріотичним номером. На сцені міжнародного конкурсу "Євробачення-2023" потужно заспівала українська співачка Марія Яремчук. Зокрема, це перший виступ виконавиці за останні п'ять років. Марія тріумфально повернулася на сцену з піснею Назарія Яремчука - "Родина". Також артистка виконала пісню Скорика, яка була саме відсилкою до Голодомору, потім долучилась до OTOY і виконала пісню "Щедрик" разом зі Златою Дзюнькою і українськими хором. Артистка була одягнена у довгу закриту білу сукню. Під час її виступу сцена підсвічувалася синьо-жовтими кольорами, а також за допомогою спецефектів показували портрети відомих українських поетів. Пісня Марії Яремчук змінилася номером танцівників, які чуттєво виконали танець під композицію "Свіча". Після цього репер OTOY потішив номером з треком на вірш Тараса Шевченка - "Садок вишневий коло хати". У кінці на сцені з'явилася виконавиця Злата Дзюнька, яка разом з іншими українськими артистами виконала всесвітньовідому композицію "Щедрик". Нагадаємо, редакція сайту ТСН.ua веде текстову онлайн-трансляцію та першою повідомить імена другої десятки фіналістів. ', '"Євробачення-2023": текстова хроніка другого півфіналу конкурсу Сьогодні свої номери показали представники 16 країн-учасниць. Сьогодні, 11 травня, у Ліверпулі, Велика Британія, відбувся другий півфінал міжнародного пісенного конкурсу "Євробачення-2023". За право пройти до грандфіналу змагалися представники 16 країн-учасниць. Серед них Данія, Вірменія, Румунія, Естонія, Бельгія, Кіпр, Ісландія, Греція, Польща, Словенія, Грузія, Сан-Марино, Австрія, Албанія, Литва, Австралія. Втім, лише 10 з них пройшли до фіналу "Євробачення-2023", який відбудеться вже цієї суботи, 13 травня. Тоді ж зі своїм номером виступлять представники України, гурт TVORCHI з піснею Heart of Steel. Сайт ТСН.ua вів текстову онлайн-трансляцію, аби наші читачі першими дізналися про перебіг подій на конкурсі та були в курсі всіх

Активація Windows
Перейдіть до розділу "Настройки", щоб активувати

4.2. Вебскрапінг (вебскрейпінг) (для сайту “hromadske.ua”)

Працювати із сайтом “hromadske.ua” було трохи складніше. Основна відмінність між сайтом “hromadske.ua” та сайтом “tsn.ua” є те, що на сайті “hromadske.ua” представлено всі новини у вигляді сторінок. Тобто, ми вже не можемо автоматично визначити наперед, за який часовий проміжок ми досліджуємо новини. Дізнатися про це можна лише шляхом аналізу дати для кожної новини або вручну передивлятися сторінку-стрічку новин. Такий нюанс щодо представлення html-сторінки, на нашу думку, є суттєвим недоліком.

Створення коду для цього етапу роботи також було оформлено у вигляді функції, яка в подальшому теж буде викликатися на етапі АОТ та автоматичного укладання ЧС. Ознайомитися з повним виглядом коду можна у Додатку Л.

Створення функції, яка автоматично видобуває текстовий вміст статей новин із сайту “hromadske.ua”, можна представити у вигляді таких кроків:

- імпортування бібліотек, необхідних для подальшої роботи;

- створення списку, де представлено перелік сторінок-стрічок новин, які необхідно взяти для подальшої роботи;
- створення циклу для видобування із сайту “hromadske.ua” посилання на кожну новину окремо;
- створення порожніх списків, які у подальшому будуть наповнюватися необхідними текстовими даними;
- створення циклу, який автоматично поетапно видобуває текст новин із отриманих посилань;
- створення двох додаткових блоків “try-except”, які необхідні для вилучення текстових даних, що належать до фотогалереї (gallery_section) або до посилання на інше джерело (iframe_section);
- створення внутрішнього циклу, який наповнюватися тілом новини без тексту з фотогалереї та без тексту з посилання на інше джерело;
- створення ще одного внутрішнього циклу задля проведення заміни окремих символів та наповнення списку основним текстом новин;
- створення ще одного циклу, який збирає докупи елементи статей та упорядковує їх в єдиний список текстів новин;
- у результаті запуску створеної функції ми отримуємо єдиний список текстів новин.

Варто зазначити, що:

❖ при видобуванні даних необхідно було провести декілька замін символів Юнікоду на порожні елементи або на необхідні уніфіковані символи (див. рядки 85, 96 та 97 у коді-додатку Л);

❖ було помічено, що створений нами код не враховує того, що слід робити, якщо класу підзаголовок (leadtext) немає. Цю проблему варто вирішити у подальшому, щоб уникнути проблем з видобуттям даних.

Один із можливих результатів роботи функції представлено нижче.

Малюнок 4.2.1. Фрагмент вигляду списку перших двох текстів новин за 17 травня 2023 року

```
[' П’яний чоловік проник у будинок радника Байдена, оминувши охорону. США розслідують, як таке могло статися  Секретна служба США розслідує, як
 нетверезий чоловік увійшов у будинок радника президента Джо Байдена з національної безпеки Джейка Салліван посеред ночі приблизно два тижні тому, і
 його не помітили агенти, які охороняли будинок. Про це повідомляє Washington Post. За даними видання, невідомий чоловік, який досі перебуває на
 свободі, увійшов у будинок Салліван близько 3 години ночі наприкінці квітня. Радник Байдена зіткнувся з цим чоловіком і наказав йому піти. Слідів
 проникнення в будинок не було. Салліван має цілодобову службу безпеки. Але агенти, які стояли біля будинку, не знали, що зловмисник проник всередину,
  аж поки той уже не пішов і Салліван не вийшов на вулицю, щоб попередити охорону. Джерела WP зазначили, що зловмисник був у стані алкогольного
 сп’яніння та не розумів, де він перебуває. Здавалося, він не знав Саллівана і не мав наміру завдати йому шкоди.  Секретна служба США повідомила, що
 почала розслідування інциденту та того, як зловмисник непоміченим проник до будинку Саллівана. Відомство вважає порушення безпеки «предметом
 серйозного занепокоєння». Також Секретна служба вжила додаткових заходів безпеки для Саллівана та навколо його будинку до завершення розслідування.
 Видання зазначає, що зазвичай будь-кого, хто проникає на територію людини, яка охороняється Секретною службою, затримують для допиту, а потім, швидше
  за все, заарештовують і звинувачують у незаконному проникненні. Але чоловік, який увійшов у будинок Саллівана, покинув місце події до того, як
 агенти дізналися про його присутність. Це не перший схожий випадок. У жовтні 2022 року невідомий скоїв злам у будинку тодішньої спікерки Палати
 представників США Ненсі Пелосі у Каліфорнії. Зловмисник кричав «Де Ненсі?», а тоді вдарив молотком чоловіка спікерки Пола Пелосі. Ненсі мала охорону,
  але її чоловік залишався без захисту, коли був не поруч із дружиною.', ' У Бахмуті загинув американський доброволець, пригожин пообіцяв передати
 його тіло з пошаною в США  У Бахмуті внаслідок російського артилерійського обстрілу загинув колишній військовослужбовець армії США Ніколас Меймер,
 який воював на боці України. Про це пише CNN. Як розповів близький друг Меймера та засновник некомерційної організації AFGFгее, яка працює в Україні,
  підполковник у відставці Перрі Блекберн, американський доброволець перебував у будівлі, що обвалилася внаслідок артобстрілу. За його словами,
 українці, які воювали разом з Меймером, вважали, що він або опинився в пастці під завалами, або був убитий «шквальним вогнем» російської артилерії.
 «Вони зайняли будівлю, артилерія почала обстріл, і будівля почала руйнуватися. Саме тоді більшість американців та українців, які перебували у
 будівлі, втекли. На жаль, Ніку не вдалося втекти», — розповів CNN американський друг Меймера в Україні. Підтвердження загибелі Меймера з’явилося 16
 травня після того, як керівник російської ПВК «Вагнер» євген пригожин показав тіло американця у своєму пропагандистському відео. У відео пригожин
 оглядає тіло американського добровольця та стверджує, що знайшов документи громадянина США. «Ми його передамо США, покладемо його в труну, накриємо
 американським прапором з повагою, тому що він не помер у своєму ліжку як дідусь, він загинув на війні. Ймовірно, це гідна [смерть], чи не так?», —
 каже пригожин. Нагадаємо, у квітні в боях за Бахмут загинув ірландський доброволець Фінбар Кафферкі. Міністр закордонних справ Ірландії Майкл Мартін
 вшанував пам’ять загиблого. У посольстві росії в Ірландії звинуватили в смерті добровольця ірландський уряд і медіа.']
```

Примітка: під першими текстами новин у цьому випадку маються на увазі ті тексти новин, які були додані нещодавно (адже цей сайт новин надає новини за спадою дати та часу публікації новини).

4.3. Автоматичне опрацювання отриманих текстів та автоматичне укладання частотних словників

Наступний етап роботи з автоматично отриманими текстами новин – це АОТ з подальшим укладанням ЧС.

Для отриманих текстів у подальшому була здійснена АОТ в межах одного файлу, тобто кодом, який є дозволяє працювати з текстами новин з обох обраних джерел.

Останнім етапом роботи є представлення частки покриття аналізованого тексту за допомогою частотного словника. Такий процес був здійснений, щоб наочно продемонструвати, наскільки достатнім є обсяг створеного тонального словника для текстів новин, які публікуються сьогодні.

АОТ та укладання ЧС було розбито на такі кроки, окремі з яких можна представити у вигляді етапів-блоків, які детально описані у самому коді (див. Додаток М):

- залучення функцій вебскрапінгу текстів новин з сайту "tsn.ua" та "hromadske.ua";
- імпортування бібліотек, необхідних для подальшої роботи;
- етап 1: робота з другою (розширеною) версією тонального словника з метою видобутку слів з їхніми можливими варіантами написання (певні кроки цього етапу збігаються з оформленням таблиці для другої версії тонального словника);
- етап 2: виклик функцій вебскрапінгу текстів новин з сайту "tsn.ua" та "hromadske.ua";
- етап 3: токенізація;
- етап 4: вилучення з текстів стоп-слів (список стоп-слів було взято з GitHub [77]);
- етап 5: лематизація (приведення слів-токенів до початкової форми);
- етап 6: автоматичне укладання ЧС та розрахунок частки покриття тексту.

  В ході створення та роботи етапу 6 було помічено, що:

➔ при укладанні ЧС не враховуються словосполучення;

➔ ЧС створюється для всіх новин (у подальшому було б краще укладати ЧС для кожної новини окремо);

➔ існує потенційна можливість у майбутньому здійснити поділ останнього етапу на декілька: створення ЧС окремо, частка покриття тексту – окремо.

Можливі результати роботи функції представлено нижче.

Малюнок 4.3.1. Фрагмент вигляду ЧС для перших п’яти текстів новин за 18 травня 2023 року (джерело: “tsn.ua”)

```
([('військовий', 18), ('крим', 14), ('це', 14), ('батут', 10), ('пляж', 8), ('україна', 8), ('місто', 8), ('травень', 7), ('нагадати', 7), ('час', 7),
 ('питання', 7), ('наталія', 7), ('дуже', 7), ('кинджал', 6), ('президент', 6), ('ракета', 6), ('зазначити', 6), ('одеса', 6), ('відкриття', 6), ('територія',
  6), ('постачання', 6), ('консультативний', 6), ('дитина', 6), ('збити', 5), ('арестович', 5), ('інформація', 5), ('розуміти', 5), ('відпочивати', 5),
 ('одеський', 5), ('море', 5), ('завдання', 5), ('окупований', 5), ('склад', 5), ('все', 5), ('київ', 4), ('російський', 4), ('сезон', 4), ('відео', 4),
 ('тривати', 4), ('місцевий', 4), ('дозволити', 4), ('узбережжя', 4), ('чорне', 4), ('колія', 4), ('експерт', 4), ('тимчасово', 4), ('реінтеграція', 4),
 ('статися', 4), ('дівчинка', 4), ('розповісти', 4), ('ганн', 4), ('так', 4), ('після', 4), ('заявити', 3), ('гучний', 3), ('дискотека', 3), ('фото', 3),
 ('заборона', 3), ('життя', 3), ('країна', 3), ('ворог', 3), ('захід', 3), ('мережа', 3), ('повинний', 3), ('працювати', 3), ('тсн', 3), ('ua', 3),
 ('рішення', 3), ('командування', 3), ('зон', 3), ('відповідний', 3), ('представник', 3), ('вибух', 3), ('свідчити', 3), ('партизан', 3), ('виконувати', 3),
 ('знищення', 3), ('амуніція', 3), ('севастополь', 3), ('зеленський', 3), ('деокупація', 3), ('автономний', 3), ('республіка', 3), ('трагедія', 3),
 ('атракціон', 3), ('аня', 3), ('дитинка', 3), ('бута', 3), ('досі', 3), ('якийсь', 3), ('також', 3), ('навіть', 3), ('вночі', 2), ('стверджувати', 2),
 ('офіс', 2), ('олексій', 2), ('сотня', 2), ('повітряний', 2), ('розвідка', 2), ('курортний', 2), ('війнути', 2), ('віддавати', 2), ('південь', 2),
 ('соціальний', 2), ('збиратися', 2), ('внутрішній', 2), ('робити', 2), ('керівниця', 2), ('коментар', 2), ('вважати', 2), ('писати', 2), ('пляжний', 2),
 ('канал', 2), ('підготовка', 2), ('муніципальний', 2), ('повідомити', 2), ('міський', 2), ('рад', 2), ('хотіти', 2), ('мерія', 2), ('підкреслити', 2),
 ('морський', 2), ('вода', 2), ('прибережний', 2), ('область', 2), ('річник', 2), ('братчук', 2), ('брифінг', 2), ('пропозиція', 2), ('купатися', 2),
 ('гуменюк', 2), ('безпека', 2), ('окупант', 2), ('підрив', 2), ('технік', 2), ('затримати', 2), ('думка', 2), ('зсу', 2), ('світан', 2), ('день', 2),
 ('вагон', 2), ('зійти', 2), ('рейка', 2), ('йти', 2), ('кримський', 2), ('міст', 2), ('мелітополь', 2), ('залізниця', 2), ('затвердити', 2), ('увійти', 2),
 ('голова', 2), ('указ', 2), ('зокрема', 2), ('заступник', 2), ('сидіти', 2), ('спин', 2), ('телефон', 2), ('розбігтися', 2), ('лєра', 2), ('миколаїв', 2),
 ('казати', 2), ('персонал', 2), ('стати', 2), ('подробиця', 2), ('загинути', 2), ('петля', 2), ('кучинська', 2), ('жінка', 2), ('лєрочок', 2), ('підбігти',
 2), ('водичка', 2), ('мама', 2), ('забігти', 2), ('рука', 2), ('г', 2), ('п', 2), ('місяць', 2), ('ніч', 2), ('проти', 2), ('дійсно', 2), ('зараза', 2),
 ('люда', 2), ('вже', 2), ('став', 2), ('разом', 2), ('збитий', 1), ('оглядач', 1), ('колишній', 1), ('радник', 1), ('ракетний', 1), ('удар', 1), ('інтерв’ю',
  1), ('юлій', 1), ('латиніна', 1), ('росія', 1), ('орієнтовно', 1), ('швидко', 1), ('закінчитися', 1), ('кинжал', 1), ('перетворитися', 1), ('носій', 1),
 ('льотчик', 1), ('підготовлений', 1), ('застосування', 1), ('вперше', 1), ('український', 1), ('небо', 1), ('сила', 1), ('ппо', 1), ('летіти', 1), ('даний',
 1), ('британський', 1), ('путін', 1), ('шокований', 1), ('вразливість', 1), ('розхвалений', 1), ('забитий', 1), ('розпочатися', 1), ('відвідувати', 1),
```

Малюнок 4.3.2. Фрагмент вигляду ЧС для перших п’яти текстів новин за 19 травня 2023 року (джерело: “hromadske.ua”)

```
([('україна', 14), ('саудівський', 11), ('аравія', 11), ('грузія', 10), ('вккс', 10), ('член', 9), ('це', 9), ('росія', 8), ('зеленський', 7), ('суд', 7),
 ('комісія', 7), ('суддів', 7), ('аеропорт', 6), ('час', 6), ('президент', 6), ('марченко', 6), ('ділянка', 6), ('російський', 5), ('прокуратура', 5),
 ('вищий', 5), ('ракета', 4), ('повідомити', 4), ('авіакомпанія', 4), ('#', 4), ('держава', 4), ('оксана', 4), ('гривня', 4), ('слідство', 4), ('фігурант',
 4), ('наразі', 4), ('кваліфікаційний', 4), ('кандидат', 4), ('зараза', 4), ('перший', 4), ('також', 4), ('понада', 4), ('скібіцький', 3), ('ракетний', 3),
 ('росіянин', 3), ('місяць', 3), ('крим', 3), ('одиниця', 3), ('літак', 3), ('москва', 3), ('протест', 3), ('тбілісі', 3), ('акція', 3), ('грузинський', 3),
 ('рф', 3), ('указ', 3), ('травень', 3), ('вперше', 3), ('йтися', 3), ('міністр', 3), ('мати', 3), ('лютий', 3), ('вартість', 3), ('особа', 3), ('справа', 3),
  ('намір', 3), ('факт', 3), ('верховний', 3), ('конкурсний', 3), ('правосуддя', 3), ('вплив', 3), ('розслідування', 3), ('посад', 3), ('врп', 3),
 ('атакувати', 2), ('зійти', 2), ('конвеєр', 2), ('російсько', 2), ('окупаційний', 2), ('військо', 2), ('значний', 2), ('озброєння', 2), ('виробництво', 2),
 ('виготовлений', 2), ('заступник', 2), ('начальник', 2), ('розвідка', 2), ('калібр', 2), ('завозити', 2), ('тимчасово', 2), ('окупований', 2), ('приблизно',
 2), ('вдаватися', 2), ('пояснити', 2), ('новий', 2), ('прилетіти', 2), ('рейс', 2), ('грузиїй', 2), ('фото', 2), ('опозиційний', 2), ('активіст', 2),
 ('учасник', 2), ('авіасполучення', 2), ('політ', 2), ('володимир', 2), ('путін', 2), ('громадянин', 2), ('влада', 2), ('вплинути', 2), ('прибути', 2),
 ('політичний', 2), ('в’язень', 2), ('виступити', 2), ('саміт', 2), ('принц', 2), ('бін', 2), ('сауд', 2), ('пріоритет', 2), ('повернення', 2), ('незаконно',
 2), ('реалізація', 2), ('наступний', 2), ('лідер', 2), ('відвідувати', 2), ('великий', 2), ('сімка', 2), ('відносини', 2), ('країна', 2), ('квітень', 2),
 ('відвідати', 2), ('прем’єр', 2), ('посольство', 2), ('даний', 2), ('візит', 2), ('порошенко', 2), ('земля', 2), ('береза', 2), ('дніпро', 2), ('передати',
 2), ('земельний', 2), ('район', 2), ('мільйон', 2), ('власність', 2), ('фінансування', 2), ('кримінальний', 2), ('отримати', 2), ('підозра', 2), ('майно',
 2), ('князєв', 2), ('обрання', 2), ('виявити', 2), ('корупція', 2), ('дан', 2), ('відповідь', 2), ('антикорупційний', 2), ('можливий', 2), ('відбір', 2),
 ('слідчий', 2), ('рад', 2), ('наголосити', 2), ('досудовий', 2), ('тривати', 2), ('певний', 2), ('орган', 2), ('відповідальний', 2), ('формування', 2),
 ('судовий', 2), ('тисяча', 2), ('тільки', 2), ('вже', 2), ('після', 2), ('процес', 2), ('вичерпати', 1), ('запас', 1), ('вдатися', 1), ('налагодити', 1),
 ('головний', 1), ('управління', 1), ('міністерство', 1), ('оборона', 1), ('вадим', 1), ('коментар', 1), ('рбк', 1), ('уточнити', 1), ('уламок', 1),
 ('застосований', 1), ('вказувати', 1), ('боєприпас', 1), ('перше', 1), ('квартал', 1), ('щонайменше', 1), ('двічі', 1), ('змога', 1), ('виготовляти', 1),
 ('-101', 1), ('кинджал', 1), ('балістичний', 1), ('м723', 1), ('іскандер', 1), ('п’ять', 1), ('розповісти', 1), ('комплектувальний', 1), ('міжнародній', 1),
 ('санкція', 1), ('гур', 1), ('вважати', 1), ('кампанія', 1), ('продовжуватися', 1), ('вистачати', 1), ('ресурс', 1), ('комбінувати', 1), ('засіб', 1),
 ('експеримент', 1), ('першочерговий', 1), ('мета', 1), ('продовжувати', 1), ('укріплювати', 1), ('ппо', 1), ('темп', 1), ('азимут', 1), ('зустрічати', 1),
 ('писати', 1), ('ехо', 1), ('кавказ', 1), ('новість', 1), ('здійснити', 1), ('прямий', 1), ('вилетіти', 1), ('внуково', 1), ('приземлитися', 1), ('столиця',
 1), ('місцевий', 1), ('наданий', 1), ('медіа', 1), ('publika', 1), ('зібратися', 1), ('гасло', 1), ('корабель', 1), ('доставити', 1), ('євросоюз', 1),
 ('чергувати', 1), ('поліція', 1), ('пройти', 1), ('зал', 1), ('приліт', 1), ('поліцейський', 1), ('затримати', 1), ('мітинг', 1), ('виступати', 1),
 ('відновлення', 1), ('полицейский', 1), ('задержати', 1), ('несколько', 1), ('оппозиционныха', 1), ('активистов', 1), ('акции', 1), ('тбилисский', 1),
```

Активація Windows
Перейдіть до розділу "Настройки", щоб активувати

Примітки щодо отриманих результатів:

- інколи трапляються службові символи (мова йде, наприклад, про символ “\n”, що означає перенесення на наступний рядок і подібні символи). Тож

потрібно врахувати такі ситуації та усунути їх в подальшому. Також ці символи можуть входити до складу слова (на початку або наприкінці слова), що теж варто взяти до уваги, тому що слова з такими службовими символами не розпізнаються як те слово, яке ми як люди можемо бачити, що впливає на отримані результати;

- інколи трапляються символи Юнікоду, з якими такі ж проблеми, як і зі службовими символами, про що вже згадано та описано вище;
- зустрічаються неправильні написання слів у текстах новин: наприклад, пропущена літера у слові або дві сусідні літери поміняні місцями (наприклад, “Перзидент” замість “Президент”) тощо;
- слова зі своїми можливими варіантами не групуються між собою, що можна було б зробити при подальшій роботі задля удосконалення вже наявних результатів;
- при токенізації слова, написані через дефіс, автоматично розбиваються на два або більше окремі слова (кількість слів залежить від кількості дефісів у слові). Це спричиняє проблеми, адже слова, написані через дефіс, аналізуються неправильно. Варіантом вирішення цієї проблеми може бути поєднання слів, які написані у тональному словнику, після лематизації. Проте практично втілити сказане досить-таки складно;
- трапляються помилки при здійсненні автоматичної лематизації шляхом залучення бібліотеки pymorphy3 (яка працює краще за pymorphy2). Тому у майбутньому хочеться здійснювати лематизацію з урахуванням частиномовної розмітки, яка була здійснена вручну та представлена у межах тонального словника.

Частотні словники укладаються автоматично при кожному запуску коду. Тому обсяг ЧС залежить від обсягів вхідних даних (текстів). У ході виконання роботи частотні словники укладалися з метою з’ясування достатності обсягу

тонального словника. Можна сказати, що реалізація роботи з ЧС проводилася поки що у тестовому режимі. Виведення ЧС слугувало наочним доведенням правильності підрахунків щодо частки покриття.

Проте можливості ЧС є значно більшими, які можна використати при подальших дослідженнях. Наприклад, можна створити функцію, яка у кінцевому результаті буде надавати інформацію щодо високочастотних слів, яких немає в тональному словнику й які варто було б включити до складу тонального словника.

4.4. Статистичні дані щодо АОТ новин

Було вирішено проаналізувати тексти новин з автоматичним укладанням частотних словників, щоб з'ясувати, чи достатньо одиниць у створеному тональному словнику задля того, щоб у подальшому представити сентимент-аналіз текстів належним чином з якомога точнішими результатами.

У зв'язку з цим було обрано різну кількість текстів новин за 3 майже рівномірних часових проміжки: за 2021, 2022 та 2023 роки.

Оскільки два обраних вебсайти новин мають різне подання новин, то й аналіз текстів новин було проведено у два різні способи:

- для сайту "tsn.ua" було обрано тексти новин за 18 травня 2021, 2022 та 2023 років;
- для сайту "hromadske.ua" було обрано тексти новин за червень місяць (один або два дні в приблизно однаковому часовому проміжку: середина місяця) за 2021, 2022 та 2023 роки.

Дані, отримані на основі роботи коду-програми, представлено нижче у таблиці 4.4.1, яка надає інформацію у відсотках щодо того, скільки одиниць тексту присутні в тональному словнику. У цій таблиці представлено частку покриття одного, трьох, п'яти, семи та всіх текстів для обраного сайту новин. Загальна кількість текстів за аналізований проміжок вказана у дужках. Також у таблиці можна ознайомитися з результатами щодо таких статистичних характеристик як

середнє значення частки покриття за обраний проміжок часу, найбільше та найменше значення відповідно.

Таблиця 4.4.1. Частка покриття текстів із сайтів новин

Дата: **18.05.21** (к-сть текстів новин)	Джерело: **tsn.ua**	Дата: **16-17.06.21** (к-сть текстів новин)	Джерело: **hromadske.ua**
1	78,16%	1	86,17%
3	61,11%	3	71,94%
5	71,43%	5	77,08%
7	67,61%	7	75,63%
всі (23)	67,7%	всі (12)	74,3%
Середнє значення (за 18.05.21):	**69,2%**	Середнє значення (за 16-17.06.21):	**77,02%**
Найбільше значення (за 18.05.21):	**78,16%**	Найбільше значення (за 16-17.06.21):	**86,17%**
Найменше значення (за 18.05.21):	**61,11%**	Найменше значення (за 16-17.06.21):	**71,94%**
Дата: **18.05.22** (к-сть текстів новин)		Дата: **13-14.06.22** (к-сть текстів новин)	
1	77,22%	1	78,57%
3	75,57%	3	78,6%

5	75,05%	5	78,5%
7	72,25%	7	75,33%
всі (23)	73,16%	всі (12)	75,38%
Середнє значення (за 18.05.22):	**74,65%**	Середнє значення (за 13-14.06.22):	**77,28%**
Найбільше значення (за 18.05.22):	**77,22%**	Найбільше значення (за 13-14.06.22):	**78,6%**
Найменше значення (за 18.05.22):	**72,25%**	Найменше значення (за 13-14.06.22):	**75,33%**
Дата: **18.05.23** (к-сть текстів новин)		Дата: **09-10.06.23** (к-сть текстів новин)	
1	78,69%	1	66,5%
3	82,02%	3	78,16%
5	74,61%	5	77,76%
7	73,42%	7	74,55%
всі (23)	72,53%	всі (12)	72,43%
Середнє значення (за 18.05.23):	**76,25%**	Середнє значення (за 09-10.06.23):	**73,88%**
Найбільше значення (за 18.05.23):	**82,02%**	Найбільше значення (за 09-10.06.23):	**78,16%**

Найменше значення (за 18.05.23):	**72,53%**	Найменше значення (за 09-10.06.23):	**66,5%**

Отримані дані було проаналізовано ще раз, узагальнено та представлено у вигляді таблиці 4.4.2.

Таблиця 4.4.2. Зведені статистичні дані

Період, статистичний показник	Джерело	
	tsn.ua	**hromadske.ua**
	Середнє значення	
за 2021	69,2%	77,02%
за 2022	74,65%	77,28%
за 2023	76,25%	73,88%
Середнє арифметичне	**73,37%**	**76,06%**
Найбільше значення	**86,17%**	
Найменше значення	**61,11%**	

Як ми можемо бачити із таблиць, які представлено вище, частка покриття досліджених текстів перебуває у межах 61-87%. Простежується така закономірність: що менший текст (у нашому випадку за текст сприймається кількість новин, які ми беремо для аналізу), то більша частка покриття тексту (див. результати щодо частки покриття для одного та трьох текстів) і навпаки. Хоча так відбувається не завжди, адже найменша частка покриття при

практичному дослідженні була розпорошеною між різними обсягами текстів: 1 раз найменша частка покриття була для одного тексту (див. 09-10.06.23 для сайту “hromadske.ua”), 2 рази – для трьох (див. результати для обох сайтів за 2021 рік), 2 рази – для семи (див. результати для обох сайтів за 2022 рік) та лише 1 раз для всіх досліджуваних текстів (див. 18.05.23 для сайту “tsn.ua”).

З урахуванням того, що для лінгвістичних досліджень достатнім є показник у 75%, отримані показники є задовільними, оскільки середня частка покриття для обох сайтів становить 74,72%.

Було помічено, що отримані результати пов’язані з багатством тематики новин, що. у свою чергу, зумовлює широку палітру добору лексики як загальномовної, так і вузькоспеціалізованої (наприклад, військова лексика).

Якщо врахувати всі ті нюанси, які необхідно покращити, та доповнити тональний словник, то у подальшому будуть усунені всі недоліки програми, що сприятиме підвищенню показнику частки покриття тексту.

Висновки до четвертого розділу

У цьому розділі представлено практичну реалізацію здійснення вебскрапінгу (вебскрейпінгу) для текстів новин із сайтів “tsn.ua” та “hromadske.ua”. Висвітлено особливості роботи з вебсайтами на прикладі двох досліджених сайтів. Згадано про можливі проблеми, з якими можна зіштовхнутися в ході автоматичного вилучення текстових даних з вебсайтів.

Також представлено першу спробу здійснити автоматичне опрацювання отриманих текстів, фінальним етапом яких є представлення частотного словника лексем зі спадом абсолютної частоти та відображення частки покриття текстів словами з тонального словника. Отримані результати засвідчили недостатність обсягів тонального словника, який у подальшому треба поповнювати лексичними одиницями.

Зокрема було проаналізовано результати роботи кодів, вказано на знайдені проблеми та недоліки, які в перспективі можна усунути задля покращення роботи програми з метою отримання більш достовірних даних.

ВИСНОВКИ

Тональний словник – це новий вид словника, який спрямований на те, щоб досліджувати емоційний аспект слова. У ході дослідження було запропоновано два варіанти визначення тонального словника. Тональний словник, або словник аналізу настроїв – 1) це той словник, що становить список слів і словосполучень зі значенням основної емоційної настроєності для кожного слова; 2) це той словник, який містить інформацію про емоції або полярність, які виражені словами, фразами або поняттями.

Тональний словник створюється з метою сентимент-аналізу тексту. У свою чергу, аналіз тональності тексту – це клас методів контент-аналізу в комп'ютерній лінгвістиці, призначений для автоматизованого виявлення в текстах емоційно забарвленої лексики й емоційної оцінки авторів (думок) по відношенню до об'єктів, про які йдеться в тексті. Було розглянуто основі етапи та типи сентимент-аналізу, які існують на сьогодні. Зокрема, було згадано про ряд переваг, які може надати сентимент-аналіз, наприклад такі як масштабованість, аналіз у реальному часі та послідовність критеріїв оцінювання.

Хоча тональних словників досить мало, проте вже існує класифікація цих словників. Виокремлюють такі два різновиди / категорії тональних словників:

1) тональний словник як публічно доступний лексичний ресурс, який є автоматизованим, або автоматизований аналіз тональності;

2) тональний словник як набір проаналізованих даних, або ручний, або аналіз тональності експертами.

Тональні словники допомагають визначити емоційне забарвлення текстів різного призначення: відгуки про готелі, банки, ресторани, коментарі про переглянуті фільми чи серіали, мультфільми, повідомлення у блогах та соціальних мережах про політичну діяльність та конкретні політичні події та ін.

Аналіз обраних п'яти тональних словників для трьох мов довів, що словник може упорядковуватися різною кількістю експертів, бути різним за обсягом та призначенням, мати різну шкалу оцінки; залучати тільки людей або комп'ютерні

технології й людей, які займаються машинним навчанням з метою подальшого опрацювання комп'ютером даних з мінімальним залученням людини. Порівняльна характеристика показала, що при створенні тональних словників використовуються два вище описані принципи укладання словників: автоматизований або ручний.

Сфери застосування сентимент-аналізу щороку лише збільшуються, що лише доводить про потребу здійснення аналізу такого роду.

У процесі роботи нами був створений та проаналізований тональний словник української мови обсягом 14 857 слів. Розроблені детальні інструкції щодо його впорядкування, з якими можна ознайомитися у додатках.

У ході роботи ми зіштовхнулися з рядом проблем, що стосувалися омонімії слів та їх форм непрямих відмінків. Частково ці проблеми були розв'язані.

Також було здійснено автоматичне зіставлення двох файлів у форматі csv задля пошуку тих слів, які варто додати до тонального словника з урахуванням слів, що вже входять до складу тонального словника.

На одному із найбільших вебсервісів для спільної розробки програмного забезпечення, GitHub, було оприлюднено дві версії створеного тонального словника у різних форматах та представленнях.

У ході цієї роботи було здійснено порівняння частотних словників (які автоматично укладалися шляхом вебскрапінгу двох вебсайтів новин: "tsn.ua" і "hromadske.ua") із тональним словником української мови з вказівкою на проблеми, які можуть виникнути у ході виконання такої роботи.

Вебскрапінг (вебскрейпінг) – це процес, який стосується видобування необхідних даних з вебсайту з метою подальшого використання отриманої інформації у зручному для дослідника чи користувача форматі. Тому вебскрапінг (вебскрейпінг) визнано одним із найефективніших та найбільш корисних способів здобуття даних з вебсайту, особливо якщо залучати комп'ютерні технології, що нами й було зроблено.

Одним із кінцевих етапів автоматичної обробки отриманих текстів новин є автоматичне укладання частотного словника зі спадом абсолютної частоти. Практичній реалізації цьому передувало теоретичне ознайомлення з особливостями створення та укладання частотних словників. Було з'ясовано, що нині існує розгалужена система щодо класифікації частотних словників за різними параметрами й перевагу надають тим частотним словникам, які укладаються автоматично. Взагалі у період діджиталізації частотні словники є одним із найголовніших інструментаріїв задля проведення різного роду лінгвістичних та статистичних досліджень, адже сфера використання таких словників є досить широкою.

Окрім представлення частотного словника, надано частку покриття текстів словами з тонального словника, яка коливається в межах від 61 до 87 відсотків. Середня частка покриття текстів для обох сайтів становить 74,72%. З урахуванням обсягів тонального словника та різноплановістю тематики у медійному стилі, цей показник є задовільним.

Усі завдання, які були поставлені нами на початку роботи, були виконані.

У планах на майбутнє є:

1) розмістити нові версії створеного словника на GitHub у вільному доступі, де будуть вилучені всі проблемні місця та розведено омонімію;

2) доповнити створений словник списком фразеологізмів, щоб програма могла аналізувати такі випадки як єдине ціле;

3) продовжити аналіз слів іншомовного походження з метою подання всіх можливих варіантів їх написання;

4) здійснити дослідження щодо можливих варіантів написання для слів з новим правописом української мови;

5) здійснювати автоматичне об'єднання слів у складні слова та словосполучення, як і для повних форм абревіатур;

6) здійснювати автоматичне об’єднання слів у групи за принципом “можливі варіанти написання”;

7) здійснювати автоматичне доповнення словника словами, якщо вони часто трапляються у досліджуваних текстах і відсутні у словнику та мають емоційне забарвлення;

8) створити програми генерування дієприкметників та дієприслівників, форм порівняння прикметників та прислівників, оскільки в словнику вони представлені не в повному обсязі;

9) удосконалювати роботу кодів, які здійснюють автоматичну обробку текстів, задля покращення роботи програми з метою отримання більш достовірних даних;

10) розпочати роботу над створенням автоматичної системи аналізу тональності.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Angoos KnowledgeREADER. URL: Angoos KnowledgeREADER in 2022 - Reviews, Features, Pricing, Comparison - PAT RESEARCH: B2B Reviews, Buying Guides & Best Practices (predictiveanalyticstoday.com) (дата звернення: 13.02.22).
2. Averbis. URL: Text Mining & Natural Language Processing analyze texts, gain answers (averbis.com) (дата звернення: 13.02.22).
3. D. Jurafsky and J. H. Martin. Speech and Language Processing (3rd ed. draft). Naive Bayes and Sentiment Classification. URL: Speech and Language Processing (stanford.edu) (дата звернення: 13.02.22).
4. Francesco Elia. Sentiment Analysis Dictionaries. URL: https://www.baeldung.com/cs/sentiment-analysis-dictionaries (дата звернення: 13.02.22).
5. Google Cloud Natural Language API. URL: Cloud Natural Language | Google Cloud (дата звернення: 13.02.22).
6. IBM Watson Natural Language Understanding. URL: IBM Watson Natural Language Understanding - Обзор - Российская Федерация | IBM (дата звернення: 13.02.22).
7. Lexalytics Salience. URL: Data Analytics with NLP & Text Analytics | Lexalytics (дата звернення: 13.02.22).
8. Meaning Cloud. URL: Text Analytics – MeaningCloud text mining solutions (дата звернення: 13.02.22).
9. Microsoft Azure Text Analytics API. URL: Анализ текста | Microsoft Azure (дата звернення: 13.02.22).
10. NetOwl. URL: NetOwl® — AI-based Text Analytics and Identity Analytics for Big Data (дата звернення: 13.02.22).
11. PrediCX. URL: Platform | Warwickanalytics (дата звернення: 13.02.22).
12. Quick Introduction to Sentiment Analysis. URL: https://towardsdatascience.com/quick-introduction-to-sentiment-analysis-74bd3dfb536c (дата звернення: 13.02.22).

13. Rosette Text Analytics. URL: Rosette Text Analytics Platform - AI for Human Language (дата звернення: 13.02.22).
14. SAS Sentiment Analysis. URL: SAS Visual Text Analytics | SAS (дата звернення: 13.02.22).
15. Sentiment Analysis Explained. URL: Sentiment Analysis | Lexalytics (дата звернення: 13.02.22).
16. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. URL: https://github.com/aesuli/SentiWordNet/blob/master/papers/LREC06.pdf (дата звернення: 13.02.22).
17. SentiWords. URL: https://hlt-nlp.fbk.eu/technologies/sentiwords (дата звернення: 15.05.21).
18. The Complete Guide to Sentiment Analysis. URL: Sentiment Analysis: Definition, Uses, Examples + Pros /Cons (getthematic.com) (дата звернення: 13.02.22).
19. Top 13 Sentiment Analysis Software. URL: Top 13 Sentiment Analysis Software in 2022 - Reviews, Features, Pricing, Comparison - PAT RESEARCH: B2B Reviews, Buying Guides & Best Practices (predictiveanalyticstoday.com) (дата звернення: 13.02.22).
20. Transformer: A Novel Neural Network Architecture for Language Understanding. URL: Google AI Blog: Transformer: A Novel Neural Network Architecture for Language Understanding (googleblog.com) (дата звернення: 13.02.22).
21. Twinword. URL: Text Analysis APIs that analyze and understand natural human text - Twinword API (дата звернення: 13.02.22).
22. Visual Text. URL: VisualText – Text Analyzer Builder (дата звернення: 13.02.22).
23. Аналіз тональності тексту. URL: https://uk.wikipedia.org/wiki/%D0%90%D0%BD%D0%B0%D0%BB%D1%96%D0%B7_%D1%82%D0%BE%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D1%81%D1%82%D1%96_%D1%82%D0%B5%D0%BA%D1%81%D1%82%D1%83 (дата звернення: 13.02.22).

24. ВікіпедіЯ. Вільна енциклопедія. URL: https://uk.wikipedia.org/wiki/Головна_сторінка (дата звернення: 13.02.22).
25. Грицай І.С. Мовні особливості сучасних засобів масової інформації. Развитие гуманитарных наук. Проблемы и перспективы. Подсекция 5. Языковедение и иностранные языки. – с. 38-40. Режим доступу до електронної версії статті (URL): konf 9_5_9.pdf (xn--e1aajfpcds8ay4h.com.ua) (дата звернення: 13.02.22).
26. Культ мови. Цирк на дроті, або Як ви пишете слово “онлайн”? URL: https://disted.edu.vn.ua/courses/learn/6883 (дата звернення: 13.02.22).
27. Міста України (за населенням). URL: https://uk.wikipedia.org/wiki/Міста_України_(за_населенням) (дата звернення: 13.02.22).
28. Новогодний датасет 2019: открытый тональный словарь русского языка. URL: https://habr.com/ru/post/482052/ (дата звернення: 13.02.22).
29. Онлайн-версія “Словника української мови” в 11 томах. URL: http://sum.in.ua/ (дата звернення: 13.02.22).
30. Поділля. URL: https://uk.wikipedia.org/wiki/Поділля (дата звернення: 13.02.22).
31. Пономарів О. Блог проф. Пономарева: україномовний чи українськомовний? URL: https://www.bbc.com/ukrainian/blog-olexandr-ponomariv-40718117 (дата звернення: 13.02.22).
32. Романюк А., Романишин М. Тональний словник української мови на основі сентимент-анотованого корпусу. Київ: Укр. мовознавство, 2013. – № 43. – с. 63-74. Режим доступу до електронної версії статті (URL): Романюк А. - Тональний словник української мови на основі сентимент-анотованого корпусу, Романишин М. (2013) (irbis-nbuv.gov.ua) (дата звернення: 13.02.22).
33. Словник. URL: https://uk.wikipedia.org/wiki/Словник (дата звернення: 13.02.22).
34. Словник. Види та типи словників. URL: https://kievpereklad.com.ua/ua/slovnik-vidi-ta-tipi-slovnikiv/ (дата звернення: 13.02.22).
35. СЛОВНИК.ua. URL: https://slovnyk.ua/ (дата звернення: 13.02.22).

36. Столиці країн Азії. URL: http://merkator.org.ua/dovidnyk/stolyci-krajin-aziji/ (дата звернення: 13.02.22).
37. Столиці країн Європи. URL: http://merkator.org.ua/dovidnyk/stolyci-krajin-jevropy/ (дата звернення: 13.02.22).
38. Столиці країн Південної Америки. URL: http://merkator.org.ua/dovidnyk/stolyci-krajin-pivdennoji-ameryky/ (дата звернення: 13.02.22).
39. Столиці країн Північної Америки. URL: http://merkator.org.ua/dovidnyk/stolyci-krajin-pivnichnoji-ameryky/ (дата звернення: 13.02.22).
40. Тональний словник української мови. URL: https://github.com/lang-uk/tone-dict-uk/blob/master/tone-dict-uk.tsv (дата звернення: 13.02.22).
41. Український правопис 2019. URL: https://mon.gov.ua/ua/osvita/zagalna-serednya-osvita/navchalni-programi/ukrayinskij-pravopis-2019 (дата звернення: 13.02.22).
42. What is Web Scraping and What is it Used For? URL: https://www.google.com/amp/s/www.parsehub.com/blog/what-is-web-scraping/amp/ (дата звернення: 15.05.23).
43. Кочерган М. П. Вступ до мовознавства: підручник. К. : Академія, 2008. 215 с.
44. Алексеев П. М. Частотные словари : учеб. Пособие. СПб. : Изд-во С.– Петерб. ун-та, 2001. 156 с.
45. Бук. С. Слов'янський досвід укладання частотних словників мови письменника. Проблеми слов'янознавства, 2011. – №60. – с. 217-224. Режим доступу до електронної версії статті (URL): Бук С. - Слов'янський досвід укладання частотних словників мови письменника (2011) (irbis-nbuv.gov.ua) (дата звернення: 19.05.23).

46. Бабак Б. Використання “частотного словника сучасної арабської мови” у викладанні арабської мови. Вісник Львівського університету. Серія філологічна : Збірник наукових праць, 2008. – №45. – с.90-97.
47. Перебийніс В. І., Сорокін В. М. Традиційна та комп’ютерна лексикографія. К. : Київський національний лінгвістичний університет, 2009. 218 с.
48. Дарчук Н.П., Грязнухіна Т.О. Частотний словник сучасної української публіцистики // Мовознавство, 1996. – №4–5. – с.15–19.
49. Перебийніс В.С. Частотний словник // Енциклопедія “Українська мова”. К., 2000. – с.724-725.
50. Карпіловська Є.А. Вступ до прикладної лінгвістики: комп’ютерна лінгвістика: підручник. Донецьк : ТОВ “Юго-Восток, Лтд”, 2006. – 188 с.
51. Перебийніс В.С., Євдокімова Т. О., Клименко Н. Ф., Комарова Л. І. та ін. Частотний словник сучасної української художньої прози в 2-х томах. Київ : “Наукова думка”, 1981. – 1716 с.
52. Обернений частотний словник сучасної української художньої прози. – К., 1998.
53. Розтяжка, град, тривога: дизайнер показав, як війна змінила значення простих слів. URL: Слова та фрази, які змінила війна в Україні - фото нової реальності (apostrophe.ua) (дата звернення: 18.11.22).
54. Абетка повномасштабної війни: які слова ввійшли в ужиток і чому. URL: Абетка повномасштабної війни: які слова ввійшли в ужиток і чому (chytomo.com) (дата звернення: 17.05.23).
55. The 10 Best Sentiment Analysis Tools in 2023. URL: https://www.bytesview.com/blog/10-best-sentiment-analysis-tools-2023/ (дата звернення: 10.06.23).
56. Top 10 AI Sentiment Analysis Tools You Should Know in 2023. URL: https://www.analyticsinsight.net/top-10-ai-sentiment-analysis-tools-you-should-know-in-2023/ (дата звернення: 10.06.23).
57. BytesView. URL: Sentiment Analysis - BytesView (дата звернення: 10.06.23).

58. MonkeyLearn. URL: https://monkeylearn.com/sentiment-analysis-online/ (дата звернення: 10.06.23).
59. Qualaroo. URL: https://qualaroo.com/features/watson/ (дата звернення: 10.06.23).
60. NICE Interaction Analytics. URL: https://www.nice.com/products/cx-analytics/interaction-analytics (дата звернення: 10.06.23).
61. Aylien. URL: https://aylien.com/product/news-api (дата звернення: 10.06.23).
62. OpenText. URL: https://www.opentext.com/ (дата звернення: 10.06.23).
63. Brand24. URL: https://brand24.com/ (дата звернення: 10.06.23).
64. ParallelDots. URL: https://www.paralleldots.com/ (дата звернення: 10.06.23).
65. Hi-Tech BPO. URL: https://www.hitechbpo.com/ (дата звернення: 10.06.23).
66. Social Mention. URL: https://www.socialmention.com/ (дата звернення: 10.06.23).
67. Social Searcher. URL: https://www.social-searcher.com/ (дата звернення: 10.06.23).
68. Sentiment Analyzer. URL: https://www.danielsoper.com/sentimentanalysis/default.aspx (дата звернення: 10.06.23).
69. Tweet Sentiment Visualization. URL: https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/?q=bcash (дата звернення: 10.06.23).
70. Ситник І.В. Нове та традиційне у дослідженнях сучасних представників філологічних наук. Частотний словник, його функції та укладання : матеріали Міжнародної науково-практичної конференції, м. Одеса, 26-27 лютого 2021 року, Одеса, 2021. с.89-92. Режим доступу до електронної версії статті (URL): I_Sytnyk_POTSPHD_conf_odesa_KUBG.pdf (дата звернення: 19.05.23).
71. Лінгвістичний портал mova.info. URL: Лінгвістичний портал MOVA.info (дата звернення: 08.06.23).

СПИСОК ВИКОРИСТАНИХ ПРОГРАМНИХ БІБЛІОТЕК

72. Microsoft Excel 2016. URL: https://www.microsoft.com/uk-ua/microsoft-365/excel (дата звернення: 19.05.23).

73. Google Таблиці. URL: Google Таблиці: онлайн-редактор електронних таблиць | Google Workspace (дата звернення: 19.05.23).

74. Welcome to Python.org. URL: https://www.python.org/ (дата звернення: 20.05.23).

75. PyCharm Edu. URL: Install PyCharm - Help | PyCharm (jetbrains.com) (дата звернення: 20.05.23).

76. Бібліотека Beautiful Soup 4 в Python. URL: beautifulsoup4 · PyPI (дата звернення: 18.05.23).

77. Список стоп-слів для української мови. URL: https://github.com/skupriienko/Ukrainian-Stopwords (дата звернення: 20.05.23).

ДОДАТКИ

Додаток А. Тональний словник №1.

У цьому додатку подано список 40 567 слововживань, з якими надалі проводилася робота.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1TUrrg-oxG10pKuZS_wr1yH55-YqarxxI?usp=sharing.

Додаток Б. Тональний словник, який відредаговано на основі додатку А.

У цьому додатку подано відредагований список 40 567 слововживань, який зменшився в обсягах та набув вигляду тонального словника.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/11JX4t6Q7wmsrKSGSwkaLGHJhXqpkWAJY?usp=sharing.

Додаток В. Тональний словник №2 та 14 списків слів, з яких сформовано тональний словник №2.

У цьому додатку подано кілька файлів:

1) теку, в якій подано 14 окремих файлів, в яких представлено списки слів. Кожен зі списків містить близько 500 слів;

2) тональний словник №2, який сформовано та відредаговано на основі поданих списків слів.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1ixa1hnJRvmcRDK3f9IsW-gYqKPZm98_D?usp=sharing.

Додаток Г. Список слів, які хочеться включити до основного масиву слів тонального словника.

У цьому додатку подано списки найчастотніших слів, які було систематизовано за темами:

1) список найбільш вживаних абревіатур;

2) список днів тижня;

3) список місяців року;

4) список пір року, куди, окрім слів на позначення пір року, було включено такі прислівники як “взимку”, “влітку”, “восени”, “навесні”, “узимку” та “улітку”;

5) список планет Сонячної системи, окрім планети Земля, оскільки це слово вже було подано у ТС №1;

6) список, куди включено найбільші міста України, де синім кольором зафарбовано ті комірки, які містять назви міст, що є обласними центрами України (таких міст 25, включаючи Сімферополь та Київ) [27];

7) від списку, який подано під номером 6, було частково утворено прикметники в називному відмінку однини чоловічого роду. Утворені прикметники й сформували сьомий список слів;

8) список слів на позначення територій, що охоплюють межі адміністративних центрів (наприклад, Чернігівщина) та історико-географічні області України (наприклад, Буковина).

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1nyKZGB-kjns09dbkcpWOQz2NRF1vCvQ0?usp=sharing.

Додаток Д. Список слів на позначення основних країн та столиць світу.

У цьому додатку подано перелік країн та столиць світу.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1QRx-QIlduqf-mt9nG6j9br2HOd9AJwrP?usp=sharing.

Додаток Е. Упорядкований список слів від додатку Д.

У цьому додатку подано відредагований список країн та столиць світу, які об’єднано в один список та відсортовано за алфавітом.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/11IAAqjo0aS49TeC9XB1gFBBLoHlvvEBB?usp=sharing.

Додаток Є. Власний міні-словник слів.

У цьому додатку подано словник слів, який, окрім власне введених слів, також надає матеріали з додатків Г та Е.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1KQUSSbzyaC7FXTNsC5nU4qi3OL7gEY37?usp=sharing.

Додаток Ж. Оцінка тональності окремих дієслів української мови.

У цьому додатку подано кілька файлів:

1) Google форма, яку заповнювали респонденти щодо тональності 11 пар дієслів за граматичним значенням “доконаний та недоконаний вид”. Опитування зроблено, щоб з’ясувати, чи є потреба вводити дієслівну пару за граматичним значенням “доконаний та недоконаний вид” у випадку її відсутності;

2) таблиця (“Додаток Ж. Діаграми”) отриманих результатів, яка сформована автоматично на основі наданих відповідей респондентів;

3) текстовий файл (“Додаток Ж. Результати”), в якому здійснено аналіз отриманих результатів та представлено отримані дані у вигляді діаграм та таблиць для кращого візуального сприйняття інформації.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1qsiHS88p4bYALSt4eQhH7Z7UI1dCcTJh?usp=sharing.

Додаток З. Код пошуку збігів у двох файлах.

У цьому додатку подано кілька файлів:

1) тека, де надано доступ до двох списків слів, які були взяті для роботи створеної міні-програми;

2) текстовий файл, де розміщено код, який створений для того, щоб зіставляти два csv-файла з метою пошуку спільних та відмінних слів.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1JADOvCFPtsOrpESbXDEK-CIk-81B5AuV?usp=sharing.

Додаток І. TEST.

У цьому додатку подано таблицю, яка ілюструє правильність виконання створеного нами коду.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1mdurD1My06um0GEtw6qfjGbJAUoS0LMf?usp=sharing.

Додаток Ї. Аналіз 13 програмних забезпечень.

У цьому додатку подано таблицю, яка демонструє здійснений нами аналіз 13 найкращих програмних забезпечень (ПЗ) станом на початок 2022 року. Аналіз ПЗ здійснювався за певними критеріями, ознайомитися з якими можна у файлі “README. Додаток Ї”, який також подано у цьому додатку.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/11O8kjeINa1vyRcIIA9diBrjE3Rd8dQjJ?usp=sharing.

Додаток Й. Власний тональний словник української мови.

У цьому додатку подано власний тональний словник як фінальний результат нашої роботи з усіма необхідними супровідними файлами, що подають деталі створення та упорядкування словника.

Ознайомитися з виглядом тонального словника української мови можна за цим посиланням: https://drive.google.com/drive/folders/1tHbTTJbZWUtVcsTmCzNABmIHCNUT27ap?usp=sharing.

Додаток К. Код вебскрапінгу (вебскрейпінгу) новин із сайту “tsn.ua”.

У цьому додатку подано код програми (у pdf-файлі), яка автоматично видобуває необхідну кількість текстів новин із сайту “tsn.ua” за обраний день.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1ZyaiHZ3oksnlHLvCdktB2THns-BkoReF?usp=share_link.

Додаток Л. Код вебскрапінгу (вебскрейпінгу) новин із сайту “hromadske.ua”.

У цьому додатку подано код програми (у pdf-файлі), яка автоматично видобуває необхідну кількість текстів новин із сайту “hromadske.ua” з урахуванням того, що першою новиною є та, яка опублікована найпізніше.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1KlKTQvxQgHox5fkr1GEsm_86vawtujiX?usp=share_link.

Додаток М. Код АОТ новин із сайту “tsn.ua” чи “hromadske.ua” (на вибір).

У цьому додатку подано код програми (у pdf-файлі), яка автоматично опрацьовує видобутий текст із сайту “tsn.ua” чи “hromadske.ua”. Для вхідного тексту здійснюється такий набір процедур: токенізація, вилучення стоп-слів,

лематизація, автоматична побудова частотного словника лексем зі спадом частоти та представлення числового значення, що стосується частки покриття тексту словами із тонального словника.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/15BtjclnhTe5idAdVxx7hita0uq4qLMc6?usp=share_link.

Додаток Н. Код, який автоматично перетворює словник у форматі csv в таблицю і зберігає створену таблицю у текстовому файлі.

У цьому додатку подано два коди програми (у двох pdf-файлах), які працюють за схожою схемою: автоматичне перетворення даних словника у форматі csv в таблицю і збереження створеної таблиці у текстовому файлі.

Потреба у другому варіанті коду стосується доповнення новою інформацією тонального словника, яка мала свої особливості.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1tT5UaDU3Y_9zJvbMP9wlaYz6MjOU1xIQ?usp=share_link.

Додаток О. Новий список 2023.

У цьому додатку подано список слів та словосполучень, який було укладено вручну шляхом залучення Інтернет-ресурсів [54]. Цей список включає в себе слова, що увійшли до активного вжитку у медійному стилі за останні півтора року.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1zWomfJa4vzwbj2CJDEfD7bVYv01G_yFB?usp=share_link.

Додаток П. Код зіставлення-порівняння.

У цьому додатку подано код програми (у pdf-файлі), який бере на вхід два файли у csv форматі, а на виході ми отримуємо два списки слів: один список надає

інформацію щодо слів, які присутні в обох вхідних файлах (збіги), інший – список слів, які варто включити до масиву тонального словника.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1n8jtvfglp3_rlUJaRSFBDVt_9wGhCBTd?usp=share_link.

Додаток Р. Зміна тональності окремих слів української мови.

У цьому додатку подано кілька файлів:

1) Google форма, яку заповнювали респонденти щодо тональності 6 пар слів, щоб з'ясувати, чи є потреба вводити ці та подібні слова з урахуванням нового семантичного значення, якого набули ці слова;

2) таблиця ("Додаток Р. Діаграми") отриманих результатів, яка сформована автоматично на основі наданих відповідей респондентів;

3) текстовий файл ("Додаток Р. Результати"), в якому здійснено аналіз отриманих результатів та представлено отримані дані у вигляді діаграм та таблиць для кращого візуального сприйняття інформації.

Ознайомитися з додатком можна за цим посиланням: https://drive.google.com/drive/folders/1GjGWscMemPcfn6pRx4kpRPAoFFDa-Us6?usp=share_link.

Всі додатки було поміщено в папку. За цим посиланням https://drive.google.com/drive/folders/1tHbTTJbZWUtVcsTmCzNABmIHCNUT27ap?usp=sharing ви можете отримати доступ до необхідних додатків, щоб детальніше ознайомитися з ними.