Магістерська кваліфікаційна робота на тему: «Інформаційна технологія аналізу англомовного тексту на наявність сталих мовних конструкцій»

Виконала ст. гр. 1КН-18м Миколюк І. О.

Науковий керівник: проф. Месюра В. І.

Мета та завдання дослідження

Метою дослідження магістерської кваліфікаційної роботи є підвищення швидкості аналізу текстової інформації.

Для досягнення наведеної мети були поставлені та вирішені наступні задачі:

Розглянути та проаналізувати методи та технології розв'язання задачі аналізу тексту;

- дослідити перелік необхідних функцій, які повинна містити інформаційна технологія;
- запропонувати математичну модель інформаційної технології аналізу англомовного тексту на наявність сталих мовних конструкцій;
- виконати програмну реалізацію запропонованої інформаційної технології;
- провести тестування програмного продукту та виконати аналіз отриманих результатів.

Об’єкт, предмет та методи дослідження

- **Об’єкт дослідження -** процес аналізу англомовного тексту на наявність сталих мовних конструкцій.

- **Предмет дослідження -** методи аналізу тексту.

- **Методи дослідження**. У роботі використані наступні методи наукових досліджень: системного аналізу для аналізу структури інформаційної системи, метод аналізу логіко-лінгвістичних моделей речень природної мови, метод пошуку текстових збігів у реченнях природної мови довільної складності, алгоритм Байєсівського класифікатора та алгоритм підрахунку TF-індексу для реалізації модуля аналізу англомовного тексту на наявність сталих мовних конструкцій, об’єктно-

Наукова новизна

- вперше запропоновано інформаційну технологію аналізу англомовного тексту на наявність сталих мовних конструкцій, яку засновано на сумісному використанні методу Байєсівського класифікатора та алгоритм підрахунку TF-індексу в поєднанні з алгоритмами логіко-лінгвістичного моделювання для аналізу текстового документу, що забезпечмло підвищення швидкості аналізу англомовного тексту.

- вдосконалено модель аналізу англомовного тексту на наявність сталих мовних конструкцій шляхом симбіозу механізмів розробки новітніх інформаційних технологій та точних методів комп'ютерної лінгвістики, що забезпечує підвищення швидкості отримання вихідних даних.

Актуальність задачі

- Необхідність володіння англійською мовою на сьогодні вважається великою, адже найбільше друкованої продукції видають саме англійською мовою. Англійська мова – це офіційна мова міжнародного бізнесу та торгівлі, Інтернету і техніки, науки і мистецтв. 80% ділового мовного простору займає саме вона.
- Усі мови світу часто використовують ідіоматичні вирази, більшість з яких мають соціокультурне, історичне чи політичне походження. Хоча багато подібних виразів можна знайти в різних мовах, багато інших не збігаються точно за своїм мовним чи семантичним значенням та вживанням. У той же час ідіоми часто є каменем спотикання для студентів другої / іноземної мови та студентів із загальноосвітніх шкіл.
- Аналіз контенту зараз використовується у багатьох сферах, починаючи від маркетингових та медіа-досліджень до літератури та риторики, етнографії та культурології, гендерних та вікових питань, соціології та політології, психології та когнітивної науки та багатьох інших галузей дослідження.

Аналіз аналогів

Програма	Призначення	Статус	Основні характеристики	Можливості та якість роботи
Textanz	Аналізує текст, надає список або словник слів, фраз і граматичних форм	Безкоштовна	Дозволяє перевірити надмірне використання або повторення слів та фраз в будь-якому документі, є можливість редагування тексту	Працює з документами формату RTF, MS Office, Open Office, HTML, XML, PDF і Дозволяє одночасний аналіз декількох документів
TextQuest	Аналізує текст	Платна але надається демо-версія	Безкоштовна версія має обмеження в 100 одиниць тексту. Надає можли-вість визначати частоту.	Використовує таблиці сортування слів, в залежності від довжини рядка символів.
HAMLET	Пошук у текстових файлах слів та розрахунок спільни х частот	Платна але надається демо-версія	Використовується як для вимірювання емпіричних властивостей тексту, так і для візуалізації отриманих даних	Дозволяє порівняти результати загальних частот, раніше отриманих з ряду текстів
Kwali-tan	Призначена для аналізу текстів, зображень, аудіо та відео фраг-ментів кодування.	Платна, але надається демоверсія на невизначений період.	Ефективно зберігає дані і має ряд функцій для якісного аналізу матеріалів, таких як кодування, отримання та класифікації кодів.	Більшість кодування здійснюється вручну, але Kwalitan також має інструмент для присвоєння кодів авто- матично.
MAX-QDA	Якісний аналіз даних, систе-матична оцінка та інтерпрета- ція текстів	Платна але надається демо-версія на 30 днів.	Зручна в користуванні, є потужним інструментом для розвитку теорій та їх перевірки	Має великі і диференцтійовані функції з хорошою візуалі-зацією, доступні для процесу кодування.

Постановка задачі дослідження

Для вирішення проблеми аналізу англомовного тексту на наявність сталих мовних конструкцій необхідно вирішити такі основні завдання:

- Обґрунтувати вибір методу аналізу тексту.
- Розробити математичну модель.
- Спроектувати структуру інформаційної технології.
- Розробити базу даних англомовних сталих конструкцій.
- Програмно реалізувати інформаційну технологію.
- Провести тестування інформаційної технології.

Математична модель

Робота алгоритму починається з пошуку конструкцій в тексті та обрахунку їх частотності

$$TF(W_i, Doc_k) = \frac{n_i}{\sum_k n_k}$$

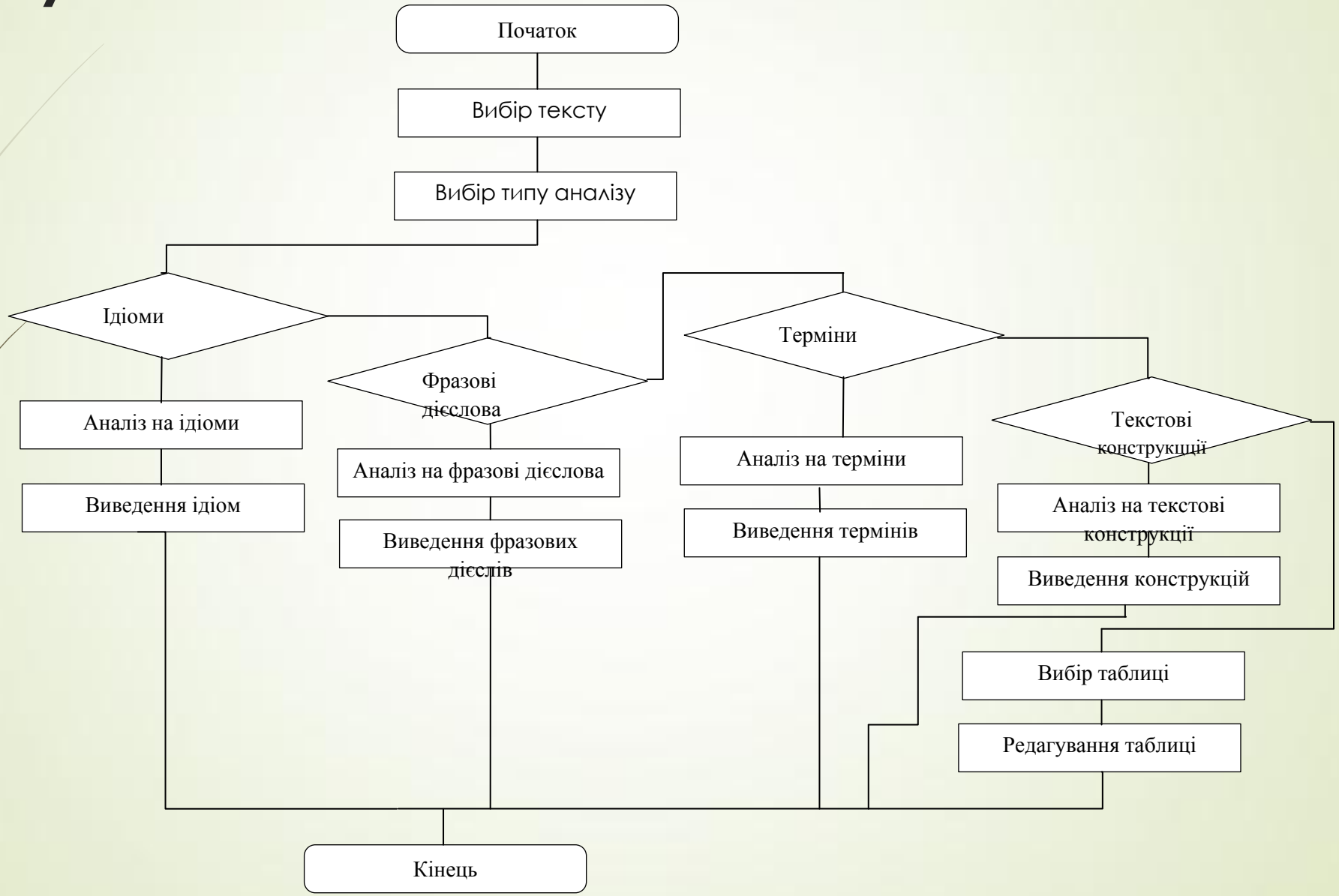
Після цього необхідно віднести текст до певної категорії. Доцільно використати алгоритм класифікації на основі методу Байєса

$$P(c|d) \approx P(c) \prod P(t_k|c).$$

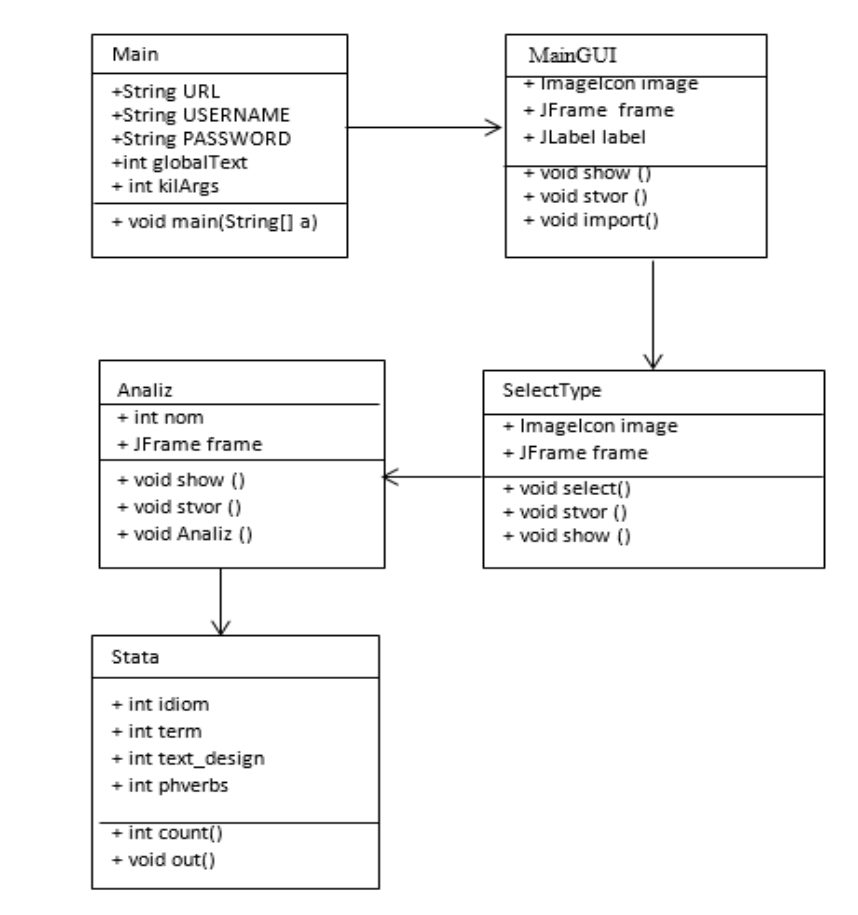
Оцінка ймовірності, по якій проводиться класифікація

$$c_{map} = \arg\max_{c \in C} \left[log \frac{Doc_c}{Doc} + \sum_{i=1}^{n} log \frac{W_{ic} + 1}{|V| + \sum_{i' \in V} W_{i'c}} \right].$$

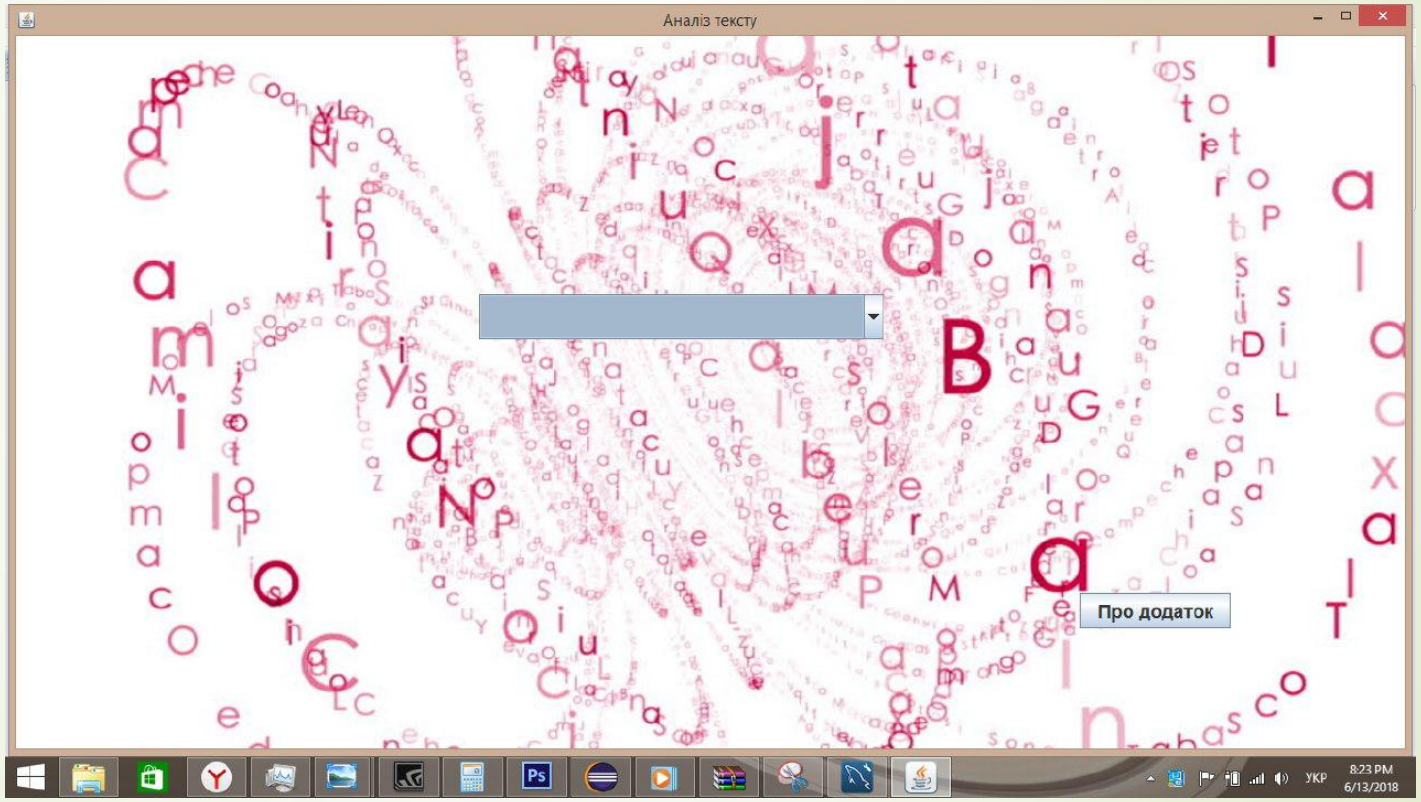
Схема загального алгоритму функціонування інтелектуальної технології

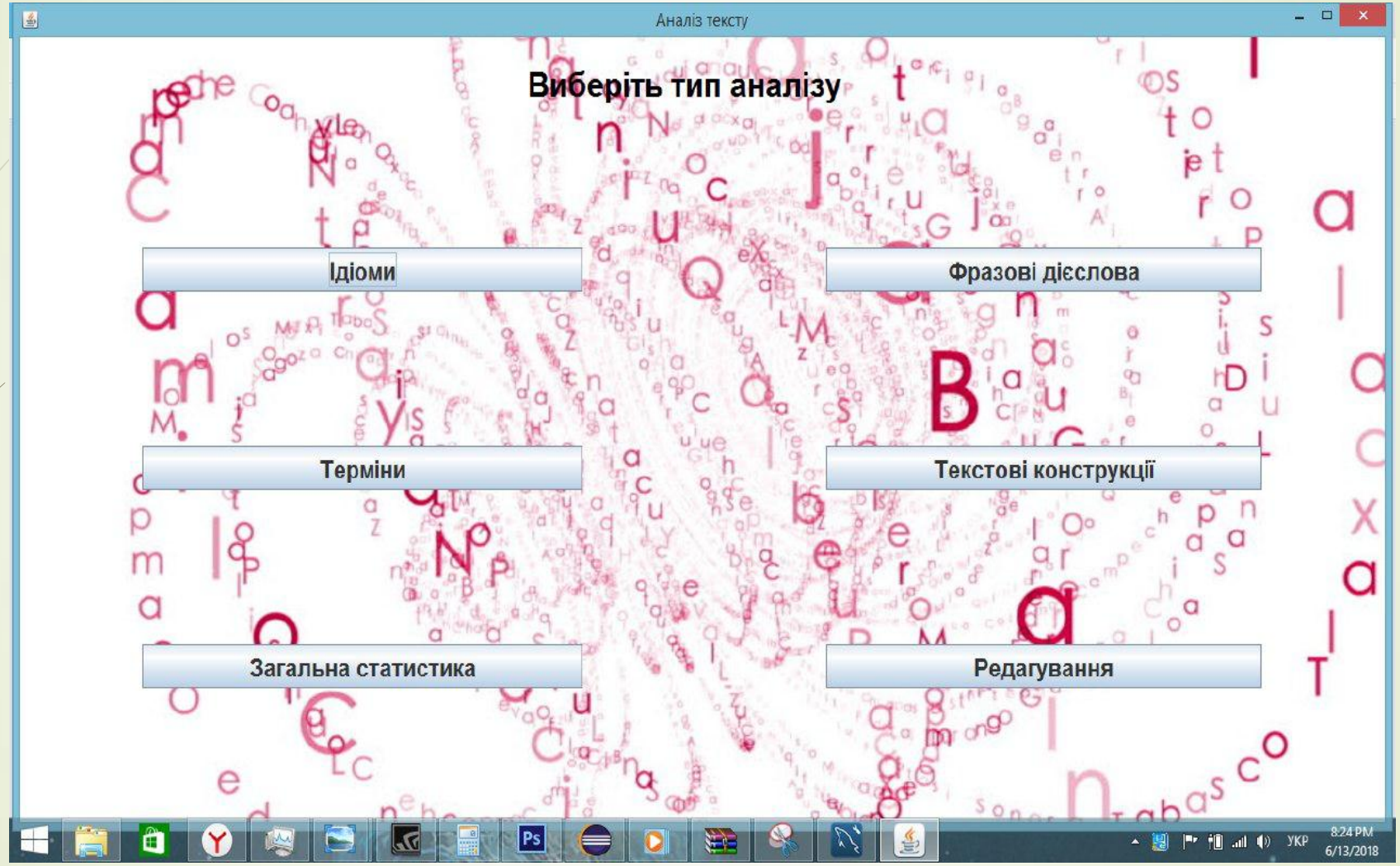


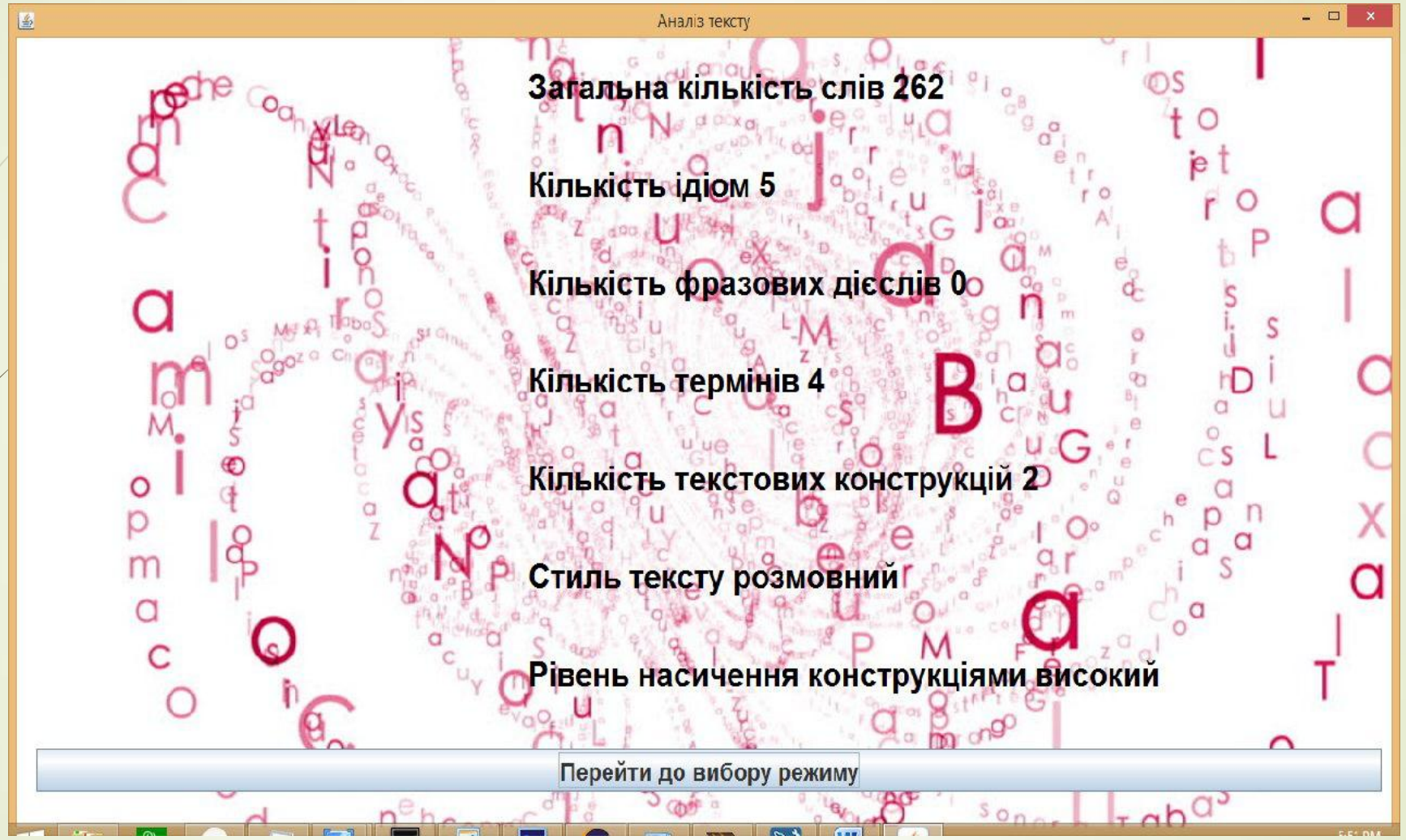
UML-діаграма класів

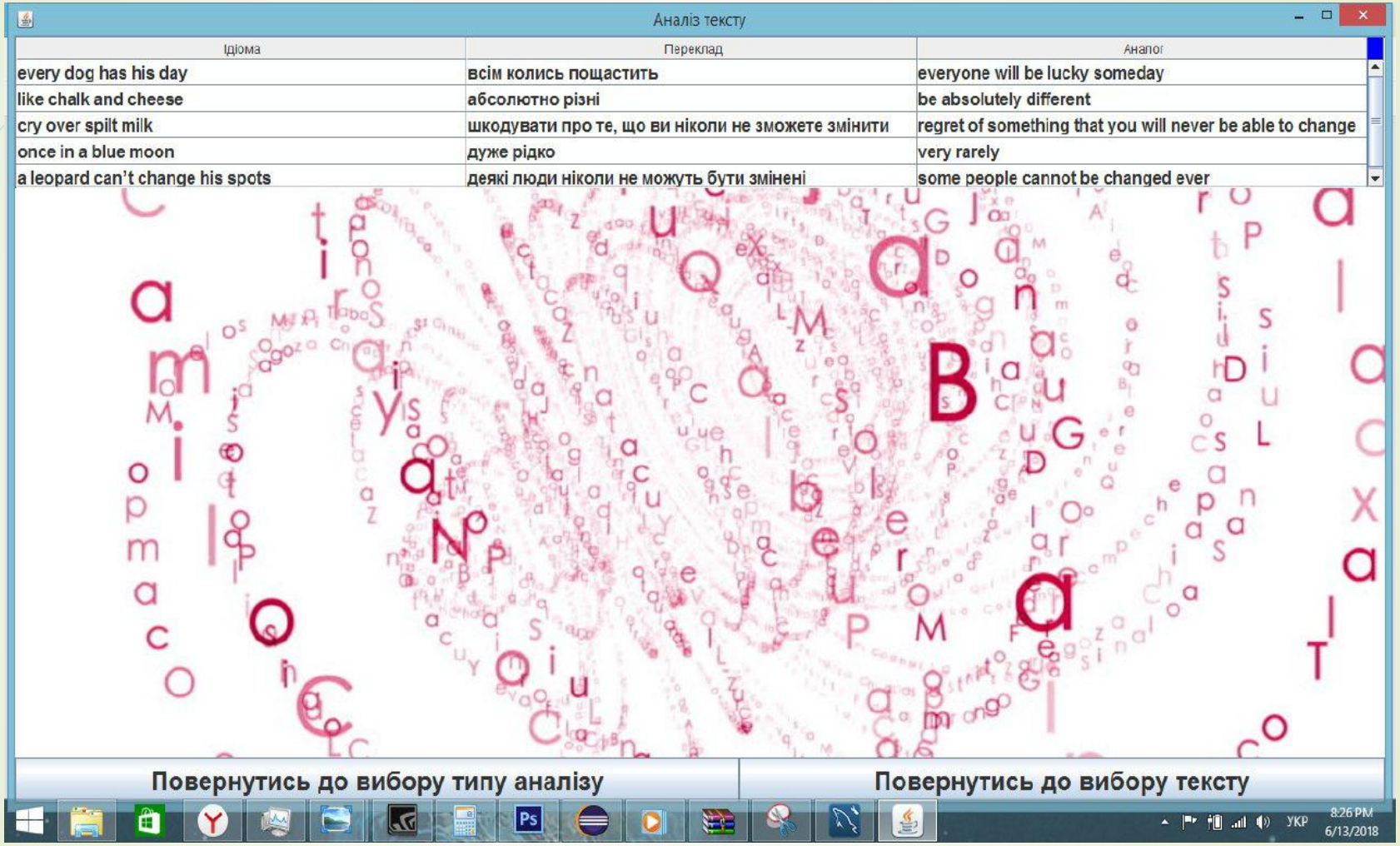


Початкова активність




Аналіз тексту
Виберіть тип аналізу
Ідіоми
Фразові дієслова
Терміни
Текстові конструкції
Загальна статистика
Редагування
УКР
8:24 PM
6/13/2018


Аналіз тексту
Загальна кількість слів 262
Кількість ідіом 5
Кількість фразових дієслів 0
Кількість термінів 4
Кількість текстових конструкцій 2
Стиль тексту розмовний
Рівень насичення конструкціями високий
Перейти до вибору режиму


Аналіз тексту
Ідіома
Переклад
Аналог
every dog has his day
всім колись пощастить
everyone will be lucky someday
like chalk and cheese
абсолютно різні
be absolutely different
cry over spilt milk
шкодувати про те, що ви ніколи не зможете змінити
regret of something that you will never be able to change
once in a blue moon
дуже рідко
very rarely
a leopard can't change his spots
деякі люди ніколи не можуть бути змінені
some people cannot be changed ever
Повернутись до вибору типу аналізу
Повернутись до вибору тексту
УКР
8:26 PM
6/13/2018

Порівняльний аналіз тестування роботи програм

Програма	Швидкість аналізу, с
Розроблений додаток	0.4163
HAMLET	0.4304
Textanz	0.4511
TextQuest	0.4857
Yoshikoder	0.4742

Економічна частина

На основі зроблених підрахунків в економічній частині магістерської кваліфікаційної роботи досягнуті наступні результати:

– визначено, що рівень комерційного потенціалу розробки є високим.

– витрати на розробку та її впровадження складають 32441,4 грн.;

– абсолютний ефект від впровадження результатів нашої розробки протягом 3-х років складе 113000 грн.

– вартість інвестицій, що можуть бути вкладені в нашу розробку становить тис. грн;

– термін окупності системи, що розробляється складає 1,11 року, що вписується в задані часові рамки та є показником доцільності розробки.

Апробація результатів роботи та публікації

Опубліковані тези доповіді на всеукраїнській науково-практичній інтернет-конференції студентів, аспірантів та молодих науковців «Молодь в науці: дослідження, проблеми, перспективи (МН 2019)» (м. Вінниця, Україна, 2019 р.) та XLVII Науково-технічній конференції факультету інформаційних технологій та комп'ютерної інженерії (2018) .

Подано заяву про реєстрацію авторського права на твір.

Висновки

Досліджено предметну область аналізу англомовного тексту на наявність сталих мовних конструкцій, що показало актуальність проблеми, яка полягає у потребі простого, багатофункціонального інструменту для обробки тексту, який забезпечить зручний та ефективний аналіз.

- У результаті дослідження методів та інформаційних технологій для аналізу тексту обрано метод Байєса, який у порівнянні з широко розповсюдженими методами, як дерев рішень, k-найближчих сусідів, забезпечує високу швидкість роботи, підтримку поступовогно навчання та відносно просту програмну реалізацію алгоритму, що робить доцільним його використання як основа аналізу.
- Обґрунтовано використання даних методів для розв'язання поставленої задачі та методів фільтрації, в наслідок чого запропоновано математичну модель для даної інформаційної технології.
- Проведено тестування розробленої інформаційної технології,що підтвердило збільшення швидкості роботи майже на 3,5%.
- Обрахувавши термін окупності даної наукової розробки визначили, що фінансування даної наукової розробки буде доцільним.

Дякую за увагу