**Міністерство освіти і науки України**

**Харківський національний університет радіоелектроніки**

Факультет комп’ютерних наук

Кафедра програмної інженерії

**АТЕСТАЦІЙНА РОБОТА**

пояснювальна записка

магістра

**Дослідження методів лексичного аналізу для оцінки складності алгоритмів в системі «Algorithms battle»**

Виконав: студент 6 курсу, групи ПЗСм-18-1

напряму підготовки (спеціальності) 8.05010301 ???

Програмне забезпечення систем

Різник О. К.

Керівник доц. Мазурова О.О.

**Допускається до захисту**

Зав. кафедри, проф. \_\_\_\_\_ Дудар З.В.

2017 р.

*Харківський національний університет радіоелектроніки*

|  |  |
| --- | --- |
| Факультет *комп’ютерних наук* | Кафедра *програмної інженерії* |
| *Спеціальність* 6.05010301  *Програмне забезпечення систем* | |
|  |  |

**ЗАВДАННЯ**

**НА АТЕСТАЦІЙНУ РОБОТУ МАГІСТРА СТУДЕНТОВІ**

*Різнику Олександру Костянтиновичу*

1. Тема проекту «*Дослідження методів лексичного аналізу для оцінки складності алгоритмів в системі «Algorithms battle»*, затверджена наказом № \_\_\_\_\_\_\_ від \_\_\_\_\_\_\_\_\_\_\_
2. Термін здазчі студентом закінченого проекту \_\_\_\_\_\_\_
3. Вихідні дані до проекту: *електронні ресурси за обраною тематикою, мінімальні вимоги до функціональності програми, загальні вимоги до архітектури системи.*
4. Зміст пояснювальної записки (перелік питань, які потрібно розробити):

*вступ, аналіз проблемної області та постановка задачі, перелік вимог до програмної системи, опис прийнятих проектних рішень, опис програмної реалізації, аналіз можливих застосувань та тестування*

1. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень): *скріншоти аналогів, схема взаємозв’язку об’єктів, UML-діаграма прецендентів, схема основних об’єктів бази даних, блок-схеми алгоритмів, схема бази даних,скріншоти інтерфейсу програмної системи, слайди презентації.*

|  |  |  |  |
| --- | --- | --- | --- |
| Розділ | Консультант | Підпис, дата | |
| Завдання видав | Завдання прийняв |
| Спецчастина | доц. Мазурова О.О. |  |  |

Календарний план

|  |  |  |  |
| --- | --- | --- | --- |
| № | Назва етапів дипломного проекту (роботи) | Термін виконання етапів проекту (роботи) | Примітка |
| 1 | Аналіз проблемної області та постановка задачи | \_\_ |  |
| 2 | Аналіз методів аналізу інформації | \_\_\_ |  |
|  | Розробка математичного описання концептуальної моделі |  |  |
|  | Розробка алгоритму підтримки концептуального моделювання |  |  |
| 3 | Розробка структури зберігання даних | \_\_\_ |  |
| 4 | Створення коду програми | \_\_\_ |  |
| 5 | Тестування і налагодження програми | \_\_\_ |  |
| 6 | Підготовка пояснювальної записки. | \_\_\_ |  |
|  | Написання статті | \_\_\_ |  |
| 7 | Підготовка презентації та доповіді | \_\_\_ |  |
| 8 | Попередній захист | \_\_\_ |  |
| 9 | Нормоконтроль, рецензування | \_\_\_ |  |
| 10 | Занесення диплома в електронний архів | \_\_\_ |  |
| 11 | Допуск до захисту у зав. кафедри | \_\_\_ |  |

Дата видачі завдання 01 вересня 2016 р.

Керівник доц. \_\_\_\_\_\_\_\_\_\_\_\_\_\_ Мазурова О. О.

Завдання прийняв до виконання\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Різник О. К.

РЕФЕРАТ/ABSTRACT

Пояснювальна записка: 77 с., 18 рис., 2 таблиці, 20 джерел, 6 додатків.

Метою роботи є дослідження методів оцінки аналізу складності алгоритмів, реалізація нового методу, створення веб-системи для онлайн змагань з вирішення алгоритмічних задач та використання нового методу оцінки складності алгоритму у системі.

Методом вирішення є концептуальне моделювання предметної області, використання об’єкто-орієнтованого підходу до створення програмного продукту. Для розробки було обрано мову програмування Java, фреймворк Spring Framework 5.0 та Spring Boot 2.0, інструмент збірки проектів Maven, черга повідомлень Apache Kafka та середовище розробки Intellij IDEA.

Результатом роботи є веб-система для вирішення алгоритмічних задач, яка оцінює складність написаного алгоритму і дає точну оцінку складності.

АЛГОРИТМИ, ВЕБ-СИСТЕМА, ВИПУСКНА РОБОТА МАГІСТРА, ЛЕКСИЧНИЙ АНАЛІЗ, СКЛАДНІСТЬ АЛГОРИТМА, BIG-O NOTATION, JAVA, MONGODB, KAFKA, SPRING BOOT, SPRING FRAMEWORK

The aim of the work is a research of algorithms complexity analysis methods, implementation of a new one, creating a web-system for online competition in solving algorithmics tasks and using the new method of algorithm complexity analysis in the system.

The solution methods are conceptual modeling of the domain, using an object-oriented approach to software application design. Java language, Spring Framework 5.0 and Spring Boot 2.0, build tool Maven, message queue Apache Kafka and development environment Intellij IDEA were chosen as the tools of development.

The resulting work is an web-application for solving algorithmics tasks which does complexity analysis of written algorithm and gives an accurate complexity estimation.

ALGORITHMS, ALGORITHMS COMPLEXITY, BIG-O NOTATION, DATABASE, FINAL MASTER WORK, JAVA, KAFKA, LEXICAL ANALYSIS, MONGODB, SPRING BOOT, SPRING FRAMEWORK, WEB-SYSTEM

ЗМІСТ

ВСТУП

У сучасному світі понад три мільярди двісті мільйонів людей мають інтернет та користуються різноманітними веб-сервісами. Тому дуже важливо мати можливість оброблювати запити всіх користувачів максимально швидко та ефективно, тому що при величезних кількостях користувачів та їх запитів – навіть найновіші сервери можуть бути недостатньо потужними аби задовільняти потреби клієнтів.

Існують архітектурні підходи, які використовуються для зменшення навантаження на сервера, такі як розширення кількості серверів які рівномірно розподіляють навантаження між собою та оброблюють запити, але наскільки б багато не було серверів, через помилки розробників – всі вони можуть не справлятись з даним навантаженням, якщо наприклад якась функція програми виконується дуже довго через неоптимальність написаного алгоритму, і якщо нею користуються велика кількість клієнтів – це може бути дуже критичним для функціонування сервісу. При правильному використанню пам’яті та часових затрат на виконання функцій – можна зберегти значну кількість ресурсів.

Для того, щоб контролювати витрати ресурсів, необхідно мати можливість точно визначати наскільки та чи інша функція/алгоритм оптимально використовує ресурси.

На сьогодні, основною мірою оптимальності алгоритму є часова складність алгоритму.

Часова складність алгоритму в комп'ютерних науках є обчислювальною складністю алгоритму, яка описує час потрібний для виконання алгоритму. Вона зазвичай визначається шляхом підрахунку кількості елементарних операцій, виконуваних алгоритмом, при цьому вважають, що кожна елементарна операція виконується за фіксовану кількість часу. [1]

Як видно з визначення – для підрахунку складності алгоритму – необхідно порахувати кількість операцій, яка виконується в програмі. Проте всі підрахунки складності на сьогодні базуються не на підрахунку кількості операцій, а на замірі часу виконання алгоритму на різних наборах даних різних розмірів та виведення формули росту функції на основі цих даних. І хоча цей спосіб оцінювання складності алгоритму досить точний, він має декілька недостатків:

* по-перше це необхідність викликати вхідну функцію велику кількість разів – чим більша кількість разів – тим більша точність складності, а при зростанні даних – це може займати багато часу, а також використовує багато ресурсів системи;
* по-друге при паралельному оцінюванні інших алгоритмів (якщо це онлайн система для вирішення алгоритмів, то паралельно велика кількість юзерів можуть відправляти свої рішення), різке зростання кількості алгоритмів для оцінювання можуть зменшити продуктивність системи – тобто збільшити час виконання алгоритму і тоді кореляція даних буде не достатньо точною.

Таким чином була поставлена задача розробити веб-систему в основі якої буде точний алгоритм оцінювання складності алгоритму, яка на основі лексичного аналізу вхідної програми та парсингу її на лексеми та завдякі подальшій обробці зможе підрахувати кількість операцій вхідної програми.

За результатами роботи було створено презентацію, що наведено в додатку А.

# **1 АНАЛІЗ ПРОБЛЕМНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ**

## Аналіз проблемної області оцінювання складності алгоритмів

На даний час усі відомі веб-сервіси, що створені для тренування роз’вязання алгоритмічних задач не мають загальноприйнятого спосібу визначення складності алгоритму.

Існує еталонний спосіб -визначення часу виконання алгоритму на наборах вхідних даних різного розміру, та з отриманої вибірки визначається формула росту часу в залежності від збільшення вхідних даних. Проте, цей спосіб не є загальновживаним оскільки неможливо передбачити чи закінчить програма своє виконання чи ні, а виконувати програму, яка потенційно має дуже велику складність виконання особливо на великих даних – дуже невигідно.

Цей спосіб є досить точним, проте його точність прямопропорційно залежить від кількості виконання програми на різних (зростаючих) наборах даних. Оскільки при паралельному навантаженні дані можуть бути недостатньо точними, а також якщо взяти до уваги те, що для досягення точного результату потрібно велику кількість разів виконувати алгоритм з великою вибіркою даних, що потребує значних затрат ресурсів – то є очевидним, що є потреба у ресурсоекономному та точному алгоритмі визначення складності алгоритму.

Головною проблемою в реалізації такого алгоритму є неможливість створення універсального алгоритму, який для довільної задачі зміг би порахувати кількість її операцій. Це твердження випливає з доведеної в 1936 році Аланом Тьюрінгом проблеми зупинки [1]. Ця проблема полягає в тому, що не може існувати загального алгоритму для розв’язання проблеми зупинки для всіх пар програма-вхідні дані. Неформально за задача звучить так: дано алгоритм та вхідні дані. Потрібно визначити, чи завершиться виконання програми у будь-який момент часу чи вона буде працювати безкінечно.

Не зважаючи на те, що універсальний алгоритм реалізувати неможливо – можна створити алгоритм, який зможе визначати складність алгоритм для визначеного набору задач (та доповнятися з часом, додаючи різні складні структури для аналізу). Корисність цього алгоритму буде полягати не в повній заміні універсального алгоритму, що працює на основі визначення часу виконання програми, а в попередньому аналізу вхідної програми оптимальним алгоритмом – і у випадку, якщо ця аналіз цієї програми не буде можливим – то перехід до універсального алгоритму. На великих вхідних даних, якщо алгоритм буде аналізувати навіть невеликий відсоток усіх алгоритмів – економія ресурсів буде значною, а точність буде більшою.

Задачею удосконаленого алгоритму (далі лексичного алгоритму), буде аналіз структур вхідної програми, розбивання її на найпростіші структури та аналіз кожної з структури на кількість виконуючихся операцій. В залежності від співвідношення розміру вхідних даних до кількості виконаних операцій програмі буде присвоєна її складність:



Рисунок 1.1 – Графік найпоширеніших функцій в аналізі алгоритмів, де N – кількість операцій, n – розмір вхідних даних для кожної з функцій.

Як видно з графіку – деякі функції мають дуже схожий графік росту на невеликах наборах даних, наприклад – [2] та [3] мають приблизно однакову кількість операцій (N) на відрізку n Є [0; 30], що показує досить схожу формулу росту функції і тому при підході визначення складності алгоритму через засікання часу потрібно дуже багато разів виконати алгоритм на великих даних щоб відрізнити дві схожі функції зростання.

На відміну від цього способу, лексичний алгоритм не потребує виконання алгоритму, а може підсталяти значення кількості вхідних даних та проводити аналіз довільну кількість разів, до тих пір поки не буде чітко видно формулу зростання функції, що набагато ресурсоекономніше та швидше ніж у загальновикористованому алгоритмі.

Проте головним недоліком лексичного алгоритму буде те, що він буде ігнорувати алгоритми, які мають рекурсивні інструкції або умовні оператори, тому що він зможе обробляти лише один варіант виконання програми (через неможливість визначення в інструкції якого умовного оператору було передано виконання програми при дійсному виконанні програми). Можливо, що є способи моделювання які б включали в себе умовні оператори, проте на даному етапі аналізу цього способу не було знайдено.

Після аналізу переваг та недоліків універсального – часового та лексичного – можна виділити наступні кроки оптимальної оцінки алгоритму:

* парсинг програми на лексеми без підрахунку кількості операцій;
* пошук інструкцій, які не підтримуються на даному етапі (що означає, що даний алгоритм неможливо оцінити лексичним алгоритмом);
* у випадку, якщо інструкції, які на даному етапі не підтримуються алгоритмом були знайдені – перехід до оцінювання алгоритму часовим алгоритмом;
* у випадку, якщо алгоритм таких інструкцій не було знайдено – оцінювання алгоритму лексичним алгоритмом (етапи алгоритму будуть описані в пункті «1.3 Постановка задачі»);

1.2 Аналіз аналогів

На даний момент існують безліч веб-сервісів, що надають задачі пов’язані з реалізацією того чи іншого алгоритму, але ні один з них не бере на себе оцінювання складності алгоритму:

* leetcode.com, hackerrank.com – одні з найвідоміших сервісів, який лише умовно порівнює час виконання алгоритму відносно інших учасників, що вирішували дану задачу. В залежності від часу доби, результати можуть буди діаметрально протилежними. У вихідні дні ввечері та у понеділок вранці виконання одного алгоритму майже точно дасть абсолютно різні результати. Через навантаженість серверу у вихідні – час виконання буде низький і може програвати гіршим алгоритмам, які були виконані в той час, коли сервер був ненавантажений і тому бистро відпрацював;

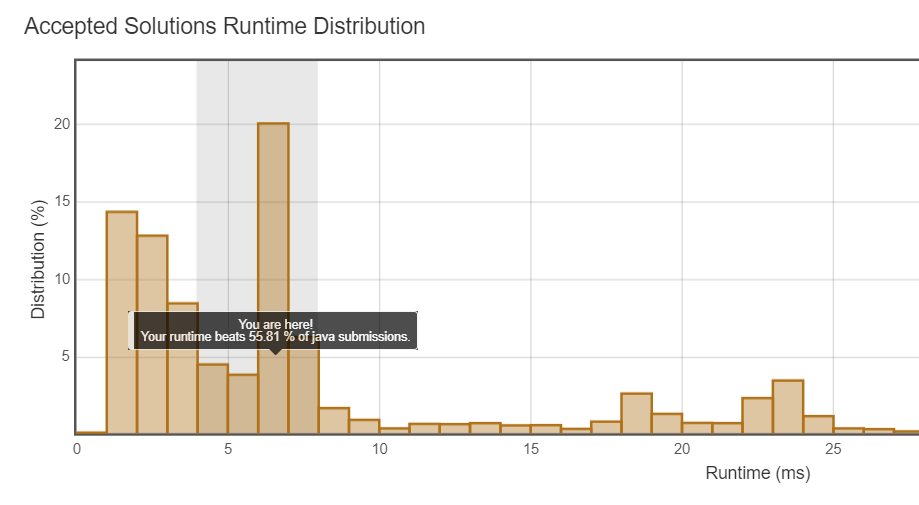


Рисунок 1.2 – Приклад оцінювання алгоритму на сайті leetcode.com.

Як видно з рисунку, оцінювання виконується лише порівняно з іншими учасниками системи, що є дуже нестабільним способом. Основною проблемою таких ресурсів є те, що для збільшення точності, якщо замірюванням часу виконання алгоритму займається декілька серверів – вони повинні бути ідентичними по обчислювальній потужності, або мати свою кореляцію часу на всі сервера. Такі ресурси мають обмеження по часу на виконання програми, після перевищення якого переривають виконання програми.

* lizard.ws – ресурс для аналізу складності коду. Цей ресурс рахує кількість вложеностей циклів, на основі чого робить висновок про складність коду.

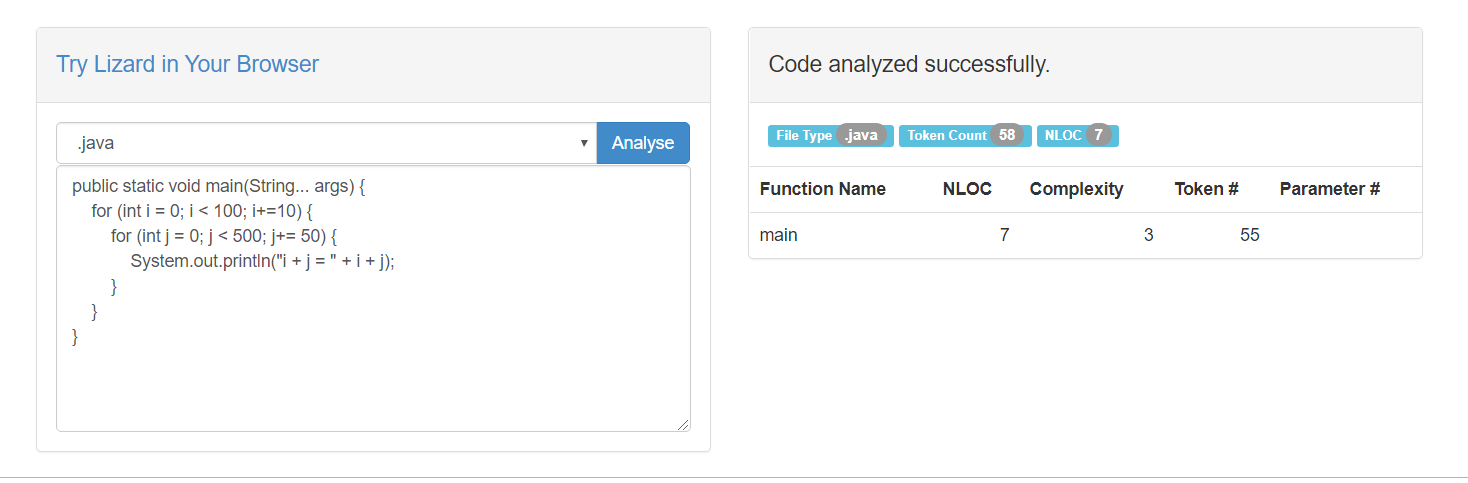


Рисунок 1.3 – Скріншот роботи сервісу lizard.ws.

Як видно зі скріншоту – програма порахувала кількість вкладених циклів і дала інформацію про складність. Проте лише з інформацією про кількість вкладеностей неможливо якісно оцінити складність коду, тому що як видно з коду на скріншоті – ітерація циклу виконується з кроком в частину від кінцевого значення індексу. Це означає, що справжня складність таких вкладених циклів буде n, при умові що n = 100.

До неділіків всіх вищеперерахованих сервісів можна віднести те, що вони не вміють вираховувати асимптотичну складність алгоритмів. Виходячи з цього – існує необхідність створити веб-систему, яка б надавала задачі для вирішення та могла б аналізувати рішення та надавати асимптотичну оцінку їх складності.

1.3 Постановка задачі

В атестаційній магістерській роботі необхідно розробити алгоритм оцінювання складності коду в рамках веб-системи, яка б надавала задачі для вирішення. Ця система може використовуватись для змагань програмістів між собою. Ця система буде корисним інструментом для розвивання аналітичного мислення та підвищення загального рівня спеціалістів. Система повинна аналізувати написаний код, розбиваючи його на лексеми та підраховуючи кількість операцій програми. У випадку, якщо роботу програми неможливо оцінити таким алгоритмом (за наявності операторів умовного переходу, рекурсії, тощо) – використати для оцінки алгоритм, який буде виконувати написану програму з різними наборами вхідних даних та аналізуватиме графік росту функції часу від розміру вхідних даних, що забезпечить велику точність оцінки.

Кінцеву задачу можна розбити на наступні етапи:

* провести аналіз етапів розробки лексичного алгоритму;
* розробити базовий лексичний алгоритм для найпростіших конструкцій (одного зовнішнього циклу);
* підібрати клас задач для даного алгоритму;
* реалізувати часовий алгоритм оцінки задач;
* створити набори даних різних розмірів але подібної структури (наприклад: усі відсортовані, або усі довільні) для більш точного аналізу алгоритму;
* з’єднати алгоритми для оцінювання коду (за допомогою GoF[4] патерна проектування Chain of Responsibility[5]);
* написати задачу, яка б не підлягала оцінюванню лексичним алгоритмом, та оцінилася часовим;
* додати задачі до бази даних;
* доповнити лексичний алгоритм для підтримки більш складних конструкцій (вкладені цикли, різні види циклів (for[6], while[7]).

До не функціональних вимог можна віднести технології розробки:

* реалізація основного функціоналу повинна бути виконана за допомогою мови програмування Java 11, фреймворку Spring Framework 5.0, Spring Boot 2.0;
* в якості сховища даних обрано NoSQL MongoDB та клауд-хостинг mLab.com;
* веб-частина додатку повинна бути реалізована за допомогою мови Javascript та фреймворку React;
* в якості API обрано RESTful API;
* в якості серверу додатків обрано Tomcat.

# **ПЕРЕЛІК ВИМОГ ДО ПРОГРАМНОЇ СИСТЕМИ**

## 2.1 Призначення розробки

Веб-система оцінювання складності алгоритму призначена для широкого спектру користувачів: англомовний інтерфейс та задачі поділені за різним рівнем складності підійдуть для людей різного рівня підготовки з усього світу. Основною категорією користувачів передбачаються бути програмісти. В першому релізі передбачається підтримка однієї мови програмування для вирішення задач – Java, проте система потенційно може підтримувати багато мов.

Проблема оптимальності написаного коду дуже гостро стоїть у сучасному світі: безліч сервісів, що забезпечують сотні тисяч транзакцій за хвилину і більше дуже обережно підходять до питання продуктивності системи: пришвидшення швидкості виконання запиту навіть на 10% може дуже сильно зменшити затрати на систему. У сучасному айті дуже багато людей, які не отримували профільної освіти і навчались самі і тому дуже часто такі люди пропускали таку важливу частину програмування – як алгоритми і структури даних, тому наявність відповідної системи необхідна для заповнення пробілу в знаннях програмістів та удосконаленню навичок висококваліфікованих спеціалістів.

## 2.2 Вимоги до програмного продукту

Загальні відомості про розробку програмної системи полягають у наступному: веб-система повинна бути побудована за клієнт-серверною архітектурою. Серверна частина повинна використовувати мову програмування Java 11 та фреймворки Spring Framework 5.0 та Spring Boot 2.0. Клієнт повинен бути написаний за допомогою мови програмування JavaScript та фреймворку React. API для взаємодії клієнта-сервера обрано RESTful[8], протокол обміну інформацією – HTTP. В якості сховища інформації використовується NoSQL MongoDB[9], як зручне сховище для зберігання неструктурованої інформації в великих об’ємах і зручному форматі. Для взаємодії з базою даних обрано Spring Data JPA та фрейморвк Hibernate для мапінгу об’єктів системи на таблиці бази даних. Система збірки проекту – Maven, система керуванян версіями – Git.

Основна функціональноість даної системи полягає у наступному:

1. виведення на екран списку доступних задач для вирішення (в основному, направлених на практику конкретних алгоритмів);
2. зчитування написаного коду та аналіз складності:
   1. пошук структур, які не підтримуються лексичним алгоритмом;
   2. розбиття коду на лексеми та обробка лексем (підрахунок кількості операцій);
   3. спроба отримати результат (у випадку невдачі – перехід до наступного пункту);
   4. оцінка коду часовим алгоритмом:
      1. виконання програми на різних наборах даних різних розмірів;
      2. підрахунок часу на виконання програми на кожному набору даних;
      3. вичислення формули росту часу від кількості даних;
      4. поверненя оцінки складності;
3. повернення результату про складність алгоритму користувачеві.

Отже, у ході розробки програмної системи необхідно проаналізувати предметну область лексичного аналізу коду, проаналізувати методи оцінювання вкладених інструкцій (використання спеціальних структур даних для зберігання стану даної лексеми), проаналізувати найпоширеніші функції складності та навчитись правильно переводити кількість операцій у функцію складності.

## 2.3 Вимоги до клієнта

Система розробляється у вигляді веб-додатку, що не потребує спеціального програмного забезпечення або високих вимог до апаратного обладнання для користування з системою. До переліку вимог можна віднести:

* наявність веб-браузеру;
* наявність доступу до мережі Internet.

# **ОПИС ПРИЙНЯТИХ ПРОЕКТНИХ РІШЕНЬ**

## Аналіз та UML-моделювання предметної області оцінювання складності алгоритмів

Для правильної розробки системи, яка зможе оцінювати асимптичну часову складність алгоритму я провів аналіз предметної області і з’ясував на які етапи треба розбити процес проектування та розробки відповідної системи:

1. аналіз етапів розробки лексичного алгоритму:
   1. аналіз всіх можливих ключових (зарезервованих) слів у мові;
   2. аналіз можливих операцій;
   3. аналіз можливих операндів;
   4. аналіз способів і методів підрахунку операцій:
      1. пошук у строці коду операцій ініціалізації змінних (операторів присвоєння);
      2. пошук у строці коду операторів (додавання, віднімання, множення, ділення, бінарні, атомарні, тощо);
      3. аналіз обробки операторів циклів (for, while);
      4. аналіз обробки вкладених циклічних операторів (наприклад: якщо всередині циклу є цикл з кількістю ітерацій n, то результатов повинна бути кількість операцій n\*m, де m – кількість операцій зовнішнього циклу);
   5. аналіз способів парсингу коду;
   6. аналіз мапінгу кількості операцій до формули складності;
2. аналіз способів реалізації часового алгоритму:
   1. аналіз методів засікання часу таким чином, щоб в рамках аналізу однієї програми було задіяно фіксовану кількість ресурсів (для забезпечення точності вичислення);
   2. підготовка наборів даних різних розмірів подібного стану (наприклад: якщо на вхід алгоритму quick sort подати набір даних повністю відсортований зі зручним опорним елементом алгоритм може відпрацювати швидше ніж алгоритм з меншим розміром вхідних даних але з невідсортованими даними та незручним опорним елементом. Тому важливо, щоб дані були подібні);
   3. аналіз мапінгу росту часу до формули складності;
3. аналіз набору задач для тестування обох алгоритмів.