

Лицей «Физико-техническая школа» Санкт-Петербургского
Академического университета

Курсовая работа (отчет по практике)

Создание программы, генерирующей кроссворды из регулярных выражений

Работу выполнили:
Данилевич Леонид (2022А)
Лельчук Александр (2022А)
Научный руководитель:
Дворкин Михаил Эдуардович
Место прохождения практики:
Лицей «ФТШ»

Санкт-Петербург, 2021

Аннотация

Мы создали программу, генерирующую кроссворды, в которых строки и столбцы описываются регулярными выражениями. В регулярных выражениях используются различные человеческочитаемые паттерны (например: палиндромы, прогрессии, словарные ключи, повторы и так далее).

Полученный кроссворд имеет единственное решение, которое может быть получено последовательным сужением круга возможных вариантов - таким образом, для отгадывания кроссворда не требуется подставлять различные буквы, после чего проверять корректность полученного предположением кроссворда.

Также программа может решить произвольный аналогичный кроссворд (поддерживается ограниченный набор конструкций, составляющих регулярные выражения).

Кроссворды из регулярных выражений

1	Введение	4
2	Постановка задачи	5
3	Методика решения задачи	6
3.1	Программное решение кроссворда	6
3.1.1	Разбор регулярных выражений	6
3.1.2	Отгадывание букв	6
3.2	Генерация буквенного заполнения кроссворда	6
3.2.1	Оценка стоимости заполнения кроссворда	7
3.2.2	Имитация отжига	7
3.3	Генерация регулярных выражений	7
3.3.1	Оценка стоимости регулярных выражений	7
3.3.2	Имитация отжига	7
3.4	Оценка сложности, проверка единственности решения	7
4	Результаты	9
4.1	Программное решение кроссворда	9
4.2	Генерация буквенного заполнения кроссворда	9
4.3	Генерация регулярных выражений	9
4.4	Оценка сложности, проверка единственности решения	13
5	Анализ результатов	14
6	Благодарности	15

Глава 1

Введение

Регулярные выражения — формальный язык поиска подстрок в тексте и манипуляций с ними. Например, регулярному выражению «.*amp(le)?» соответствуют строки «Sample», «example», «Iamp» и некоторые другие. С помощью регулярных выражений можно достаточно легко искать в тексте подстроки определённого формата и заменять их на соответствующие им другие подстроки. В 2013 году, как задание конкурса «MIT Mystery Hunt», был создан кроссворд из регулярных выражений (рис. 1.1). Мы решили написать приложение, автоматически генерирующее подобные кроссворды, а также позволяющее их решать. Мы считаем, что такое приложение будет полезно многим людям, изучающим регулярные выражения, а для уже знакомых с ними оно будет просто интересно.

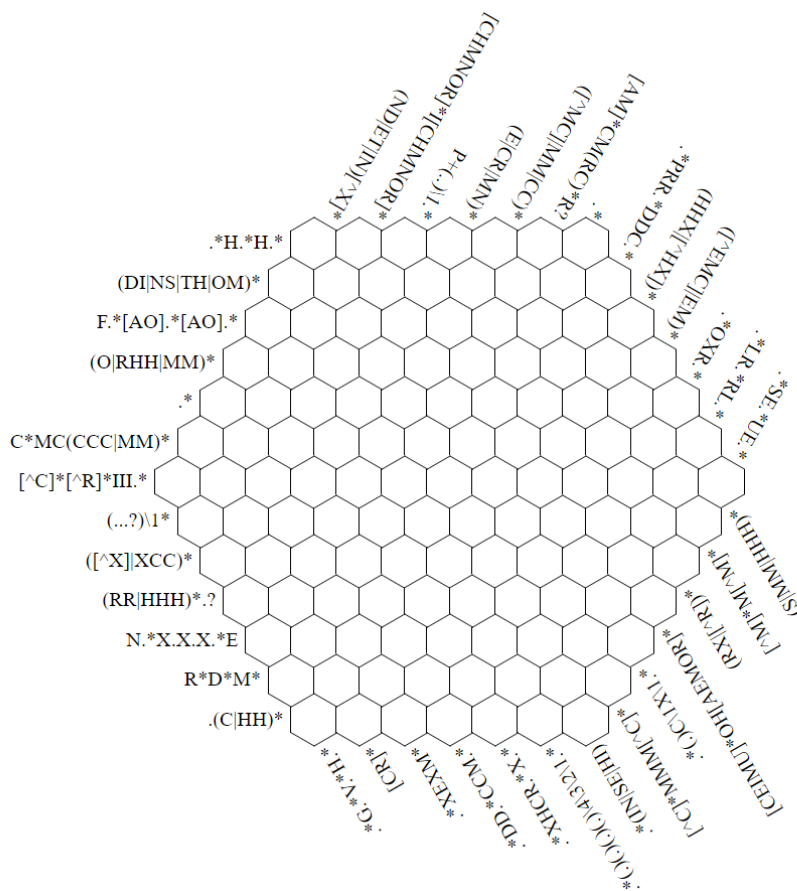


Рис. 1.1: Кроссворд с «MIT Mystery Hunt» 2013 года

Глава 2

Постановка задачи

Создать приложение на базе Android, функционалом которого является генерация, отображение и поддержка решения пользователем красивых кроссвордов из регулярных выражений.

Кроссворд из регулярных выражений – клеточная сетка правильной формы (прямоугольник, шестиугольник и т. д.). Решением кроссворда называется такое сопоставление символа каждой клетке, что линии, образованные этими символами (например, в прямоугольнике это столбцы и строки) формируют буквенные строчки, закодированные соответственными регулярными выражениями.

Мы называем кроссворд из регулярных выражений красивым, если он имеет единственное решение; в нём регулярные выражения, использующиеся для кодирования линий символов, разнообразны (используется большое количество конструкций: символьные классы, группы, перечисления, квантификаторы, обратные ссылки и т. д.); в полученных линиях символов должны наблюдаться паттерны, которые легко воспринимаются человеком.

Глава 3

Методика решения задачи

Так как задача состоит в написании программы, то решается программно. Задача дробится на части:

3.1 Программное решение кроссворда

Собственно, решение уже имеющегося кроссворда. Понадобится для оценки сложности кроссворда и удостоверения единственности решения.

3.1.1 Разбор регулярных выражений

Для решения произвольного кроссворда необходимо разобрать («распарсить») регулярные выражения, кодирующие его строчки. Для этого написана рекурсивная функция с запоминанием ответа (техника «мемоизации» динамического программирования), которая для заданного регулярного выражения и требуемой длины строчки находит в компактном виде всевозможные строчки, подходящие под данное выражение. Компактный вид обусловлен тем, что строчки состоят не из символов, а из битовых масок, где из любой маски можно выбрать любой символ и получить корректную строчку. Обратные ссылки в регулярных выражениях обрабатываются следующей техникой: два символа, запрашиваемые быть одинаковыми, соответствуют одному и тому же битовому объекту-маске.

3.1.2 Отгадывание букв

После разбора регулярных выражений кроссворд решается методом инкрементальных улучшений. Изначально каждой букве сопоставлена битовая маска, соответствующая всем символам алфавита. Алгоритм рассматривает все ещё не разгаданные буквы по порядку, сужая для каждой множество возможных значений. Существует вероятность встретить кроссворд, имеющий решение, но нерешаемый данной техникой. Однако для его разгадывания человеку потребуется долгий и утомительный перебор. Создание таких кроссвордов не входит в нашу задачу. В случае полного разгадывания кроссворда можно утверждать, что его решение существует и единственно.

3.2 Генерация буквенного заполнения кроссворда

Первым шагом для генерации произвольного кроссворда является генерация «буквенного заполнения» кроссворда, а именно, нахождение поля, которое получит пользователь, правильно разгадавший кроссворд.

3.2.1 Оценка стоимости заполнения кроссворда

Для красоты кроссворда необходимо образование строк, подходящих под различные регулярные выражения. Максимизируется красота заполнения — субъективное свойство, включающее в себя регулярность кроссворда. Именно, максимизируется количество и длина входящих в кроссворд подстрок, являющихся шаблонами следующих типов: палиндром (строка, читающаяся одинаково слева направо и справа налево); повтор (строка, состоящая из одинаковой части, повторённой несколько раз); прогрессия (строка, в которой на нечётных (в 1-индексации) местах стоят одинаковые буквы, а на чётных - разные; словарный ключ - существующее английское слово.

3.2.2 Имитация отжига

Для генерации случайного поля, максимизирующего потенциальную стоимость кроссворда, а именно, разнообразие типов регулярных выражений, разгадка необходимой степени сложности, и субъективная оценка красоты, применён метод имитации отжига («simulated annealing»). Каждая итерация отжига заменяет одну букву на изначально равномерно случайно сгенерированном поле, после чего в некотором случае изменение принимается, поле обновляется.

3.3 Генерация регулярных выражений

Вторым шагом является «восстановление» регулярных выражений, описывающих линии - строки и столбцы - кроссворда. Кроссворд должен иметь единственное решение, но не быть очевидным.

3.3.1 Оценка стоимости регулярных выражений

Для оценки стоимости анализируются регулярные выражения в строковом представлении, максимизируется разнообразие паттернов.

3.3.2 Имитация отжига

Здесь также используется метод имитации отжига. Начальные регулярные выражения побуквенно (следовательно, однозначно) задают соответствующие линии кроссворда. На каждой итерации метода имитации отжига регулярное выражение, соответствующее случайной линии кроссворда изменяется. После этого в случае улучшения цены кроссворда (при условии, что кроссворд по-прежнему решается единственным образом) изменение принимается.

3.4 Оценка сложности, проверка единственности решения

Оценка сложности и проверка единственности решения осуществляется путём программного решения.

Для решения кроссворда в случайном порядке выбираются клетки, буква в которых ещё не определена однозначно, после чего сужается круг возможных значений. Пока неясно, что за буква в ней стоит, одна и та же клетка может быть выбрана сколь угодно много раз. Сложность кроссворда считается как усреднённое количество клеток, которые надо рассмотреть,

чтобы полностью решить кроссворд. Очевидно, минимальная сложность равна суммарному количеству всех клеток в кроссворде, но сложность нетривиального кроссворда всегда больше.

Глава 4

Результаты

4.1 Программное решение кроссворда

Получившаяся программа успешно разгадывает кроссворды и составляет собственные. При вводе регулярных выражений из кроссворда, упоминавшегося во введении (рис. 1.1), программа выдаёт следующее решение (рис. 4.1)

```
S E C U E M C
M L R C R L M C
M M X O X R X M H
H E M H E M H E M H
H H X M I R H H X D C
H P R R M I O H H D D C
S T X M C M I E C R X R G
A M A M M C M R C R C R
H O X M M C C O X R N
E M N M N C R E C R
P O X O X C X R V
H I O M C M R O
N D F M M C H
```

Рис. 4.1: Решение кроссворда (рис. 1.1)

4.2 Генерация буквенного заполнения кроссворда

Алгоритм получилось написать. Результатом является довольно хорошо удовлетворяющие условиям кроссворда (рис. 4.2): Рассмотрим подробнее результат работы отжига. Выделим некоторые существенные паттерны в одном из кроссвордов (рис. 4.3)

4.3 Генерация регулярных выражений

Регулярные выражения также получилось сгенерировать. На картинке (рис. 4.4) изображён кроссворд с описывающими строки регулярными выражениями, полученный в результате работы программы.

G R E E T I N G	
O N H N O R O W O	
T D A D I D O D O A	
C O O L V L O R S W P	G R E E T I N G
T A T A P Z S T O E N A	N L L P O P L L
O O W O O O A R O R S W R	O S E E B E B F
P C T C P O S S S S O O N C	O L C P O P C L
S P A O V S A R A R D R H I S	N E T E T F T E
C S C P C S O N Z A M N R C	O L I P H P I L
S P E R E R D R V R H A S	O S H R I F T F
S P C S P N O N D N D C	N E W K W K W L
E S E R E R T G H A S	
O A L N P N O N R C	G R E E T I N G
G R I R I C H I S	S U E P D Q D P T
L G P G P G O D	H T Z E U U S O B T
	S U I T U Q F W G I G
G R E E T I N G	T R X S Z D A T A U Y B
U A U A U A U A	R F Q T U U T Q F R F O B
M R M R A I N W	I E X E X E I T A U P O N G
E V O K E A E A	P P R E S S Z W D Q D B F N P
N R N R N R N R	F T E T A X O O U U E A G A
E A O E E O E E	P F W F C P D R D P C N L
M R M R A P N J	E I E T G S U R G E G E
U E U E R E E L	T L R F D I E A L N L
	S T S T T W I N G E
	M Z D E D Z M O L
	U K M M K U T V

Рис. 4.2: Сгенерированные кроссворды

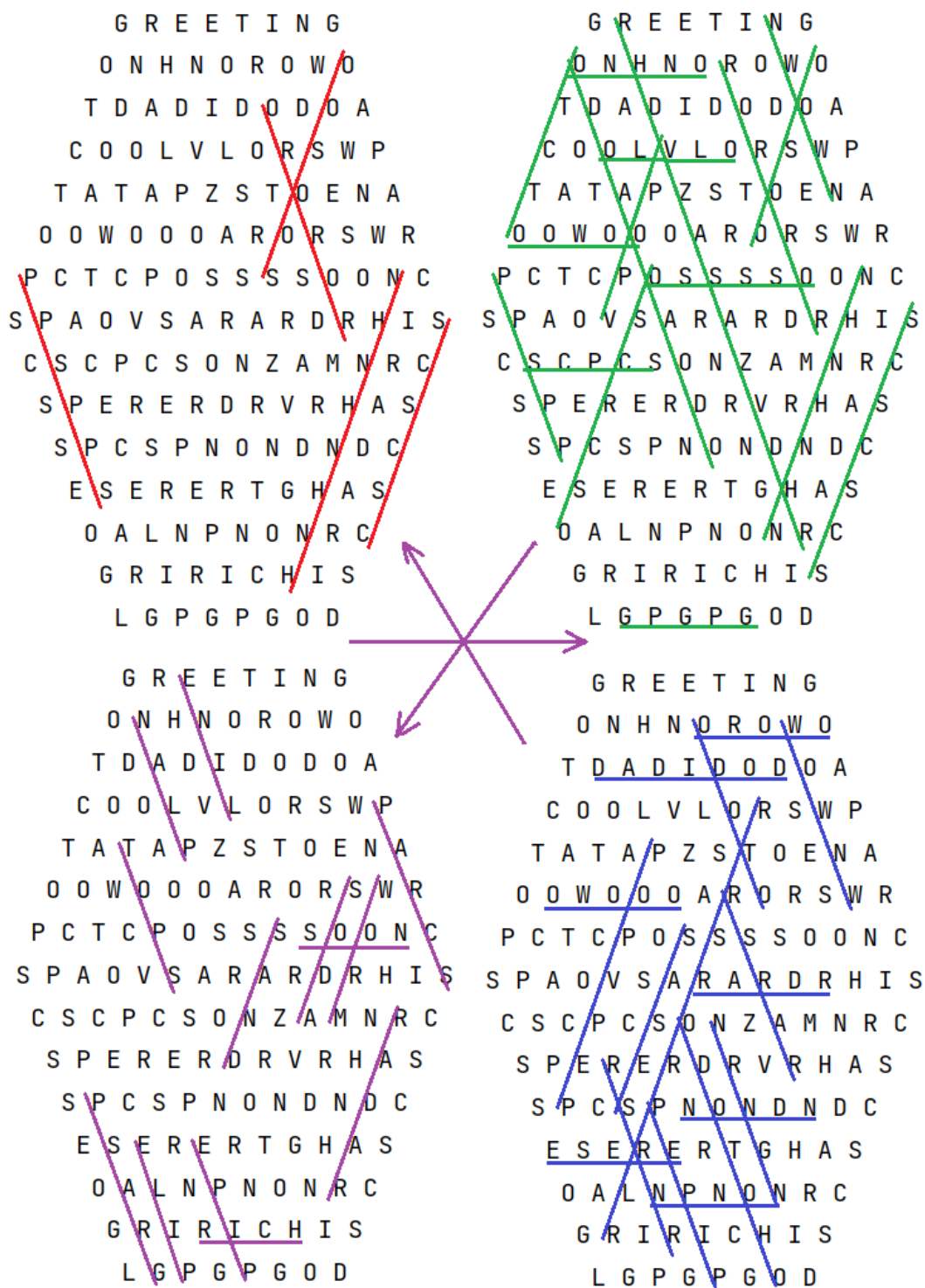


Рис. 4.3: Слева направо, сверху вниз: повторы, палиндромы, слова, прогрессии

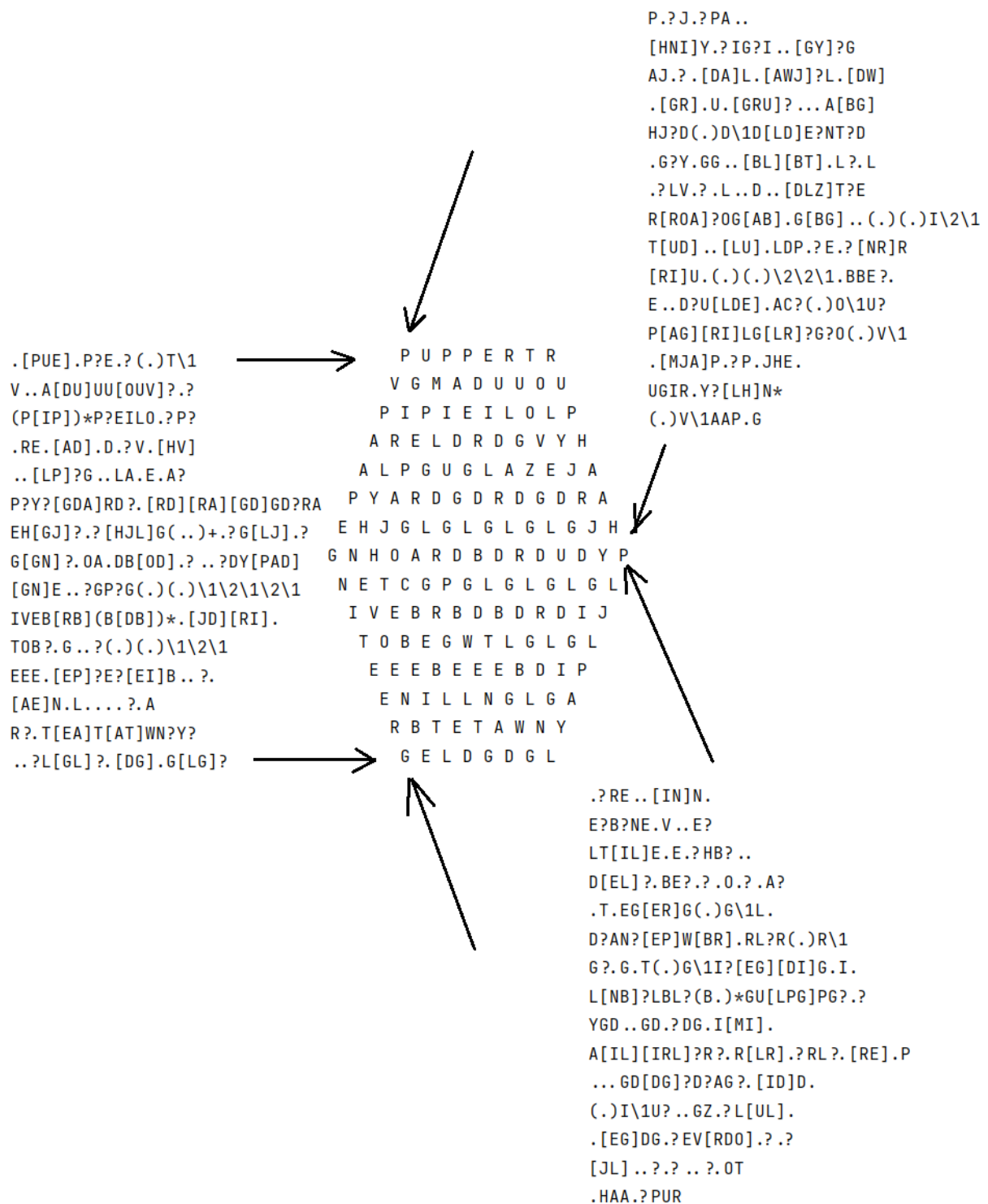


Рис. 4.4: Кроссворд и регулярные выражения

4.4 Оценка сложности, проверка единственности решения

Решение кроссворда заодно проверяет и единственность данного решения. Реализовано, как и написано в методике. Оказалось, что в текущих условиях даже наивная тактика перебора случайных клеток даёт очень неплохие результаты. Так, усреднённое количество клеток, которые нужно выбрать по 10 запускам, чтобы полностью решить приведённый на рисунке (рис. 4.4) кроссворд из 169 клеток в среднем равно 205. Для сравнения, кроссворд MIT (рис. 1.1) решается за 320 итераций, что намного лучше.

Глава 5

Анализ результатов

Написать программу под Android не удалось. Основная логика программы, а именно, решение и генерация кроссвордов написаны. Она подлежит доработке, так как получившиеся кроссворды решаются слишком быстро.

Глава 6

Благодарности

Мы благодарим нашего научного руководителя **Дворкина Михаила Эдуардовича** за научное руководство и направление теоретической части нашей работы.

