Самостоятельная работа 12

Тема. Поиск в тесте образца

Цель:

- получить знания, умения и навыки по применению алгоритмов поиска в тексте подстроки (образца);
- получить знания, умения и навыки по формированию регулярного выражения;
- получить знания, умения и навыки по применению аппарата поиска вхождения подстроки в текст посредством регулярного выражения.

1. Требования к выполнению практической работы

В данной практической работе требуется выполнить два задания. Задачи для каждого задания представлены в табл. 26.

1. В первом задании требуется разработать и реализовать алгоритм поиска заданной подстроки (подстрок) в некотором тексте, используя алгоритм, указанный в варианте 1 и исследовать алгоритм на тестах.
2. Во втором задании требуется познакомиться с технологией построения регулярных выражений, разработать и реализовать алгоритм поиска заданной подстроки (подстрок) в некотором тексте, используя механизм регулярных выражений.
3. Оформить отчет.

2. Задание 1

Разработать и реализовать алгоритм поиска образца в тексте, согласно первой задаче варианта (табл. 26).

2.1. Требования к выполнению задания

1. Изучить алгоритм, предложенный в варианте. Подготовить пять тестов для разных исходов поиска. Исследовать алгоритм на подготовленных тестах.
2. Разработать и реализовать алгоритм. Определить функцию (или несколько функций) для реализации алгоритма.
3. Выполнить его тестирование и убедиться в его корректности, используя инвариантность.
4. Выполнить тестирование алгоритма.

5. Оценить практическую сложность алгоритма в зависимости от длины текста и длины образца и отобразить результаты в таблицу (табл. 25) (для отчета).

3. Задание 2

Разработать алгоритм и функцию поиска образца в тексте с применением регулярных выражений для второй задачи варианта (табл. 26)

3.1. Требования к выполнению задания 2

1. Разработать регулярное выражение в соответствии задачей варианта.
2. Разработать функцию, реализующую проверку входной строки на соответствие регулярному выражению и в некоторых вариантах дополнительного действия со строкой.
3. Выполнить тестирование на разработанных тестах.
4. Оформить отчет, представив результаты по пунктам задания.

Таблица 25. Сводная таблица результатов задания 1

n	m	T(n, m)	$T_т$=f(C)	$T_п$=Cф
100				
1000				
10000				
100000				
1000000				

Примечание. m – длина образца. Значение m подберите исходя из алгоритма для заданной длины n (длина текста).

4. Варианты задач к заданиям 1 и 2

Таблица 26. Варианты задач к заданиям 1 и 2 практической работы 12

№	Условие задачи	Метод поиска
1.	1. Дан текст. Найти первое вхождение заданной подстроки в текст, с указанием позиции места размещения подстроки в тексте.	Бойера-Мура-Хорспула.

	2. Дан текст, содержащий адреса сайтов. Заменить в тексте в доменных имена России ru на рус.	
2.	1. Дан текст. Найти количество вхождений заданной подстроки в текст. 2. Дан текст. Удалите из текста все значения времени из промежутка 00:00 до 02:00. До и после этих значений может быть по одному пробелу.	Бойера-Мура.
3.	1. Дан текст. Найти индекс последнего вхождения заданной подстроки в текст, с указанием позиции места размещения подстроки в тексте. 2. Дан текст. Замените все даты в американском формате, датами в русском формате.	Кнута-Мориса-Пратта.
4.	1. Проверка на плагиат. Даны два текста, разбитые на слова. Из исходного текста выбирается слово и проверяется, входит ли слово в проверяемый текст. Сформировать список частоты появления слова в текст. 2. Дан текст. Сформировать список всех email-адресов, содержащихся в этом тексте.	Робина-Карпа.
5.	3. Дан текст и подстрока. Определить, количество вхождений подстроки в строку. 4. В тексте найти все строки со значением IP-адреса по формату v6 и заменить его элементы числами в двоичной системе счисления.	Кнута-Мориса-Пратта.
6.	1. Дан текст и подстрока - IP-адрес по формату v6. Определить, сколько раз упомянут этот IP-адрес в исходном тексте. 2. В тексте найти все строки со значением IP-адреса по формату v6 и заменить его элементы числами в 16-тиричной системе счисления.	Кнута-Мориса-Пратта.
7.	1. Строка S была записана много раз подряд, после чего из получившейся строки взяли произвольную часть строки - подстроку и передали, как входные данные. Необходимо определить минимально возможную длину исходной строки S. 2. Дан список придуманных пользователем паролей. Проверить, какие из паролей корректно составлены, т.е. удовлетворяют требованиям: в паролях могут быть только английские буквы (строчные или прописные) и цифры. Пробелы, подчеркивания и другие знаки препинания не допускаются.	Кнута-Мориса-Пратта.

Продолжение табл. 26

№	Условие задачи	Метод поиска
8.	1. Даны две строки a и b. Требуется найти максимальную длину префикса строки a, которая входит как подстрока в строку b. При этом считать, что пустая строка является подстрокой любой строки. 2. Дан текст, в тексте имеются данные по ценам на товары, причем в различных форматах отображения, найти цену на заданный товар в этом тексте. Цена на товар может быть указана в различных форматах т.е.: в ней могут встречаться запятые и символы валюты.	Кнута-Мориса-Пратта.
9.	1. Дано предложение, состоящее из слов. Сформировать словарь слов на основе префиксного дерева Ахо-Карасик. Используя дерево, вывести все слова текста, которые начинаются с заданной буквы или сочетания букв. 2. Дан текст, содержащий артикулы товаров, заключенные в кавычки. Найти все артикулы текста. Формат артикула: три латинские прописные буквы, дефис, трехзначное число.	Ахо-Корасик.
10.	1. Дан текст и список слов. Определить, сколько раз каждое слово входит в текст. 2. Дан текст, содержащий стандартные российские автомобильные номера, формируемые по формату: «буква - три цифры - две буквы - код региона». Причем код региона может быть двух или трехзначным, а в качестве букв применяются только те, что похожи внешне на латиницу. Номер в тексте ограничен с двух сторон пробелом. Найти автомобильный номер в исходном тексте. Например, дан текст: БМВ X3 B123AУ777 черн.E83 2.0d, результат B123AУ777.	Робина-Карпа.
11.	1. Назовем строку палиндромом, если она одинаково читается слева направо и справа налево. Примеры палиндромов: "abcba", "55", "q", "xyzzyx". Требуется для заданной строки найти максимальную по длине ее подстроку, являющуюся палиндромом. 2. Дан полный путь к файлу. Определить только имя файла.	Кнута-Мориса-Пратта.

Продолжение табл. 26

№	Условие задачи	Метод поиска
12.	1. Дан текст и множество подстрок-образцов. Сформировать таблицу, содержащую информацию о том, сколько раз каждый из образцов входит в исходный текст. 2. Дано предложение. Определить содержит ли оно одно из имен: Анна, Антонина, Алевтина, Алла.	Бойера и Мура.
13.	1. Дана подстрока – образец длиной не более 17 символов и строка с текстом, не ограниченная по длине. Применяя алгоритм Бойера-Мура вывести индексы строки, на которые смещается алгоритм при поиске вхождения образца. 2. Дан адрес места жительства, содержащий название города. Информация по городу начинается символами «г.» и завершается запятой. Определить, название города. Примеры адресов: 1) 123456, г.Москва, ул. Строителей, д.25 2) Московская область, г.Красногорск, ул. Широкая, д.13. 3) г.Москва, 123456, ул. Строителей, д.2.5.	Бойера-Мура-Хорспула.
14.	1. Назовем строку палиндромом, если она одинаково читается слева направо и справа налево. Примеры палиндромов: "abcba", "55", "q", "xyzzyx". Требуется для заданной строки найти максимальную по длине ее подстроку, являющуюся палиндромом. 2. Дан текст, содержащий имена файлов. Имена разделяются запятыми. Найти имена текстовых файлов, в имени которых последний символ цифра, а первый символ – заданная буква.	Бойера-Мура.
15.	1. Дан текст и множество подстрок образцов. Определить сколько раз каждый из образцов входит в исходный текст. Примечание. Для всех образцов создать хеш-таблицу. 2. Дан текст, содержащий имена файлов. Имена разделяются запятыми. Найти имена текстовых файлов, кроме тех, которые заканчиваются цифрой, а первый символ – заданная буква.	Робина и Карпа.
16.	1. Дан текст (строка) S длиной не более 100 000 букв ‘a’-‘z’ и ‘A’-‘Z’ и K (K < 1000) запросов (строки) , где каждый запрос содержит строку T максимальной длины 1000 только из букв ‘a’-‘z’ и ‘A’-‘Z’). Определить, какие из строк T из набора K являются подстрокой стоки S. 2. Дан текст. Вывести все email-адреса, содержащиеся в тексте.	Ахо-Корасик.
17.	1. Игра Словомания. Поле 4x4, заполненном буквами, необходимо найти как можно больше слов, составленных из этих букв. Из каждой клетки можно передвигаться в следующую по вертикали, горизонтали и диагоналям. 2. Дан текст. Определить, входит ли в этот текст заданная фраза из нескольких слов.	Цифровой поиск (по бору)https://habr.com/ru/post/216845.

Продолжение табл. 26

№	Условие задачи	Метод поиска
18.	1. Игра определения слов из заданного набора букв. Дано: 4 картинки, длина угадываемого слова и набор букв, выбирать буквы можно в любом порядке. 2. Дан текст в форме пронумерованного списка. Заменить номера на символ тире.	Цифровой поиск (по бору) https://habr.com/ru/post/216845.
19.	1. Дан код программы на языке C++. Определить, использовались ли операторы цикла в этом коде. 2. Определить, является ли строка корректным IPv4-адресом	Кнута-Мориса-Пратт.а
20.	1. Дан текст книги некоторого автора и множество слов (10). Авторы часто в своих произведениях используют любимые слова. Определить сколько раз каждое слово встретилось в тексте автора. Примечание. По количеству тех-или иных слов иногда определяют авторам. 2. Определить, является ли строка алгебраическим уравнением. Пример: y=x+2*k.	Робина-Карпа.
21.	1. Имеется два больших (100 000+) списка строк, и требуется отфильтровать первый список (SourceList) таким образом, чтобы в нем остались только строки, содержащие подстроки из второго списка (SearchList). 2. Дан текст с заданиями по арифметике, содержащий арифметические выражения. Арифметическое выражение состоит из двух чисел и операции между ними, например: 1 + 2; 1.2 *3.4; -3/ -6; -2-2. Вокруг оператора и чисел могут присутствовать пробелы. Найти арифметическое действие (операции) и их два операнда.	Цифровой поиск (по бору).
22.	1. Дан пакет из n документов. Каждый документ – это текст протокола регистрации ДТП. В протоколе указан номер автомобиля, участвующего в ДТП. Российские автомобильные номера, формируемые по формату: «буква - три цифры - две буквы - код региона». Причем код региона может быть двух или трехзначным, а в качестве букв применяются только те, что похожи внешне на латиницу. Определить сколько нарушений у владельца автомобиля с заданным номером. Например, дан текст: БМВ Х3 В123АУ777 черн. Е83 2.0d, результат В123АУ777 – 1. 2. Определить, является ли строка корректной датой с 1000 года. Учесть количество дней в месяцах. Считать, что в феврале всегда 29 дней.	Кнута-Мориса_Пратта.

Продолжение табл. 26

№	Условие задачи	Метод поиска
23.	1. Дан пакет из n документов. Каждый документ = это текст протокола собрания коллектива. В протоколе есть фраза: Слушали сообщение: после которой через пробел следует фамилия и инициалы (записаны по формату: Иванов И.И.) выступившего. Сформировать массив данных по выступившим для каждого протокола. 2. Определить, является ли строка номером телефона в формате +7-000-000-00-00.	Кнута Мориса-Пратта ю
24.	1. Дан пакет из n документов. Каждый документ – это текст протокола регистрации ДТП. В протоколе указан номер автомобиля, участвующего в ДТП. Российские автомобильные номера, формируемые по формату: «буква - три цифры - две буквы - код региона». Причем код региона может быть двух или трехзначным, а в качестве букв применяются только те, что похожи внешне на латиницу. Например, дан текст: БМВ Х3 В123АУ777 черн. Е83 2.0d, результат В123АУ777 2. Дан текст программы. Заменить все объявления переменных (в том числе с инициализацией) объявлением указателей того же типа, инициализируя их значением NULL.	Бойера-Мура-Хорспу ла.
25.	1. Дан текст с данными о выданных читателям книгах. Читателей много. Каждая книга характеризуется уникальным для книги значением (ISBN) – это 10 -ти значная строка из цифр в формате: X-XXXX-XXXX-X. Найти количество выданных книг с одним и тем же ISBN. 2. Определить, является ли строка объявлением с инициализацией переменной в C++, включая целочисленные массивы. Пример: char c=0; int x[3]={5, 5, 5};	Кнута-Мориса-Пратта.
26.	1. Дан текст с данными о заболеваниях пациентов поликлиники за некоторый промежуток времени. Каждый пациент идентифицируется номером медицинской карты. Номер карты представлен текстом, содержащим три прописные буквы (по букве от фамилии имени и отчества) и семь строчных букв. Определить, сколько раз пациент с заданным номером карты обращался в поликлинику. 2. Дано алгебраическое выражение. Левый операнд каждой арифметической операции заменить нулем	Бойера-Мура-Хорспула.
27.	1. Дан текст из нескольких предложений и слово. Определить в каком предложении из представленных заданное слово входит большее число раз. 2. Если после слова, начинающегося с большой буквы, идет слово, также начинающееся с большой буквы, отделить первое точкой так, будто это отдельное предложение.	Бойера-Мура.

Окончание табл. 26

№	Условие задачи	Метод поиска
28.	1. Дан текст с данными о заболеваниях пациентов поликлиники за некоторый промежуток времени. Каждый пациент идентифицируется номером медицинской карты. Номер карты представлен текстом, содержащим три прописные буквы (по букве от фамилии имени и отчества) и семь строчных букв. Определить, сколько раз пациент с заданным номером карты обращался в поликлинику. 2. В HTML тексте цвета задаются тегом с символом # и кодом цвета: это три или шесть шестнадцатеричных цифр. Найдите не валидные определения цветов в документе.	Кнута-Мориса-Пратта.
29.	1. Данные касс торгового зала при сбое работы базы данных были сохранены как текст. Операции, проводимые кассой, содержали ее номер – это строка из не более чем 15 символов, содержащая буквы, цифры в произвольном порядке. Определить, сколько операций провела касса с заданным номером. 2. В институте объединили и переименовали все учебные группы направлений бакалавров (ИКБО, ИНБО, ИВБО) в ИИБО с сохранением номера группы и года поступления, исправить шифры групп в тексте.	Кнута =Мориса-Пратта..
30.	1. Данные касс торгового зала при сбое работы базы данных были сохранены как текст. В тексте сохранилась информация, проводимая кассами торговой точки. Каждая операция содержала информацию о товаре, идентифицируя его некоторым кодом длиной не более 13 символов, включающим цифры и буквы. Определить, сколько единиц заданного товара продано торговой точкой за период работы кассы и учтено в текстовом файле. 2. Если 2 и более одинаковых гласных идут подряд, заменить все, кроме последней символом '_'.	Бойера-Мура-Хорспула.

5. Список источников по теории регулярных выражений

https://habr.com/ru/post/545150/ регулярные выражения (17.02)

https://regex101.com/, - инструмент построения регулярных выражений и проверки их корректности (17.02)

https://www.softwaretestinghelp.com/regex-in-cpp/ применение регулярных выражений в C++ (17.02)

http://website-lab.ru/article/regexp/shpargalka_po_regulyarnyim_vyirajeniyam/ - шпаргалка по символам регулярных выражений

6. Структура отчета

Титульный лист.

Оглавление.

1. Отчет по заданию 1.
   1.1. Условие задания и задание варианта.
   1.2. Описание подхода к решению задачи. Определить структуру элемента таблицы и определение таблицы.
      1.2.1. Описать особенности алгоритмов, реализуемых в задании.
      1.2.2. Привести таблицу тестов для тестирования алгоритма.
      1.2.3. Код программы.
      1.2.4. Скрины результатов тестирования.
   1.3. Представить таблицу (табл. 25) с указанием времени выполнения алгоритма, его фактическую и теоретическую вычислительную сложность.
2. Отчет по заданию 2.
   2.1. Условие задания и задание варианта.
   2.2. Привести регулярное выражения для задачи. Коротко описать ваш подход к построению выражения для решения задачи.
   2.3. Описать требующийся аппарат языка C++ для выполнения задачи.
   2.4. Привести тесты для успешного и безуспешного поиска.
   2.5. Реализовать алгоритм поиска. Отобразить код реализации.
   2.6. Привести скрины результатов тестирования.
3. Выводы по полученным знаниям, умениям и навыкам.

7. Контрольные вопросы

1. Что называют, строкой?
2. Что называют подстрокой?
3. Что называют префиксом строки?
4. Что называют суффиксом строки?
5. Приведите пример строки и укажите все ее суффиксы и префиксы.
6. Асимптотическая сложность последовательного поиска подстроки в строке?
7. В чем особенность поиска образца алгоритмом Бойера –Мура?
8. Приведите асимптотическую сложность алгоритма Бойера –Мура поиска подстроки в строке по времени и памяти.
9. Приведите пример входных данных для реализации эффективного метода прямого поиска подстроки в строке.
10. Приведите пример строки, для которой поиск подстроки "aaabaaa" будет более эффективным, если делать его методом Кнута, Морриса и Пратта, чем, если делать его методом Бойера и Мура. И наоборот.

11. Объясните, как влияет размер таблицы кодов в алгоритме Бойера и Мура на скорость поиска.
12. За счет чего в алгоритме Бойера и Мура поиск оптимален в большинстве случаев?
13. Поясните влияние префикс-функции в алгоритме Кнута, Морриса и Пратта (КМП) на организацию поиска подстроки в строке.
14. Приведите пример префикс-функции для поиска образца в тексте для алгоритма КМП.
15. В чем особенность поиска образца алгоритмом Рабина и Карпа?
16. Приведите асимптотическую сложность алгоритма Рабина и Карпа поиска подстроки в строке.
17. Что такое бор?
18. Какие структуры хранения данных используются для реализации простого бора?
19. Приведите пример бора и реализуйте его одним из способов. Объясните алгоритм поиска образца с использованием бора.
20. Поясните применение алгоритма Ахо – Корасик. Приведите его вычислительную и емкостную сложность.
21. Какие функции языка C++ используются при поиске подстроки через регулярные выражения?