# **Отчет**

Вариант 1

**i) Создайте регулярное выражение, которое выберет слова только с латинскими буквами и цифрами.**

Regex re = new Regex(@"\b[a-zA-Z0-9]+\b");

\b – границы слова.

[a-zA-Z0-9]+ - любые символы из группы символов (1 или > раз)

**j) Создайте регулярное выражение, которое выберет слова, имеющие хотя бы одну латинскую букву.**

Regex re = new Regex(@"\w\*[a-zA-Z]+\w\*");

\w\* - любой символ (0 или > раз)

[a-zA-Z]+ - любые символы из группы символов (0 или > раз)

**l) Напишите регулярное выражение, которое выделяет все листовые теги, т.е. теги, в которых нет других тегов, а содержится только текст. Вам придется использовать обратные ссылки.**

Regex re = new Regex(@"<([a-z]+)>[^<>]\*</\1>");

**<([a-z]+)> -**любой символ из группы символов (0 или > раз)

[^<>]\* - любой символ кроме <> (0 или > раз)

</\1> - закрывающий тег. \1 - обратная ссылка на единственную группу на которую можно сослаться.

**m) Ответьте на следующие вопросы**:

1. Что означает символ @ в начале строкового литерала в C#? Как будет выглядеть литерал @"(\W\*)(\w+)(\W+)(\w+)(\W+)(\w+)(\W\*)" без использования символа @?

@ - дает возможность использовать \. То есть преобразовывает текст в дословный. Используются для специальной интерпретации этой строки.

Без символа @, получим - "(\\W\*)(\\w+)(\\W+)(\\w+)(\\W+)(\\w+)(\\W\*)"

1. Что означают шаблоны: \w\*, \W\*, \w+ и \W+?

\w\* - любой символ слова – цифра, буква, знак подчеркивания 0 или более

\W\* - не символ слова 0 или более

\w+ - любой символ слова – цифра, буква, знак подчеркивания 1 или более

\W+ - не символ слова 1 или более

1. Что означает шаблон [^<>]\*?

любой символ кроме <> (0 или > раз)

1. Как сделать именованную группу?

(?<name> что-нибудь.)

1. Где в использованных выше выражениях использовались:
   1. Исчислители? - [^<>]\* \*- встречается 0 или более раз.
   2. Классы символов? – "\w\*[a-zA-Z]+\w\*" \w
   3. Обычные символы? [^<>]\*обычные символы <>
   4. Обратные ссылки? <([a-z]+)>[^<>]\*</\1> \1 – обратная ссылка на ([a-z]+)
   5. Подстановки? "$1$4$3$2$5$6$7"

**Практическая задача.**

Сначала были выявлены все одиночные заглавные буквы, которые окружены тегами.

Регулярное выражение для этого:

@"((<(?<open>([\w-]+))>)\s\*<(((?!italic).)\*)>\s\*(?<letter>[А-Я])\s\*</\4>\s\*(?(\k<open>)|)(</(\k<open>)>)\s\*)|(\s\*<(((?!italic).)\*)>\s\*(?<letter>[А-Я])\s\*</\9>\s\*)"

Были выявлены некоторые проблемы, например выводилась одиночная буква П под тегом italic, так как это единственный случай, то в регулярном выражения просто исключается данный тег.

Так или иначе все разделы написаны с ошибками, то есть:

Например неверные теги - <article>Х</article>, или лишние символы переноса строки, например для букв Б и В (хотя теги – верные)

Для исправления и приведения всех букв раздела в единый стиль была разработана следующая подстановка: "\n<letter><bold>${letter}</bold></letter>\n"

**Следующая ошибк**а это не <main-word> за буквой начала раздела теги, которые при этом пустые (ничего незначащие). Для выявления таких ошибок была создана следующая подстановка:@"<letter><bold>(?<letter>[А-Я])</bold></letter>(.\*|\s\*)<((?!<main-word>)([a-z]+))></\2>". А для исправления этой ошибки использовалась таже подстановка что и в предыдущем случае.

Такая ошибка была не найдена для букв Г, Д, Х. В букве Г после буквы раздела едет вынесенный тег <blod> после которого идет <main-word>. А после буквы X и Д идет сразу <main-word>. Этот результат был выявлен специальной подстановкой и выведен для дальнейшего изучения в консоль.