Основы и методология программирования

Семинар ***

Основы регулярных выражений.

Задача 1

В тексте, представленном строкой содержится температура в градусах Фаренгейта нужно заменить ее в этом тексте на температуру в градусах Цельсия.

Пример, «У него жар, на градуснике 99.5F. Надо что-то делать.»

Вывод:

У него жар, на градуснике 37.5C. Надо что-то делать.

Регулярные выражения

Регулярные выражения — это самостоятельный встроенный язык, широко используемой в огромном диапазоне программ. Данный мини-язык программирования имеет одно специфическое назначение: находить подстроки в больших строковых выражениях.

Пример РВ**:** `\b(\w+)(\s)(\w+)\b`

Язык РВ служит специально для **обработки строк**. Он включает две части:

- **Набор управляющих символов** для идентификации специфических типов знаков (метасимволы)
- **Система для группирования** частей подстрок и промежуточных результатов таких действий

РВ позволяют выполнять достаточно сложные действия **над строками,** например:

- Находить все повторяющиеся слова в строке
- Изменять заглавные буквы слов на строчные и наоборот
- Обеспечить правильную капитализацию предложений
- и др.

Метасимволы

[...]	Любой из символов, указанных в скобках, например [a-z]
[^...]	Любой из символов, не указанных в скобках [^0-9]
.	Любой символ, кроме перевода строки или другого разделителя Unicode-строки
\w	Любой текстовый символ, не являющийся пробелом, символом табуляции и т.п.
\W	Любой символ, не являющийся текстовым символом
\s	Любой пробельный символ из набора Unicode
\S	Любой непробельный символ из набора Unicode. Обратите внимание, что символы \w и \S - это не одно и то же
\d	Любые цифры. Эквивалентно [0-9]
\D	Любой символ, отличный от цифр. Эквивалентно [^0-9]

Символы повторения (квантификаторы)

{n,m}	Соответствует предшествующему шаблону, повторенному не менее n и не более m раз	s{2,4}	"Press", "ssl", "progressss"
{n,}	Соответствует предшествующему шаблону, повторенному n или более раз	s{1,}	"ssl"
{n}	Соответствует в точности n экземплярам предшествующего шаблона	s{2}	"Press", "ssl", но не "gasss"
?	Соответствует нулю или одному экземпляру предшествующего шаблона	Эквивалентно {0,1}	
+	Соответствует одному или более экземплярам предшествующего шаблона	Эквивалентно {1,}	
*****	Соответствует нулю или более экземплярам предшествующего шаблона	Эквивалентно {0,}	

Якорные символы регулярных выражений

^	Соответствует началу строкового выражения или началу строки при многострочном поиске.	^Hello	"**Hello, world**", но не "Ok, Hello world" т.к. в этой строке слово "Hello" находится не в начале
$	Соответствует концу строкового выражения или концу строки при многострочном поиске.	Hello$	"World, Hello"
\b	Соответствует границе слова, т.е. соответствует позиции между символом \w и символом \W или между символом \w и началом или концом строки.	\b(my)\b	В строке "Hello my world" выберет слово "my"
\B	Соответствует позиции, не являющейся границей слов.	\B(ld)\b	Соответствие найдется в слове "World", но не в слове "ld"

Использование регулярных выражений

Главный компонент обработки текста с помощью регулярных выражений — механизм регулярных выражений, представленный в библиотеке **re**.

Для обработки данных механизму регулярных выражений необходимо предоставить **два элемента**:

1. Шаблон регулярного выражения для определения текста.
2. Текст, который будет проанализирован на соответствие шаблону регулярного выражения.