РЕГУЛЯРНІ ВИРАЗИ

Метасимвол (англ. *metacharacter*) – це символ або комбінація символів, яка має спеціальне значення, якщо вона зустрічається у складі регулярного виразу. Найпростіший приклад метасимволів – символи узагальнення.
Символ узагальнення (англ. *wildcard character*) – символ, який дає змогу робити у рядку підстановку інших символів, щоб в одному запиті шукати, наприклад, одразу багато файлів. Зазвичай з цією метою вживають символи «*» (заміняє довільне число інших символів) та «?» (заміняє один довільний символ).
Регулярний вираз (англ. *regular expression, regexp*) – це зразок (англ. *pattern*) (послідовність символів та метасимволів), який відповідає (або не відповідає) послідовностям символів у тексті. Як правило, регулярні вирази використовуються для того, щоб задати правило пошуку підрядка у тексті.
Регулярні вирази використовуються не самі по собі, а разом з якоюсь мовою програмування або застосуванням. Синтаксис регулярних виразів може дещо відрізнятися, в залежності від того, як саме вони використовуються

Деякі метасимволи:
. (крапка) – один довільний символ (крім символу переходу на новий рядок)
\t – табуляція
\n – новий рядок
\\ – власне \ (backslash)
\s – один довільний пробільний символ (пробіл, табуляція, новий рядок)
\S – один довільний символ, який не входить у перелічені для \ s
\d – одна довільна цифра (digit)
\D – один довільний символ, який не може бути цифрою
\w – одна довільна літера або цифра або знак підкреслювання (word character), те саме що [A-Za-z0-9_]
\W – один довільний символ, який не входить у перелічені для \w
- – позначає діапазон(у класі символів) або власне цей символ, якщо це перший символ у класі ([-abc])
$ – кінець рядка
^ – початок рядка або заперечення, якщо це перший символ у класі символів ([^abc])
^$ – пустий рядок
| - логічне АБО (використовується у групі символів)
\< – початок слова
\> – кінець слова
\b – межа слова (початок або кінець) або символ backspace, якщо він знаходиться у класі символів
\B – позиція, що не є межею слова
\. – власне крапка
\$ – власне символ $
\[, \] – власне квадратні дужки
\(, \) – власне круглі дужки
\{, \} – власне фігурні дужки

Групування та повторення

* – нуль або більше разів повторений попередній символ (або група символів)
+ – один або більше разів повторений попередній символ (або група символів)
? – нуль або один раз повторений попередній символ (або група символів)\
() – групують символи (всі, що присутні у дужках, можливе застосування логічного АБО (символ |))
[] – визначають *клас* (або *множину*) *символів* – неупорядковану групу символів, з якої для відповідного регулярного виразу обирається один довільний
{n} – рівно n разів повторений попередній символ (або група символів)
{n, m} – попередній символ (або група символів) повторений від n до m разів
{n, } – n або більше разів повторений попередній символ (або група символів)

У випадках типу (a|(ab)) можливо, що другий варіант взагалі не знаходиться, хоча він насправді є (тому що алгоритм пошуку використовує скорочену форму АБО – як тільки значення логічного виразу встановлене, подальша обробка не проводиться – тобто знайшли перший (коротший) варіант і заспокоїлися). Вихід: «довший» варіант треба писати на початку.

При визначенні діапазонів символів слід враховувати, що вважається, що символи беруться з таблиці ASCII або Unicode (з якої саме, буде визначатися конкретним випадком використання).

Приклади:

[A-Za-z] – один довільний символ латинської абетки, незалежно від регістру (тут не можна писати [A-z], бо між літерами у великому та малому регістрах знаходяться інші символи)
[0-9] або (0|1|2|3|4|5|6|7|8|9) або \d – одна довільна цифра
\(\d{3}\) \d{3}-\d{4} – номер телефону у форматі (044) 123-4567
#[0-9a-fA-F]{6} – шістнадцятковий код кольору (наприклад, #12CCAA)
([0-9]|[1-9][0-9]|1[1-9][0-9]|2[0-4][0-9]|25[0-5]) – довільне число з діапазону 0-255 (спробуйте виправити і скоротити цей вираз самостійно!)

Завдання. Записати *<u>найкоротші</u>* регулярні вирази, які будуть відповідати

1) усім формам вашого прізвища;
2) усім формам вашого імені (можна врахувати і зменшувальні або інші варіанти);
3) усім телефонним кодам *обласних центрів та інших великих міст* тієї частини України, де живуть ваші батьки (Північ, Південь, Схід, Захід, Центр – поділ приблизний, можна подивитися у прогнозах погоди), – або аналогу такої інформації для вашої країни;
4) кодам IATA авіакомпаній України (пасажирські перевезення, сертифікат А1);
5) кодам IATA *міжнародних* аеропортів України;
6) 12 Інтернет-доменам першого рівня (top-level domains) на ваш вибір, крім національних та географічних;
7) діапазону чисел від *m* до *n**1000, де *mn* – ваш номер за списком групи;
8) двійковим числам у межах від 1 до 11111111;
9) шістнадцятковим числам у межах від 1*m* до AF*n*, де *mn* – ваш номер за списком групи;
10) номерам телефонів мобільних операторів України.

Для завдань 3-7, 9-10 у звіті навести дані, для яких будується регулярний вираз.