THEORETICAL COMPUTER SCIENCE
Jagiellonian University

wyrażenia regularne i grep

Wyrażenie regularne to sekwencja znaków opisująca wzorzec tekstowy, czyli zbiór słów odpowiadających podanemu opisowi.

Na wyrażenie regularne składają się znaki trzech różnych typów: *znaki pozycjonujące*, *zbiory znaków* oraz *znaki modyfikujące*.

Znaki pozycjonujące to:

- `^` początek linii
- `$` koniec linii
- `\<` początek słowa
- `\>` koniec słowa

Powyższe znaki pełnią funkcję pozycjonujących tylko wtedy, gdy są w odpowiednim miejscu wyrażenia regularnego.

Przykłady:

wyrażenie	**znaczenie**
`^A`	"A" na początku linii
`A$`	"A" na końcu linii
`A^`	"A^" gdziekolwiek w linii
`$A`	"$A" gdziekolwiek w linii
`^^`	"^" na początku linii
`$$`	"$" na końcu linii

Najprostszym wyrażeniem regularnym są ciągi liter. Dla przykładu wyrażenie "a" odpowiada po prostu literze "a", natomiast wyrażenie "krowa" odpowiada słowu "krowa".

Zbiory znaków (z przykładami użycia):

wyrażenie	**znaczenie**
`.`	linia, w której występuje dowolny znak (kropka oznacza zbiór wszystkich znaków)
`^.$`	linia zawierająca dokładnie jeden znak
`^[0123456789]$`	linia zawierająca dokładnie jedną cyfrę
`^[0-9]$`	to samo, co wyżej, ale krócej zapisane
`[A-Za-z0-9_]`	linia z co najmniej jedną literą, cyfrą lub podkreślnikiem
`^T[a-z][aeiou]`	linia zaczynająca się od "T", potem dowolnej litery i samogłoski łacińskiej
`[]`	linia zawierająca nawiasy kwadratowe
`[0]`	linia zawierająca 0
`[^0-9]`	linia zawierająca znak, który nie jest cyfrą
`[-0-9]`	linia zawierająca znak, który jest cyfrą lub minusem
`[0-9-]`	to samo, co wyżej
`[^-0-9]`	linia zawierająca znak, który nie jest cyfrą ani minusem
`[]0-9]`	linia zawierająca znak, który jest cyfrą albo]
`[0-9]]`	linia zawierająca cyfrę, a po niej]
`[0-9-z]`	linia zawierająca cyfrę lub znak z zakresu od 9 do z
`[0-9\-a\]]`	linia zawierająca cyfrę, minus, "a" lub]

Istnieją także alternatywne oznaczenia dla zbiorów - klasy znaków POSIX:

`[:alnum:]`	równoważny: `A-Za-z0-9`
`[:alpha:]`	równoważny: `A-Za-z`
`[:blank:]`	spacja lub tabulator
`[:cntrl:]`	znaki kontrolne
`[:digit:]`	równoważny: `0-9`
`[:graph:]`	znaki graficzne, zakres kodów ASCII 33-126
`[:lower:]`	równoważny: `a-z`
`[:print:]`	znaki drukowalne (znaki graficzne + spacja)
`[:space:]`	białe znaki
`[:upper:]`	równoważny: `A-Z`
`[:xdigit:]`	równoważny: `0-9A-Fa-f`

Uwaga! Te klasy wymagają ujęcia w dodatkową parę nawiasów kwadratowych, np.:

`[[:digit:]]`	zbiór oznaczający dowolną cyfrę
`[[:alnum:]_]`	zbiór oznaczający dowolną literę, cyfrę bądź podkreślnik

Znaki modyfikujące:

`\`	zmienia znaczenie następnego znaku ze specjalnego na zwykły i na owdrót
`*`	następując po znaku (lub wyrażeniu w nawiasach) sprawia, że ten występuje 0 lub więcej razy
`\+`	następując po znaku (lub wyrażeniu w nawiasach) sprawia, że ten występuje 1 lub więcej razy
`\?`	następując po znaku (lub wyrażeniu w nawiasach) sprawia, że ten występuje 0 lub 1 raz

Co więcej, można nakazać danej sekwencji znaków pojawiać się dokładnie określoną liczbę razy przy pomocy wstawienia przedziału wewnątrz `\{` i `\}`.

Przykłady:

wyrażenie	**znaczenie**
`*`	dowolna linia z gwiazdką
`\*`	to samo, co wyżej
`\\`	dowolna linia z backslashem
`^*`	dowolna linia zaczynająca się od gwiazdki
`^A*`	dowolna linia
`^A\*`	dowolna linia zaczynająca się od "A*"
`^AA*`	dowolna linia zaczynająca się jednym "A"
`^A*B`	dowolna linia zaczynająca się od "B" lub jednym lub więcej "A", poprzedzającym "B"
`^AA*B`	dowolna linia zaczynająca się jednym lub więcej "A", poprzedzającym "B"
`^A\+B`	to samo, co wyżej
`^A\?B`	dowolna linia zaczynająca się od "B" lub od "AB"

`^A\{4,8\}B`	dowolna linia zaczynająca się od 4, 5, 6, 7 lub 8 "A", poprzedzających "B"
`^A\{4,\}B`	dowolna linia zaczynająca się od co najmniej 4 "A", poprzedzających "B"
`^A\{4\}B`	dowolna linia zaczynająca się od "AAAAB"
`^\{4,8\}`	dowolna linia z "{4,8}"
`^A{4,8}`	dowolna linia z "A{4,8}"

Istnieje również wzorzec pozwalający wyszukiwać słowa z powtórkami. Przykładowo, wyrażenie `[a-z][a-z]` wskaże nam dwie sąsiednie małe litery, ale niekoniecznie takie same (podobnie jak `[a-z]*` wskaże nam dowolnej długości ciąg małych liter, lecz niekoniecznie różnych). Wyjściem jest zastosowanie nawiasów: `\(` i `\)`, w które zamykamy interesujący nas podwzorzec, by później odwołać się do niego przy pomocy cyfry poprzedzonej backslashem. Dysponujemy 9 takimi slotami pamięci, oznaczonymi od `\1` od `\9`. Przykładowo, pięcioliterowych palindromów (takich jak np. "radar") możemy szukać przez następujące wyrażenie regularne:

```
\([a-z]\)\([a-z]\)[a-z]\2\1
```

Polecenie `grep` jest dobrym przykładem polecenia wykorzystującego wyrażenia regularne. Służy ono do wypisywania na standardowe wyjście tych linijek pliku, które pasują do podanego wyrażenia. Przykładowe polecenie, które wyszukuje linie pliku "tekst.in", które zaczynają się od dużej litery:

```
grep '^[A-Z]' tekst.in
```

Kilka przydatnych opcji dla polecenia `grep` w przykładach:

`grep -c '^[A-Z]' tekst.in`	wypisuje tylko liczbę linii dla podanego wzorca
`grep -n '^[A-Z]' tekst.in`	wypisuje dopasowane linie wraz z ich numerami
`grep -v '^[A-Z]' tekst.in`	wypisuje linie, w których nie występuje wzorzec (w tym przypadku: linie, które nie zaczynają się od dużej litery)
`grep -w '^[A-Z]' tekst.in`	wypisuje linie, w których podany wzorzec tworzy całe słowo (w tym przypadku: linie zaczynające się od pojedynczej dużej litery)

Wyrażenia regularne wykorzystywane w `grep`ie pozwalają na używanie alternatywy słów, a nie jedynie alternatywy znaków. W przypadku wyszukiwania linii ze słowami "alfa" lub "beta" możemy postąpić tak, jak poniżej:

```
grep '\(alfa\|beta\)' tekst.in
```

Zadanie 1. Ile słów w słowniku scrabblisty kończy się na literę "g"?

Zadanie 2. Ile słów w słowniku scrabblisty nie zawiera litery "a" i litery "e"?

Zadanie 3. Ile słów w słowniku scrabblisty zawiera "ab" jako kolejne dwie litery?

Zadanie 4. Ile słów w słowniku scrabblisty zawiera literę "a" oraz literę "b" występującą w słowie po literze "a" lecz nie koniecznie bezpośrednio po niej?

Zadanie 5. Ile słów w słowniku scrabblisty zawiera "ab" lub "ef" jako kolejne dwie litery?

Zadanie 6. Ile słów w słowniku scrabblisty zawiera "ab" jako kolejne dwie litery i "ef" jako dwie kolejne litery?

Zadanie 7. Ile jest 15-literowych słów w języku scrabblisty?

Zadanie 8. Ile jest słów o co najwyżej 4 literach w słowniku scrabblisty?

Zadanie 9. Ile jest linii z słowami co najmniej 30-literowymi w słowniku języka polskiego? A czy umiesz policzyć ile jest słów co najmniej 30-literowych w tym słowniku?

Zadanie 10. Ile jest słów zawierających powtórzenie (czyli słowo składające się z dwóch indentycznych bloków) w słowniku scrabblisty?

Zadanie 11. Ile jest słów będących powtórzeniem w słowniku scrabblisty?

Zadanie 12. Znajdź numery linii słów "kara" i "kadra" przy pomocy jednego wywołania programu grep.

Zadanie 13. Wypisz wszystkie słowa ze słownika scrabblisty składające się jedynie z nie łacińskich liter (czyli, ąćęłńóśźż).

Zadanie 14. Wypisz wszystkie palindromy 9-literowe ze słownika scrabblisty. Wypisz wszystkie słowa zawierające 9-literowy palindrom jako podsłowo.