Laboratorium 7 - Filtrowanie strumienia tekstu i wyrażenia regularne

1. Teoria

http://www.regexr.com/ - testowanie wyrażeń regularnych online

http://regexcrossword.com/ - krzyżówki gdzie hasłami są wyrażenia regularne

1.1. `grep`

Wypisuje linie pliku, które pasujące do wzorca.

Argument	Opis
`-o`	wypisuje tylko dopasowanie
`-c`	zlicza dopasowania.
`-i`	ignoruje wielkość liter

`grep` obsługuje różne rodzaje wyrażeń regularnych:

- `-F` fixed
- `-G` basic (BRE) - domyślne
- `-E` extended (ERE)
- `-P` Perl (PCRE)

1.2. Proste wzorce (BRE)

- `.` zastępuje dowolny znak
- `*` poprzedzający element zostanie dopasowny zero lub więcej razy
- `^` reprezentuje początek linii
- `$` reprezentuje koniec linii
- `[…]` klasa znaków
- `[^…]` zaprzeczona klasa znaków

1.3. Rozszerzone wzorce (ERE)

To co BRE, a dodatkowo:

- `+` poprzedzający element zostanie dopasowny jeden lub więcej razy
- `?` poprzedzający element zostanie dopasowny zero lub jeden raz
- `(…)` nawiasy grupują wyrażenia; jeśli po nawiasie zamykającym wystąpi jeden ze znaków `?`, `*`, `+` to ten operator odnosi do wyrażenia w nawiasie
- `|` alternatywa wyrażeń
- `{m,n}` poprzedzający element zostanie dopasowny od `m` do `n` razy

- `{m,}` poprzedzający element zostanie dopasowny przynajmniej `m` razy
- `{,n}` poprzedzający element zostanie dopasowny co najwyżej `n` razy
- `{m}` poprzedzający element zostanie dopasowny `m` razy

1.4. Praktyczne zadania

1.4.1. Zadanie

Skopiowałem z PDFa poniższy tekst. Chciałbym aby słowa były oddzielone pojedynczą spacją.

```
Grep    umożliwia       użycia  rozszerzonej     (extended)     składni,
która   udostępnia mechanizm    wyrażeń
regularnych.    Wyrażenie       regularne       służą do        opisu
formalnego      łańcuchów       symboli.        Jeśli   symbole
wykazują        jakieś  regularności    (np.    w       wyrazach
aba,    abba,   abbba,  abbbbba,        … regularność   polega
na      tym,    że      na      początku        i       końcu   znajduje
się     litera  a,      a       w       środku  dowolna niezerowa
liczba  liter   b)
to      można   utworzyć        zapis,  który   opisze  tą      regularność
(np.    ab+a).
```

W jaki sposób wykonać zadanie jeśli między słowami byłyby pomieszane spacje i tabulatory?

1.4.2. Zadanie

Mamy raport finansowy. Chcielibyśmy usunąć wszystkie występujące w nim liczby.

```
Aktywa obrotowe stanowiły 23,1% sumy aktywów. W stosunku do
grudnia 2012 roku zwiększyły się o 2 005,5 mln zł (tj. o 57,7%) w
następstwie
wzrostu inwestycji krótkoterminowych o 1 752,9 mln zł (tj.
o 93,9%), należności krótkoterminowych o 296 mln zł (tj. o 22,2%) oraz
krótkoterminowych rozliczeń międzyokresowych o 13,7 mln zł (tj.
o 17,7%).
```

1.4.3. Zadanie

Mamy plik podobny w strukturze do `/etc/passwd`, ale separatorem jest ciąg + / +. W jaki sposób wyciąć ostatnie pole?

```
root / x / 0 / 0 / administrator / /root / /bin/bash
daemon / x / 1 / 1 / daemon / /usr/sbin / /bin/sh
bin / x / 2 / 2 / bin / /bin / /bin/sh
sys / x / 3 / 3 / sys / /dev / /bin/sh
sync / x / 4 / 65534 / sync / /bin / /bin/sync
games / x / 5 / 60 / games / /usr/games / /bin/sh
man / x / 6 / 12 / man / /var/cache/man / /bin/sh
```

1.4.4. Zadanie

W jaki sposób przerobić poniższą listę

```
raz
dwa
trzy
```

na wersję HTML?

```
<li>raz</li>
<li>dwa</li>
<li>trzy</li>
```

2. Praktyka

2.1. Zadanie

`unzip`

Pobierz z moodle i rozpakuj plik ***slowa.txt.zip***

2.2. Zadanie

Przy użyciu `grep` i `ls` odnajdź wszystkie pliki lub katalogi w `/etc` które zawierają słowo `conf`.

2.3. Zadanie

Przeglądnij zawartość pliku `slowa.txt` za pomocą poleceń head i tail. Sprawdź czy twoja konsola poprawnie wyświetla polskie znaki diakrytyczne. Kodowanie plików to UTF-8.

2.4. Zadanie

`^`, `$`

Wyświetl oraz policz wszystkie wyrazy, które:

- posiadają ciąg `maku`,
- zaczynają się od `maku`,
- kończą się na `maku`.

2.5. Zadanie

`[…]`, `{…}`

Napisz klasę reprezentującą wszystkie samogłoski w języku polskim. Sprawdź jaki wyraz (wyrazy) posiadają zbitkę o największej długości samogłosek w swojej treści. Np. wyraz `boeing` posiada zbitkę o długości 3 `oei`.

2.6. Zadanie

`[^…]`

Wykonaj analogiczne zadanie, ale znajdź wyrazy o najdłuższej zbitce spółgłosek. Ile ich jest?

2.7. Zadanie

`.*`

Znajdź wyrazy, które zaczynają się na `ma` a kończą na `my`.

2.8. Zadanie

Znajdź wszystkie wyrazy, które zawierają w sobie polskie znaki diakrytyczne.

2.9. Zadanie

`(…)`

Znajdź wszystkie wyrazy, które zawierają w sobie przynajmniej 5 polskich znaków diakrytycznych.

2.10. Zadanie

`|`

Znajdź za pomocą jednego wyrażenia regularnego wyrazy zaczynające się albo od `pod` albo od `nie` i kończące się na `śmy`, zawierające w środku pomiędzy tymi członami zbitkę dwóch samogłosek.

2.11. Zadanie

Pobierz plik z moodle ***plan_roku.txt***

2.12. Zadanie

Napisz wyrażenie regularne, które wyszuka wszystkie daty w tekście (wyświetli na wyjściu same daty). Skorzystaj w tym celu z opcji `-o` polecenia `grep`.

2.13. Zadanie

Napisz wyrażenie regularne, które dopasuje się do wzorca numeru uchwały (powinno być uniwersalne dla różnych numerów). Numer uchwały w tym tekście to `42/VI/2011`.

2.14. Zadanie

Pobierz plik z moodle ***email.txt*** i napisz wzorzec który zwróci wszystkie adresy e-mail. Przetestuj go na innych znanych Ci adresach e-mail dopisując je do pliku.