

Przetwarzanie tekstu 2

Operacje na plikach tekstowych w systemie Linux

Operacje na plikach tekstowych w systemie Linux

filtry programy przetwarzające pliki (w szczególności tekstowe) w taki sposób, że odczytują dane (plik wejściowy) ze standardowego wejścia i zapisują wynik ich przetworzenia (plik wyjściowy) na standardowe wyjście

Operacje na plikach tekstowych w systemie Linux

wybrane programy pakietu textutils (Text Utilities)

- cat** wypisywanie i łączenie plików
- wc** zliczanie znaków, słów i linii
- cut** wypisywanie wybranych części linii
- tr** zamiana znaków, wymazywanie znaków, usuwanie powtórzeń znaków
- uniq** usuwanie powtórzeń linii w posortowanym pliku
- sort** sortowanie linii pliku
- head** wypisywanie początku pliku
- tail** wypisywanie końca pliku
- nl** numerowanie linii
- split** dzielenie pliku na części
- comm** porównywanie posortowanych plików linia po linii

Operacje na plikach tekstowych w systemie Linux

inne kluczowe programy

grep wybieranie z pliku linii zawierających zadane wyrażenie

sed edytor tekstowy wsadowy (strumieniowy)

awk interpreter języka AWK

Operacje na plikach tekstowych w systemie Linux

wybrane programy pakietu textutils: `cat`

`cat` [opcje] [plik]...

- ▶ przepisywanie z wejścia na wyjście, łączenie plików (nie tylko tekstowych)
- ▶ opcje:
 - n numerowanie linii
 - s usuwanie powtórzeń pustych linii
 - ... (patrz dokumentacja)

Operacje na plikach tekstowych w systemie Linux

wybrane programy pakietu textutils: `nl`

`nl [opcje] [plik]...`

- ▶ numerowanie linii
- ▶ opcje: ... (patrz dokumentacja)

przykład

polecenie: `nl`

we:

```
Ala ma 22 koty.  
Ola ma  
44 psy.  
Karol nie ma  
kota.
```

wy:

```
1 Ala ma 22 koty.  
2 Ola ma  
3 44 psy.  
4 Karol nie ma  
5 kota.
```

Operacje na plikach tekstowych w systemie Linux

wybrane programy pakietu textutils: `split`

`split` [opcje] [plik [prefiks]]

- ▶ podział pliku tekstowego na równe fragmenty
- ▶ opcje:
 - l *n* po *n* linii
 - b *n* po *n* bajtów
 - ... (patrz dokumentacja)

przykład

polecenie: `split -l 2`

we:

```
Ala ma 22 koty.  
Ola ma  
44 psy.  
Karol nie ma  
kota.
```

wy:

```
Ala ma 22 koty.  
Ola ma
```

```
44 psy.  
Karol nie ma
```

```
kota.
```

Operacje na plikach tekstowych w systemie Linux

wybrane programy pakietu textutils: `wc`

`wc` [opcje] [plik]...

- ▶ zliczanie lini, słów i znaków
- ▶ opcje:
 - l zliczaj tylko linie
 - w zliczaj tylko słowa
 - c zliczaj tylko znaki

przykład

polecenie: `wc -l`

we:

```
Ala ma 22 koty.  
Ola ma  
44 psy.  
Karol nie ma  
kota.
```

wy:

```
5
```


Operacje na plikach tekstowych w systemie Linux

wybrane programy pakietu textutils: `sort`

`sort` [opcje] [plik]...

- ▶ sortowanie linii pliku
- ▶ opcje:
 - u usuwaj duplikaty
 - r sortuj malejąco
 - n sortuj numerycznie
- ▶ linie mogą być traktowane, jako całość lub jako ciąg pól oddzielonych separatorem
- ▶ opcje dotyczące sortowania z uwzględnieniem pól:
 - t *c* użyj znaku *c* jako separatora pól
 - k *m, n* sortuj wg fragmentu linii od pola *m* do pola *n*
 - k *m* sortuj wg fragmentu linii od pola *m* do końca
 - ... (patrz dokumentacja)

Operacje na plikach tekstowych w systemie Linux

wybrane programy pakietu textutils: `sort`

przykład

polecenie: `sort -r`

we:

```
Ala ma 22 koty.  
Ola ma  
44 psy.  
Karol nie ma  
kota.
```

wy:

```
Ola ma  
kota.  
Karol nie ma  
Ala ma 22 koty.  
44 psy.
```

Operacje na plikach tekstowych w systemie Linux

wybrane programy pakietu textutils: **uniq**

uniq [opcje] [plikwe [plikwy]]

- ▶ usuwa, zlicza, drukuje powtarzające się linie w posortowanym pliku (domyślnie usuwa duplikaty linii)
- ▶ opcje:
 - c zliczaj liczbę wystąpień każdej linii
 - d drukuj tylko powtarzające się linie
 - u drukuj tylko niepowtarzające się linie
 - ... (patrz dokumentacja)

Operacje na plikach tekstowych w systemie Linux

wybrane programy pakietu textutils: `comm`

`textutils: comm [opcje] plik1 plik2`

- ▶ porównanie dwóch posortowanych plików linia po linii
- ▶ w pliku wynikowym linie obecne w obu plikach są poprzedzone dwoma znakami tabulacji, linie obecne tylko w drugim pliku – jednym znakiem tabulacji, a linie obecne tylko w pierwszym pliku nie są niczym poprzedzone.

Operacje na plikach tekstowych w systemie Linux

wybrane programy pakietu textutils: `cut`

cut [opcje] [plik]...

► drukowanie części linii

► opcje:

- c *lista_znaków* drukuj tylko znaki na podanych pozycjach
- f *lista_pól* drukuj tylko pola o podanych numerach
- d *c* użyj znaku *c* jako separatora pól
- ... (patrz dokumentacja)

Operacje na plikach tekstowych w systemie Linux

wybrane programy pakietu `textutils`: `tr`

`tr` [opcje] *zbiór_znaków1* [*zbiór_znaków2*]

- ▶ zamiana znaków, wymazywanie znaków, usuwanie powtórzeń znaków; domyślnie (bez opcji) podstawia za znaki ze *zbióru_znaków1* odpowiednie znaki ze *zbióru_znaków2*
- ▶ opcje:
 - c zamiast *zbióru_znaków1* użyj jego dopełnienia
 - s likwiduj duplikaty znaków ze *zbióru_znaków1* (znaczenie tej opcji)
 - d usuń znaki ze *zbióru_znaków1*
 - ... (patrz dokumentacja)

Operacje na plikach tekstowych w systemie Linux

wybrane programy pakietu textutils: `tr` – przykłady

przykłady:

- ▶ zamiana wielkich liter na małe

```
tr A-Z a-z
```

lub

```
tr '[:upper:]' '[:lower:]'
```

- ▶ usuwanie pustych linii

```
tr -s '\n'
```

- ▶ stworzenie listy słów zawartych w tekście (zastąpienie wszystkich ciągów znaków różnych od liter jednym znakiem nowej linii)

```
tr -cs '[:alpha:]' '\n'
```

Operacje na plikach tekstowych w systemie Linux

grep

grep [opcje] wzorzec [plik]...

- ▶ wybieranie linii pasujących do wzorca (domyślnie tylko one są wypisywane na wyjście)
- ▶ opcje:
 - c zliczaj linie pasujące do wzorca zamiast je drukować
 - v drukuj niepasujące linie zamiast pasujących
 - E interpretuj wzorzec jako *rozszerzone/extended* wyr. reg. (UWAGA: domyślnie *podstawowe/basic*)
 - ... (patrz dokumentacja)

Operacje na plikach tekstowych w systemie Linux

`sed` – edytor wsadowy

`sed`

- ▶ `sed` przetwarza plik zgodnie z zadaniem skryptem linia po linii
 - ▶ skrypt może być podany w linii polecenia lub w pliku
 - ▶ `sed` przechodzi plik tylko raz i może być używany w roli filtra
 - ▶ skrypt jest sekwencją komend oddzielonych średnikami lub znakami nowej linii
 - ▶ komenda składa się z (opcjonalnego) adresu, bądź pary adresów i komendy, oddzielonych odstępem
 - ▶ obszarem działania polecenia jest linia tekstu
 - ▶ adres ogranicza zakres stosowalności komendy do wybranych linii
- ▶ opcje:
 - r używaj odmiany *rozszerzonej/extended* wyrażeń regularnych (UWAGA: domyślnie *basic*)
 - n nie drukuj domyślnie linii na wyjście
 - ... (patrz dokumentacja)

Operacje na plikach tekstowych w systemie Linux

sed – komendy

[adres1 , adres2]] funkcja [argumenty]

► adres

<i>n</i>	numer linia numer <i>n</i>
\$	ostatnia linia
/wzorzec/	linie pasujące do wzorca
adres !	negacja adresu
...	(patrz dokumentacja)

► funkcja

s /wzorzec/ podstawienie /	zastąp pierwsze wystąpienie napisu pasującego do wzorca przez podstawienie
s /wzorzec/ podstawienie /g	j.w., ale zastąp wszystkie wystąpienia
d	wymaż linię
...	(patrz dokumentacja)

Operacje na plikach tekstowych w systemie Linux

sed – przykłady

- ▶ najprostsza postać wywołania: **sed** skrypt
- ▶ przykłady:
 - ▶ usuń wszystkie znaki odstepu na końcach linii
`sed -r 's/[]+$' //`
 - ▶ zamień wszystkie wystąpienia słowa *alfa* na *beta*, ale tylko w liniach nie rozpoczynających się od znaku *#*.
`sed -r '/^#/! s/alfa/beta/g'`

Operacje na plikach tekstowych w systemie Linux

łączenie filtrów

- ▶ łącząc omówione programy za pomocą potoków możemy dokonywać całkiem skomplikowanych operacji na tekście
- ▶ schemat:

```
cat plik_we | filtr1 | filtr2 | ... | filtrn > plik_wy
```

Operacje na plikach tekstowych w systemie Linux

łączenie filtrów: obliczanie liczby słów w tekście

```
cat text | tr -sc '[:alpha:]' '\n' | wc -l
```

```
Ala ma 22 koty.  
Ola ma  
44 psy.  
Karol nie ma  
kota.
```

→

```
Ala  
ma  
koty  
Ola  
ma  
psy  
Karol  
nie  
ma  
kota
```

→

```
10
```

Operacje na plikach tekstowych w systemie Linux

łączenie filtrów: wypisanie 3 najczęstszych słów występujących w tekście

```
cat text | tr -sc '[:alpha:]' '\n' | sort | uniq -c \
| sort -nr | head -3 | sed -r 's/^ *[0-9]+ +//'
```

