Pracownia Technik Obliczeniowych

Przetwarzanie plików tekstowych – textutils

Paweł Daniluk

Wydział Fizyki

Wiosna 2015

System operacyjny GNU

GNU (GNU is Not Unix) to podobny do Unixa system operacyjny składający się w całości z wolnego oprogramowania (*free software*)

System operacyjny GNU

GNU (GNU is Not Unix) to podobny do Unixa system operacyjny składający się w całości z wolnego oprogramowania (*free software*)

Free as in speech, not free as in beer.

System operacyjny GNU

GNU (GNU is Not Unix) to podobny do Unixa system operacyjny składający się w całości z wolnego oprogramowania (*free software*)

Free as in speech, not free as in beer.

System operacyjny składa się z jądra i programów narzędziowych. Hurd – jądro GNU nie jest gotowy. Dlatego używa się Linuxa. Narzędzia GNU są dostępne również na innych systemach operacyjnych (nawet na Windows).

Niektóre elementy systemu GNU

- GNU Compiler Collection (GCC)
- GNU Binary Utilities (binutils)
- powłoka bash
- biblioteka standardowych funkcji języka C GNU C Library (glibc)
- GNU Coreutils (coreutils)

GNU Coreutils – trzy rodziny programów narzędziowych

fileutils

chcon, chgrp, chown, chmod, cp, dd, df, dir, dircolors, install, ln, ls, mkdir, mkfifo, mknod, mktemp, mv, rm, rmdir, shred, sync, touch, truncate, vdir

textutils

base64, cat, cksum, comm, csplit, cut, expand, fmt, fold, head, join, md5sum, nl, od, paste, ptx, pr, sha1sum, sha224sum, sha256sum, sha384sum, sha512sum, shuf, sort, split, sum, tac, tail, tr, tsort, unexpand, uniq, wc

shellutils

arch, basename, chroot, date, dirname, du, echo, env, expr, factor, false, groups, hostid, id, link, logname, nice, nohup, pathchk, pinky, printenv, printf, pwd, readlink, runcon, seq, sleep, stat, stty, su, tee, test, timeout, true, tty, uname, unlink, uptime, users, who, whoami, yes

Przetwarzanie potokowe

Zamiast tworzyć skomplikowane programy do wyspecjalizowanych zadań, lepiej posiadać małe narzędzia i metodę łączenia ich w bardziej złożone konstrukcje.

Przetwarzanie potokowe

Standardowe wyjście jednego programu jest przekierowywane na standardowe wejście programu następnego w potoku.

- | – przekierowuje wyjście jednego programu na wejście drugiego
- >, >> – przekierowuje wyjście do pliku
- < – przekierowuje plik na wejście
- 2> – przekierowuje standardowy błąd do pliku
- 2>&1 – przekierowuje standardowy błąd na standardowe wyjście

Przetwarzanie potokowe – przykład

Zadanie

Ile w plików w bieżącym katalogu ma w nazwie napis “ala”.

```
ls -1 | grep ala | wc -l
```

Przetwarzanie potokowe – przykład

Zadanie

Ile w plików w bieżącym katalogu ma w nazwie napis “ala”.

```
ls -1 | grep ala | wc -l
```

`ls -1`

Zwraca nazwy plików w bieżącym katalogu po jednej w każdym wierszu.

Przetwarzanie potokowe – przykład

Zadanie

Ile w plików w bieżącym katalogu ma w nazwie napis “ala”.

```
ls -1 | grep ala | wc -l
```

`ls -1`

Zwraca nazwy plików w bieżącym katalogu po jednej w każdym wierszu.

`grep ala`

Wybiera wiersze z napisem `ala`.

Przetwarzanie potokowe – przykład

Zadanie

Ile w plików w bieżącym katalogu ma w nazwie napis “ala”.

```
ls -1 | grep ala | wc -l
```

`ls -1`

Zwraca nazwy plików w bieżącym katalogu po jednej w każdym wierszu.

`grep ala`

Wybiera wiersze z napisem `ala`.

`wc -l`

Podaje liczbę linii.

Przeglądanie plików

Concatenate – `cat`

Wyprowadza na stardardowe wyjście zawartość podanych plików.

`head`, `tail`

- `head -n <i>` – wyprowadza na standardowe wyjście *i* początkowych linii
- `tail -n <i>` – wyprowadza na standardowe wyjście *i* końcowych linii

`tail -f` pozwala śledzić na bieżąco przyrosty pliku.

Operacje na kolumnach

cut

Wycina kolumny z pliku

- `cut -c list [file ...]` – wycina kolumny znaków określone listą
- `cut -f list [-d delim] [file ...]` – wycina pola oddzielone znakiem `delim` (domyślnie białymi znakami)

`paste [-s] [-d list] file ...`

Skleja linie podanych plików. Opcja `-d` pozwala określić separatory. `-` oznacza standardowe wejście.

Przykład

```
$> seq 9 | paste - - -
1 2 3
4 5 6
7 8 9
```

Operacje na kolumnach

`join [-a file_number | -v file_number] [-1 field] [-2 fi`

Skleja linie, które mają jednakową wartość wskazanych pól.

- `-1`, `-2` – pozycje pól, według których następuje sklejaniem, w pierwszym i drugim pliku
- `-a` – dołącza niedopasowane wiersze ze wskazanego pliku (“złączenie zewnętrzne”)
- `-v` – zwraca tylko niedopasowane linie

Uwaga

Sklejanie odbywa się w porządku, w jakim wiersze występują w plikach.

Proste zamiany

expand, unexpand

Zamieniają tabulacje na spacje i *vice versa*.

Translate – tr

- `tr string1 string2` – zamienia wystąpienia znaków ze `string1` na odpowiadające im znaki ze `string2`.
- `tr -d string1` – usuwa wystąpienia znaków ze `string1`.
- `tr -s string1` – usuwa powtórzenia znaków ze `string1`.

Operacje na wierszach

`sort`

Sortuje wiersze w pliku. Niektóre opcje:

- `-r` – odwraca kolejność (sortowanie malejące)
- `-n` – sortowanie numeryczne (domyślnie sortuje leksykograficznie)
- `-k` – kolumna, według której odbywa się sortowanie
- `-u` – pomiń powtarzające się wiersze

`uniq`

Zwraca niepowtarzające się wiersze. Niektóre opcje:

- `-c` – dodaje do każdej linii liczbę powtórzeń
- `-d` – zwraca tylko powtarzające się linie
- `-u` – zwraca tylko niepowtarzające się linie
- `-f num` – pomija pierwsze `num` pól
- `-s chars` – pomija pierwsze `chars` znaków

Przeszukiwanie plików – grep

Protoplasta programów narzędziowych.

Możliwości

- przeszukuje podane pliki lub standardowe wejście
- wspiera wyrażenia regularne
- zwraca linie pasujące lub niepasujące do wzorca, albo nazwy plików
- wyrażenia regularne
- podawanie listy wzorców w pliku
- rekursywne przeszukiwanie katalogów

grep – przykłady

- `grep ala lista.txt` – wiersze zawierające napis "ala" w pliku
- `grep ^ala lista.txt` – wiersze zaczynające się od "ala"
- `grep -v ala` – wiersze nie zawierające napisu
- `grep -l ala *` – pliki zawierające napis

Ćwiczenie

Cel

Dla każdej rodziny białek w bazie SCOP znaleźć reprezentanta o najwyższej jakości.

Dane źródłowe

- SCOP – `http://scop.berkeley.edu/downloads/parse/dir.des.scope.2.05-stable.txt`
- ASTRAL – `http://scop.berkeley.edu/astral/subsets/?ver=2.05&get=sc&seqOption=0&item=verbose&cut=sp`

Kryterium oceny ćwiczenia są pliki zdeponowane w repozytorium SVN.

Zadanie 0

Pobierz dane wejściowe.

Podpowiedź

Użyj programu `wget`, albo `curl`.

Adres zawierający znaki & warto umieścić w cudzysłowach.

Zadanie 1

Wybierz wszystkie nazwy domen z obydwu plików i je porównaj.

Zadanie 1

Wybierz wszystkie nazwy domen z obydwu plików i je porównaj.

Podpowiedzi

- Puste linie można usunąć poleceniem `grep -v '^$'`

Zadanie 1

Wybierz wszystkie nazwy domen z obydwu plików i je porównaj.

Podpowiedzi

- Puste linie można usunąć poleceniem `grep -v '^$'`
- Warto posortować pliki wynikowe.

Zadanie 2

Sprawdź czy listy domen są jednakowe. Jeżeli nie są znajdź różnice.

Zadanie 2

Sprawdź czy listy domen są jednakowe. Jeżeli nie są znajdź różnice.

Podpowiedzi

- Najpierw warto porównać liczbę wierszy (`wc -l`).

Zadanie 2

Sprawdź czy listy domen są jednakowe. Jeżeli nie są znajdź różnice.

Podpowiedzi

- Najpierw warto porównać liczbę wierszy (`wc -l`).
- Do porównywania posortowanych plików znakomicie nadaje się `join`.

Zadanie 3

Przygotuj pliki zawierające odpowiednio nazwę domeny i jej rodzinę SCOP oraz nazwę domeny i jej score ASTRAL.

Zadanie 3

Przygotuj pliki zawierające odpowiednio nazwę domeny i jej rodzinę SCOP oraz nazwę domeny i jej score ASTRAL.

Podpowiedzi

- Jeszcze raz przyda się `cut`.

Zadanie 3

Przygotuj pliki zawierające odpowiednio nazwę domeny i jej rodzinę SCOP oraz nazwę domeny i jej score ASTRAL.

Podpowiedzi

- Jeszcze raz przyda się `cut`.
- Można go użyć kreatywnie, żeby pozbyć się).

Zadanie 3

Przygotuj pliki zawierające odpowiednio nazwę domeny i jej rodzinę SCOP oraz nazwę domeny i jej score ASTRAL.

Podpowiedzi

- Jeszcze raz przyda się `cut`.
- Można go użyć kreatywnie, żeby pozbyć się).
- Do sortowania po konkretnej kolumnie służy opcja `-k`.

Zadanie 4

Zrób plik zawierający dla każdej domeny informację o rodzinie i score ASTRAL. Posortuj malejąco według score.

Zadanie 4

Zrób plik zawierający dla każdej domeny informację o rodzinie i score ASTRAL. Posortuj malejąco według score.

Podpowiedzi

- Jeszcze raz przyda się `join`.
- Do sortowania numerycznego służy opcja `-n`. Opcja `-r` odwraca kolejność.

Zadanie 5

Wybierz reprezentanta dla każdej rodziny.

Zadanie 5

Wybierz reprezentanta dla każdej rodziny.

Podpowiedzi

- Jeszcze raz przyda się `join`.

Zadanie 5

Wybierz reprezentanta dla każdej rodziny.

Podpowiedzi

- Jeszcze raz przyda się `join`.
- Do sortowania numerycznego służy opcja `-n`. Opcja `-r` odwraca kolejność. Opcja `-s` wymusza sortowanie stabilne (elementy równe nie są zamieniane miejscami).

Zadanie 5

Wybierz reprezentanta dla każdej rodziny.

Podpowiedzi

- Jeszcze raz przyda się `join`.
- Do sortowania numerycznego służy opcja `-n`. Opcja `-r` odwraca kolejność. Opcja `-s` wymusza sortowanie stabilne (elementy równe nie są zamieniane miejscami).
- `uniq -f` wieńczy dzieło.