Unix-like Data Processing 2025(?)

Finding and using simple (Unix-like) shell commands to perform data processing tasks

Oleks Shturmov (https://oleks.info, oleks@oleks.info)

Slides & Data:

https://codeberg.org/oleks/uldp25

https://uldp25.osvita.io

Finding and using simple (Unix-like) shell commands to perform complex data processing tasks

Contents

1. Introductions
2. Unix-like? (Unix Philosophy)
3. Case 1: Web Access Logs
4. Case 2: Word Count
5. Outro

Introductions

About Oleks

- Born in **Kyiv, Ukraine** in 1990
- Moved to Denmark at the age of 10
- DIKU alumni (**BSc** 2009-2013, **MSc** 2012-2015)
- **Ex**-DIKU **PhD student** (2015-2017)
- **Ex**-Datalogisk **K@ntineforening formand og kasserer** (2013-2021-ish)
- **Ex**-DIKU **Assistant Lecturer** (2022-2024)
- **PhD student** at the University of Oslo (2017-present)
- Research Interests:
  - Programming Language Design and Implementation
  - Multi-agent Systems
  - Teaching

Oleks - The Man Behind

Some of the stuff I worked on at / in collaboration with DIKU:

- Original designs of the **Software Development** course
- Original designs of the **Computer Systems** course
- Elements of the **Advanced Programming** course
- Original designs of **OnlineTA** (e.g., as used in Advanced Programming)
- **staffeli** - a Python-based REST API wrapper for Canvas
- **remarks** - a DSL for marking student work

Unix-Like Data Processing @ DIKU

- 2015: Akademiet for Talentfulde Unge (1 day course)
  - Oleks with Morten Brøns Pedersen (br0ns from pwnies)
- 2016: https://github.com/DIKU-EDU/uldp16 (1 week course)
  - Oleks with Troels Henriksen
  - Lecture Notes: https://github.com/DIKU-EDU/uldp16/blob/master/material/unix-like-data-processing.md
  - YouTube (search “uldp16”):
    - https://www.youtube.com/results?search_query=uldp16
- 2017: https://github.com/DIKUNIX/uldp17 (1 week course)
  - Oleks with Troels Henriksen
  - Lecture Notes: https://github.com/DIKUNIX/uldp17/tree/master/notes
  - Also here: https://uldp25.osvita.io/notes/

Your Turn!

Hello, Neighbour!

Turn to your neighbour, and ask:

1. Their name, if you don't already know them
2. What they most like about the programming language they chose (i.e., C#, F#, or Python)
3. What they expect to get out of today's session

Time: **3 min**

Questionnaire

https://nettskjema.no/a/553804

Time: **10 min**

Discuss with your neighbour as you progress.

Unix-like? (Unix Philosophy)

Unix-like?

- “Unix” colloquially refers to a family of operating systems designed at Bell Labs in 1970s
- Most notorious descendants: Linux, BSD, macOS, Android, iOS
- Most notably not a descendant: Windows

Unix Philosophy

Peter H. Salus in A Quarter-Century of Unix (1994):

1. Write programs that **do one thing and do it well**.
2. Write programs to work together.
3. Write programs to handle text streams, because **text is a universal interface**.

Unix Philosophy

Doug McIlroy in the Bell System Technical Journal (1978):

1. Make each program **do one thing well**. To do a new job, build afresh rather than complicate old programs by adding new “features”.
2. **Expect the output of every program to become the input to another**, as yet unknown, program. Don’t clutter output with extraneous information. Avoid stringently columnar or binary input formats. Don’t insist on interactive input.
3. **Design and build software, even operating systems, to be tried early**, ideally within weeks. Don’t hesitate to throw away the clumsy parts and rebuild them.
4. **Use tools in preference to unskilled help** to lighten a programming task, even if you have to detour to build the tools and expect to throw some of them out after you’ve finished using them.

Case 1: Web Access Logs

A Web Access Log

https://uldp25.osvita.io/access.txt

- Access log of https://uldp25.osvita.io/
- Especially crafted to be read as CSV (Character-Separated Values)
- Format:

```
$remote_addr"$time_iso8601"$request"$http_user_agent`
```

- Example line:

```
167.99.182.39"2025-10-07T11:08:20+00:00"GET /login.action HTTP/1.1"Go-http-client/1.1
```

- Collected over less than 24 hours of having a web server up

Case 2: Word Count

Word Count

E-mail from potential employer at **14:01**:

> For the technical interview, please prepare a solution to the following task. You may use any programming language you like.
>
> **Write a program that counts the number of words in a file.**

Reponse at **14:03**:

> You must be joking.

```
$ cat sherlock.txt | tr '[:upper:]' '[:lower:]' | \
  tr -d '[:digit:]' | sed 's/--/ /g' | \
  tr -d ".,\"'?\!()\;:\`[]\*" | tr ' ' '\n' | \
  tr -s '\n' | sort | uniq | wc -l
```

A Large(ish) Text File

https://uldp25.osvita.io/sherlock.txt

Outro

Review

https://nettskjema.no/a/554992

Time: **5 min**