University of Sussex
Sussex Humanities Lab

Sussex Humanities Lab

CITY UNIVERSITY LONDON

software carpentry

BRITISH LIBRARY

# Library Carpentry

## Week Two: Controlling Data

The Software Sustainability Institute

www.software.ac.uk

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. Exceptions: logos, embeds to and from external sources and direct quotations

# Schedule

Week 1: Some Basics

**Week 2: Controlling Data (with the Shell)**

Week 3: Versioning Data (with Git)

Week 4: Cleaning Data (with Open Refine)

CC BY SA

# Where to go for help

Stickers
Helpers
Sticky notes
github.com/LibraryCarpentry

BY SA

# Week 2: Controlling Data (with the Shell)

17:45-18:25 Basics
18:30-19:25 Counting and Mining
19:30-20:25 Cleaning and Transforming

CC BY SA

The query: *I want to know the number of articles published in 2009 in journals whose title contains the word 'International'*

The file: **2014-01_JA.tsv**

The code: **grep 2009 2014-01_JA.tsv | grep INTERNATIONAL | awk -F'\t' '{print $5}' | sort | uniq -c**

CC BY SA

# Basics (navigation)

pwd

ls -lh

cd

BY SA

# Basics (file interaction)

mv

cp

cat

rm

*

BY SA

# Counting and Mining

wc -w -l

>

grep -c

grep -i

grep -v

grep –w

grep –file=list.txt

BY SA

# Cleaning and Transforming

## Free text exercise

CC BY SA

The Software Sustainability Institute

www.software.ac.uk

@j_w_baker

# Cleaning and Transforming

```
tr ' ' '\n' < gulliver-clean.txt | sort
    | uniq -c > gulliver-final.txt
```

CC BY SA

# Cleaning and Transforming

```
grep 2009 2014-01_JA.tsv |
grep INTERNATIONAL | awk -F'\t'
'{print $5}' | sort | uniq -c
```

CC BY SA

# Where to go next...

Ray & Ray, *Unix and Linux: visual quickstart guide*, 4th edition (2009)

Invaluable reference guide

The Command Line Crash Course

http://cli.learncodethehardway.org/book/

'Learn Code the Hard Way'

Al Sweigart, Automate the Boring Stuff with Python (2015)

http://automatetheboringstuff.com/

'Practical Programming for Total Beginners'

CC BY SA

# Where to go next...

## Coursera Computer Science 101

https://www.coursera.org/course/cs101

'essential ideas of CS for a zero-prior-experience audience'

## Programming for Everybody (Python)

https://www.coursera.org/course/pythonlearn

'The basics of programming computers using Python'

## The Programming Historian

http://programminghistorian.org/

'a bridge between broad ‘getting started’ portals and generic 'programming’ resources'

CC BY SA

# NER (Named Entity Recognition)

**Step 1...** stanford-ner/ner.sh gulliver-noheadfootpunct.txt > gulliver_ner.txt

**Step 2...** sed 's/\/O / /g' < gulliver_ner.txt > gulliver_ner-clean.txt

**Step 3...** egrep -o -f personpattr gulliver_ner-clean.txt | sed 's/\/PERSON//g' | sort | uniq -c | sort -nr > gulliver_ner-pers-freq.txt

CC BY SA

# Next Week

Week 3: Versioning Data (with Git)

You will need a computer
Set-up instructions on Github
Log an issue if you have trouble

See you next week!

CC BY SA

University of Sussex
Sussex Humanities Lab

Sussex Humanities Lab

CITY UNIVERSITY LONDON

software carpentry

BRITISH LIBRARY

# Library Carpentry

## Week Two: Controlling Data

The Software Sustainability Institute

www.software.ac.uk

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. Exceptions: logos, embeds to and from external sources and direct quotations