

CHASE Digital Texts Workshop

James Baker
Lecturer in Digital History

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

Where to go for help

Your neighbour! And if that fails...
Sticky note up!

Schedule

1100-1110 Welcome

1110-1145 Digital Texts as Data

1145-1230 Shell (an approach to working with data)

1230-1315 Lunch

1315-1400 Counting and Mining Texts

1400-1445 Ripping a Text Apart

1445-1530 Finding People and Places

1530-1600 Next Steps

Digital Texts as Data

The why (Adam Crymble: Big Data + Old History)

Digital Texts as Data

The what

- Digitisation of texts (or increasingly born-digital texts)
- Metadata about texts
- Data that represents texts

Digital Texts as Data

The how

- Trends in the metadata about the texts
- Trends in representations of the texts
- Trends in the scans of the texts

... all of which *complements* our close reading

Basics (navigation)

pwd

ls -lh

cd

Basics (file interaction)

mv

cp

cat

rm

*

Counting and Mining

`WC -w -l`

`>`

`grep -c`

`grep -i`

`grep -v`

`grep -w`

Ripping a text apart

Free text exercise



Ripping a text apart

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

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 personpatr
gulliver_ner-clean.txt | sed 's/\PERSON//g'
| sort | uniq -c | sort -nr > gulliver_ner-pers-
freq.txt`

Where to go next...

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

Invaluable reference guide

The Command Line Crash Course

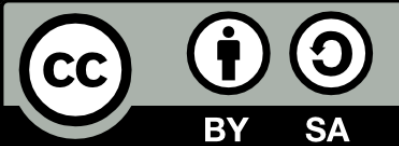
<http://cli.learncodethehardway.org/book/>

'Learn Code the Hard Way'

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

<http://automatetheboringstuff.com/>

'Practical Programming for Total Beginners'



The Software Sustainability
Institute



www.software.ac.uk

@j_w_baker

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'

CHASE Digital Texts Workshop

James Baker
Lecturer in Digital History

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