

The ABC of Computational Text Analysis

Supplements

Alex Flückiger

Faculty of Humanities and Social Sciences
University of Lucerne

25 February 2021

Faculty of Humanities and Social Sciences
University of Lucerne

Purpose

Here I present some stuff that we did not cover in class.

Tasks

- find various ngrams with wildcards
- check gender specific language

what follows she/he or her/his

Forms of Data

- content data

clean, plain text data

preferable as .txt

- metadata ~ information about the actual data

publishing date, authors, source, version

preferable as .csv

Key Word in Context (KWIC)

```
ptx -f -w 50 */*.txt > ptx.txt  
egrep -i "[a-z] word" ptx.txt
```

Select Column in Dataset

```
cut -d\t -f1      # extract the 2nd column from a tab-separated file
```

Extract texts from tsv:

- <http://www.theunixschool.com/2012/05/shell-read-text-or-csv-file-and-extract.html>

Variables

```
echo "Starting program at $(date)"
```


Better Tokenization

- tokenization ~ splitting into words

```
# new, improved approach
cat text.txt | tr -sc "[a-zäöüA-ZÄÖÜ0-9-]" "\n"

# old approach
cat text.txt | tr ' ' '\n'
```

Batch Processing

```
for file in *.txt; do          # loop over all text files
  cat "$file" | pipe commands > "proc_$file"
done
```

Batch Renaming

```
rename " " "_" *.txt    # replace spaces with underscores
# since there are different versions, if this doesn't work try:
# rename 's/ /_/' *.txt
```

```
i=1
for file in *.txt; do    # loop over all text files
  mv -- "$file" "text_$i.txt" # rename each file with a sequential number
  i=$((i+1))
done
```

Data Cleaning

In-class: Exercises I

Grid Example

- text processing
- ⋮
- existing resources
 - creating new resources

COL 1

⋮

COL 2

⋮