The ABC of Computational Text Analysis

Supplements

Alex Flückiger
Faculty of Humanities and Social Sciences
University of Lucerne

April 14, 2023

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)

```
1 ptx -f -w 50 */*.txt > ptx.txt
2 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

```
1 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

```
1 for file in *.txt; do                  # loop over all text files
2  cat "$file" | pipe commands > "proc_$file"
3 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 sequentia
 i=$((i+1))
done
```

Imperfect Data: A Tail of Bias

- **social bias**

  view from somewhere, stereotypes

- **data/archive holes**

  lost, uncollected

- **corpus curation**

  supposition that key-word indicates topic

- **noise in data**

  OCR errors, inconsistent spelling, non-content

👉 **think about the data and mitigate issues**

Outlook: NLP is on Fire 🔥

- supervised machine learning
- **you can do basically anything with modern NLP**

  train on human-annotated data
- **effort, insights and quality may differ**

  for better or worse

Mind your Data

- Who has a voice in your data?

  social context

- bigger is not necessarily better

  more vs. more diverse data

- clean your data thoroughly

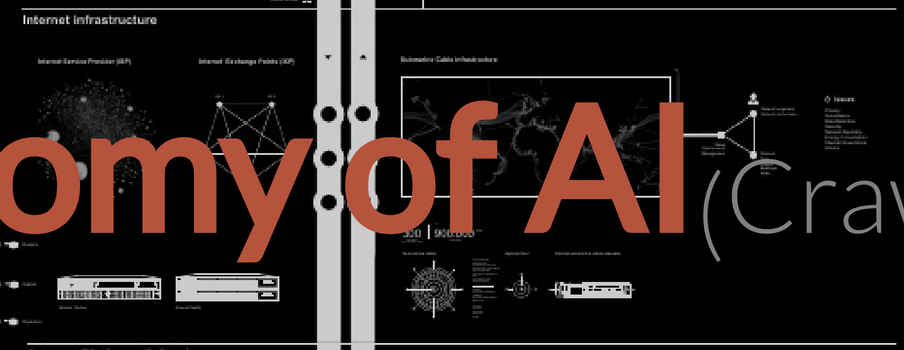
  noisy vs. clean data

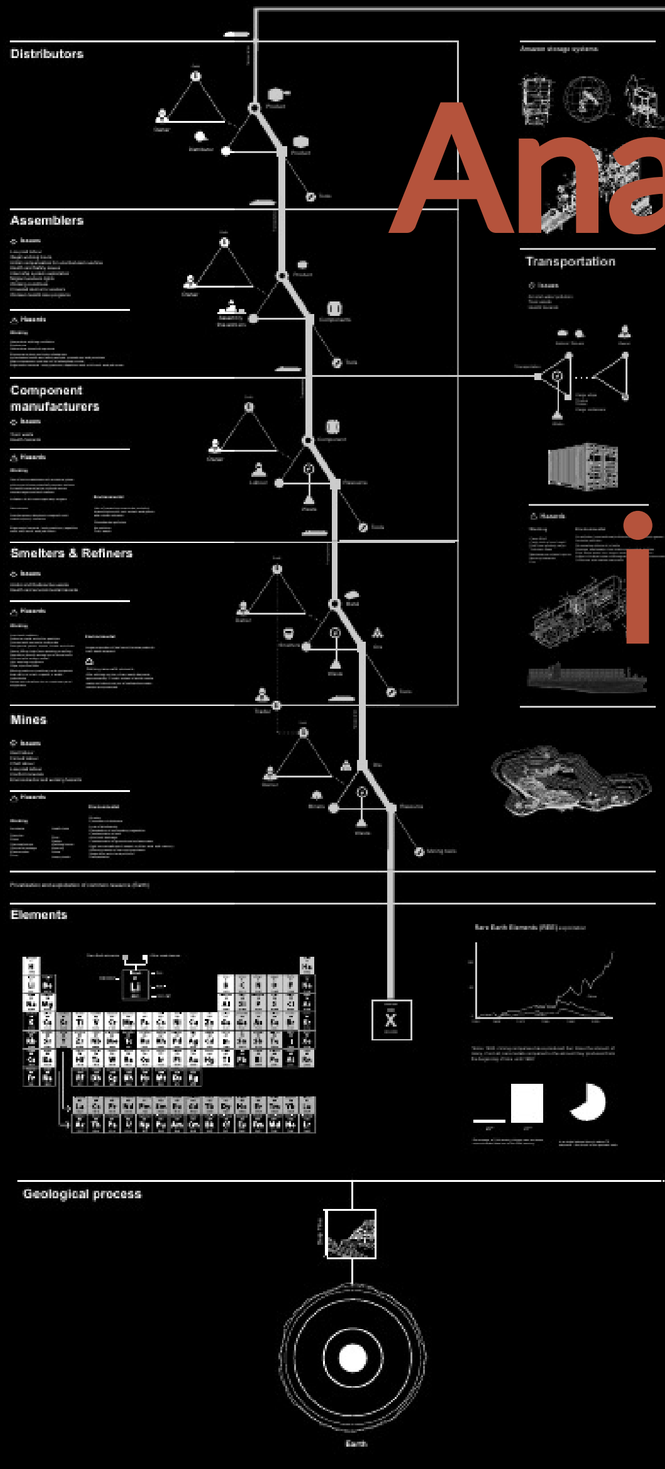
Anatomy of an AI system

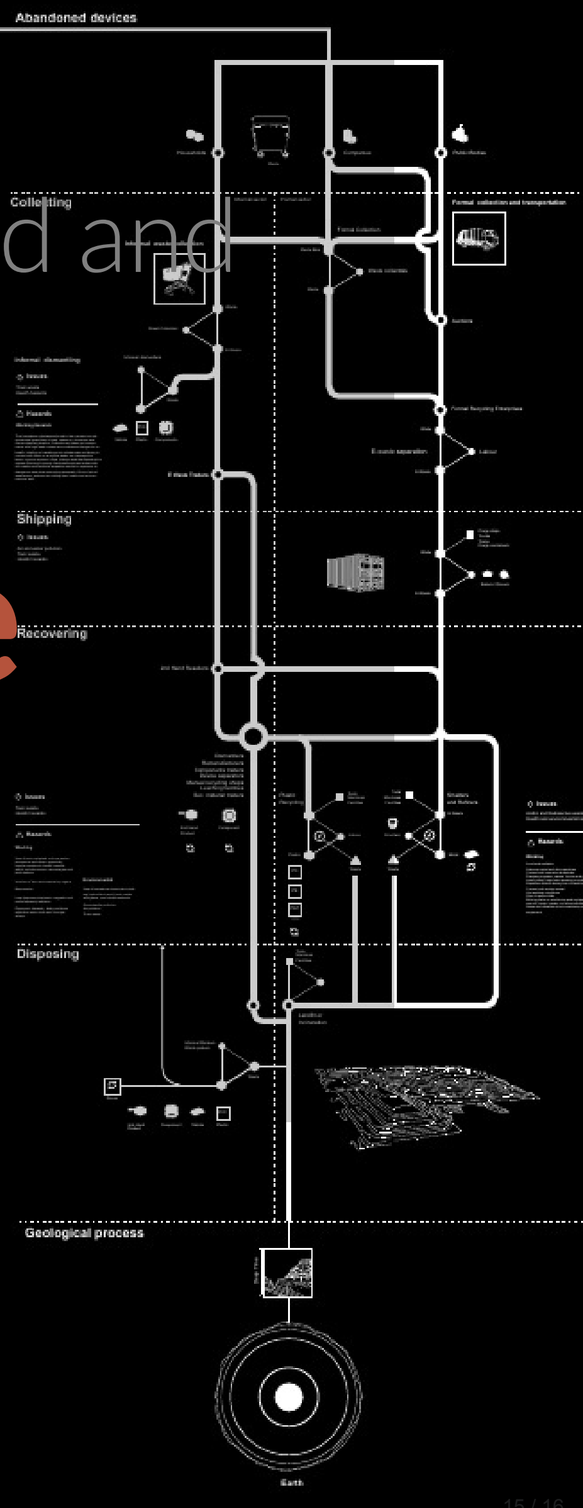
A vast machine resource map

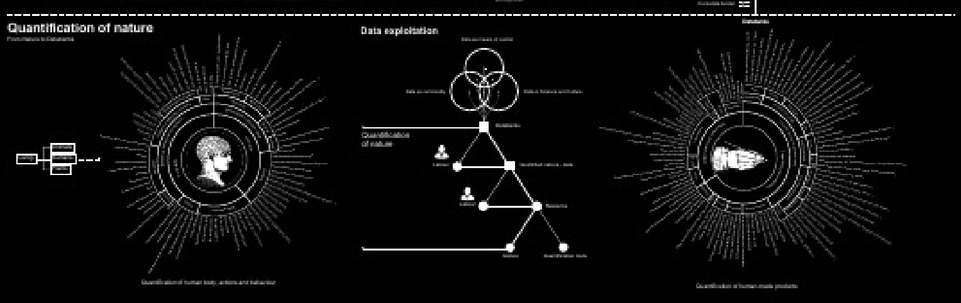
Marx's dialectic of subject and object in economy

Fractal chains of production and exploitation









Anatomy of AI (Crawford and Joler, 2018)

illustrated by the Amazon Echo

Grid Example

:::

- text processing

:::

- existing resources
- creating new resources

COL 1

:::

COL 2

:::

Crawford, Kate, and Vladan Joler. 2018. “Anatomy of an AI System.” Anatomy of an AI System. 2018. http://www.anatomyof.ai.