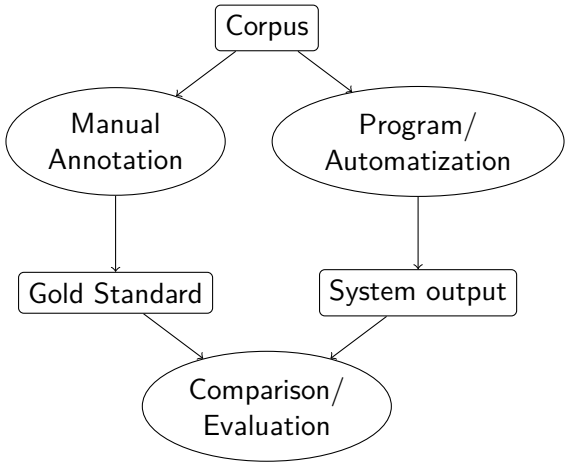
Tools, Resources, Infrastruktur

HS Sprachtechnologie für eine bessere Welt (Winter semester 2022/23)

Nils Reiter,
nils.reiter@uni-koeln.de

January 24, 2023

Landscape


Corpus
Manual Annotation
Program/ Automatization
Gold Standard
System output
Comparison/ Evaluation

Section 1

Annotation

Annotation Task Design Options

Options

- Task Type
  - Span annotation (used for sequence labeling tasks)
  - Text classification (used for text classification tasks)
- Number and kind of annotators
  - Many with little to no training
  - A few, using extensive guidelines
- Language
- Context knowledge

Tool Options

- LimeSurvey: Online questionnaires
- Inception: Web-based annotation
- Paper: Offline questionnaires
  - Ok, but need to be digitized at some point

Tool Options

LimeSurvey

- Hosted by the RRZK on its own hardware
- Open/Closed questionnaires
- Export in various formats: XML, CSV

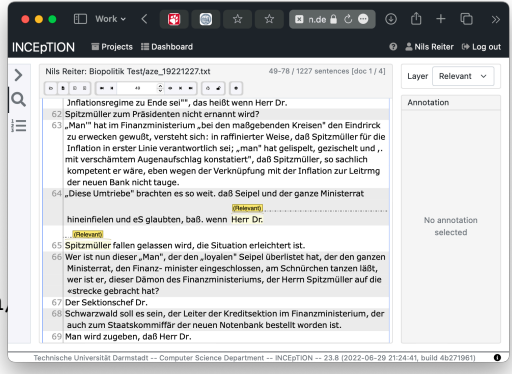
https://rrzk.uni-koeln.de/internetzugang-web/bausteine-fuer-webseiten/online-umfragen/limesurvey

Tool Options

Inception

- Web-based tool for span annotation
- Highly configurable (e.g., one can attach a database and do named entity linking)
- Supports annotation-adjudication-machine-learing-workflow
- Open source, based on Java
- One version hosted by IDH

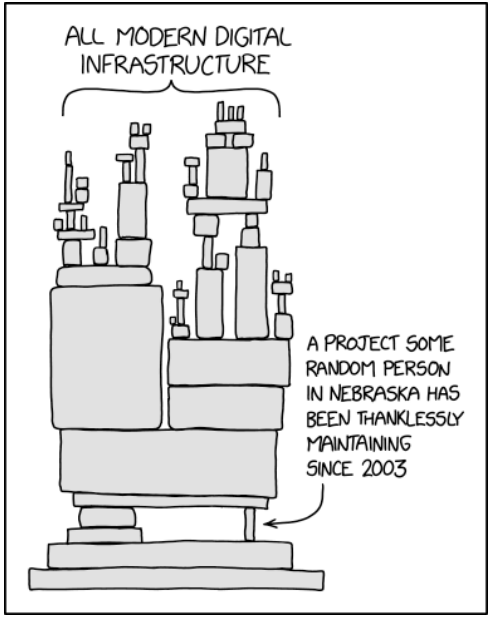
`https://www.spinfo.uni-koeln.de/inception/`



demo

Section 2

Automatization


ALL MODERN DIGITAL INFRASTRUCTURE
A PROJECT SOME RANDOM PERSON IN NEBRASKA HAS BEEN THANKLESSLY MAINTAINING SINCE 2003

Automatization

- Countless ways of doing that
- Core tasks are already taken care of
  - E.g., machine learning, data analysis
- Our tasks: “Glue code” to connect existing software/libraries

Automatization

- Countless ways of doing that
- Core tasks are already taken care of
  - E.g., machine learning, data analysis
- Our tasks: "Glue code" to connect existing software/libraries
- Python: `scikit-learn`, `numpy`, `pandas`, `tensorflow/keras` oder `pytorch`, ggf. `transformers`
- Java: `Weka`, `Deeplearning4j`, `Mallet`
- R: `data.table`, `caret`, `keras`

Core Components

Two GUIs

- Weka — cs.waikato.ac.nz/ ml/weka/
  - Written in Java (extensible with Java, also usable as an API)
  - Open source software, developed at the University of Waikato
  - Make ML experiments
- Orange — orangedatamining.com
  - Written in Python (extensible with Python)
  - Open source software, developed at the University of Ljubljana
  - Design and run data processing workflows

Example

- Data set: ucberkeley-dlab/measuring-hate-speech
- Source: `https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech`
- $39\,565$ comments, 7912 annotators

demo

References I