Center for Humanities Computing Aarhus
A Tutorial and Selected Applications in Arts

Kristoffer L Nielbo
kln@cas.dk
knielbo.github.io

Center for Humanities Computing|chcaa.io
Aarhus University, Denmark
Outline

1 Background
   State of affairs
   Origin

2 Organization
   Core Principles
   Task Domains
   Agile Structure

3 Tutorial
   Submit a ticket
   Project life-cycle

4 Applications
   NFSG Sandbox
   RDM service
   Anonymization service
   Auto-transcription
   Affective computing
   Change detection
   Advertising
A need for Humanities Computing

*the data deluge* is transforming knowledge discovery and understanding in every domain of human inquiry.

*a large part of these data are soft and unstructured* ⇒ to get value from these data, humanities (and social sciences) must utilize automation.

**humanities computing - automatic information processing in the humanities**

Where are we coming from?

from 392 slim nodes, 64 fat nodes, 72 GPU nodes & 40+ researchers to center for humanities computing aarhus
Core principles

Vision
“Develop state-of-the-art research software for arts and humanities in a dedicated Arts Research & Development unit”

RSEs & Developers
“Developers are always on the critical path designing, coding, and unit testing. They have no time for non-development tasks. If these tasks can be given to someone else to validate and cull down to their essence, then the developers can spend their time to produce better code faster.”

Project manager
“A perfectly built doghouse is a waste of time and money if the dog won’t sleep in it.” → ensuring that the stakeholders will sleep in their house
Tasks Domains

1. Support Center – Receive work and send it back
2. Mobility Pool – Work 'upstream' in research groups for short period of time
3. Community of Practice – Facilitate knowledge sharing and common standards
4. Enablers/teachers – Educate and supervise students on analytics methods and platforms

human analytics
APM structure
Project life-cycle

Project can have any size or shape, e.g., simple web page or data hosting service to large-scale AI applications.

- **Project preparation** (submit ticket, build business case, identify stakeholders, project charter)
- **Project start-up** (solution architecture + dev/delivery plan, prioritize requirements, project abstract, communication plan)
- **Development** [ITERATE] (timeboxing with delivery on 1-2 week sprint cadence, hack-days)
- **Deployment** [ITERATE] (assemble, review and deploy)
- **Project finalization** (assessment and transference of rights/project to project owner)
NFSG Sandbox Environment makes *Grundtvigs Værker* accessible to researchers with a Parser and DB that allow sub-corpus construction.

NFSG Sandbox offers a range of services through CHCAA’s cluster og data hosting (eg., entity extraction, syntactic parsing, text re-use &c).
Concept Nucleus is the first sandbox application based on a language model of Grundtvigs Værker.
RDM service for Arts

Define Data Field

Field Name
title

Field data type
String

Brief description of the field
The title of your document in your filesystem

- is text searchable: true
- is in default index: true
- results can be sorted by this: true

> Build Lucene Query

Lucene Query String
(title:"foo bar" AND body:"quick brown") OR title:fox

Logic Plane Editor

```
AND
<<

  title:"foo bar"

>>

body:"quick brown"
```

```
DeepAnon anonymization service

Data De-Identification for Analysis of Unstructured Data

Categories of Surface Features. SF only rely on shallow language and statistical analysis of natural language:

- **Orthographic features** capitalization, word length, common bit information about the word form (contains a digit or not, has uppercase characters inside the word, has punctuation marks inside the word, has digit inside the word, the word is roman or Arabic number) and several regular expressions that describe the common surface characteristics of AGE, DATE, ID and PHONE classes.

- **Frequency information** Extract frequencies of tokens from https://korpus.dsl.dk/resources.html and use the frequency of the token, the ratio of the token’s capitalized and lowercase occurrences, the ratio of capitalized and sentence beginning frequencies of the token.

- **Phrasal information** a forecasted class of several preceding words (online evaluation) and common phrase suffixes (e.g. “Hospital”) seen in the training set.

- **Dictionaries** first names, geographical locations in the DK, names of countries, world’s largest cities; a list containing non-targets tokens from the training data and a list containing non named entity tokens from an external corpus.

- **Contextual information** sentence position, the closest section heading, trigger words from the train text that often precede or follow target word, whether the word fell between quotes, whether the word fell between brackets, the whole context is in uppercase.
Auto-transcription service
Close reading & affective computing

– Combine fractal theory and affective computing to automate assessment of text quality
– solve more “proper” humanities problems that relate to only a few data points (e.g., a single novel)

Change detection in Saxo

- Debate regarding the bipartite composition *Gesta Danorum*
  
  1. is the transition between the old mythical and new historical parts located in book eight, nine, or ten?
  2. is this transition gradual (continuous) or sudden (point-like)?

Effects of advertisements

Articles and advertisements from *De Tijd* (1890-1974) and *De Telegraaf* 1893-1989, \( N \approx 30E^6 \).

**Shaping:** advertisements \(\rightarrow\) articles

**Reflecting:** articles \(\rightarrow\) advertisements

**Complex:** advertisements \(\leftrightarrow\) articles

---

THANKS

kln@au.dk
knielbo.github.io
chcaa.io