CLSINFRA COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE

Computational Literary Studies

Een verkenning van de infrastructurele mogelijkheden met Labs

Julie M. Birkholz & Tess Dejaeghere

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

Digital Research Lab

A lab for text and data mining research on KBR's digitised and born-digital collections.

Tags

Open data, Open Science, Research

The KBR Digital Research Lab serves to facilitate text and data mining research on KBR's diverse, multilingual digitised and born-digital collections. This includes supporting the digital access of textual sources, stimulating the (re)use and research of these digital sources, data and metadata of these collections.

Through a unique long-term cooperation with the Ghent Centre for Digital Humanities, Ghent University, the Lab seeks to serve as a site of **research and experimentation** for providing advice and support for realising these digital projects, specifically the use of computational tools such as **text and data mining**, and **digital**

https://www.kbr.be/en/projects/digital-research-lab/

Overzicht

CLSINFRA

Introductie

CLS infra

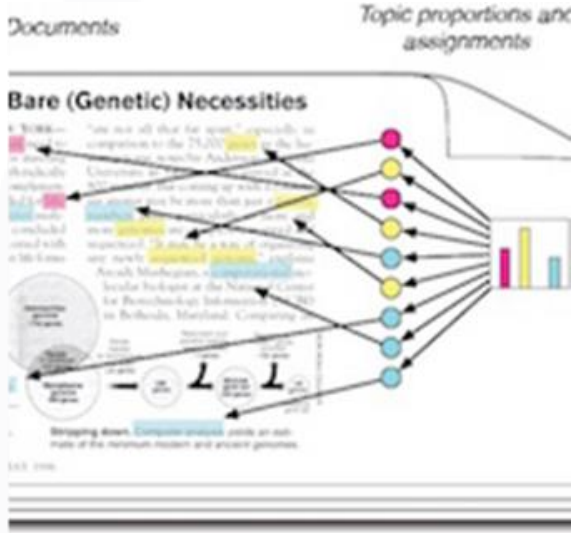
WP8: ons onderzoek

Mogelijkheden

Vragen

CLSINFRA


relationship
over the longer ter
that the supplier who
r best supplier for som
ry of the first project
hort, don't cut ex


Documents
Topic proportions and assignments
Bare (Genetic) Necessities

Introductie: waarom CLS?

- Meer digitalisatie = meer data.
- Uitrol van computationele technieken.
- Meer inzichten genereren in grote collecties d.m.v. “**distant reading**”.

- **Huidige uitdagingen**
  - Steeds meer onderzoek, steeds minder overzicht.
  - Tools zijn niet afgestemd op historische en/of literaire teksten.
  - Veelvoud aan tools beschikbaar.
  - Gebrek aan technische kennis bij historici en letterkundigen.

➢ **Nood aan infrastructuur!**

Introductie: wat is CLS?

- **Wat is Computational Literary Studies (CLS)?**
  - **Definitie**: De toepassing van computationele technieken en algoritmen op literair-historische tekstdata.
  - https://clsinfra.io/

- Enkele toepassingen ter illustratie:
  - De aanleg van corpora
  - annotaties - TEI
  - Auteursherkenning (*authorship attribution*)
  - Topic modelling
  - Emotion detection
  - (Character) netwerkanalyse
  - …

CLS infra: introductie

Computational Literary Studies Infrastructure project (CLS infra)

- **Wat**: infrastructure grant
- **Duur**: 4 jaar
- **Funding**: Europese Commissie
- **Wie**: collectief van Europese universiteiten
  - Ghent University (België)
  - University of Galway (Ierland)
  - Austrian Academy of Sciences (Oostenrijk)
  - Charles University (Tsjechië)
  - University of Potsdam (Duitsland)
  - Trier University (Duitsland)
  - UNED (Madrid)

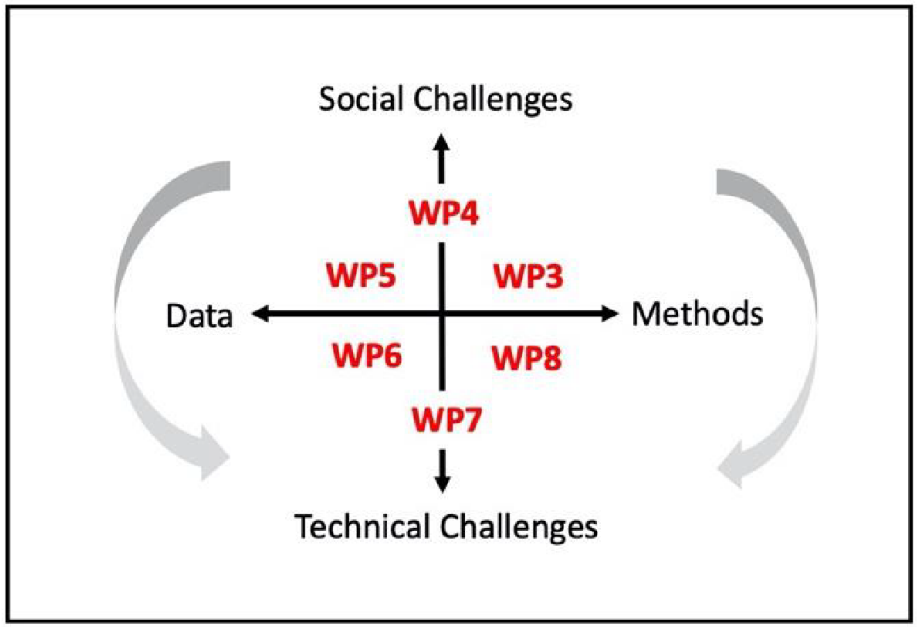
CLS infra: introductie

Computational Literary Studies Infrastructure project (CLS infra)

- **Doelen**:
  - Inzicht in de **noden** van de CLS-gemeenschap.
  - Bouwen aan een gedeelde en duurzame **infrastructuur** om aan digitale literatuurwetenschappen te doen.
  - Nieuwe inzichten genereren in de mogelijkheden van bestaande **tools.**
  - De creatie van **nieuwe tools**.
  - Meer aandacht voor en inzichten in ons gedeelde **Europees cultureel erfgoed over de nationale grenzen heen**.

CLS infra: werkpakketten

- **Hoe:** CLS infra werkpakketten
  - **Sociale uitdagingen**
    - Wie doet aan CLS en waarom?
  - **Data**
    - Wat voor data hebben we nodig?
  - **Methodologieën**
    - Hoe kunnen we aan CLS doen?
  - **Technische uitdagingen**
    - Welke tools zijn beschikbaar en wat zijn hun lacunes?



CLS infra: werkpakketten

- **Hoe:** CLS infra werkpakketten

WP1	Management, coördinatie en innovatieplanning van CLS infra
WP2	Disseminatie en communicatie m.b.t. CLS infra
WP3	Methodologische noden binnen CLS
WP4	Training en vaardigheden voor CLS
WP5	Uitdagingen aangaan m.b.t. datacuratie en selectie
WP6	Data klaarmaken voor CLS-toepassingen
WP7	Maken van ecosystemen en programmeerbare corpora
WP8	**Corpusverrijking en NLP toolchains**
WP9	Transnational Access Fellowships

WP2: Disseminatie en communicatie m.b.t. CLS infra

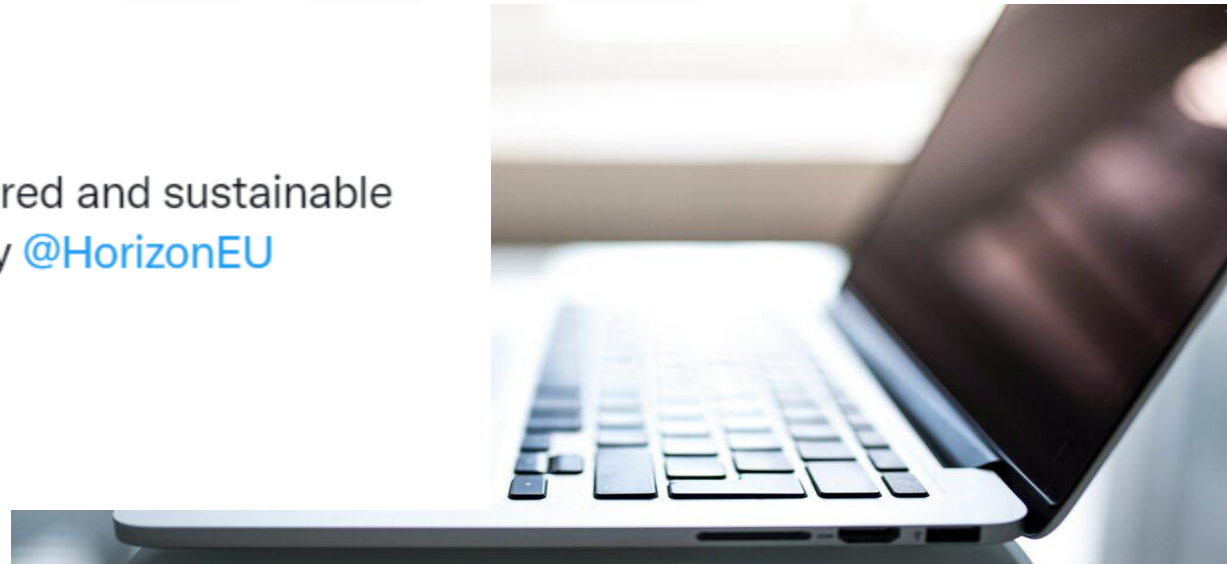
Its objective is to build awareness for CLS INFRA with appropriate measures according to the specific stakeholder groups, to generate materials and instruments by which to promote project activities and results and build the wider project network, and to plan and document project communication, dissemination and exploitation activities.

CLS INFRA
@CLSinfra

Computational Literary Studies Infrastructure: building a shared and sustainable infrastructure for literary studies in the digital age. Funded by @HorizonEU

Europe clsinfra.io Joined November 2020

157 Following **321** Followers



WP3: Methodologische noden binnen CLS

It aims to consolidate the CLS community by establishing user requirements and by documenting and disseminating methodological best practices. It also aims to raise awareness among the wider community of CLS researchers and beyond regarding key issues that might hinder the progressive development and uptake of shared research infrastructures as well as showcase successful areas of application.

WP4: Training en vaardigheden voor CLS

Its function is to inform CLS INFRA development with a baseline analysis of the skills required for CLS, the existing resources for acquiring these skills, and the gaps that exist. It also develops and delivers materials and curricula for the acquisition of CLS skills for researchers with a variety of goals and experience baselines.

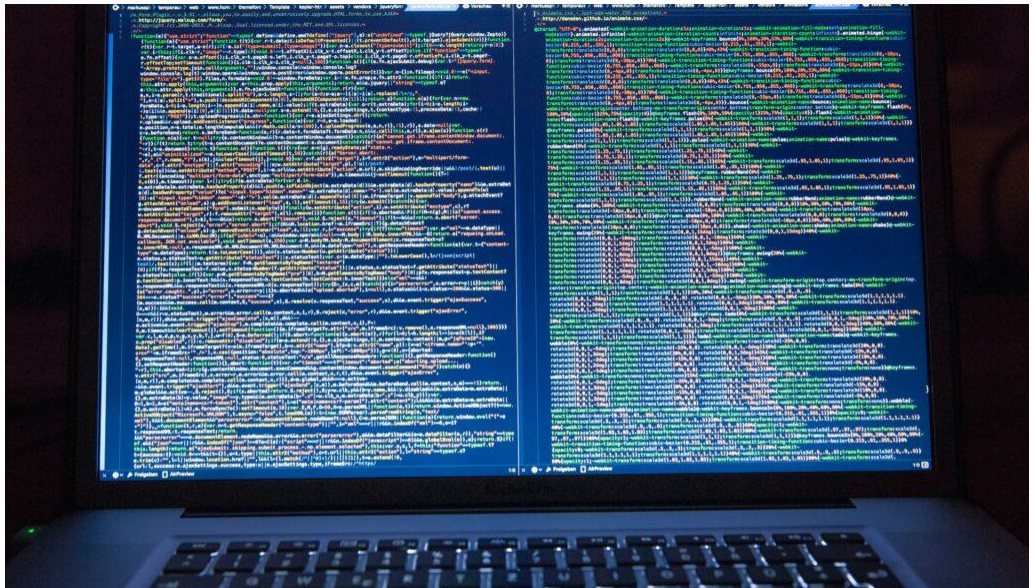
WP5: Uitdagingen aangaan m.b.t. datacuratie en selectie

CLSINFRA

It reviews the data landscape for literary studies, including issues of both what is available and how it can be accessed. It also examines the institutional perspective on literary data sharing for research use, producing case studies to illustrate best practice and policy instruments to inform it.

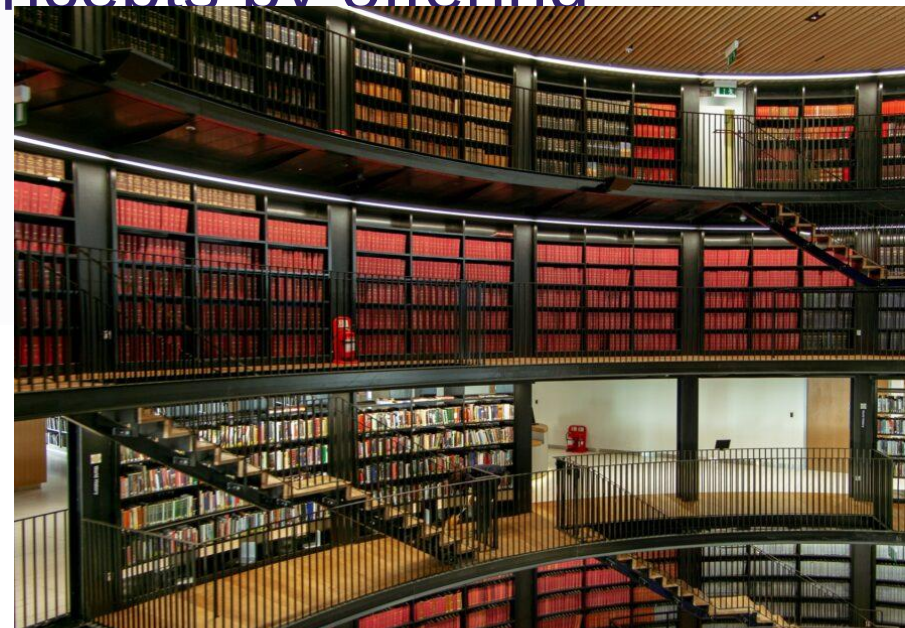
WP6: Data klaarmaken voor CLS-toepassingen

It aims to build an inventory to capture the heterogeneous landscape of data sources and available formats defined by WP 5 in relation to the tools and services in the ecosystem to process these. It will also establish a toolkit to facilitate comparability and integration among corpora, including conversion tools for transformations between existing formats.



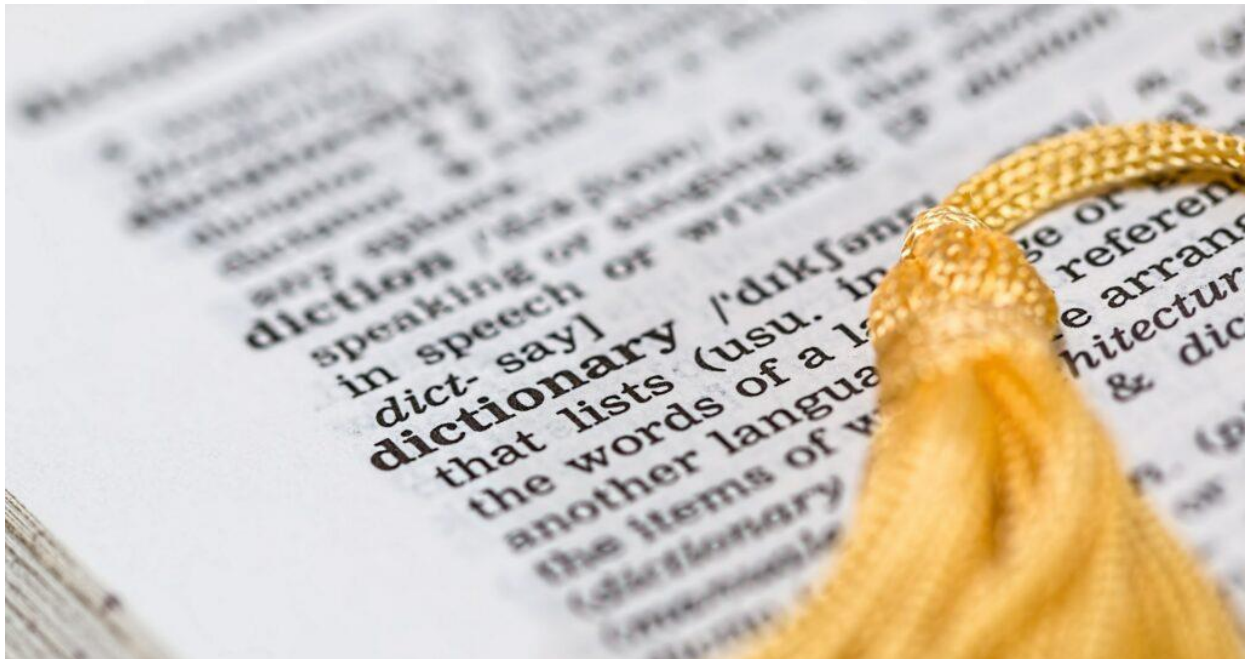
WP7: Maken van ecosystemen en programmeerbare corpora

Its objective is to integrate and federate purpose-specific research data in an open distributed ecosystem to stabilise and facilitate access to data (related to the Linked Open Data cloud), thereby broadening the CLS research community and fostering comparative research on transnational corpora. It also aims to: establish interoperability concepts by offering APIs to tailor the access to information.



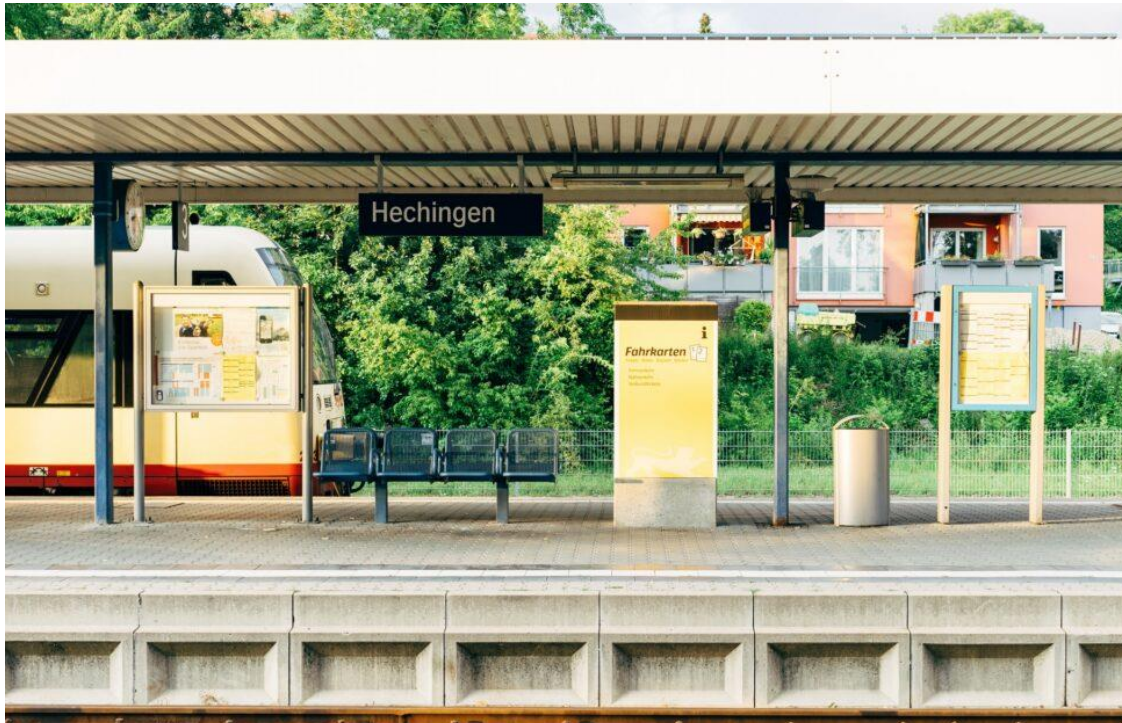
WP8: Corpusverrijking en NLP toolchains

It aims to optimise the availability of fundamental NLP tools within a workflow for literary texts, to support wider use of NLP toolchains in literary research with enhanced multilingual workflows, and to develop a suite of workflows and prototypes to facilitate common research tasks within literary studies using NLP enrichments.



WP9: Transnational Access Fellowships

Its function is to ensure that the CLS INFRA TNA programme is accessible via a transparent and scientifically assured process and to support the needs of all CLS INFRA TNA Fellows before, during and after their access periods.



WP8: ons onderzoek

Doel: Toepassing van Natural Language Processing (NLP) toolchains op literair-historische data.

Natural Language Processing (NLP)

- Onderzoekstak binnen de artificiële intelligentie en de linguïstiek.
- **Definitie:** computationele verwerking van menselijke taal.
- **Soorten taalmodellen:** rule-based, machine learning, deep learning.
- **Populaire toepassingen**:
  - Machinevertaling (Google Translate)
  - Spraakherkenning (Siri, Google Assistant)
  - Chatbots

Zin	Label
“Ik ben blij”	Positief
“Ik ben boos”	Negatief
“ik ben verdrietig”	Negatief

WP8: ons onderzoek

Named Entity Recognition (NER)

- Automatisch extraheren en classificeren van entiteiten uit een tekst.
- Commerciële toepassing in bv. *text summarization* en *question answering systems.*

Voorbeeld: “Julie werkt bij GhentCDH en drinkt graag thee.”

PER ORG

Sentiment Analysis (SA)

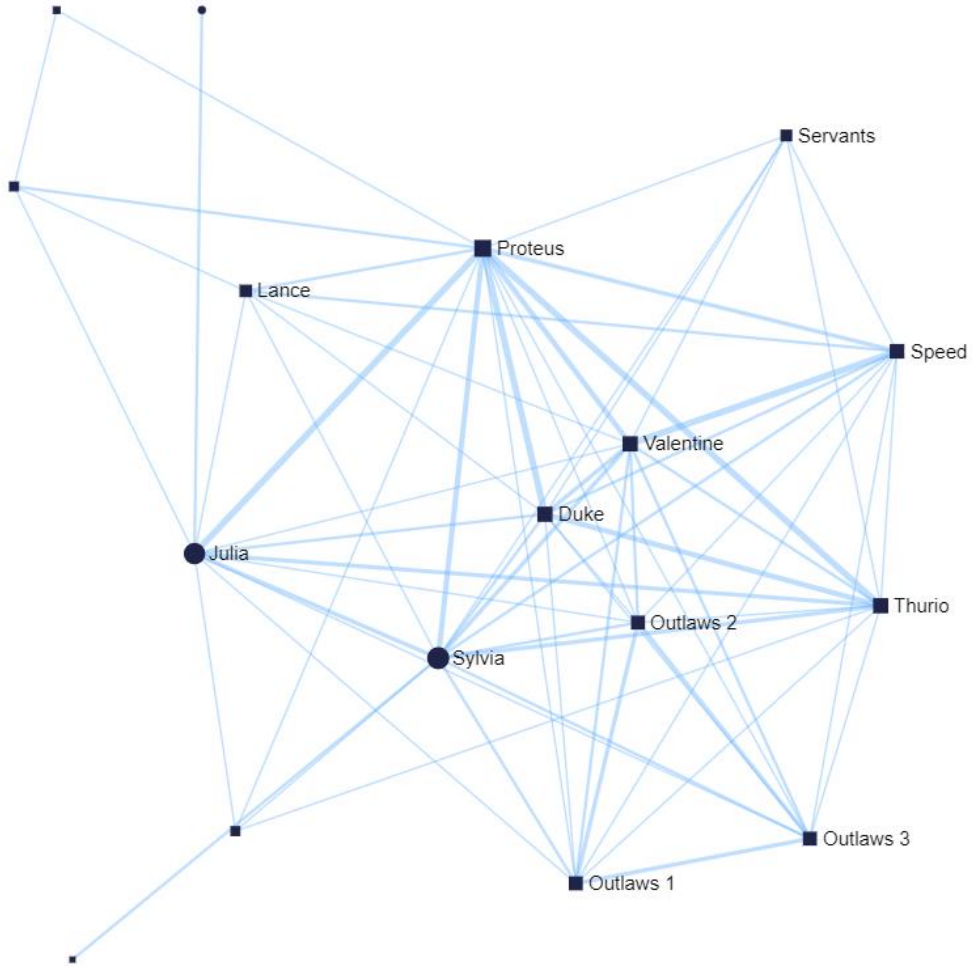
- Automatisch herkennen van sentiment in een tekst.
- Commerciële toepassing in bv. *opinion mining* en *social media mining.*

Voorbeeld: “Julie was erg tevreden met haar warme kop thee!” □ “positief”

WP8: ons onderzoek

Het potentieel van NLP-technieken in onderzoek

- Automatisch netwerkvisualisaties maken van personages in een boek (NER) (Ehrmann et al., 2021).
- Exploratie-interfaces maken voor historische documenten zoals kranten (NER) (Ehrmann et al., 2021).
- Voorspellen of een verhaal goed eindigt (SA) (Zehe et al., 2016).
- Emotionele dynamieken tussen personages automatisch in kaart brengen (SA) (Nalisnick & Baird, 2013).
- …



WP8: ons onderzoek

	Natural Language Processing	(Digital) Humanities
Einddoelen	• Taalmodellen verbeteren. • Linguïstische vraagstukken beantwoorden.	• Meta-tekstuele vraagstukken beantwoorden.
Tools	• Tools getraind op moderne talen. • Tools getraind op reviews.	• Historische teksten • Literair taalgebruik • Meertaligheid • OCR-fouten
Technische kennis	• Kennis over machine learning, deep learning en programmeren.	• Vaak niet voldoende kennis over NLP om modellen aan te passen.

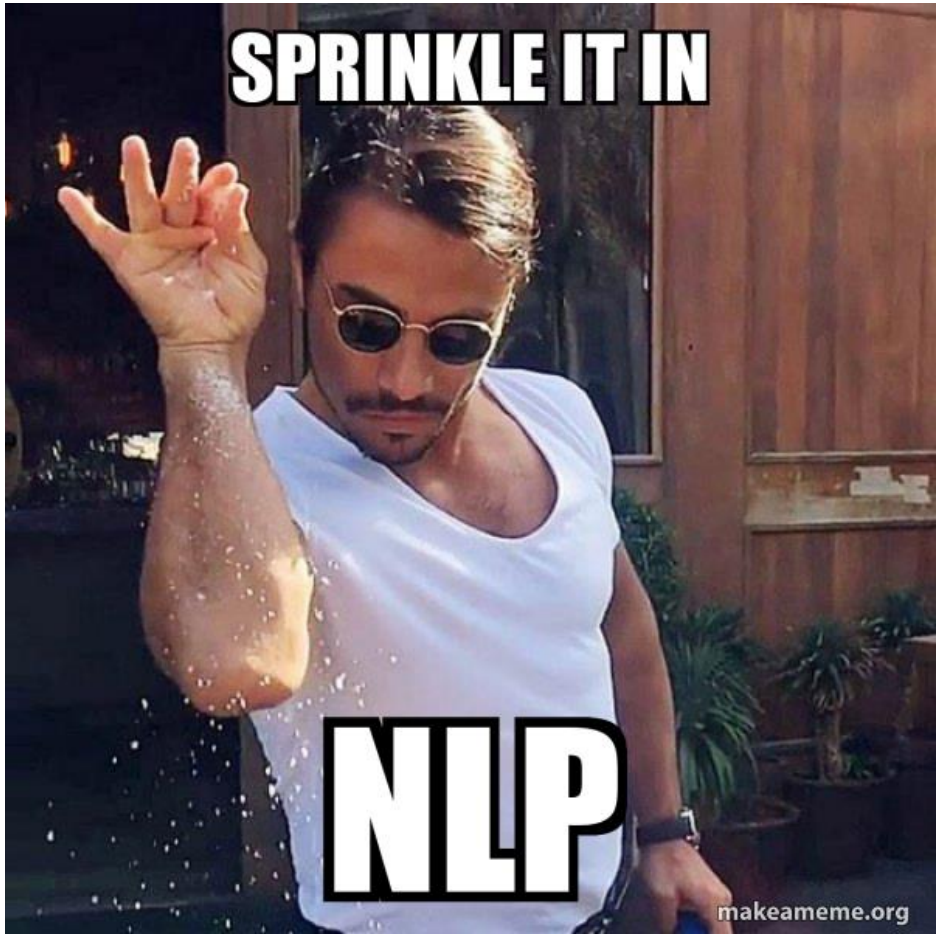
WP8: ons onderzoek

Doelstellingen binnen WP8

! Nood aan evaluatie van off-the-shelf NLP-tools binnen literair-historische context.
! Nood aan evaluatie van de output van NLP-tools binnen literair-historische context.

→ **begrijpelijke**, **transparante**, **herbruikbare** en **duurzame** workflows/toolchains.

- Pipeline voor ***NER*** op literair-historische corpora in Nederlands, Frans, Duits, Tsjechisch en Engels.
- Pipeline voor ***SA*** op literair-historische corpora in Nederlands, Frans, Duits en Engels.



Mogelijkheden: CLS infra

- **Uitwisseling**
  - **TNA (Transnational Access Activities)**
    - **Doel:** Mogelijkheid om een beurs aan te vragen om een CLS-gerelateerd project uit te werken bij één van de deelnemende onderzoeksinstituten!
    - **Data**: 2 keer per jaar, in totaal 6 calls. Duur onderzoeksproject is variabel.
    - **Schrijf je in:** https://clsinfratna.sciencescall.org/
- **Trainings**
  - **Data and Annotation summer school**
    - **Doel:** leer om literaire data te structureren (XML) en te verrijken (NLP)!
    - **Data:** 7-9 juni
    - **Schrijf je in:** https://clsinfra.io/events/training-school/

Vragen? Opmerkingen? Ideeën?

UNIVERSITEIT GENT

Contacteer ons @GhentCDH

Geospatial Humanities
Collaborative database building
Optical Character Recognition
Digital Text Analysis
Algemene ondersteuning bij DH-projecten

@GhentCDH

Lise.foket@ugent.be

Contacteer @CLSinfra

CLS infra project
TNA programma's
Trainings

@CLSinfra

clsinfratna@sciencesconf.org

Bedankt voor jullie aandacht!

Julie M. Birkholz & Tess Dejaeghere

Julie.Birkholz@ugent.be & Tess.Dejaeghere@ugent.be