UNIVERSITEIT GENT

Studiefiche

Vanaf academiejaar 2024-2025

Data- en corpusbeheer met Python: Theorie en toepassingen (A005869)

Cursusomvang *(nominale waarden; effectieve waarden kunnen verschillen per opleiding)*

Studiepunten 5.0 **Studietijd 150 u**

Aanbodsessies in academiejaar 2024-2025

A (semester 2) Nederlands Gent

Lesgevers in academiejaar 2024-2025

Goethals, Patrick	LW22	Verantwoordelijk lesgever
Swaelens, Colin	LW22	Medewerker
Van Hee, Cynthia	LW22	Medelesgever

Aangeboden in onderstaande opleidingen in 2024-2025	stptn	aanbodsessie
Bachelor of Arts in de toegepaste taalkunde: combinatie van ten minste twee talen(afstudeerrichting Nederlands, Duits, taaltechnologie)	5	A
Bachelor of Arts in de toegepaste taalkunde: combinatie van ten minste twee talen(afstudeerrichting Nederlands, Engels, taaltechnologie)	5	A
Bachelor of Arts in de toegepaste taalkunde: combinatie van ten minste twee talen(afstudeerrichting Nederlands, Frans, taaltechnologie)	5	A
Voorbereidingsprogramma tot Master of Arts in de meertalige communicatie: combinatie van ten minste twee talen	5	A
Voorbereidingsprogramma tot Master of Arts in het vertalen: combinatie van ten minste twee talen	5	A

Onderwijstalen

Nederlands

Trefwoorden

Situering

Dit vak bouwt verder op het vak Inleiding Programmeren voor de humanities en gaat specifiek in op Python-ondersteunde programmeertechnieken om data te verzamelen, te beheren en te annoteren in verschillende formaten, te bevragen en weer te geven.
De studenten maken ook kennis met verschillende technologieplatformen ter ondersteuning van corpusannotatie en leren hoe de data te importeren, annoteren en beschikbaar te maken voor analyse.

Inhoud

Python Project Management: GIT, Virtual Envs, PyCharm, Google Colab, API
dataverzameling:
- inlezen van bestaande corpora in verschillende formats
- aanmaken van nieuwe corpora met gebruik van API, webcrawling, eenvoudige GUI

databeheer en databevraging
- data opslaan in verschillende datatypes (json dictionaries, Pickle-bestanden, csv, conllu corpus format)
- inlezen verschillende types gestructureerde bestanden
- sql queries

data-annotatie tools
- WebAnno, Inception

data-uitvoer
- documentgeneratie in verschillende formats: csv, Word, Excel-bestanden
- visualiseringstechnieken

Begincompetenties

Een goede basiskennis Python programmeren (zie inhouden Inleiding Programmeren voor de

humanities)
Vertrouwdheid met technieken uit het domein van de digitale tekstanalyse (studenten dienen zich tegelijkertijd in te schrijven voor het vak Inleiding tot de Digitale Tekstanalyse)

Eindcompetenties

1 Studenten moeten corpora van verschillende bestandstypes kunnen inlezen, opschonen en voorbereiden voor analyse, inclusief tokenisering en omgaan met coderingsproblemen.
2 Bekwaam zijn in het gebruik van Python voor webscraping en corpusbeheer.
3 In staat zijn om corpora te creëren, te annoteren en voor te bereiden voor verdere analyse.
4 In staat zijn om teksten in verschillende formaten uit te voeren en basis corpusstatistieken te visualiseren met behulp van de correcte Python libraries.

Creditcontractvoorwaarde

De toegang tot dit opleidingsonderdeel via creditcontract is open: de student houdt zelf rekening met voorkennis uitgedrukt in begincompetenties

Examencontractvoorwaarde

Dit opleidingsonderdeel kan niet via examencontract gevolgd worden

Didactische werkvormen

Werkcollege

Studiemateriaal

Type: Slides

Naam: slides
Richtprijs: Gratis of betaald door opleiding
Optioneel: nee

Referenties

Vakinhoudelijke studiebegeleiding

Evaluatiemomenten

periodegebonden en niet-periodegebonden evaluatie

Evaluatievormen bij periodegebonden evaluatie in de eerste examenperiode

Schriftelijke evaluatie

Evaluatievormen bij periodegebonden evaluatie in de tweede examenperiode

Evaluatievormen bij niet-periodegebonden evaluatie

Werkstuk

Tweede examenkans in geval van niet-periodegebonden evaluatie

Niet van toepassing

Eindscoreberekening

Werkstuk: codeerproject 40%
Schriftelijk examen: 60%
De studenten kunnen maximaal 9/20 halen indien zij minder dan 8/20 scoren op het schriftelijk examen.