

## Introduction to Machine Learning and Feature Engineering for NLP (A005871)

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

**Studiepunten 5.0** **Studietijd 150 u**

### Aanbodsessies in academiejaar 2026-2027

A (semester 2)	Nederlands	Gent
B (semester 1)	Engels	Gent

### Lesgevers in academiejaar 2026-2027

Hoste, Veronique	LW22	Verantwoordelijk lesgever
Maladry, Aaron	LW22	Medewerker
Lefever, Els	LW22	Medelesgever

### Aangeboden in onderstaande opleidingen in 2026-2027

	stptn	aanbodsessie
<a href="#">Bachelor of Arts in de toegepaste taalkunde: combinatie van ten minste twee talen (afstudeerrichting Nederlands, Duits, taaltechnologie)</a>	5	A
<a href="#">Bachelor of Arts in de toegepaste taalkunde: combinatie van ten minste twee talen (afstudeerrichting Nederlands, Engels, taaltechnologie)</a>	5	A
<a href="#">Bachelor of Arts in de toegepaste taalkunde: combinatie van ten minste twee talen (afstudeerrichting Nederlands, Frans, taaltechnologie)</a>	5	A
<a href="#">Master of Arts in Advanced Studies in Linguistics (afstudeerrichting Natural Language Processing: Theory and Practice)</a>	6	A
<a href="#">Postgraduate Certificate Computer-Assisted Language Mediation</a>	5	B
<a href="#">Vorbereidingsprogramma tot Master of Arts in de meertalige communicatie: combinatie van ten minste twee talen</a>	5	A
<a href="#">Vorbereidingsprogramma tot Master of Arts in het vertalen: combinatie van ten minste twee talen</a>	5	A

### Onderwijstalen

Engels, Nederlands

### Trefwoorden

data-annotatie, machine learning, feature-gebaseerde systemen

### Situering

In deze cursus gaan we in op de architectuur van datagebaseerde machine learning (ML) of zelflerende systemen. De volledige workflow van een machine learning systeem wordt besproken, gaande van annotatie van data tot het trainen en evalueren van verschillende featuregebaseerde machine learning-architecturen.

Deze cursus bouwt verder op het vak NLP and Linguistic Analysis en wordt gedoceerd gedurende 4 lesweken (weken 9-12).

### Inhoud

De volgende onderwerpen komen aan bod:

- Data verrijken met behulp van verschillende types data-annotatie (expertannotatie, games with a purpose, crowdsourcing).
- De betrouwbaarheid van annotaties evalueren aan de hand van verschillende metrieken (Cohen's & Fleiss's Kappa, Krippendorff's Alpha).
- Verschillende tekstenmerken (lexicaal, syntactisch, semantisch) automatisch uit tekst extraheren en omzetten naar featurevectoren en deze ook verder uitbreiden met informatie uit externe bronnen (woordenlijsten, corpora, treebanks, taalkundige databanken).
- De onderliggende algoritmes en technieken van zelflerende systemen.

- Trainen en evalueren van de verschillende systemen aan de hand van verschillende taaltechnologische toepassingen (bijv. leesbaarheidsvoorspelling, sentimentanalyse).
- Ethische vraagstukken die komen kijken bij datagebaseerde machine learning.

### **Begincompetenties**

Basiskennis NLP en Python.

### **Eindcompetenties**

- 1 Studenten beschikken over basiskennis over en inzicht in zelflerende systemen.
- 2 Studenten beschikken over de nodige praktische kennis en vaardigheden in alle stappen van machine learning, gaande van data-annotatie, bouwen van feature vectoren tot het trainen en evalueren van zelflerende systemen.

### **Creditcontractvoorwaarde**

Dit opleidingsonderdeel kan niet via creditcontract gevolgd worden

### **Examencontractvoorwaarde**

Dit opleidingsonderdeel kan niet via examencontract gevolgd worden

### **Didactische werkvormen**

Werkcollege, Hoorcollege, Zelfstandig werk

### **Toelichtingen bij de didactische werkvormen**

Het interactief karakter van de lessen laat niet toe om kwalitatief hoogstaande opnames te voorzien.

### **Studiemateriaal**

Geen

### **Referenties**

- Daniel Jurafsky & James Martin, "Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition" (ed3).
- Mitchell, T. (1997). "Machine learning". McGraw-hill New York.
- Python Software Foundation. Official Python documentation. <http://www.python.org/doc/>

### **Vakinhoudelijke studiebegeleiding**

Ondersteunend discussieforum via de Ufora-leeromgeving  
Individuele begeleiding na afspraak via e-mail met de lesgever

### **Evaluatiemomenten**

periodegebonden en niet-periodegebonden evaluatie

### **Evaluatievormen bij periodegebonden evaluatie in de eerste examenperiode**

Mondelinge evaluatie

### **Evaluatievormen bij periodegebonden evaluatie in de tweede examenperiode**

Mondelinge evaluatie

### **Evaluatievormen bij niet-periodegebonden evaluatie**

Participatie, Werkstuk

### **Tweede examenkans in geval van niet-periodegebonden evaluatie**

Examen in de tweede examenperiode is mogelijk

### **Toelichtingen bij de evaluatievormen**

Eerste zittijd:

- Aanwezigheid en participatie in de les (10%)
- Werkstuk (30%)
- Mondeling examen waaronder bespreking van werkstuk en toetsing van theorie (60%)

Tweede zittijd:

- Werkstuk (40%)
- Mondeling examen (60%)

### **Eindscoreberekening**

Een combinatie van niet-periodegebonden (40%) en periodegebonden evaluatie (60%).

Eerste zittijd: Om te slagen moet de student gemiddeld 10 of meer behalen, met een minimumscore van 40% voor zowel de niet-periodegebonden als periodegebonden evaluatie. Indien deze minimumscore niet gehaald wordt op elk van beide onderdelen, kan een student maximaal 9/20 halen voor het totaal van het opleidingsonderdeel.

Tweede zittijd: Beide onderdelen moeten opnieuw worden afgelegd. Om te slagen voor het opleidingsonderdeel moet de student gemiddeld 10 of meer behalen, met een minimumscore van 40% voor zowel de niet-periodegebonden als periodegebonden evaluatie. Indien deze minimumscore niet gehaald wordt op elk van beide onderdelen, kan een student maximaal 9/20 halen voor het totaal van het opleidingsonderdeel.

#### **Faciliteiten voor werkstudenten**

Aanwezigheid in de lessen is ten sterkste aanbevolen.

Beperkte mogelijkheid tot feedback per e-mail voor concrete vragen.