UNIVERSITEIT GENT

Studiefiche

Vanaf academiejaar 2024-2025

Natuurlijke taalverwerking (E061341)

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

Studiepunten 6.0 **Studietijd 180 u**

Aanbodsessies en werkvormen in academiejaar 2024-2025

A (semester 2)	Engels	Gent	groepswerk	10.0u
			hoorcollege	35.0u
			practicum	15.0u

Lesgevers in academiejaar 2024-2025

Develder, Chris	TWO5	Verantwoordelijk lesgever
Demeester, Thomas	TWO5	Medelesgever

Aangeboden in onderstaande opleidingen in 2024-2025	**stptn**	**aanbodsessie**
Brugprogramma Master of Science in Bioinformatics(afstudeerrichting Engineering)	6	A
Master of Science in Bioinformatics(afstudeerrichting Engineering)	6	A
Master of Science in Computer Science Engineering	6	A
Master of Science in de informatica	6	A
Master of Science in de ingenieurswetenschappen: computerwetenschappen	6	A
Master of Science in Statistical Data Analysis	6	A
Uitwisselingsprogramma informatica (niveau master)	6	A

Onderwijstalen

Engels

Trefwoorden

Natuurlijke taalverwerking, machinaal leren, artificiële neurale netwerken, statistische methoden

Situering

In verschillende toepassingsdomeinen bestaat een substantieel deel van de gegevens uit tekst in natuurlijke taal, bv. webpagina's, nieuwsartikels, magazines, blogs, tweets, Facebook aankondigingen, tekstberichtjes, etc. Deze puur tekstuele informatie omvat bruikbare info die heel wat waardevoller zou zijn mocht die interpreteerbaar zijn voor verdere verwerking door geautomatiseerde computerverwerking (bv. omzetting naar gestructureerde gegevens in databanken). De waarde van die informatie ontsluiten door het ontwikkelen van technieken die menselijke taal interpreteren aan de hand van computeralgoritmen is precies wat natuurlijke taalverwerking (Eng. Natural Language Processing, NLP) beoogt.
NLP is een onderzoeksdomein dat een combinatie vormt van computerwetenschappen, artificiële intelligentie en taalkunde, en globaal tot doel heeft computers in staat te stellen taken op te lossen die het begrijpen en/of genereren van menselijke taal vereisen. Deze taken die met menselijke taal omgaan, zijn alomtegenwoordig in ons dagelijks leven, en gaan van basiszoektaken (met internet-zoekmachines) tot het automatisch beantwoorden van vragen of automatisch vertalen.
In deze cursus focussen we op het verwerken van taal in schriftelijke, tekstuele vorm (en zullen dus bv. spraakverwerking niet behandelen). Meer specifiek zullen we voornamelijk ingaan op (1) Klassieke NLP, waarnaar ook verwezen wordt met de term Statistiche NLP (SNLP), gezien die hoofdzakelijk bouwt op statistiek en machinaal leren, alsook (2) Neurale NLP, d.w.z. methoden gebaseerd op neurale

netwerken. In SNLP worden computers niet rechtstreeks geprogrammeerd om taal te verwerken, maar zullen ze typisch leren hoe taal te verwerken (of te genereren) gebaseerd op de statistische eigenschappen van een (doorgaans zeer groot) corpus van natuurlijke taal. Meer recente modellen gebaseerd op neurale netwerken zijn meer rechtstreeks data-gedreven, en vermijden doorgaans specifieke feature engineering en/of het expliciet definiëren van subtaken (bv. PoS tagging, dependency parsing) als deelstappen van de uiteindelijke eind-toepassing. Gegeven de extreme snelle evolutie van het domein van neurale NLP, zowel op vlak van onderzoekspublicaties als open source modellen en software, zullen studenten aangemoedigd worden de recente vakliteratuur te raadplegen en hun inzichten te delen met studiegenoten. Meer specifiek, zullen studenten, na een inleiding van de basis-bouwstenen van transformers en de toonaangevende model-architecturen, autonoom een recent onderzoeksartikel lezen en de kernpunten daarvan in een korte presentatie aan hun medestudenten voorstellen.
Het doel van de cursus is om studenten zowel theoretische als praktische kennis bij te brengen van de belangrijkste concepten en technieken in NLP, zodat ze vertrouwd zijn met (1) de belangrijkste NLP-taken kennen (inclusief tekstclassificatie, sequentietagging, syntactische ontleding, taalmodellering, machinevertaling), (2) de essentiële methoden en raamwerken (voor zowel statistische NLP als op neurale netwerken gebaseerde methoden), evenals (3) de praktische implementaties daarvan (inclusief moderne softwaretools, bv. Pytorch).

Inhoud

- Inleiding: wat is NLP, basistaken in NLP, voorbeeld-toepassingen;
- Klassieke NLP:
  - Technieken voor tekstclassificatie (inclusief Naïeve Bayes, logistische regressie)
  - N-gram taalmodellen
  - Sequentie tagging (inclusief Part-of-speech tagging)
  - Constituency en dependency parsing
- Neurale NLP
  - Recurrente sequentie-modellen
  - Neurale taalmodellen
  - Vectoriële woord- en zinsrepresentaties
  - Sequence-to-sequence modellen (inclusief die voor automatische vertaling)
  - Transformers: bouwstenen, toonagevende basismodellen, recente ontwikkelingen
  - Recente paradigmas (bv. pre-training en finetuning, prompt engineering)

Begincompetenties

- Basisprogrammeren in Python
- Basisconcepten machinaal leren:
  - Supervised learning (op basis van train/dev/test sets)
  - Basis neurale netwerken (multilayer perceptron, gradiënt-gebaseerde training)

Eindcompetenties

1 De basis-NLP-taken kennen, alsook methoden om ze op te lossen (bv. (voor) verwerken van tekstdocumenten, taalmodellering, parsing, sequence tagging, tekstclassificatie, sequence-to-sequence taken).
2 Methoden voor NLP-gebaseerde toepassingen kunnen uitleggen, toepassen en evalueren, bv. named entity recognition (NER), automatische vertaling, zinsclassificatie, informatie-extractie.
3 Inzicht hebben in modellen gebaseerd op geleerde vectoriële representaties (gaande van statische woord-embeddings tot vooraf getrainde transformer modellen) en compatibele neurale netwerkbouwstenen, om hiermee specifieke NLP problemen aan te pakken.
4 De verschillende types (bv. intrinsiek vs. extrinsiek) van evaluatie, en de gangbare evaluatiemetrieken begrijpen en kunnen uitleggen.
5 Een NLP-toepassing kunnen implementeren en evalueren met Python.
6 In staat zijn recente vakliteratuur op te volgen en nieuwe ter beschikking gestelde modellen kunnen gebruiken.

Creditcontractvoorwaarde

Toelating tot dit opleidingsonderdeel via creditcontract is mogelijk mits gunstige beoordeling van de competenties

Examencontractvoorwaarde

Dit opleidingsonderdeel kan niet via examencontract gevolgd worden

Didactische werkvormen

Groepswerk, Hoorcollege, Practicum, Zelfstandig werk

Toelichtingen bij de didactische werkvormen

- De cursus wordt aangeboden onder de vorm van wekelijkse theoriecolleges over onderwerpen in Klassieke en Neurale NLP. Aan het einde van het semester worden de studenten in kleine groepen verdeeld om de belangrijkste ideeën van een recent onderzoeksartikel voor te stellen.
- Practica zullen georganiseerd worden als begeleide zelfstudie sessies (d.w.z. met ondersteuning op afstand via MS Teams).

Studiemateriaal

Type: Slides

Naam: cursus slides
Richtprijs: Gratis of betaald door opleiding
Optioneel: nee
Beschikbaar op Ufora : Ja

Referenties

- Speech and Language Processing, D. Jurafsky and J.H. Martin
- Neural Network Models in Natural Language Processing, Y. Goldberg
- Natural Language Processing - A Machine Learning Perspective, Y. Zhang, Z. Teng
- Deep Learning, I. Goodfellow, Y. Bengio and A. Courville
- Foundations of Statistical Natural Language Processing, C.D. Manning and H. Schütze
- Wetenschappelijke artikels uit de recente literatuur; op aanvraag verkrijgbaar indien nodig.

Vakinhoudelijke studiebegeleiding

De docenten en hun assistent(en) zijn tijdens en tussen de colleges beschikbaar voor extra uitleg (op afspraak).

Evaluatiemomenten

periodegebonden en niet-periodegebonden evaluatie

Evaluatievormen bij periodegebonden evaluatie in de eerste examenperiode

Schriftelijke evaluatie

Evaluatievormen bij periodegebonden evaluatie in de tweede examenperiode

Schriftelijke evaluatie

Evaluatievormen bij niet-periodegebonden evaluatie

Werkstuk

Tweede examenkans in geval van niet-periodegebonden evaluatie

Examen in de tweede examenperiode is niet mogelijk

Toelichtingen bij de evaluatievormen

- Periodegebonden evaluatie: schriftelijk examen, met het oog op het evalueren van het begrip en de praktische toepassing van de concepten zoals behandeld in de cursus.
- Permanente evaluatie:
  - Practica: bedoeld om de theorie in de praktijk toe te passen. Studenten zullen oplossingen voor NLP-taken implementeren, alsook de resultaten ervan evalueren en interpreteren. Hierbij kan het nodig zijn da studenten wetenschappelijke artikels moeten lezen ter voorbereiding van het practicum.
  - Presentatie van een wetenschappelijk artikel: dit zal beoordeelde worden op wetenschappelijke inhoud en de helderheid van de presentatie; de score op de presentatie krijgt hetzelfde gewicht als 1 practicum sessie.
  - Er is geen tweedekans voor de permanente evaluatie; de score van de practica zal worden overgedragen in het geval van tweede zit (die dan enkel doorgaat voor het schriftelijk examen). Merk op dat gezien de scoreberekening hieronder een totale score van meer dan 8/20 behaald moet

worden, voor alle practica samen, om te kunnen slagen voor het vak.

Eindscoreberekening

- Schriftelijk examen: 75%
- Permanente evaluatie (practica en presentatie onderzoeksartikel): 25%
- Bijkomend criterium om te slagen: een score van minstens 9/20 voor beide onderdelen (schriftelijk examen en permanente evaluatie) is vereist om te slagen. Wanneer de student/e 8/20 of minder behaalt voor één van de onderdelen, zal een totaalscore van tien of meer op twintig worden teruggebracht tot het hoogste niet-geslaagd cijfer (9/20).

Faciliteiten voor werkstudenten

Mogelijkheid tot aanpassing van de timing van de practica.