



Gefinancierd door de Europese Unie. De geuite standpunten en meningen zijn echter alleen die van de auteur(s) en komen niet noodzakelijkerwijs overeen met die van de Europese Unie of het Europees Uitvoerend Agentschap onderwijs en cultuur (EACEA). Noch de Europese Unie, noch het EACEA kan hiervoor verantwoordelijk worden gesteld.

STEAME ACADEMY

TEACHING FACILITATION LEARNING & CREATIVITY PLAN (L&C PLAN) - NIVEAU 2

SERVICE DOCENTEN: Text Mining: zijn deze documenten hetzelfde?

S

T

Eng

A

M

Ent



1. Overzicht

Titel	Text Mining: zijn deze documenten hetzelfde?
Drijvende vraag of onderwerp	Hoe kunnen zoekmachines resultaten vinden voor een zoekopdracht van een gebruiker op basis van trefwoorden? Hoe identificeren computers tekstdocumenten die zich op dezelfde onderwerpen richten? Hoe modelleren computeralgoritmen ongestructureerde gegevens voor digitale verwerking?
Leeftijden, cijfers, ...	16 tot 18 jaar 10e tot 12e leerjaar
Duur, tijdlijn, activiteiten	18 uur 9 sessies van elk 2 uur 21 activiteiten
Afstemming van het curriculum	Datamining, machine learning, ongestructureerde datamodellering, computerprogrammering
Medewerkers, partners	
Samenvatting - Synopsis	Studenten maken kennis met datamining en machine learning, waarbij de nadruk ligt op de kernonderwerpen van digitale verwerking van tekst. Tekstgelijkenis wordt onderzocht met behulp van de wiskundige grondslagen, de doorsnede van verzamelingen en cosinus tussen twee vectoren. Studenten werken in teams om een eenvoudige tool te implementeren om de gelijkenis tussen twee tekstdocumenten te meten. In de laatste sessies worden studenten uitgedaagd voor een wedstrijd om de beste implementatie te identificeren. Tijdens alle sessies maken studenten kennis met de belangrijkste methoden voor tekstvoorbewerking, zoals stopwoorden en stammen. De laatste sessie wordt afgesloten door de studenten te stimuleren om de overeenkomsten tussen hun implementatie en een zoekmachine te bespreken en te identificeren en van daaruit een zoekmachine te ontwerpen met behulp van hun eerdere implementatie voor tekstovereenkomst.
Referenties, Dankbetuigingen	

2. STEAME ACADEMY-kader*

Samenwerking van leerkrachten	Leraar 1 (Wetenschap) Machine learning en datamining: overzicht van het veld, algemene architectuur, veelvoorkomende toepassingen in het dagelijks leven, veelvoorkomende problemen
-------------------------------	---

- Datamodellering voor digitale verwerking: gestructureerde datarepresentatie voor machine learning en datamining, ongestructureerde datamodellering voor machine learning en datamining
- Text mining: overzicht van het vakgebied, kernconcepten, modellering (Booleaans model, TF-model, TFxIDF-model), gelijkenis van tekstdocumenten, voorbewerking, hoofdtaken en toepassingen

Leraar 2 (Techniek)

- Python-programmering voor text mining
- Vectoralgebra in Excel

Leraar 3 (Wiskunde)

- Cosinusfunctie, vectoralgebra
- Operatoren instellen, snijpunt

Leerkracht 1 werkt samen met Leerkracht 2 en Leerkracht 3 om:

- de Python-bibliotheken identificeren die moeten worden gebruikt voor text mining
- de Excel-functies identificeren die moeten worden gebruikt voor vectoralgebra
- Creëer de oefeningen en de uitdaging

Leerkracht 1 werkt samen met Leerkracht 2 om:

- Verzamel de benodigde corpora voor de praktische activiteiten
- Annoteer de benodigde corpora voor elke oefening

Leerkracht 1 werkt samen met Leerkracht 3 om:

- Introduceer set-operatoren in een tekstverwerkingsomgeving
- Introduceer vectoralgebra in een tekstverwerkingsomgeving.

STEAME in Life (SiL) organisatie

De laatste sessie wordt afgesloten door de studenten te stimuleren om de overeenkomsten tussen hun implementatie en een zoekmachine, zoals Google, te bespreken en te identificeren, en van daaruit een zoekmachine te ontwerpen met behulp van hun eerdere implementatie voor tekstovereenkomst.

Formulering van het actieplan

Vorbereidingsfase

1. Onderzoek, datamining en machine learning: traditionele en state-of-the-art toepassingen; link naar zoekmachines en generatieve AI; verwijzen naar gevallen van gestructureerde en ongestructureerde gegevens; Bekijk de belangrijkste uitdagingen van datamining en machine learning (datamodellering, multidimensionaliteit, overfitting, ontbrekende data, datavolume, big data, verklaren versus voorspellen, ...)
2. Verzamel en annoteer corpora voor oefeningen
3. Stel de Python-programmeeromgeving in (docker, repository in Github om te klonen, andere)

Structuur van de workshop

1. Introductie
 - 1.1. Overzicht van Data Mining en Machine Learning: historisch perspectief, taken/problemen, toepassingen, richt het gesprek op datamodellering (begin met datasets in tabelvorm en toon voorbeelden, vraag vervolgens naar oplossingen voor ongestructureerde gegevens zoals afbeeldingen en tekst). Bespreek deze oplossingen met studenten.
2. Tekst modelleren

- 2.1. Leg de corpora uit die we zullen gebruiken (moet kleine documenten hebben met een zeer beperkt lexicon; voeg documenten toe waar TF een verschil kan maken in vergelijking met het Booleaanse model).
- 2.2. Verwijs opnieuw naar tekstmodellering en begin met het Booleaanse model, toon voorbeelden, vraag hoe overeenkomsten tussen documenten te vinden?
- 2.3. Introduceer setoperatoren (vereniging, kruispunt, enz.).
- 2.4. Vraag studenten in teams van 3 of 4 om een tool in Excel te implementeren om hun oplossingen te implementeren en te testen; Elke groep legt uit welke set-operators ze gebruiken en waarom.
- 2.5. Introduceer vectoralgebra: intern product en cosinus. Studenten discussiëren en proberen relaties te vinden tussen verzamelingoperatoren en vectoralgebra (doorsnede en intern product, vereniging en som). Herstructureer hun Excel-implementatie om nu een combinatie van set- en vectoroperatoren te gebruiken.
- 2.6. Toon de impact/relevantie van termfrequentie in vergelijking met Booleaans; studenten herstructureren om het TF-model te implementeren en overeenkomsten te vinden met behulp van dezelfde corpora als voorheen. Het zou nu duidelijk moeten zijn dat de cosinus prima werkt.
- 2.7. Bespreek de wijzigingen en het probleem dat zich kan voordoen wanneer gelijkenis alleen op TF is gebaseerd (termen die in alle documenten voorkomen, hebben geen discriminerende kracht); Vraag de cursisten om oplossingen voor dit probleem te vinden.
- 2.8. Introduceer IDF en TFxIDF; bespreek, ontwerp en implementeer in Excel een gelijkenismaat op basis van het TFxIDF-model die moet worden gebruikt voor speelgoeddocumenten met een beperkt lexicon.
3. Implementatie
 - 3.1. Introduceer R, Python of andere bibliotheken voor text mining.
 - 3.2. Implementeer een functie om de gelijkenis tussen twee documenten te berekenen met behulp van R, Python of andere. Aan het einde van deze fase hebben alle studententeams een lopende functie; of anders zorgen voor een basisimplementatie voor iedereen.
4. Wiskundige verkenning
 - 4.1. Betrek leerlingen bij praktische activiteiten waarbij ze setoperatoren en vectoralgebra verkennen om de gelijkenis tussen tekstdocumenten in een corpus te berekenen.
 - 4.2. Faciliteren van discussies over de wiskundige principes achter tekstverwerking.
5. Culminerende projecten
 - 5.1. De teams van studenten worden uitgedaagd om het beste algoritme voor een similariteitsfunctie te vinden, te herstructureren en te implementeren.
 - 5.2. Geef de studenten een corpus dat eerder is geannoteerd op basis van gelijkenis tussen de statische testquery's/documenten (elk document in het corpus wordt geannoteerd met de similariteitsrangschikking voor elk van de testcases), reserveer een validatieset (ook geannoteerd; dit kunnen twee of drie tekstdocumenten zijn, elk met een paar termen alsof het een query voor een zoekmachine is; geef voor elk van deze "query's" de rangschikking van de documenten in het studentencorpus, dit zal worden gebruikt om uiteindelijk de F1 te berekenen en de winnaar aan te kondigen).
 - 5.3. Geef studenten de tijd om hun gelijkenisfunctie te implementeren en te verfijnen om de beste resultaten op hun corpus te krijgen, met

feedback en begeleiding van facilitators die studenten kennis laten maken met voorverwerkingstechnieken terwijl de discussie vordert (stopwoord verwijderen, stammen, enz.).

6. Link naar zoekmachines

- 6.1. Een "document" kan een zoekopdracht zijn zoals die we gebruiken in zoekmachines zoals Google, zoals "camélias porto" of "computer vision". Toon in het echt met Google. Geef de studenten de validatieset, d.w.z. drie documenten, elk met een paar trefwoorden, alsof het een zoekopdracht voor een zoekmachine is; Voer de gelijkenisfuncties uit om een rangschikking van de documenten in het corpus te geven en studenten de beste rangschikking te geven. Bereken de winnaar met behulp van de F1-meting. Leg studenten uit wat F1, Recall en Precision is.

Evaluatie en reflectie

1. Beoordeel het begrip en de toepassing van set- en vectoralgebra-operatoren en -concepten door studenten door middel van projectgebaseerde beoordelingen, presentaties en schriftelijke reflecties.
2. Moedig studenten aan om na te denken over hun leerervaringen en benadruk de relatie tussen wiskunde en text mining.

Vraag de leerlingen om een zoekmachinemethodologie en een niet-functioneel prototype te ontwerpen en te presenteren met behulp van wat ze hebben geleerd.

** in ontwikkeling van de laatste elementen van het kader*

3. Doelstellingen en methodologieën

Leerdoelen en doelstellingen

1. Inzicht krijgen in de algemene modellerings- en verwerkingsconcepten en -technieken die worden gebruikt bij text mining
2. Verken de interdisciplinaire verbanden tussen text mining, zoekmachines en vectoralgebra
3. Illustreer de gelijkenis tussen tekstdocumenten en andere ongestructureerde datasets als toepassingen van vectoralgebra

Leerresultaten en verwachte resultaten

Leerresultaten

- A. Bespreek onderwerpen op hoog niveau met betrekking tot tekstmining en zoekmachines
- B. Beschrijf de relatie tussen vectoralgebra, verzamelingenleer en text mining
- C. Pas basistechnieken van text mining toe om eenvoudige gebruiksscenario's aan te pakken

Verwachte resultaten

1. Functie voor gelijkenis van tekstdocumenten in R, Python of een andere programmeertaal

Voorkennis en vereisten

1. Fundamentele kennis van de vectoralgebra
2. Bekendheid met Excel
3. Basisvaardigheden voor het programmeren van software
4. Bekwaam gebruik van IT-tools

Motivatie, methodologie, strategieën, steigers

1. Wijs studenten toe aan kleine teams (3 of 4 studenten).
2. Ontwerp een oplossing, implementeer, test en verfijn op een iteratieve manier. Gebruik een iteratieve ontwikkelingsmethodologie.
3. Markeer de verbanden tussen vectoralgebra, documentmodellen, cosinus en gelijkenis.

4. Verken zoekmachines om de relaties tussen tekstovereenkomst en de resultaten van zoekmachines te laten zien.
5. Leid studenten door een evolutionair pad van de meest eenvoudige modellen (Booleaans) naar complexere benaderingen (TFxIDF), waarbij stap voor stap uitdagingen worden geïntroduceerd terwijl ze in Excel essay8ng met basisimplementaties voor kleine documenten met een of twee korte zinnen en een paar verschillende termen, uit een lexicon van 10 of 20 termen.

4. Voorbereiding en middelen

Voorbereiding, Ruimte-instelling, *Tips voor het oplossen van problemen*

De workshop wordt gehouden in een klaslokaal voor ongeveer 20 studenten, in groepen van 3 tot 4 studenten. Idealiter is de indeling van het klaslokaal georganiseerd in 5 tot 7 groepen tafels waar studenten van elk team tegenover elkaar kunnen zitten. De kamer heeft een beamer nodig en een muur voor presentaties aan iedereen en een whiteboard met pennen om ideeën te bespreken.

Middelen, gereedschappen, materiaal, bijlagen, uitrusting

Een repository in GDrive, Teams, Github of een andere provider moet op voorhand worden voorbereid met alle programmeeromgeving (R, Python, ...) en de corpora die nodig zijn voor de hands-on sessies.

Er moet een document worden verstrekt om studenten tijdens de hele cursus/workshop te begeleiden, met uitleg over details, verwachte resultaten, beoordeling en leerresultaten per sessie.

Gezondheid en veiligheid

5. Uitvoering

Educatieve activiteiten, procedures, reflecties

Structuur van de workshop

1. Introductie **[Sessie 1: 2 uur, 3 activiteiten]**
 - 1.1. Overzicht van Data Mining en Machine Learning: historisch perspectief, taken/problemen, methodologieën, technieken en tools (Weka, R, ...), taken (classificatie, clustering, ...), toepassingen, richt het gesprek op datamodellering (begin met datasets in tabelvorm en toon voorbeelden). **[40 minuten debat]**
 - 1.2. Debat met studenten over het modelleren van ongestructureerde gegevens (zoals afbeeldingen en tekst) voor automatische verwerking. Bespreek alternatieve oplossingen met studenten en introduceer relevante zaken (zakje met woorden, synoniemen, locatie van de tekst: titel, abstract, ...). **[40 minuten, debat]**
 - 1.3. Introduceer documentgelijkenis en de relevantie ervan in text mining. **[40 minuten, debat]**
2. Tekstmodellering 1 **[Sessie 2: 2 uur, 4 activiteiten]**
 - 2.1. Verwijs naar tekstmodellering en begin met het demonstreren van een tekstgelijkenisfunctie met het Booleaanse model, toon voorbeelden, vraag hoe overeenkomsten tussen documenten te vinden? **[40 minuten demonstratie, hands-on]**
 - 2.2. Leg de corpora uit die we zullen gebruiken (moet kleine documenten hebben met een zeer beperkt lexicon; voeg documenten toe waar TF een verschil kan maken in vergelijking met het Booleaanse model). **[20 minuten, verklarende sessie]**

- 2.3. Introduceer setoperatoren (vereniging, kruispunt, enz.). [20 minuten, verklarend, demonstratie]
- 2.4. Vraag studenten in teams van 3 of 4 om een tekstgelijkenisfunctie in Excel te implementeren en te testen; elke groep legt uit welke set-operatoren ze gebruiken en waarom. [40 minuten, hands-on]
3. Tekstmodellering 2 [Sessie 3: 2 uur, 4 activiteiten]
 - 3.1. Introduceer vectoralgebra: intern product en cosinus. Faciliteren van discussies over de wiskundige principes achter tekstverwerking. [20 minuten, verklarend, demonstratie]
 - 3.2. Betrek leerlingen bij praktische activiteiten waarbij ze setoperatoren en vectoralgebra verkennen om de gelijkenis tussen tekstdocumenten in een corpus te berekenen. Studenten debatteren in teams om relaties tussen verzamelingoperatoren en vectoralgebra (doorsnede en intern product, vereniging en som) te bespreken/vinden. Herstructureer Excel-implementaties van de gelijkenisfunctie om een combinatie van set- en vectoroperatoren te gebruiken. [40 minuten, verklarende sessie]
 - 3.3. Toon de impact/relevantie van termfrequentie in vergelijking met Booleaanse bij het berekenen van gelijkenis tussen documenten. [20 minuten, verklarende sessie]
 - 3.4. Studenten herstructureren hun implementatie om het TF-model te implementeren en vinden overeenkomsten met behulp van dezelfde corpora als voorheen. Het zou nu duidelijk moeten zijn dat de cosinus prima werkt. [40 minuten, verklarende sessie]
4. Tekstmodellering 3 [Sessie 4: 2 uur, 3 activiteiten]
 - 4.1. Bespreek de wijzigingen en het probleem dat zich kan voordoen wanneer gelijkenis alleen op TF is gebaseerd (termen die in alle documenten voorkomen, hebben geen discriminerende kracht). [20 minuten, verklarende sessie]
 - 4.2. Studenten vinden oplossingen voor dit probleem. [60 minuten, verklarende sessie]
 - 4.3. Introduceer IDF en TFxIDF. Bespreek, ontwerp en implementeer met de leerlingen een gelijkenismaat op basis van het TFxIDF-model die moet worden gebruikt voor speelgoeddocumenten met een beperkt lexicon. [40 minuten, demonstratie]
5. Implementatie [Sessie 5: 2 uur, 3 activiteiten]
 - 5.1. Introduceer R, Python of andere bibliotheken voor text mining. [30 minuten, verklarende sessie]
 - 5.2. Stel de ontwikkelomgeving in voor text mining. Leraren hebben een gids opgesteld en aan de studenten gegeven. [30 minuten, projectwerk]
 - 5.3. Studenten implementeren in teams een functie om de gelijkenis tussen twee documenten te berekenen met behulp van R, Python of andere. Aan het einde van deze fase moeten alle studententeams een lopende functie hebben (docenten bereiden van tevoren een implementatie voor om indien nodig in alles te voorzien). [60 minuten, hands-on]
6. Afsluitend project 1 [Sessie 6: 2 uur, 1 activiteit]
 - 6.1. De teams van studenten worden uitgedaagd om het beste algoritme voor een similariteitsfunctie te vinden, te herstructureren en te implementeren. [120 minuten, onderzoek, projectwerk]
7. Afsluitend project 2 [Sessies 7 en 8: 4 uur, 1 activiteit]
 - 7.1. Geef de studenten een corpus dat eerder is geannoteerd op basis van gelijkenis tussen de statische testquery's/documenten (elk document in het corpus wordt geannoteerd met de similariteitsrangschikking voor elk van de testcases), reserveer een validatieset (ook geannoteerd; dit kunnen twee of drie tekstdocumenten zijn, elk met een paar termen alsof het een query voor een zoekmachine is; geef voor elk van deze

	"query's" de rangschikking van de documenten in het studentencorpus, dit zal worden gebruikt om uiteindelijk de F1 te berekenen en de winnaar aan te kondigen). Geef studenten de tijd om hun gelijkenisfunctie te implementeren en te verfijnen om de beste resultaten op hun corpus te krijgen, met feedback en begeleiding van facilitators die studenten kennis laten maken met voorverwerkingstechnieken terwijl de discussie vordert (stopwoord verwijderen, stammen, enz.). [240 minuten, projectwerk, hands-on]
	8. Link naar zoekmachines [Sessie 9: 2 uur, 2 activiteiten] 8.1. Een "document" kan een zoekopdracht zijn zoals die we gebruiken in zoekmachines zoals Google, zoals "camélias porto" of "computer vision". Toon in het echt met Google. Geef de studenten de validatieset, d.w.z. drie documenten, elk met een paar trefwoorden, alsof het een zoekopdracht voor een zoekmachine is; Voer de gelijkenisfuncties uit om een rangschikking van de documenten in het corpus te geven en studenten de beste rangschikking te geven. Bereken de winnaar met behulp van de F1-meting. Leg studenten uit wat F1, Recall en Precision is. [Verklarende sessie van 60 minuten] 8.2. Debat, reflectie en conclusies. [Verklarende sessie van 60 minuten]
Beoordeling - Evaluatie	Evaluatie en reflectie 1. Beoordeel het begrip en de toepassing van set- en vectoralgebra-operatoren en -concepten door studenten door middel van projectgebaseerde beoordelingen, presentaties en schriftelijke reflecties. 2. Moedig studenten aan om na te denken over hun leerervaringen en benadruk de relatie tussen wiskunde en text mining. 3. Vraag de leerlingen om een zoekmachinemethodologie en een niet-functioneel prototype te ontwerpen en te presenteren met behulp van wat ze hebben geleerd.
Presentatie - Rapportage - Delen	1. Een functie die de overeenkomst tussen twee tekstdocumenten berekent. 2. Een presentatie in PowerPoint waarin een methodologie en een niet-functioneel prototype van een nieuwe zoekmachine worden beschreven die door de studenten worden voorgesteld op basis van wat we in de workshop hebben ontwikkeld (de functie voor documentgelijkenis).
Extensies - Overige informatie	

Bronnen voor de ontwikkeling van het STEAME ACADEMY Learning and Creativity Plan Template

In het geval van leren door middel van projectmatige activiteit

STEAME ACADEMY Prototype/Gids voor Leren & Creativiteit Aanpak Formulering van het actieplan

Belangrijke stappen in de STEAME-leeraanpak:

FASE I: Voorbereiding door een of meer docenten

1. Formuleren van eerste gedachten over de thematische sectoren/gebieden die aan bod moeten komen
2. Betrokkenheid bij de wereld van de bredere omgeving / werk / bedrijf / ouders / maatschappij / milieu / ethiek
3. Doelgroep van studenten - Associëren met het officiële curriculum - Doelen en doelstellingen stellen
4. Organisatie van de taken van de betrokken partijen - Aanwijzing van de coördinator - Werkplekken etc.

FASE II: Formulering van het actieplan (stappen 1-18)

Vorbereiding (door docenten)

1. Relatie tot de echte wereld – reflectie
2. Incentive – Motivatie
3. Formulering van een probleem (eventueel in fasen of fasen) als gevolg van het bovenstaande

Ontwikkeling (door studenten) – Begeleiding & Evaluatie (in 9-11, door docenten)

4. Achtergrond creatie - Zoeken / Informatie verzamelen
5. Vereenvoudig het probleem - Configureer het probleem met een beperkt aantal vereisten
6. Case Making - Ontwerpen - materialen identificeren voor bouwen / ontwikkelen / creëren
7. Bouw - Workflow - Uitvoering van projecten
8. Observatie-experimenten - Eerste conclusies
9. Documentatie - Zoeken naar thematische gebieden (AI-velden) die verband houden met het bestudeerde onderwerp - Uitleg op basis van bestaande theorieën en / of empirische resultaten
10. Verzamelen van resultaten / informatie op basis van de punten 7, 8, 9
11. Eerste groepspresentatie door studenten

Configuratie & Resultaten (door studenten) – Begeleiding & Evaluatie (door docenten)

12. Configureer STEAME-modellen om de resultaten te beschrijven / weergeven / illustreren
13. De resultaten in 9 bestuderen en conclusies trekken aan de hand van 12
14. Toepassingen in het dagelijks leven - Suggesties voor het ontwikkelen van 9 (Ondernemerschap - SIL Days)

Beoordeling (door docenten)

15. Bekijk het probleem en bekijk het onder meer veeleisende omstandigheden

Afronding van het project (door studenten) – Begeleiding en evaluatie (door docenten)

16. Herhaal stap 5 tot en met 11 met aanvullende of nieuwe eisen zoals geformuleerd in 15
17. Onderzoek - Casestudies - Uitbreiding - Nieuwe theorieën - Nieuwe conclusies testen
18. Presentatie van conclusies - Communicatietactieken.

FASE III: STEAME ACADEMY Acties en samenwerking in creatieve projecten voor scholieren

Titel van het project: _____

Korte beschrijving/overzicht van organisatorische regelingen / verantwoordelijkheden voor actie

PODIUM	Activiteiten/Stappen Leraar 1 (T1) Samenwerking met T2 en studentenbegeleiding	Activiteiten /Stappen Door studenten Leeftijdsgroep: ____	Activiteiten /Stappen Leraar 2 (T2) Samenwerking met T1 en Begeleiding van studenten
Een	Vorbereiding van de stappen 1,2,3		Samenwerking in stap 3
B	Begeleiding in stap 9	4,5,6,7,8,9,10	Ondersteuning begeleiding in stap 9
C	Creatieve evaluatie	11	Creatieve evaluatie
D	Begeleiding	12	Begeleiding
E	Begeleiding	13 (9+12)	Begeleiding
F	Organisatie (SIL) STEAME in het leven	14 Ontmoeting met vertegenwoordigers van het bedrijfsleven	Organisatie (SIL) STEAME in het leven
G	Vorbereiding van stap 15		Samenwerking in stap 15
H	Begeleiding	16 (herhaling 5-11)	Ondersteuning Begeleiding
Ik	Begeleiding	17	Ondersteuning Begeleiding
K	Creatieve evaluatie	18	Creatieve evaluatie