Co-funded by the European Union

Финансиран от Европейския съюз. Изразените възгледи и мнения обаче са само на автора(ите) и не отразяват непременно тези на Европейския съюз или Европейската изпълнителна агенция за образование и култура (EACEA). Нито Европейският съюз, нито EACEA могат да носят отговорност за тях.

STEAME АКАДЕМИЯ
УЛЕСНЯВАНЕ НА ПРЕПОДАВАНЕТО - ПЛАН ЗА ОБУЧЕНИЕ И ТВОРЧЕСТВО (L&C P) - НИВО 1: СТУДЕНТИ- УЧИТЕЛИ
ТЕМА: Извличане на текст: тези документи еднакви ли са?

S T E ng A M E nt

● ● ●

1. Преглед

Заглавие	Text Mining: тези документи еднакви ли са?
Въпрос или тема	Как търсачките могат да намерят резултати за потребителско търсене въз основа на ключови думи? Как компютрите идентифицират текстови документи, фокусирани върху едни и същи теми? Как компютърните алгоритми моделират неструктурирани данни за цифрова обработка?
Възраст, степени, …	16 до 18 години 10 до 12 клас
Продължителност, график, дейности	132 часа 17 дейности
Съгласуване на учебната програма	Извличане на данни, машинно обучение, моделиране на неструктурирани данни, компютърно програмиране
Сътрудници, партньори	
Резюме – Синопсис	Учениците се запознават с извличането на данни и машинното обучение, като се фокусират върху основните теми на цифровата обработка на текст. Изследва се сходството на текста, като се показват неговите математически основи, пресичане на множества и косинус между два вектора. Учениците работят в екипи, за да внедрят прост инструмент за измерване на приликата между два текстови документа. В последните сесии учениците са предизвикани за състезание за определяне на най-доброто изпълнение. По време на всички сесии учениците се запознават с ключовите методи за предварителна обработка на текст, като стоп-думи и корен. Последната сесия приключва, като настоява учениците да обсъдят и идентифицират приликите между тяхната реализация и търсачка и оттам да проектират търсачка, използвайки предишната си реализация за текстово сходство.
Използвана литература, благодарности	

2. Рамка на STEAME ACADEMY *

Учителско сътрудничество	Учител 1 (Наука) • Машинно обучение и извличане на данни: преглед на областта, обща архитектура, общи приложения в ежедневието, често срещани проблеми

	• Моделиране на данни за цифрова обработка: структурирано представяне на данни за машинно обучение и извличане на данни, моделиране на неструктурирани данни за машинно обучение и извличане на данни • Копаене на текст : преглед на областта, основни концепции, моделиране (булев модел, TF модел, TFxIDF модел), сходство на текстови документи, предварителна обработка, основни задачи и приложения Учител 2 (Инженерство) • Програмиране на Python за копаене на текст • Векторна алгебра в Excel Учител 3 (Математика) • Функция косинус, векторна алгебра • Множество оператори, пресичане Учител 1 си сътрудничи с Учител 2 и Учител 3 за: - идентифицирайте библиотеките на Python, които да използвате за копаене на текст - идентифицирайте функциите на Excel, които да използвате за векторна алгебра - създайте упражненията и предизвикателството Учител 1 си сътрудничи с Учител 2 за: - събират необходимите корпуси за практическите дейности - анотирайте необходимите корпуси за всяко упражнение Учител 1 си сътрудничи с Учител 3 за: - въвеждане на оператори за множество в среда за обработка на текст - въвеждане на векторна алгебра в среда за обработка на текст.
на STEAME в живота (SiL).	Последната сесия завършва, като настоява учениците да обсъдят и идентифицират приликите между тяхната реализация и търсачка, като Google, и оттам да проектират търсачка, използвайки предишната им реализация за сходство на текста.
Формулиране на план за действие	**Подготвителна фаза** 1. Проучване на традиционни и най-съвременни приложения за извличане на данни и машинно обучение; връзка към търсачки и генериращ AI; отнасят се до случаи на структурирани и неструктурирани данни; преглед на основните предизвикателства на извличането на данни и машинното обучение (моделиране на данни, многоизмерност, пренастройване, липсващи данни, обем на данни, големи данни, обяснение срещу прогнозиране, …) 2. Съберете и анотирайте корпуси за упражнения 3. Настройте средата за програмиране на Python (докер, хранилище в Github за клониране, друго) **Структура на цеха** 1. Въведение 1.1. Преглед на извличането на данни и машинното обучение: историческа перспектива, задачи/проблеми, приложения, моделиране на неструктурирани данни като изображения и текст). 2. Текстово моделиране

2.1. Съберете корпусите, които ще се използват за упражненията и работата по проекта (трябва да има малки документи с много ограничен лексикон; включете документи, при които TF може да направи разликата в сравнение с булевия модел).
2.2. Модели за представяне на текст, примери: Булев модел, TF, TFxIDF , други. Обърнете се към Торба с думи.
2.3. Оператори за множество (обединение, пресичане и др.).
2.4. Приложете няколко версии на инструмент в Excel за изчисляване на приликата между два текстови документа въз основа на оператори за множество и/или векторна алгебра (вътрешен продукт, косинус, пресечна точка и обединение на множества, …).
2.5. Въведете векторна алгебра: вътрешно произведение и косинус.
2.6. Покажете въздействието/уместността на честотата на термина в сравнение с булево. Изяснете, че косинусът работи добре.
2.7. Предизвикателства на сходството, базирани само на TF (термините, които присъстват във всички документи, нямат дискриминационна сила); алтернативни решения.
2.8. Представяне на IDF и TFxIDF ; обсъждане, проектиране и внедряване в Excel на мярка за сходство, базирана на модела TFxIDF , която да се използва за документи играчки с намален лексикон.

3. Внедряване
3.1. Проучете R, Python или други библиотеки за копаене на текст. Подгответе хранилище и ръководство за настройка за учениците, за да инсталират тази среда за разработка.
3.2. Приложете функция за изчисляване на приликата между два документа с помощта на R, Python или друг. Осигурете внедряване за всички ученици, ако е необходимо (ако учениците не могат да разработят свои собствени навреме).

4. Математическо изследване
4.1. Ангажирайте учениците в практически дейности, изследващи оператори на множество и векторна алгебра, за да изчислят сходството между текстови документи в корпус.
4.2. Улеснявайте дискусията върху математическите принципи зад обработката на текст.

5. Кулминационни проекти
5.1. Проектирайте проекта (най-добра функция за сходство за текстови документи) и начертайте ръководството.
5.2. Съберете и анотирайте корпус и набор от заявки, за да оцените прецизността и припомнянето, произведени от реализациите на учениците на функциите за сходство на текст. Прецизността и извикването ще бъдат оценени от статичните тестови заявки/документи (всеки документ в корпуса ще бъде анотиран с класирането на сходство за всеки един от тестовите случаи), резервирайте набор за валидиране (също анотиран; това може да бъде два или три текстови документа, всеки с няколко термина, сякаш е заявка за търсачка; за всяка една от тези „заявки“ осигурете класирането на документите в учениците корпус, това ще се използва за изчисляване на F1 в крайна сметка и за обявяване на победителя).
5.3. Предоставете информация за техниките за предварителна обработка (премахване на стоп-дума, произтичане и т.н.) и тяхното прилагане с помощта на избраните за проекта библиотеки за копаене на текст.

	6. Линк към търсачките 6.1. Подгответе сценария за валидиране. „Документ“ може да бъде заявка като тези, които използваме в търсачки като Google, като „ camélias porto ” или „компютърно зрение”. Показвайте в реално време с помощта на Google. Дайте на учениците комплекта за валидиране, т.е. три документа, всеки с няколко ключови думи, сякаш са заявка за търсачка; стартирайте функциите за сходство, за да осигурите класиране на документите в корпуса и да дадете на учениците най-доброто класиране. Изчислете победителя с помощта на мярка F1. Обяснете на учениците какво е F1, Recall и Precision. **Оценка и рефлексия** 1. Оценявайте разбирането и прилагането на операторите и концепциите на множествената и векторната алгебра от учениците чрез оценяване, базирано на проекти, презентации и писмени разсъждения. 2. Насърчете учениците да разсъждават върху своите учебни преживявания, подчертавайки връзката между математиката и извличането на текст. Помолете учениците да проектират и представят методология на търсачката и нефункционален прототип, използвайки това, което са научили.

[*] *в процес на разработка на крайните елементи на рамката*

3. Цели и методологии

Цели и задачи на обучението	1. Разберете общите концепции и техники за моделиране и обработка, използвани при копаене на текст 2. Изследвайте интердисциплинарните връзки между копаене на текст, търсачки и векторна алгебра 3. Илюстрирайте приликата между текстови документи и други неструктурирани набори от данни като приложения на векторната алгебра
Резултати от обучението и очаквани резултати	**Резултати от обучението** A. Обсъждайте теми на високо ниво, свързани с полетата на копаене на текст и търсачките B. Опишете връзката между векторната алгебра, теорията на множествата и извличането на текст C. Приложете основни техники за копаене на текст, за да адресирате прости случаи на употреба **Очаквани резултати** 1. Функция за сходство на текстови документи в R, Python или друг език за програмиране
Предварителни знания и предпоставки	1. Фундаментални познания по векторна алгебра 2. Познаване на Excel 3. Основни умения за програмиране на софтуер 4. Умело използване на ИТ инструменти
Мотивация, Методология, Стратегии	1. Разпределете учениците в малки екипи (3 или 4 ученика). 2. Проектирайте решение, внедрете, тествайте и усъвършенствайте по итеративен начин. Използвайте итеративна методология за разработка. 3. Подчертайте връзките между векторна алгебра, модели на документи, косинус и подобие.

4. Изследвайте търсачките, за да покажете връзките между приликата на текста и резултатите от търсачките.
5. Насочете учениците през еволюционен път от най-простите модели (булеви) до по-сложни подходи (TFxIDF), като въвеждате предизвикателства стъпка по стъпка, докато есе в Excel с основни реализации за малки документи с едно или две кратки изречения и няколко различни термина, от лексикон от 10 или 20 термина.

4. Подготовка и средства

Подготовка, настройка на пространството, *съвети за отстраняване на неизправности*	Семинарът ще се проведе в класна стая за приблизително 20 ученика, в групи от 3 до 4 ученика. В идеалния случай класната стая ще бъде организирана в 5 до 7 групи от маси, където учениците от всеки отбор могат да седят един срещу друг. Стаята се нуждае от проектор и стена за презентации за всички и бяла дъска с химикалки за обсъждане на идеи.
Ресурси, инструменти, материали, приставки, оборудване	хранилище в GDrive , Teams, Github или всеки друг доставчик с цялата среда за програмиране (R, Python, …) и корпусите, необходими за практическите сесии. Трябва да бъде предоставен документ, който да насочва учениците през целия курс/работилница, като обяснява подробности, очаквани резултати, оценка и резултати от обучението на сесия.
Здраве и безопасност	

5. Внедряване

Обучителни дейности, процедури, рефлексии

Структура на цеха

1. Въведение [1 дейност, 16 часа]
   1.1. Преглед на извличането на данни и машинното обучение: историческа перспектива, задачи/проблеми, приложения, моделиране на неструктурирани данни като изображения и текст). [16 часа]

2. Моделиране на текст [8 дейности, 52 часа]
   2.1. Съберете корпусите, които ще се използват за упражненията и работата по проекта (трябва да има малки документи с много ограничен лексикон; включете документи, при които TF може да направи разликата в сравнение с булевия модел). [16 часа]
   2.2. Модели за представяне на текст, примери: Булев модел, TF, TFxIDF , други. Обърнете се към Торба с думи. [8 часа]
   2.3. Опишете релевантни множество оператори (обединение, пресичане и т.н.). [4 часа]
   2.4. Приложете няколко версии на инструмент в Excel, за да изчислите сходството между два текстови документа въз основа на оператори за множество и/или векторна алгебра (вътрешен продукт, косинус, пресечна точка и обединение на множества, …). [4 часа]
   2.5. Въведете векторна алгебра: вътрешно произведение и косинус. [4 часа]
   2.6. Покажете въздействието/уместността на честотата на термина в сравнение с булево. Демонстрирайте с примери. Изяснете, че косинусът работи добре. [4 часа]
   2.7. Опишете предизвикателствата на сходството, когато се основава само на TF (термините, които присъстват във всички документи, нямат дискриминационна сила); алтернативни решения. Демонстрирайте с примери. [4 часа]

2.8. Представяне на IDF и TFxIDF ; обсъждане, проектиране и внедряване в Excel на мярка за сходство, базирана на модела TFxIDF , която да се използва за документи играчки с намален лексикон. [8 часа]

3. Изпълнение [2 дейности, 32 часа]
   3.1. Проучете R, Python или други библиотеки за копаене на текст. Подгответе хранилище и ръководство за настройка за учениците, за да инсталират тази среда за разработка. [16 часа]
   3.2. Приложете функция за изчисляване на приликата между два документа с помощта на R, Python или друг. Осигурете внедряване за всички ученици, ако е необходимо (ако учениците не могат да разработят свои собствени навреме). [16 часа]

4. Математическо изследване [2 дейности, 4 часа]
   4.1. Ангажирайте учениците в практически дейности, изследващи оператори на множество и векторна алгебра, за да изчислят сходството между текстови документи в корпус. [2 часа]
   4.2. Улеснявайте дискусията върху математическите принципи зад обработката на текст. [2 часа]

5. Кулминационни проекти [3 дейности, 20 часа]
   5.1. Проектирайте проекта (най-добра функция за сходство за текстови документи) и начертайте ръководството. [8 часа]
   5.2. Съберете и анотирайте корпус и набор от заявки, за да оцените прецизността и припомнянето, произведени от реализациите на учениците на функциите за сходство на текст. Прецизността и извикването ще бъдат оценени от статичните тестови заявки/документи (всеки документ в корпуса ще бъде анотиран с класирането на сходство за всеки един от тестовите случаи), резервирайте набор за валидиране (също анотиран; това може да бъде два или три текстови документа, всеки с няколко термина, сякаш е заявка за търсачка; за всяка една от тези „заявки“ осигурете класирането на документите в учениците корпус, това ще се използва за изчисляване на F1 в крайна сметка и за обявяване на победителя). [8 часа]
   5.3. Предоставете информация за техниките за предварителна обработка (премахване на стоп-дума, произтичане и т.н.) и тяхното прилагане с помощта на избраните за проекта библиотеки за копаене на текст. [4 часа]

6. Връзка към търсачки [1 дейност, 8 часа]
   6.1. Подгответе сценария за валидиране. „Документ“ може да бъде заявка като тези, които използваме в търсачки като Google, като „ camélias porto ” или „компютърно зрение”. Показвайте в реално време с помощта на Google. Дайте на учениците комплекта за валидиране, т.е. три документа, всеки с няколко ключови думи, сякаш са заявка за търсачка; стартирайте функциите за сходство, за да осигурите класиране на документите в корпуса и да дадете на учениците най-доброто класиране. Изчислете победителя с помощта на мярка F1. Обяснете на учениците какво е F1, Recall и Precision. [8 часа]

Оценка

Оценка и рефлексия

1. Оценявайте разбирането и прилагането на операторите и концепциите на множествената и векторната алгебра от учениците чрез оценяване, базирано на проекти, презентации и писмени разсъждения.

	2. Насърчете учениците да разсъждават върху своите учебни преживявания, подчертавайки връзката между математиката и извличането на текст. 3. Помолете учениците да проектират и представят методология на търсачката и нефункционален прототип, използвайки това, което са научили.
Представяне - Отчитане - Споделяне *Разширения - друга информация*	1. Функция, която изчислява приликата между два текстови документа. Презентация в PowerPoint, описваща методология и нефункционален прототип на нова търсачка, предложена от учениците, използвайки това, което сме разработили в семинара (функция за сходство на документи).

Прототип/Ръководство на STEAME ACADEMY за подход за обучение и творчество
Формулиране на план за действие

Основни стъпки в подхода за обучение на STEAME:

I ЕТАП: Подготовка от един или повече учители

1. Формулиране на първоначални мисли относно тематичните сектори/области, които да бъдат обхванати
2. Ангажиране на света на по-широката среда / работа / бизнес / родители / общество / среда / етика
3. Целева възрастова група ученици - Свързване с официалната учебна програма - Поставяне на цели и задачи
4. Организация на задачите на участващите страни - Определяне на координатор - Работни места и др.

ЕТАП II: Формулиране на план за действие (стъпки 1-18)

Подготовка (от учители)

1. Отношение към реалния свят – Отражение
2. Стимул – Мотивация
3. Формулиране на проблем (възможно на етапи или фази), произтичащ от горното

Развитие (от ученици) – Насоки и оценка (в 9-11, от учители)

4. Създаване на фон - Търсене / Събиране на информация
5. Опростете проблема – Конфигурирайте проблема с ограничен брой изисквания
6. Изработка на случай - Проектиране - идентифициране на материали за изграждане / разработване / създаване
7. Строителство - Работен процес - Изпълнение на проекти
8. Наблюдение-Експериментиране - Първоначални заключения
9. Документация - Търсене в тематични области (AI полета), свързани с изучавания предмет - Обяснение въз основа на съществуващи теории и/или емпирични резултати
10. Събиране на резултати / информация въз основа на точки 7, 8, 9
11. Първа групова презентация от ученици

Конфигуриране и резултати (от ученици) – Насоки и оценка (от учители)

12. Конфигурирайте моделите на STEAME, за да опишете/представите/илюстрирате резултатите
13. Проучване на резултатите в 9 и правене на заключения, като се използва 12
14. Приложения в ежедневието - Предложения за развитие 9 (Предприемачество - SIL Days)

Преглед (от учители)

15. Прегледайте проблема и го прегледайте при по-взискателни условия

Завършване на проекта (от ученици) – Насоки и оценка (от учители)

16. Повторете стъпки от 5 до 11 с допълнителни или нови изисквания, както са формулирани в 15
17. Разследване - Казуси - Разширяване - Нови теории - Тестване на нови заключения
18. Представяне на заключения - тактика на общуване.

ЕТАП III: STEAME ACADEMY Действия и сътрудничество в творчески проекти за ученици

Заглавие на проекта: _____
Кратко описание/Очертание на организационните договорености/Отговорности за действие

ЕТАП	**Дейности/Стъпки** Учител 1(Т1) Сътрудничество с Т2 и ръководство на учениците	**Дейности/Стъпки** **От ученици** Възрастова група: ____	**Дейности/Стъпки** Учител 2 (Т2) Сътрудничество с Т1 и ръководство на учениците
А	Подготовка на стъпки 1,2,3		Сътрудничество в стъпка 3
Б	Насоки в стъпка 9	4,5,6,7,8,9,10	Насоки за поддръжка в стъпка 9
В	Творческа оценка	11	Творческа оценка
Г	Насоки	12	Насоки
Д	Насоки	13 (9+12)	Насоки
Е	Организация (SIL) STEAME в живота	14 Среща с представители на бизнеса	Организация (SIL) STEAME в живота
Ж	Подготовка на стъпка 15		Сътрудничество в стъпка 15
З	Насоки	16 (повторение 5-11)	Ръководство за поддръжка
И	Насоки	17	Ръководство за поддръжка
К	Творческа оценка	18	Творческа оценка