



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 ПРЕПОДАВАНЕ ПЛАН ЗА ОБУЧЕНИЕ И ТВОРЧЕСТВО (L&C PLAN) - НИВО 2 УЧИТЕЛИ: Извличане на текст: тези документи еднакви ли са?

**S**

**T**

**E<sub>ng</sub>**

**A**

**M**

**E<sub>nt</sub>**



#### 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    | 18 часа   | 9 сесии по 2 часа | 21 дейности |
| Съгласуване на учебната програма     | Извличане на данни, машинно обучение, моделиране на неструктурирани данни, компютърно програмиране  |                   |             |
| Сътрудници, партньори                |   |                   |             |
| Резюме – Синопис                     | 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 основните методи за предварителна обработка на текст, като стоп-думи и произволни думи. 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 ) <ul style="list-style-type: none"> <li>Машинно обучение и извличане на данни: преглед на областта, обща архитектура, общи приложения в ежедневието, често срещани проблеми</li> </ul> |
|--------------------------|--|

- 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
- аотирайте 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 аотирайте корпуси за упражнения
3. Настройте средата за програмиране на Python (докер, хранилище в Github за клониране, друго)

#### **Структура на цеха**

1. Въведение

- 1.1. Общ преглед на извличането на данни и машинното обучение: историческа перспектива, задачи/проблеми, приложения, насочете разговора към моделиране на данни (започнете с таблични набори от данни и покажете примери, след това попитайте за решения за

неструктурирани данни като изображения и текст). Обсъдете тези решения с учениците.

## 2. Текстово моделиране

- 2.1. Обяснете корпусите, които ще използваме (трябва да има малки документи с много ограничен лексикон; включете документи, при които TF може да има значение в сравнение с булевия модел).
- 2.2. Обърнете се отново към моделирането на текст и започнете с булевия модел, покажете примери, попитайте как да намерите прилики между документи?
- 2.3. Въвеждане на множествени оператори (обединение, пресичане и т.н.).
- 2.4. Помолете учениците в екипи от 3 или 4 да внедрят инструмент в Excel, за да приложат своите решения и да ги тестват; всяка група обяснява кои набор оператори използва и защо.
- 2.5. Въведете векторна алгебра: вътрешно произведение и косинус. Учениците обсъждат и се опитват да намерят връзки между операторите на множество и векторната алгебра (пресечна точка и вътрешно произведение, обединение и сума). Преработете имплементацията на Excel, за да използвате сега комбинация от множествени и векторни оператори.
- 2.6. Покажете въздействието/уместността на честотата на термина в сравнение с булево; учениците преработват, за да внедрят TF модел и да намерят прилики, използвайки същите корпуси, както преди. Сега трябва да е ясно, че косинусът работи добре.
- 2.7. Обсъдете промените и проблема, който може да възникне, когато сходството се основава само на TF (термините, които присъстват във всички документи, нямат дискриминационна сила); помолете учениците да намерят решения на този проблем.
- 2.8. Представяне на IDF и  $TF \times IDF$ ; обсъждане, проектиране и внедряване в Excel на мярка за сходство, базирана на модела  $TF \times IDF$ , която да се използва за документи играчки с намален лексикон.

## 3. Внедряване

- 3.1. Представете R, Python или други библиотеки за копаене на текст.
- 3.2. Приложете функция за изчисляване на приликата между два документа с помощта на R, Python или друг. Всички ученически екипи трябва да имат текуща функция в края на тази фаза; или пък осигурете основно изпълнение за всички.

## 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 дискусиата върху математическите принципи зад обработката на текст.

## 5. Кулминационни проекти

- 5.1. Ученическите екипи са изправени пред предизвикателството да намерят, преработят и внедрят най-добрия алгоритъм за функция за подобие.
- 5.2. Дайте на учениците предварително анотиран корпус за сходство между статичните тестови заявки/документи (всеки документ в корпуса ще бъде анотиран с класирането на сходството 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 ученицкия корпус, това ще се използва за изчисляване на F1 в крайна сметка и за обявяване на победителя).

5.3. Осигурете време на учениците да внедрят и фино настроят своята функция за сходство, за да получат най-добри резултати в техния корпус, включвайки обратна връзка и насоки от фасилитатори, които ще запознаят учениците с техниките за предварителна обработка, докато дискусиата напредва (премахване на стоп-дума, корен и т.н.).

6. Линк към търсачките

6.1. „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                  | <ol style="list-style-type: none"> <li>1. Разберете общите концепции и техники за моделиране и обработка, използвани при копаене на текст</li> <li>2. Изследвайте интердисциплинарните връзки между копаене на текст, търсачки и векторна алгебра</li> <li>3. Илюстрирайте приликата между текстови документи и други неструктурирани набори от данни като приложения на векторната алгебра</li> </ol>  |
| Резултати от обучението и очаквани резултати | <p><b>Резултати от обучението</b></p> <ol style="list-style-type: none"> <li>A. Обсъждайте теми на високо ниво, свързани с полетата на копаене на текст и търсачките</li> <li>B. Опишете връзката между векторната алгебра, теорията на множествата и извличането на текст</li> <li>C. Приложете основни техники за копаене на текст, за да адресирате прости случаи на употреба</li> </ol> <p><b>Очаквани резултати</b></p> <ol style="list-style-type: none"> <li>1. Функция за сходство на текстови документи в R, Python или друг език за програмиране</li> </ol> |
| Предварителни знания и предпоставки          | <ol style="list-style-type: none"> <li>1. Фундаментални познания по векторна алгебра</li> <li>2. Познаване на Excel</li> </ol>  |

|   |  |
|---|--|
| Мотивация,<br>Методология,<br>Стратегии | <ol style="list-style-type: none"> <li>3. Основни умения за програмиране на софтуер</li> <li>4. Умело използване на ИТ инструменти</li> </ol>  |
|   | <ol style="list-style-type: none"> <li>1. Разпределете учениците в малки екипи (3 или 4 ученика).</li> <li>2. Проектирайте решение, внедрете, тествайте и усъвършенствайте по итеративен начин. Използвайте итеративна методология за разработка.</li> <li>3. Подчертайте връзките между векторна алгебра, модели на документи, косинус и подобие.</li> <li>4. Изследвайте търсачките, за да покажете връзките между приликата на текста и резултатите от търсачките.</li> <li>5. Насочете учениците през еволюционен път от най-простите модели (булеви) до по-сложни подходи (TFxIDF), като въведете 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.</li> </ol> |

#### 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.<br>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 | <p><b>Структура на цеха</b></p> <ol style="list-style-type: none"> <li>1. Въведение [Сесия 1: 2 часа, 3 дейности] <ol style="list-style-type: none"> <li>1.1. Преглед на извличането на данни и машинното обучение: историческа перспектива, задачи/проблеми, методологии, техники и инструменти (Weka, R, ...), задачи (класификация, групиране, ...), приложения, насочете разговора към моделиране на данни (започнете с таблични набори от данни и покажете примери). [40 минути дебат]</li> <li>1.2. Дебат със учениците относно моделирането на неструктурирани данни (като изображения и текст) за автоматична обработка. Обсъдете алтернативни решения с учениците, докато въведете подходящи въпроси (пакет от думи, синоними, местоположение на текста: заглавие, резюме, ...). [40 минути, дебат]</li> <li>1.3. Въведете сходството на документи и неговата уместност при копаене на текст. [40 минути, дебат]</li> </ol> </li> <li>2. Текстово моделиране 1 [Сесия 2: 2 часа, 4 дейности] <ol style="list-style-type: none"> <li>2.1. Обърнете се към моделирането на текст и започнете да демонстрирате функция за сходство на текст с булевия модел,</li> </ol> </li> </ol> |
|---|---|

- покажете примери, попитайте как да намерите прилики между документи? [40 минути демонстрация, практически]
- 2.2. Обяснете корпусите, които ще използваме (трябва да има малки документи с много ограничен лексикон; включете документи, при които TF може да има значение в сравнение с булевия модел). [20 минути, разяснителна сесия]
  - 2.3. Въвеждане на множествени оператори (обединение, пресичане и т.н.). [20 минути, изложение, демонстрация]
  - 2.4. Помолете учениците в екипи от 3 или 4 да внедрят функция за сходство на текст в Excel и да я тестват; всяка група обяснява кои набор оператори използва и защо. [40 минути, практически]
3. Текстово моделиране 2 [Сесия 3: 2 часа, 4 дейности]
- 3.1. Въведете векторна алгебра: вътрешно произведение и косинус. Улеснявайте дискусията върху математическите принципи зад обработката на текст. [20 минути, изложение, демонстрация]
  - 3.2. Ангажирайте учениците в практически дейности, изследващи оператори на множество и векторна алгебра, за да изчислят сходството между текстови документи в корпус. Учениците дебатира в екипи, за да обсъдят/открият връзките между операторите на множество и векторната алгебра (пресечна точка и вътрешно произведение, обединение и сума). Преработете реализациите на Excel на функцията за подобие, за да използвате комбинация от множествени и векторни оператори. [40 минути, разяснителна сесия]
  - 3.3. Показване на въздействието/уместността на честотата на термина в сравнение с булевото при изчисляване на сходство между документи. [20 минути, разяснителна сесия]
  - 3.4. Учениците преработват внедряването си, за да приложат TF модел и да намерят прилики, използвайки същите корпуси, както преди. Сега трябва да е ясно, че косинусът работи добре. [40 минути, разяснителна сесия]
4. Текстово моделиране 3 [Сесия 4: 2 часа, 3 дейности]
- 4.1. Обсъдете промените и проблема, който може да възникне, когато сходството се основава само на TF (термините, които присъстват във всички документи, нямат дискриминационна сила). [20 минути, разяснителна сесия]
  - 4.2. Учениците намират решения на този проблем. [60 минути, разяснителна сесия]
  - 4.3. Представете IDF и TFxIDF . Обсъдете, проектирайте и приложете с учениците мярка за сходство, базирана на модела TFxIDF , която да се използва за документи играчки с намален лексикон. [40 минути, демонстрация]
5. Изпълнение [Сесия 5: 2 часа, 3 дейности]
- 5.1. Представете R, Python или други библиотеки за копаене на текст. [30 минути, разяснителна сесия]
  - 5.2. Настройте средата за разработка за копаене на текст. Учителят е подготвил ръководство и го е предоставил на учениците. [30 минути, работа по проект]
  - 5.3. Учениците в екипи изпълняват функция за изчисляване на приликата между два документа с помощта на R, Python или друг. Всички ученически екипи трябва да имат текуща функция в края на тази фаза (учителите подготвят предварително изпълнение, за да осигурят всички, ако е необходимо). [60 минути, практически]
6. Кулминационен проект 1 [Сесия 6: 2 часа, 1 дейност]

- 6.1. Ученицките екипи са изправени пред предизвикателството да намерят, преработят и внедрят най-добрия алгоритъм за функция за подобие. [120 минути, проучване, работа по проект]
7. Кулминационен проект 2 [Сесии 7 и 8: 4 часа, 1 дейност]
- 7.1. Дайте на учениците предварително аотиран корпус за сходство между статичните тестови заявки/документи (всеки документ в корпуса ще бъде аотиран с класирането на сходството за всеки един от тестовите случаи), резервирайте набор за валидиране (също аотиран; 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 ученицкия корпус, това ще се използва за изчисляване на F1 в крайна сметка и за обявяване на победителя). Осигурете време на учениците да внедрят и фино настроят своята функция за сходство, за да получат най-добри резултати в техния корпус, включвайки обратна връзка и насоки от фасилитатори, които ще запознаят учениците с техниките за предварителна обработка, докато дискусиата напредва (премахване на стоп-дума, корен и т.н.). [240 минути, работа по проект, практически]
8. Връзка към търсачки [Сесия 9: 2 часа, 2 дейности]
- 8.1. „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. [60 минути разяснителна сесия]
- 8.2. Дебат, размисъл и заключения. [60 минути разяснителна сесия]

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

1. Функция, която изчислява приликата между два текстови документа.
2. 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).

Разширения - друга  
информация

**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 STEAME моделите, 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 - Комуникационни тактики.

### **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<br>Учител 1(T1)<br>Сътрудничество с T2<br>и ръководство на<br>учениците | Дейности/Стъпки<br>От ученици<br>Възрастова група: _____ | Дейности/Стъпки<br>Учител 2 (T2)<br>Сътрудничество с T1 и<br>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)<br>STEAME в живота  | 14<br>Среща с представители на<br>бизнеса                | Организация (SIL)<br>STEAME в живота  |
| Ж    | Подготовка на стъпка 15   |  | Сътрудничество в стъпка 15  |
| з    | Насоки  | 16 (повторение 5-11)                                     | Ръководство за поддръжка  |
| аз   | Насоки  | 17   | Ръководство за поддръжка  |
| К    | Творческа оценка  | 18   | Творческа оценка  |