Co-funded by the European Union

STEAME TEACHER ACADEMY

Financiado por la Unión Europea. Sin embargo, los puntos de vista y opiniones expresados son únicamente los del autor o autores y no reflejan necesariamente los de la Unión Europea o de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser consideradas responsables de ellos.

ACADEMIA DE VAPOR
FACILITACIÓN DE LA ENSEÑANZA PLAN DE APRENDIZAJE Y CREATIVIDAD (PLAN L&C) - NIVEL 1 ESTUDIANTES DE MAGISTERIO: Minería de textos: ¿son estos documentos lo mismo?

C T I A M E

1. Descripción general

Título	Minería de textos: ¿son iguales estos documentos?
Pregunta o tema de conducción	¿Cómo pueden los motores de búsqueda encontrar resultados para una búsqueda de usuario basada en palabras clave? ¿Cómo identifican los ordenadores los documentos de texto centrados en los mismos temas? ¿Cómo modelan los algoritmos informáticos los datos no estructurados para su procesamiento digital?
Edades, grados, ...	De 16 a 18 años Grados 10º a 12º
Duración, cronograma, actividades	132 horas 17 actividades
Contenidos curriculares	Minería de datos, aprendizaje automático, modelado de datos no estructurados, programación informática
Colaboradores, Socios	
Resumen - Sinopsis	Los estudiantes se introducen en la minería de datos y el aprendizaje automático centrándose en los temas centrales del procesamiento digital de texto. Se explora la similitud del texto mostrando sus fundamentos matemáticos, la intersección de conjuntos y el coseno entre dos vectores. Los estudiantes trabajan en equipos para implementar una herramienta simple para medir la similitud entre dos documentos de texto. En las últimas sesiones, los estudiantes son desafiados a una competencia para identificar la mejor implementación. Durante todas las sesiones, se presenta a los estudiantes los métodos clave para el preprocesamiento de textos, como las palabras vacías y la lematización. La última sesión concluye empujando a los estudiantes a discutir e identificar las semejanzas entre su implementación y un motor de búsqueda y, a partir de ahí, diseñar un motor de búsqueda utilizando su implementación anterior para la similitud de texto.

Referencias, Agradecimientos	

2. Marco de STEAME ACADEMY*

Cooperación de los docentes	Profesor 1 (Ciencias) • Aprendizaje automático y minería de datos: visión general del campo, arquitectura general, aplicaciones comunes a la vida cotidiana, problemas comunes • Modelado de datos para el procesamiento digital: representación de datos estructurados para el aprendizaje automático y la minería de datos, modelado de datos no estructurados para el aprendizaje automático y la minería de datos • Minería de textos: visión general del campo, conceptos básicos, modelado (modelo booleano, modelo TF, modelo TFxIDF), similitud de documentos de texto, preprocesamiento, principales tareas y aplicaciones Profesor 2 (Ingeniería) • Programación en Python para minería de textos • Álgebra vectorial en Excel Profesor 3 (Matemáticas) • Función de coseno, álgebra vectorial • Establecer operadores, intersección El Maestro 1 coopera con el Maestro 2 y el Maestro 3 para: - identificar las bibliotecas de Python que se van a utilizar para la minería de texto - identificar las funciones de Excel que se usarán para el álgebra vectorial - Crea los ejercicios y el reto El Maestro 1 coopera con el Maestro 2 para: - Reunir los corpus necesarios para las actividades prácticas - Anote los corpus necesarios para cada ejercicio El Maestro 1 coopera con el Maestro 3 para: - Introducción de operadores de conjunto en un entorno de procesamiento de texto - Introducir el álgebra vectorial en un entorno de procesamiento de textos.

Relación con el contexto	La última sesión concluye empujando a los estudiantes a discutir e identificar las semejanzas entre su implementación y un motor de búsqueda, como Google, y, a partir de ahí, diseñar un motor de búsqueda utilizando su implementación anterior para la similitud de texto.
Plan de Acción	**Fase Preparatoria** 1. Investigación, minería de datos y aprendizaje automático: aplicaciones tradicionales y de última generación; enlace a motores de búsqueda e IA generativa; se refieren a casos de datos estructurados y no estructurados; Revisión de los principales retos de la minería de datos y el machine learning (modelado de datos, multidimensionalidad, sobreajuste, datos faltantes, volumen de datos, big data, explicación versus predicción, ...) 2. Reunir y anotar corpus para los ejercicios 3. Configurar el entorno de programación Python (docker, repositorio en Github para clonar, otro) **Estructura del taller** 1. Introducción 1.1. Visión general de la minería de datos y el aprendizaje automático: perspectiva histórica, tareas/problemas, aplicaciones, modelado de datos no estructurados como imágenes y texto). 2. Modelado de texto 2.1. Reúna los corpus que se utilizarán para los ejercicios y el trabajo del proyecto (debe tener documentos pequeños con un léxico muy reducido; incluir documentos en los que TF pueda marcar la diferencia en comparación con el modelo booleano). 2.2. Modelos para representar texto, ejemplos: Modelo booleano, TF, TFxIDF, otros. Refiérase a la Bolsa de Palabras. 2.3. Establecer operadores (unión, intersección, etc.). 2.4. Implementar varias versiones de una herramienta en Excel para calcular la similitud entre dos documentos de texto en base a operadores de conjuntos y/o álgebra vectorial (producto interno, coseno, intersección y unión de conjuntos, ...). 2.5. Introducción al álgebra vectorial: producto interno y coseno. 2.6. Mostrar el impacto/relevancia de la frecuencia de los términos en comparación con los booleanos. Deja en claro que el coseno funciona bien. 2.7. Impugnaciones de similitud basadas únicamente en el TF (los términos que están presentes en todos los documentos no tienen poder discriminatorio); soluciones alternativas. 2.8. Presentar IDF y TFxIDF; discutir, diseñar e implementar en Excel una medida de similitud basada en el modelo TFxIDF para ser utilizada en documentos de juguete con un léxico reducido. 3. Implementación 3.1. Investiga R, Python u otras bibliotecas para la minería de textos.

Prepare un repositorio y una guía de configuración para que los estudiantes instalen este entorno de desarrollo.

3.2. Implemente una función para calcular la similitud entre dos documentos utilizando R, Python u otro. Proporcionar una implementación para todos los estudiantes si es necesario (si los estudiantes no pueden desarrollar la suya a su debido tiempo).

4. Exploración matemática
   4.1. Involucre a los estudiantes en actividades prácticas que exploren los operadores de conjuntos y el álgebra vectorial para calcular la similitud entre los documentos de texto en un corpus.
   4.2. Facilitar la discusión sobre los principios matemáticos detrás del procesamiento de textos.

5. Proyectos culminantes
   5.1. Diseñar el proyecto (mejor función de similitud para documentos de texto) y redactar la guía.
   5.2. Reúna y anote un corpus y un conjunto de consultas para evaluar la precisión y el recuerdo producidos por las implementaciones de los estudiantes de las funciones de similitud de textos. La precisión y la recuperación se evaluarán a partir de las consultas/documentos de prueba estática (cada documento del corpus se anotará con el ranking de similitud para cada uno de los casos de prueba), reservar un conjunto de validación (también anotado; puede ser dos o tres documentos de texto, cada uno con unos términos como si fuera una consulta para un motor de búsqueda; para cada una de estas "consultas" proporcionar la clasificación de los documentos en el corpus de estudiantes, esto se utilizará para calcular F1 al final y anunciar al ganador).
   5.3. Proporcionar información sobre las técnicas de preprocesamiento (eliminación de palabras vacías, lematización, etc.) y su implementación utilizando las bibliotecas de minería de texto seleccionadas para el proyecto.

6. Enlace a los motores de búsqueda
   6.1. Prepare el escenario de validación. Un "documento" puede ser una consulta como las que utilizamos en buscadores como Google, como "camélias porto" o "computer vision". Mostrar en real usando Google. Entregue a los estudiantes el conjunto de validación, es decir, tres documentos, cada uno con algunas palabras clave, como si fueran una consulta para un motor de búsqueda; Ejecute las funciones de similitud para proporcionar una clasificación de los documentos en el corpus y proporcionar a los estudiantes la mejor clasificación. Calcule el ganador usando la medida F1. Explique a los estudiantes qué es F1, Recuerdo y Precisión.

Evaluación y reflexión

1. Evaluar la comprensión y aplicación de los estudiantes de operadores y

	conceptos de álgebra vectorial y de conjuntos a través de evaluaciones basadas en proyectos, presentaciones y reflexiones escritas. 2. Anime a los estudiantes a reflexionar sobre sus experiencias de aprendizaje, destacando la relación entre las matemáticas y la minería de textos. Pida a los estudiantes que diseñen y presenten una metodología de motor de búsqueda y un prototipo no funcional utilizando lo que han aprendido.

[*] *En desarrollo Los elementos finales del marco*

3. Objetivos y metodologías

Objetivos de aprendizaje	1. Comprender los conceptos y técnicas genéricas de modelado y procesamiento utilizados en la minería de textos. 2. Explora las conexiones interdisciplinarias entre la minería de textos, los motores de búsqueda y el álgebra vectorial 3. Ilustrar la similitud entre documentos de texto y otros conjuntos de datos no estructurados como aplicaciones del álgebra vectorial
Resultados de aprendizaje	**Resultados de aprendizaje** A. Discutir temas de alto nivel relacionados con los campos de la minería de textos y los motores de búsqueda B. Describir la relación entre el álgebra vectorial, la teoría de conjuntos y la minería de textos. C. Aplicar técnicas básicas de minería de textos para abordar casos de uso sencillos **Resultados esperados** 1. Función de similitud de documentos de texto en R, Python u otro lenguaje de programación
Conocimientos y requisitos previos	1. Conocimientos fundamentales de álgebra vectorial 2. Familiaridad con Excel 3. Habilidades básicas de programación de software 4. Uso competente de las herramientas informáticas
Motivación, Metodología, Estrategias, Andamiaje	1. Asigne a los estudiantes a equipos pequeños (3 o 4 estudiantes). 2. Diseñe una solución, implemente, pruebe y perfeccione de forma iterativa. Utilice una metodología de desarrollo iterativa. 3. Destaca las conexiones entre el álgebra vectorial, los modelos de documentos, el coseno y la similitud. 4. Explore los motores de búsqueda para mostrar las relaciones entre la similitud del texto y los resultados de los motores de búsqueda. 5. Guíe a los estudiantes a través de un camino evolutivo desde los modelos más simples (booleanos) hasta los enfoques más complejos (TFxIDF), presentando desafíos paso a paso mientras ensayan en Excel con implementaciones básicas para documentos pequeños con una o dos oraciones cortas y algunos términos distintos, de un léxico de 10 o 20

	términos.

4. Preparación y medios

Preparación, configuración del espacio, *consejos para la resolución de problemas*	El taller se llevará a cabo en un aula para aproximadamente 20 estudiantes, en grupos de 3 a 4 estudiantes. Lo ideal es que la distribución del aula se organice en 5 a 7 grupos de mesas donde los alumnos de cada equipo puedan sentarse uno frente al otro. La sala necesita un proyector y una pared para presentaciones a todos y una pizarra blanca con bolígrafos para discutir ideas.
Recursos, Herramientas, Material, Accesorios, Equipos	Se preparará previamente un repositorio en GDrive, Teams, Github o cualquier otro proveedor con todo el entorno de programación (R, Python, ...) y los corpus necesarios para las sesiones prácticas. Se debe proporcionar un documento para guiar a los estudiantes a lo largo de todo el curso/taller, explicando los detalles, los resultados esperados, la evaluación y los resultados de aprendizaje por sesión.
Salud y seguridad	

5. Implementación

Actividades	**Estructura del taller** 1. Introducción [1 actividad, 16 horas] 1.1. Visión general de la minería de datos y el aprendizaje automático: perspectiva histórica, tareas/problemas, aplicaciones, modelado de datos no estructurados como imágenes y texto). [16 horas] 2. Modelado de textos [8 actividades, 52 horas] 2.1. Reúna los corpus que se utilizarán para los ejercicios y el trabajo del proyecto (debe tener documentos pequeños con un léxico muy reducido; incluir documentos en los que TF pueda marcar la diferencia en comparación con el modelo booleano). [16 horas] 2.2. Modelos para representar texto, ejemplos: Modelo booleano, TF, TFxIDF, otros. Refiérase a la Bolsa de Palabras. [8 horas] 2.3. Describir los operadores de conjunto de relevancia (unión, intersección, etc.). [4 horas] 2.4. Implementar varias versiones de una herramienta en Excel para calcular la similitud entre dos documentos de texto en base a operadores de conjuntos y/o álgebra vectorial (producto interno, coseno, intersección y unión de conjuntos, ...). [4 horas] 2.5. Introducción al álgebra vectorial: producto interno y coseno. [4 horas] 2.6. Mostrar el impacto/relevancia de la frecuencia de los términos en comparación con los booleanos. Demuéstralo con ejemplos. Deja en claro que el coseno funciona bien. [4 horas] 2.7. Describa los desafíos de la similitud cuando se basa únicamente en TF

(los términos que están presentes en todos los documentos no tienen poder discriminatorio); soluciones alternativas. Demuéstralo con ejemplos. [4 horas]

2.8. Presentar IDF y TFxIDF; discutir, diseñar e implementar en Excel una medida de similitud basada en el modelo TFxIDF para ser utilizada en documentos de juguete con un léxico reducido. [8 horas]

3. Implementación [2 actividades, 32 horas]
   3.1. Investiga R, Python u otras bibliotecas para la minería de textos. Prepare un repositorio y una guía de configuración para que los estudiantes instalen este entorno de desarrollo. [16 horas]
   3.2. Implemente una función para calcular la similitud entre dos documentos utilizando R, Python u otro. Proporcionar una implementación para todos los estudiantes si es necesario (si los estudiantes no pueden desarrollar la suya a su debido tiempo). [16 horas]

4. Exploración Matemática [2 actividades, 4 horas]
   4.1. Involucre a los estudiantes en actividades prácticas que exploren los operadores de conjuntos y el álgebra vectorial para calcular la similitud entre los documentos de texto en un corpus. [2 horas]
   4.2. Facilitar la discusión sobre los principios matemáticos detrás del procesamiento de textos. [2 horas]

5. Proyectos culminantes [3 actividades, 20 horas]
   5.1. Diseñar el proyecto (mejor función de similitud para documentos de texto) y redactar la guía. [8 horas]
   5.2. Reúna y anote un corpus y un conjunto de consultas para evaluar la precisión y el recuerdo producidos por las implementaciones de los estudiantes de las funciones de similitud de textos. La precisión y la recuperación se evaluarán a partir de las consultas/documentos de prueba estática (cada documento del corpus se anotará con el ranking de similitud para cada uno de los casos de prueba), reservar un conjunto de validación (también anotado; puede ser dos o tres documentos de texto, cada uno con unos términos como si fuera una consulta para un motor de búsqueda; para cada una de estas "consultas" proporcionar la clasificación de los documentos en el corpus de estudiantes, esto se utilizará para calcular F1 al final y anunciar al ganador). [8 horas]
   5.3. Proporcionar información sobre las técnicas de preprocesamiento (eliminación de palabras vacías, lematización, etc.) y su implementación utilizando las bibliotecas de minería de texto seleccionadas para el proyecto. [4 horas]

6. Enlace a buscadores [1 actividad, 8 horas]
   6.1. Prepare el escenario de validación. Un "documento" puede ser una consulta como las que utilizamos en buscadores como Google, como "camélias porto" o "computer vision". Mostrar en real usando Google.

	Entregue a los estudiantes el conjunto de validación, es decir, tres documentos, cada uno con algunas palabras clave, como si fueran una consulta para un motor de búsqueda; Ejecute las funciones de similitud para proporcionar una clasificación de los documentos en el corpus y proporcionar a los estudiantes la mejor clasificación. Calcule el ganador usando la medida F1. Explique a los estudiantes qué es F1, Recuerdo y Precisión. [8 horas]
Valoración - Evaluación	**Evaluación y reflexión** 1. Evaluar la comprensión y aplicación de los estudiantes de operadores y conceptos de álgebra vectorial y de conjuntos a través de evaluaciones basadas en proyectos, presentaciones y reflexiones escritas. 2. Anime a los estudiantes a reflexionar sobre sus experiencias de aprendizaje, destacando la relación entre las matemáticas y la minería de textos. 3. Pida a los estudiantes que diseñen y presenten una metodología de motor de búsqueda y un prototipo no funcional utilizando lo que han aprendido.
Presentación - Informes - Compartir	1. Función que calcula la similitud entre dos documentos de texto. Una presentación en PowerPoint que describe una metodología y un prototipo no funcional de un motor de búsqueda novedoso propuesto por los estudiantes utilizando lo que hemos desarrollado en el taller (la función de similitud de documentos).
Extensiones - Otra información	

STEAME ACADEMY Prototipo/Guía para el Aprendizaje y la Creatividad
Formulación del Plan de Acción

Principales pasos en el enfoque de aprendizaje de STEAME:

ETAPA I: Preparación por parte de uno o más profesores

1. Formulación de reflexiones iniciales sobre los sectores/áreas temáticas que se van a abarcar
2. Involucrarse en el mundo del medio ambiente / trabajo / empresa / padres / sociedad / medio ambiente / ética
3. Grupo de edad objetivo de los estudiantes - Asociación con el currículo oficial - Establecimiento de metas y objetivos
4. Organización de las tareas de las partes involucradas - Designación de Coordinador - Lugares de trabajo, etc.

ETAPA II: Formulación del Plan de Acción (Pasos 1-18)

Preparación (por parte de los profesores)

1. Relación con el Mundo Real – Reflexión
2. Incentivo – Motivación
3. Formulación de un problema (posiblemente en etapas o fases) que resulte de lo anterior

Desarrollo (por parte de los estudiantes) – Orientación y Evaluación (en 9-11, por los profesores)

4. Creación de antecedentes - Buscar / Recopilar información
5. Simplifique el problema: configure el problema con un número limitado de requisitos
6. Fabricación de casos - Diseño - identificación de materiales para la construcción / desarrollo / creación
7. Construcción - Flujo de trabajo - Implementación de proyectos
8. Observación-Experimentación - Conclusiones Iniciales
9. Documentación - Búsqueda de Áreas Temáticas (campos de IA) relacionadas con el tema en estudio - Explicación basada en Teorías Existentes y/o Resultados Empíricos
10. Recopilación de resultados / información basada en los puntos 7, 8, 9
11. Primera presentación grupal de los estudiantes

Configuración y resultados (por parte de los estudiantes) – Orientación y evaluación (por parte de los profesores)

12. Configurar modelos STEAME para describir/representar/ilustrar los resultados
13. Estudiar los resultados en 9 y sacar conclusiones, utilizando 12
14. Aplicaciones en la vida cotidiana - Sugerencias para desarrollar 9 (Emprendimiento - Días SIL)

Revisión (por parte de los profesores)

15. Revisar el problema y revisarlo en condiciones más exigentes

<u>Finalización del proyecto (por parte de los estudiantes) – Orientación y evaluación (por parte de los profesores)</u>

16. Repita los pasos 5 a 11 con requisitos adicionales o nuevos tal como se formularon en 15
17. Investigación - Estudios de caso - Expansión - Nuevas teorías - Prueba de nuevas conclusiones
18. Presentación de Conclusiones - Tácticas de Comunicación.

ETAPA III: STEAME ACADEMY Acciones y Cooperación en Proyectos Creativos para estudiantes de la escuela

Título del proyecto: _____
Breve descripción/esbozo de los arreglos organizacionales/responsabilidades para la acción

ETAPA	**Actividades/Pasos** Profesor 1(T1) Cooperación con T2 y orientación estudiantil	**Actividades / Pasos** **Por los estudiantes** Grupo de edad: ____	**Actividades / Pasos** Profesor 2 (T2) Cooperación con T1 y Orientación al estudiante
Un	Preparación de los pasos 1,2,3		Cooperación en la etapa 3
B	Orientación en el paso 9	4,5,6,7,8,9,10	Guía de soporte en el paso 9
C	Evaluación creativa	11	Evaluación creativa
D	Orientación	12	Orientación
E	Orientación	13 (9+12)	Orientación
F	Organización (SIL) STEAME en la vida	14 Reunión con representantes de las empresas	Organización (SIL) STEAME en la vida
G	Preparación de la etapa 15		Cooperación en la etapa 15
H	Orientación	16 (repetición 5-11)	Orientación de soporte
Yo	Orientación	17	Orientación de soporte
K	Evaluación creativa	18	Evaluación creativa