Co-funded by the European Union

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões.

ACADEMIA STEAME
FACILITAÇÃO DO ENSINO APRENDIZAGEM -&PLANO DE CRIATIVIDADE (PLANO L&C) - NÍVEL 2 PROFESSORES DE SERVIÇO: Mineração de texto: estes documentos são os mesmos?

S	T	Eng	A	M	Emp
●		●		●	

1. Síntese

Título	Mineração de texto: esses documentos são os mesmos?		
Questão ou Tema orientador	Como podem os motores de busca encontrar resultados para uma pesquisa de utilizador com base em palavras-chave? Como é que os computadores identificam documentos de texto focados nos mesmos tópicos? Como os algoritmos computacionais modelam dados não estruturados para processamento digital?		
Idades, Níveis, ...	16 a 18 anos	10º ao 12º ano	
Duração, Cronograma, Atividades	18 horas	9 sessões de 2 horas cada	21 atividades
Alinhamento Curricular	Mineração de dados, aprendizado de máquina, modelagem de dados não estruturados, programação de computadores		
Colaboradores, Parceiros			
Resumo - Sinopse	Os alunos são apresentados à Data Mining e Machine Learning com foco nos principais tópicos do processamento digital de texto. A semelhança textual é explorada mostrando seus fundamentos matemáticos, intersecção de conjuntos e cosseno entre dois vetores. Os alunos trabalham em equipa para implementar uma ferramenta simples para medir a semelhança entre dois documentos de texto. Nas últimas sessões os alunos são desafiados para um concurso para identificar a melhor implementação. Durante todas as sessões, os alunos estão a ser apresentados aos principais métodos para o pré-processamento de texto, como stop-words e stemming. A última sessão é concluída incentivando os alunos a discutir e identificar as semelhanças entre a sua implementação e um motor de busca e, a partir daí, a projetar um motor de busca usando a sua implementação anterior para semelhança de texto.		

Referências, Agradecimentos	

2. Estrutura da STEAME ACADEMY*

Cooperação dos professores	Professor 1 (Ciências) • Machine learning e mineração de dados: visão geral do campo, arquitetura geral, aplicações comuns à vida diária, problemas comuns • Modelação de dados para processamento digital: representação de dados estruturados para aprendizagem automática e prospeção de dados, modelação de dados não estruturados para aprendizagem automática e prospeção de dados • Mineração de texto: visão geral do campo, conceitos centrais, modelagem (modelo booleano, modelo TF, modelo TFxIDF), semelhança de documentos de texto, pré-processamento, principais tarefas e aplicações Professor 2 (Engenharia) • Programação Python para mineração de texto • Álgebra vetorial no Excel Professor 3 (Matemática) • Função cossina, álgebra vetorial • Operadores de conjunto, intersecção O Professor 1 coopera com o Professor 2 e o Professor 3 para: - identificar as bibliotecas Python a serem usadas para mineração de texto - identificar as funções do Excel a serem usadas para álgebra vetorial - criar os exercícios e o desafio O Professor 1 coopera com o Professor 2 para: - reunir os corpora necessários para as atividades práticas - anotar os corpora necessários para cada exercício O Professor 1 coopera com o Professor 3 para: - Introduzir operadores de conjunto em um ambiente de processamento de texto - introduzir álgebra vetorial em um ambiente de processamento de texto.
Organização STEAME in Life (SiL)	A última sessão é concluída incentivando os alunos a discutir e identificar as semelhanças entre a sua implementação e um motor de busca, como o Google, e, a partir daí, a projetar um motor de busca usando a sua implementação

	anterior para semelhança de texto.
Formulação do Plano de Ação	**Fase preparatória** 1. Pesquisa de mineração de dados e machine learning aplicações tradicionais e de última geração; link para motores de busca e IA generativa; referir-se a casos de dados estruturados e não estruturados; rever os principais desafios da mineração de dados e aprendizagem automática (modelação de dados, multidimensionalidade, sobreajuste, dados em falta, volume de dados, big data, explicação versus previsão, ...) 2. Reunir e anotar corpora para exercícios 3. Configurar o ambiente de programação Python (docker, repositório no Github para clonar, outros) **Estrutura do Workshop** 1. Introdução 1.1. Visão geral de Data Mining e Machine Learning: perspetiva histórica, tarefas/problemas, aplicativos, direcione a conversa para a modelagem de dados (comece por conjuntos de dados tabulares e mostre exemplos, depois pergunte sobre soluções para dados não estruturados, como imagens e texto). Discuta estas soluções com os alunos. 2. Modelação de texto 2.1. Explique os corpora que usaremos (deve ter pequenos documentos com um léxico muito reduzido; incluir documentos onde o TF pode fazer a diferença quando comparado com o modelo booleano). 2.2. Consulte novamente a modelagem de texto e comece com o modelo booleano, mostre exemplos, pergunte como encontrar semelhanças entre documentos? 2.3. Introduzir operadores de conjunto (união, intersecção, etc.). 2.4. Pedir aos alunos em equipas de 3 ou 4 pessoas que implementem uma ferramenta em Excel para implementar as suas soluções e testá-las; Cada grupo explica quais operadores de conjunto estão usando e por quê. 2.5. Introduzir álgebra vetorial: produto interno e cosseno. Os alunos discutem e tentam encontrar relações entre operadores de conjuntos e álgebra vetorial (intersecção e produto interno, união e soma). Refatore sua implementação do Excel para usar agora uma combinação de operadores de conjunto e vetor. 2.6. Mostrar o impacto/relevância da frequência de termos em comparação com Booleano; os alunos refatoram para implementar o modelo de TF e encontram semelhanças usando os mesmos corpora de antes. Agora deve ficar claro que o cosseno funciona bem. 2.7. Discutir as alterações e o problema que pode surgir quando a semelhança se baseia apenas no FT (termos que estão presentes em todos os documentos não têm poder discriminatório); Peça aos alunos que encontrem soluções para este problema. 2.8. Introduzir IDF e TFxIDF; discutir, projetar e implementar em Excel uma

medida de similaridade baseada no modelo TFxIDF para ser usado para documentos de brinquedo com léxico reduzido.

3. Execução
   3.1. Introduza R, Python ou outras bibliotecas para mineração de texto.
   3.2. Implemente uma função para calcular a semelhança entre dois documentos usando R, Python ou outro. Todas as equipas de alunos terão uma função de funcionamento no final desta fase; ou então fornecer uma implementação básica para todos.

4. Exploração Matemática
   4.1. Envolva os alunos em atividades práticas explorando operadores de conjunto e álgebra vetorial para calcular a semelhança entre documentos de texto em um corpus.
   4.2. Facilitar a discussão sobre os princípios matemáticos por trás do processamento de texto.

5. Projetos culminantes
   5.1. As equipas dos alunos são desafiadas a encontrar, refatorar e implementar o melhor algoritmo para uma função de similaridade.
   5.2. Dê aos alunos um corpus previamente anotado para semelhança entre as consultas/documentos estáticos do teste (cada documento do corpus será anotado com a classificação de semelhança para cada um dos casos de teste), reserve um conjunto de validação (anotado também; pode ser dois ou três documentos de texto, cada um com alguns termos como se fosse uma consulta para um motor de busca; para cada uma dessas "consultas" forneça a classificação do documentos no corpus dos alunos, isso será usado para calcular F1 no final e anunciar o vencedor).
   5.3. Dar tempo para que os alunos implementem e ajustem sua função de semelhança para obter os melhores resultados em seu corpus, incorporando feedback e orientação de facilitadores que apresentarão aos alunos técnicas de pré-processamento enquanto a discussão progride (remoção de palavras-paradas, derivação, etc.).

6. Link para motores de busca
   6.1. Um "documento" pode ser uma consulta como as que usamos em motores de busca como o Google, como "camélias porto" ou "visão computacional". Mostrar em real usando o Google. Dê aos alunos o conjunto de validação, ou seja, três documentos, cada um com algumas palavras-chave, como se fossem uma consulta para um motor de busca; executar as funções de semelhança para fornecer uma classificação dos documentos no corpus e dar aos alunos a melhor classificação. Calcule o vencedor usando a medida F1. Explique aos alunos o que é F1, Recall e Precisão.

Avaliação e Reflexão

	1. Avaliar a compreensão dos alunos e a aplicação de operadores e conceitos de álgebra vetorial e de conjunto através de avaliações baseadas em projetos, apresentações e reflexões escritas. 2. Incentivar os alunos a refletir sobre as suas experiências de aprendizagem, destacando a relação entre a matemática e a mineração de texto. Peça aos alunos que projetem e apresentem uma metodologia de mecanismo de pesquisa e um protótipo não funcional usando o que aprenderam.

** em desenvolvimento os elementos finais do quadro*

3. Objetivos e metodologias

Metas e Objetivos de Aprendizagem	1. Compreender os conceitos e técnicas genéricas de modelagem e processamento usados na mineração de texto 2. Explore as conexões interdisciplinares entre mineração de texto, mecanismos de busca e álgebra vetorial 3. Ilustrar a semelhança entre documentos de texto e outros conjuntos de dados não estruturados como aplicações da álgebra vetorial
Resultados de Aprendizagem e Resultados Esperados	**Resultados de Aprendizagem** A. Discutir tópicos de alto nível relacionados com os campos de mineração de texto e motores de busca B. Descrever a relação entre álgebra vetorial, teoria dos conjuntos e mineração de texto C. Aplicar técnicas básicas de mineração de texto para lidar com casos de uso simples **Resultados esperados** 1. Função de similaridade de documento de texto em R, Python ou outra linguagem de programação
Conhecimentos Prévios e Pré-requisitos	1. Conhecimento fundamental de álgebra vetorial 2. Familiaridade com o Excel 3. Conhecimentos básicos de programação de software 4. Utilização eficiente de ferramentas informáticas
Motivação, Metodologia, Estratégias, Apoios pedagógicos	1. Atribua alunos a pequenas equipas (3 ou 4 alunos). 2. Projete uma solução, implemente, teste e refine de forma iterativa. Use uma metodologia de desenvolvimento iterativo. 3. Destaque as conexões entre álgebra vetorial, modelos de documentos, cosseno e similaridade. 4. Explore os motores de busca para mostrar as relações entre a semelhança de texto e os resultados dos motores de busca. 5. Guie os alunos através de um caminho evolutivo desde os modelos mais simples (booleanos) até abordagens mais complexas (TFxIDF), introduzindo desafios passo a passo enquanto ensaio8ng em Excel com implementações básicas para pequenos documentos com uma ou duas frases curtas e alguns

	termos distintos, a partir de um léxico de 10 ou 20 termos.

4. Preparação e meios

Preparação, configuração de espaço, orientações para resolução de problemas	A oficina será realizada em sala de aula para aproximadamente 20 alunos, em grupos de 3 a 4 alunos. Idealmente, a disposição da sala de aula será organizada em 5 a 7 grupos de mesas onde os alunos de cada equipa podem sentar-se de frente uns para os outros. A sala precisa de um beamer e uma parede para apresentações para todos e um quadro branco com canetas para discutir ideias.
Recursos, Ferramentas, Material, Anexos, Equipamento	Um repositório no GDrive, Teams, Github ou qualquer outro provedor deve ser preparado com antecedência com todo o ambiente de programação (R, Python, ...) e os corpora necessários para as sessões práticas. Deve ser fornecido um documento para orientar os alunos ao longo de todo o curso/workshop, explicando detalhes, resultados esperados, avaliação e resultados de aprendizagem por sessão.
Saúde e Segurança	

5. Execução

Atividades de ensino, Procedimentos, Reflexões	**Estrutura do Workshop** 1. Introdução [Sessão 1: 2 horas, 3 atividades] 1.1. Visão geral de Data Mining e Machine Learning: perspetiva histórica, tarefas/problemas, metodologias, técnicas e ferramentas (Weka, R, ...), tarefas (classificação, clustering, ...), aplicações, direcione a conversa para a modelagem de dados (comece por conjuntos de dados tabulares e mostre exemplos). [40 minutos de debate] 1.2. Debate com os alunos sobre a modelação de dados não estruturados (como imagens e texto) para processamento automático. Discutir soluções alternativas com os alunos, introduzindo assuntos relevantes (saco de palavras, sinónimos, localização do texto: título, resumo, ...). [40 minutos, debate] 1.3. Introduza a semelhança de documentos e sua relevância na mineração de texto. [40 minutos, debate] 2. Modelação de texto 1 [Sessão 2: 2 horas, 4 atividades] 2.1. Consulte a modelagem de texto e comece a demonstrar uma função de semelhança de texto com o modelo booleano, mostre exemplos, pergunte como encontrar semelhanças entre documentos? [40 minutos de demonstração, hands-on] 2.2. Explique os corpora que usaremos (deve ter pequenos documentos com um léxico muito reduzido; incluir documentos onde o TF pode fazer a diferença quando comparado com o modelo booleano). [20

minutos, sessão expositiva]

2.3. Introduzir operadores de conjunto (união, intersecção, etc.). [20 minutos, expositivo, demonstração]

2.4. Peça aos alunos em equipes de 3 ou 4 pessoas que implementem uma função de similaridade de texto no Excel e testem-na, cada grupo explica quais operadores de conjunto estão usando e por quê. [40 minutos, hands-on]

3. Modelação de texto 2 [Sessão 3: 2 horas, 4 atividades]

3.1. Introduzir álgebra vetorial: produto interno e cosseno. Facilitar a discussão sobre os princípios matemáticos por trás do processamento de texto. [20 minutos, expositivo, demonstração]

3.2. Envolva os alunos em atividades práticas explorando operadores de conjunto e álgebra vetorial para calcular a semelhança entre documentos de texto em um corpus. Os alunos debatem em equipas para discutir/encontrar relações entre operadores de conjuntos e álgebra vetorial (intersecção e produto interno, união e soma). Refatore implementações do Excel da função de semelhança para usar uma combinação de operadores de conjunto e vetor. [40 minutos, sessão expositiva]

3.3. Mostre o impacto/relevância da frequência de termos em comparação com Boolean ao calcular a semelhança entre documentos. [20 minutos, sessão expositiva]

3.4. Os alunos refatoram sua implementação para implementar o modelo de TF e encontram semelhanças usando os mesmos corpora de antes. Agora deve ficar claro que o cosseno funciona bem. [40 minutos, sessão expositiva]

4. Modelação de texto 3 [Sessão 4: 2 horas, 3 atividades]

4.1. Discuta as alterações e o problema que pode surgir quando a semelhança se baseia apenas no FT (os termos que estão presentes em todos os documentos não têm poder discriminatório). [20 minutos, sessão expositiva]

4.2. Os alunos encontram soluções para este problema. [60 minutos, sessão expositiva]

4.3. Introduza IDF e TFxIDF. Discutir, projetar e implementar com os alunos uma medida de similaridade baseada no modelo TFxIDF para ser usada para documentos de brinquedos com léxico reduzido. [40 minutos, demonstração]

5. Implementação [Sessão 5: 2 horas, 3 atividades]

5.1. Introduza R, Python ou outras bibliotecas para mineração de texto. [30 minutos, sessão expositiva]

5.2. Configure o ambiente de desenvolvimento para mineração de texto. O professor preparou um guia e forneceu-o aos alunos. [30 minutos, trabalho de projeto]

5.3. Os alunos, em equipa, implementam uma função para calcular a semelhança entre dois documentos usando R, Python ou outro. Todas as equipas de alunos terão uma função de funcionamento no final desta fase (os professores preparam com antecedência uma implementação para prover a todos, se necessário). [60 minutos, hands-on]

	6. Projeto culminante 1 [Sessão 6: 2 horas, 1 atividade] 6.1. As equipas dos alunos são desafiadas a encontrar, refatorar e implementar o melhor algoritmo para uma função de similaridade. [120 minutos, investigação, trabalho de projeto] 7. Culminando o projeto 2 [Sessões 7 e 8: 4 horas, 1 atividade] 7.1. Dê aos alunos um corpus previamente anotado para semelhança entre as consultas/documentos estáticos do teste (cada documento do corpus será anotado com a classificação de semelhança para cada um dos casos de teste), reserve um conjunto de validação (anotado também; pode ser dois ou três documentos de texto, cada um com alguns termos como se fosse uma consulta para um motor de busca; para cada uma dessas "consultas" forneça a classificação do documentos no corpus dos alunos, isso será usado para calcular F1 no final e anunciar o vencedor). Dar tempo para que os alunos implementem e ajustem sua função de semelhança para obter os melhores resultados em seu corpus, incorporando feedback e orientação de facilitadores que apresentarão aos alunos técnicas de pré-processamento enquanto a discussão progride (remoção de palavras-paradas, derivação, etc.). [240 minutos, trabalho prático, hands-on] 8. Link para motores de busca [Sessão 9: 2 horas, 2 atividades] 8.1. Um "documento" pode ser uma consulta como as que usamos em motores de busca como o Google, como "camélias porto" ou "visão computacional". Mostrar em real usando o Google. Dê aos alunos o conjunto de validação, ou seja, três documentos, cada um com algumas palavras-chave, como se fossem uma consulta para um motor de busca; executar as funções de semelhança para fornecer uma classificação dos documentos no corpus e dar aos alunos a melhor classificação. Calcule o vencedor usando a medida F1. Explique aos alunos o que é F1, Recall e Precisão. [60 minutos de sessão expositiva] 8.2. Debate, reflexão e conclusões. [60 minutos de sessão expositiva]
Avaliação das aprendizagens – Avaliação do ensino	**Avaliação e Reflexão** 1. Avaliar a compreensão dos alunos e a aplicação de operadores e conceitos de álgebra vetorial e de conjunto através de avaliações baseadas em projetos, apresentações e reflexões escritas. 2. Incentivar os alunos a refletir sobre as suas experiências de aprendizagem, destacando a relação entre a matemática e a mineração de texto. 3. Peça aos alunos que projetem e apresentem uma metodologia de mecanismo de pesquisa e um protótipo não funcional usando o que aprenderam.
Apresentação - Relatórios - Partilha	1. Uma função que calcula a semelhança entre dois documentos de texto. 2. Uma apresentação em PowerPoint descrevendo uma metodologia e um protótipo não funcional de um novo motor de busca proposto pelos alunos usando o que desenvolvemos no workshop (a função de semelhança de documentos).

Extensões - Outras Informações	

STEAME ACADEMY Protótipo/Guia para Aprendizagem e Abordagem da Criatividade
Formulação do Plano de Ação

Principais passos na abordagem de aprendizagem SATEAME:

ETAPA I: Preparação por um ou mais professores

1. Formular reflexões iniciais sobre os sectores/áreas temáticas a abranger
2. Envolver o mundo do ambiente em geral / trabalho / negócios / pais / sociedade / meio ambiente / ética
3. Faixa Etária Alvo dos Alunos - Associando-se ao Currículo Oficial - Definição de Metas e Objetivos
4. Organização das tarefas das partes envolvidas - Designação do Coordenador - Locais de trabalho, etc.

ETAPA II: Formulação do Plano de Ação (Etapas 1-18)

Preparação (pelos professores)

1. Relação com o Mundo Real – Reflexão
2. Incentivo – Motivação
3. Formulação de um problema (possivelmente em fases ou fases) resultante do acima exposto

Desenvolvimento (pelos alunos) – Orientação e Avaliação (em 9-11, pelos professores)

4. Criação de Background - Pesquisa / Recolha de Informação
5. Simplifique o problema - Configure o problema com um número limitado de requisitos
6. Case Making - Designing - identificação de materiais para construção / desenvolvimento / criação
7. Construção - Workflow - Implementação de projetos
8. Observação-Experimentação - Conclusões Iniciais
9. Documentação - Pesquisa de Áreas Temáticas (campos de IA) relacionadas com o tema em estudo – Explicação baseada em Teorias Existentes e/ou Resultados Empíricos
10. Recolha de resultados/informações com base nos pontos 7, 8 e 9
11. Primeira apresentação em grupo pelos alunos

Configuração e Resultados (pelos alunos) – Orientação e Avaliação (pelos professores)

12. Configurar modelos STEAME para descrever/representar/ilustrar os resultados
13. Estudar os resultados em 9 e tirar conclusões, usando 12
14. Aplicações no Quotidiano - Sugestões para o Desenvolvimento 9 (Empreendedorismo - SIL Days)

Revisão (por professores)

15. Reveja o problema e reveja-o em condições mais exigentes

Conclusão do Projeto (pelos alunos) – Orientação e Avaliação (pelos professores)

16. Repita as etapas 5 a 11 com requisitos adicionais ou novos, conforme formulado em 15
17. Investigação - Estudos de Caso - Expansão - Novas Teorias - Testando Novas Conclusões
18. Apresentação de Conclusões - Táticas de Comunicação.

ETAPA III: STEAME ACADEMY Ações e Cooperação em Projetos Criativos para alunos da escola

Título do Projeto: _____
Breve Descrição/Esboço dos Arranjos Organizacionais / Responsabilidades pela Ação

PALCO	**Atividades/Passos** Professor 1(T1) Cooperação com o T2 e orientação estudantil	**Atividades / Passos** **Por Estudantes** Grupo etário: ____	**Atividades / Passos** Professor 2 (T2) Cooperação com T1 e orientação estudantil
Um	Preparação das etapas 1,2,3		Cooperação na etapa 3
B	Orientação na etapa 9	4,5,6,7,8,9,10	Orientação de suporte na etapa 9
C	Avaliação Criativa	11	Avaliação Criativa
D	Orientações	12	Orientações
E	Orientações	13 (9+12)	Orientações
F	Organização (SIL) STEAME na Vida	14 Reunião com representantes empresariais	Organização (SIL) STEAME na Vida
G	Preparação da etapa 15		Cooperação na etapa 15
H	Orientações	16 (repetição 5-11)	Orientações de Suporte
Eu	Orientações	17	Orientações de Suporte
K	Avaliação Criativa	18	Avaliação Criativa