Co-funded by the European Union

Financé par l'Union européenne. Les points de vue et opinions exprimés n'engagent toutefois que leurs auteurs et ne reflètent pas nécessairement ceux de l'Union européenne ou de l'Agence exécutive européenne pour l'éducation et la culture (EACEA). Ni l'Union européenne ni l'EACEA ne peuvent en être tenus responsables.

STEAME ACADEMY
FACILITATION PÉDAGOGIQUE PLAN D'APPRENTISSAGE ET DE CRÉATIVITÉ (PLAN L&C) - NIVEAU 2 SERVICE ENSEIGNANTS :
Text Mining : ces documents sont-ils les mêmes ?

S	T	Eng	Un	M	ORL
●		●		●	

1. Vue d'ensemble

Titre	Text Mining : ces documents sont-ils les mêmes ?		
Question ou sujet moteur	Comment les moteurs de recherche peuvent-ils trouver des résultats pour une recherche d'utilisateur basée sur des mots-clés ? Comment les ordinateurs identifient-ils les documents texte portant sur les mêmes sujets ? Comment les algorithmes informatiques modélisent-ils les données non structurées pour le traitement numérique ?		
Âges, grades, ...	16 à 18 ans	De la 10e à la 12e année	
Durée, chronologie, activités	18 heures	9 séances de 2 heures chacune	21 activités
Alignement du programme d'études	Exploration de données, apprentissage automatique, modélisation de données non structurées, programmation informatique		
Contributeurs, Partenaires			
Résumé - Synopsis	Les étudiants sont initiés à l'exploration de données et à l'apprentissage automatique en se concentrant sur les sujets fondamentaux du traitement numérique du texte. La similarité du texte est explorée en montrant ses fondements mathématiques, l'intersection des ensembles et le cosinus entre deux vecteurs. Les étudiants travaillent en équipe pour mettre en place un outil simple permettant de mesurer la similitude entre deux documents texte. Au cours des dernières sessions, les étudiants sont mis au défi de participer à un concours afin d'identifier la meilleure mise en œuvre. Au cours de toutes les sessions, les étudiants sont initiés aux méthodes clés de prétraitement de texte, comme les mots vides et le stemming. La dernière session se termine en poussant les étudiants à discuter et à identifier les ressemblances entre leur		

	implémentation et un moteur de recherche et, à partir de là, à concevoir un moteur de recherche en utilisant leur ancienne implémentation pour la similarité de texte.
Références, remerciements	

2. Cadre de la STEAME ACADEMY*	
Coopération des enseignants	Enseignant 1 (Sciences) • Machine learning et data mining : vue d'ensemble du domaine, architecture générale, applications courantes à la vie quotidienne, problématiques courantes • Modélisation de données pour le traitement numérique : représentation de données structurées pour l'apprentissage automatique et l'exploration de données, modélisation de données non structurées pour l'apprentissage automatique et l'exploration de données • Fouille de texte : aperçu du domaine, concepts de base, modélisation (modèle booléen, modèle TF, modèle TFxIDF), similitude des documents texte, prétraitement, tâches principales et applications Enseignant 2 (Ingénierie) • Programmation Python pour l'exploration de texte • Algèbre vectorielle dans Excel Enseignant 3 (Mathématiques) • Fonction cosinus, algèbre vectorielle • Opérateurs de plateau, intersection L'enseignant 1 coopère avec l'enseignant 2 et l'enseignant 3 pour : - identifier les bibliothèques Python à utiliser pour l'exploration de texte - identifier les fonctions Excel à utiliser pour l'algèbre vectorielle - Créer les exercices et le défi L'enseignant 1 coopère avec l'enseignant 2 pour : - rassembler les corpus nécessaires aux activités pratiques - Annoter les corpus nécessaires pour chaque exercice L'enseignant 1 coopère avec l'enseignant 3 pour : - Introduire des opérateurs d'ensemble dans un environnement de traitement de texte - Introduire l'algèbre vectorielle dans un environnement de traitement de texte.
Organisation STEAME in Life (SiL)	La dernière session se termine en poussant les étudiants à discuter et à identifier les ressemblances entre leur implémentation et un moteur de recherche, tel que Google, et, à partir de là, à concevoir un moteur de recherche utilisant leur ancienne implémentation pour la similarité de texte.

Formulation du plan d'action

Phase préparatoire

1. Exploration de données de recherche et apprentissage automatique : applications traditionnelles et de pointe ; lien vers les moteurs de recherche et l'IA générative ; faire référence à des cas de données structurées et non structurées ; Passer en revue les principaux défis du Data Mining et du Machine Learning (modélisation des données, multidimensionnalité, surapprentissage, données manquantes, volume de données, Big Data, expliquer versus prévoir, ...)
2. Rassembler et annoter des corpus pour les exercices
3. Configurer l'environnement de programmation Python (docker, dépôt dans Github à cloner, autre)

Structure de l'atelier

1. Introduction
   1.1. Vue d'ensemble du Data Mining et du Machine Learning : perspective historique, tâches/problèmes, applications, orientez la discussion vers la modélisation des données (commencez par des jeux de données tabulaires et montrez des exemples, puis demandez des solutions pour les données non structurées telles que les images et le texte). Discutez de ces solutions avec les élèves.

2. Modélisation de texte
   2.1. Expliquer les corpus que nous utiliserons (doit avoir de petits documents avec un lexique très réduit ; inclure des documents où TF peut faire une différence par rapport au modèle booléen).
   2.2. Référez-vous à nouveau à la modélisation de texte et commencez par le modèle booléen, montrez des exemples, demandez comment trouver des similitudes entre les documents ?
   2.3. Introduire des opérateurs de décor (union, intersection, etc.).
   2.4. Demandez aux élèves en équipe de 3 ou 4 d'implanter un outil dans Excel pour implanter leurs solutions et les tester ; Chaque groupe explique quels opérateurs de set ils utilisent et pourquoi.
   2.5. Introduisez l'algèbre vectorielle : produit interne et cosinus. Les élèves discutent et tentent de trouver des relations entre les opérateurs d'ensemble et l'algèbre vectorielle (intersection et produit interne, union et somme). Refactorisez leur implémentation Excel pour utiliser maintenant une combinaison d'opérateurs d'ensemble et de vecteur.
   2.6. Montrer l'impact/la pertinence de la fréquence des termes par rapport au booléen ; les étudiants refactorisent pour mettre en œuvre le modèle TF et trouvent des similitudes en utilisant les mêmes corpus qu'auparavant. Il devrait maintenant être clair que le cosinus fonctionne correctement.
   2.7. Discutez des changements et du problème qui pourrait survenir lorsque la similitude est basée uniquement sur le TF (les termes présents dans tous les documents n'ont pas de pouvoir discriminatoire) ; Demandez aux élèves de trouver des solutions à ce problème.
   2.8. Présenter IDF et TFxIDF ; discuter, concevoir et mettre en œuvre dans Excel une mesure de similarité basée sur le modèle TFxIDF à utiliser

pour les documents jouets à lexique réduit.

3. Implémentation
   3.1. Introduisez R, Python ou d'autres bibliothèques pour l'exploration de texte.
   3.2. Implémentez une fonction pour calculer la similarité entre deux documents à l'aide de R, Python ou autre. À la fin de cette phase, toutes les équipes d'étudiants auront une fonction de course. ou bien fournir une mise en œuvre de base pour tous.

4. Exploration mathématique
   4.1. Faites participer les élèves à des activités pratiques explorant les opérateurs d'ensemble et l'algèbre vectorielle pour calculer la similitude entre les documents texte d'un corpus.
   4.2. Animer la discussion sur les principes mathématiques qui sous-tendent le traitement de texte.

5. Projets culminants
   5.1. Les équipes d'étudiants sont mises au défi de trouver, de refactoriser et d'implémenter le meilleur algorithme pour une fonction de similarité.
   5.2. Donnez aux étudiants un corpus préalablement annoté pour la similarité entre les requêtes/documents de test statique (chaque document du corpus sera annoté avec le classement de similarité pour chacun des cas de test), réservez un ensemble de validation (annoté également ; il peut s'agir de deux ou trois documents textes, chacun avec quelques termes comme s'il s'agissait d'une requête pour un moteur de recherche ; pour chacune de ces « requêtes », fournissez le classement de la documents dans le corpus des étudiants, cela sera utilisé pour calculer F1 à la fin et annoncer le gagnant).
   5.3. Donnez aux élèves le temps de mettre en œuvre et d'affiner leur fonction de similarité afin d'obtenir les meilleurs résultats sur leur corpus, en intégrant les commentaires et les conseils des animateurs qui initieront les élèves aux techniques de prétraitement au fur et à mesure que la discussion progresse (suppression des mots vides, racinisation, etc.).

6. Lien vers les moteurs de recherche
   6.1. Un « document » peut être une requête comme celles que nous utilisons dans les moteurs de recherche comme Google, comme « camélias porto » ou « computer vision ». Afficher en vrai à l'aide de Google. Donnez aux étudiants l'ensemble de validation, c'est-à-dire trois documents, chacun avec quelques mots-clés, comme s'il s'agissait d'une requête pour un moteur de recherche ; Exécutez les fonctions de similarité pour fournir un classement des documents du corpus et donner aux étudiants le meilleur classement. Calculez le gagnant à l'aide de la mesure F1. Expliquez aux élèves ce qu'est la F1, la mémorisation et la précision.

	Évaluation et réflexion 1. Évaluez la compréhension et l'application par les élèves des opérateurs et des concepts de l'algèbre ensembliste et vectorielle au moyen d'évaluations basées sur des projets, de présentations et de réflexions écrites. 2. Encouragez les élèves à réfléchir à leurs expériences d'apprentissage, en soulignant la relation entre les mathématiques et l'exploration de textes. Demandez aux élèves de concevoir et de présenter une méthodologie de moteur de recherche et un prototype non fonctionnel à l'aide de ce qu'ils ont appris.

[*] *en cours d'élaboration, les derniers éléments du cadre*

3. Objectifs et méthodologies

Buts et objectifs d'apprentissage	1. Comprendre les concepts et techniques génériques de modélisation et de traitement utilisés dans le text mining 2. Explorer les liens interdisciplinaires entre l'exploration de texte, les moteurs de recherche et l'algèbre vectorielle 3. Illustrer la similitude entre les documents texte et d'autres ensembles de données non structurées en tant qu'applications de l'algèbre vectorielle
Résultats d'apprentissage et résultats attendus	**Résultats d'apprentissage** A. Discuter de sujets de haut niveau liés aux domaines de l'exploration de texte et des moteurs de recherche B. Décrire la relation entre l'algèbre vectorielle, la théorie des ensembles et la fouille de texte C. Appliquer des techniques de base de text mining pour répondre à des cas d'utilisation simples **Résultats attendus** 1. Fonction de similarité de document texte en R, Python ou autre langage de programmation
Connaissances préalables et prérequis	1. Connaissances fondamentales de l'algèbre vectorielle 2. Familiarité avec Excel 3. Compétences de base en programmation de logiciels 4. Utilisation pratique des outils informatiques
Motivation, méthodologie, stratégies, échafaudages	1. Répartissez les élèves en petites équipes (3 ou 4 élèves). 2. Concevez une solution, mettez-la en œuvre, testez et affinez de manière itérative. Utilisez une méthodologie de développement itérative. 3. Mettez en évidence les liens entre l'algèbre vectorielle, les modèles de documents, le cosinus et la similitude. 4. Explorez les moteurs de recherche pour mettre en évidence les relations entre la similarité de texte et les résultats des moteurs de recherche. 5. Guidez les étudiants à travers un chemin évolutif des modèles les plus

	simples (booléens) aux approches plus complexes (TFxIDF), en introduisant les défis étape par étape tout en dissertant 8ng dans Excel avec des implémentations de base pour de petits documents avec une ou deux phrases courtes et quelques termes distincts, à partir d'un lexique de 10 ou 20 termes.

4. Préparation et moyens

Préparation, configuration de l'espace, *conseils de dépannage*	L'atelier se déroulera dans une salle de classe d'environ 20 élèves, par groupes de 3 à 4 élèves. Idéalement, l'aménagement de la salle de classe sera organisé en 5 à 7 groupes de tables où les élèves de chaque équipe pourront s'asseoir face à face. La salle a besoin d'un projecteur et d'un mur pour les présentations à tous, ainsi que d'un tableau blanc avec des stylos pour discuter des idées.
Ressources, outils, matériel, pièces jointes, équipement	Un dépôt dans GDrive, Teams, Github ou tout autre fournisseur doit être préparé à l'avance avec tout l'environnement de programmation (R, Python, ...) et les corpus nécessaires aux sessions pratiques. Un document doit être fourni pour guider les étudiants tout au long du cours/atelier, expliquant les détails, les résultats attendus, l'évaluation et les résultats d'apprentissage par session.
Santé et sécurité	

5. Mise en œuvre

Activités pédagogiques, procédures, réflexions	**Structure de l'atelier** 1. Introduction [Séance 1 : 2 heures, 3 activités] 1.1. Vue d'ensemble du Data Mining et du Machine Learning : perspective historique, tâches/problèmes, méthodologies, techniques et outils (Weka, R, ...), tâches (classification, clustering, ...), applications, orienter l'intervention vers la modélisation des données (commencer par des jeux de données tabulaires et montrer des exemples). [Débat de 40 minutes] 1.2. Débat avec les élèves sur la modélisation de données non structurées (telles que des images et du texte) pour le traitement automatique. Discutez avec les étudiants des solutions alternatives tout en introduisant les éléments pertinents (sac de mots, synonymes, emplacement du texte : titre, résumé, ...). [40 minutes, débat] 1.3. Présentez la similitude des documents et leur pertinence dans l'exploration de texte. [40 minutes, débat] 2. Modélisation de texte 1 [Séance 2 : 2 heures, 4 activités] 2.1. Reportez-vous à la modélisation de texte et commencez à démontrer une fonction de similarité de texte avec le modèle booléen, montrez

des exemples, demandez comment trouver des similitudes entre les documents ? [Démonstration de 40 minutes, pratique]

2.2. Expliquer les corpus que nous utiliserons (doit avoir de petits documents avec un lexique très réduit ; inclure des documents où TF peut faire une différence par rapport au modèle booléen). [20 minutes, séance d'exposition]

2.3. Introduire des opérateurs de décor (union, intersection, etc.). [20 minutes, exposé, démonstration]

2.4. Demandez aux élèves en équipes de 3 ou 4 d'implémenter une fonction de similarité de texte dans Excel et de la tester ; chaque groupe explique quels opérateurs de jeu ils utilisent et pourquoi. [40 minutes, pratique]

3. Modélisation de texte 2 [Séance 3 : 2 heures, 4 activités]

3.1. Introduisez l'algèbre vectorielle : produit interne et cosinus. Animer la discussion sur les principes mathématiques qui sous-tendent le traitement de texte. [20 minutes, exposé, démonstration]

3.2. Faites participer les élèves à des activités pratiques explorant les opérateurs d'ensemble et l'algèbre vectorielle pour calculer la similitude entre les documents texte d'un corpus. Les élèves débattent en équipe pour discuter/trouver des relations entre les opérateurs d'ensembles et l'algèbre vectorielle (intersection et produit interne, union et somme). Refactorisez les implémentations Excel de la fonction de similarité pour utiliser une combinaison d'opérateurs d'ensemble et de vecteur. [40 minutes, séance d'exposition]

3.3. Montrez l'impact/la pertinence de la fréquence des termes par rapport au booléen lors du calcul de la similarité entre les documents. [20 minutes, séance d'exposition]

3.4. Les étudiants refactorisent leur implémentation pour implémenter le modèle TF et trouvent des similitudes en utilisant les mêmes corpus qu'auparavant. Il devrait maintenant être clair que le cosinus fonctionne correctement. [40 minutes, séance d'exposition]

4. Modélisation de texte 3 [Séance 4 : 2 heures, 3 activités]

4.1. Discutez des changements et du problème qui pourrait survenir lorsque la similitude est basée uniquement sur le TF (les termes présents dans tous les documents n'ont pas de pouvoir discriminatoire). [20 minutes, séance d'exposition]

4.2. Les élèves trouvent des solutions à ce problème. [60 minutes, séance d'exposition]

4.3. Présentation de IDF et TFxIDF. Discuter, concevoir et mettre en œuvre avec les élèves une mesure de similarité basée sur le modèle TFxIDF à utiliser pour les documents jouets à lexique réduit. [40 minutes, démonstration]

5. Mise en œuvre [Séance 5 : 2 heures, 3 activités]

5.1. Introduisez R, Python ou d'autres bibliothèques pour l'exploration de texte. [30 minutes, séance d'exposition]

5.2. Configurer l'environnement de développement pour l'exploration de texte. L'enseignant a préparé un guide et l'a fourni aux élèves. [30 minutes, travail de projet]

5.3. Les étudiants, en équipe, implémentent une fonction pour calculer la similitude entre deux documents à l'aide de R, Python ou autre. À la fin de cette phase, toutes les équipes d'élèves auront une fonction de

	fonctionnement (les enseignants préparent à l'avance une mise en œuvre pour prévoir tous les besoins en cas de besoin). [60 minutes, pratique] 6. Projet culminant 1 [Séance 6 : 2 heures, 1 activité] 6.1. Les équipes d'étudiants sont mises au défi de trouver, de refactoriser et d'implémenter le meilleur algorithme pour une fonction de similarité. [120 minutes, recherche, travail de projet] 7. Projet culminant 2 [Séances 7 et 8 : 4 heures, 1 activité] 7.1. Donnez aux étudiants un corpus préalablement annoté pour la similarité entre les requêtes/documents de test statique (chaque document du corpus sera annoté avec le classement de similarité pour chacun des cas de test), réservez un ensemble de validation (annoté également ; il peut s'agir de deux ou trois documents textes, chacun avec quelques termes comme s'il s'agissait d'une requête pour un moteur de recherche ; pour chacune de ces « requêtes », fournissez le classement de la documents dans le corpus des étudiants, cela sera utilisé pour calculer F1 à la fin et annoncer le gagnant). Donnez aux élèves le temps de mettre en œuvre et d'affiner leur fonction de similarité afin d'obtenir les meilleurs résultats sur leur corpus, en intégrant les commentaires et les conseils des animateurs qui initieront les élèves aux techniques de prétraitement au fur et à mesure que la discussion progresse (suppression des mots vides, racinisation, etc.). [240 minutes, travail de projet, pratique] 8. Lien vers les moteurs de recherche [Séance 9 : 2 heures, 2 activités] 8.1. Un « document » peut être une requête comme celles que nous utilisons dans les moteurs de recherche comme Google, comme « camélias porto » ou « computer vision ». Afficher en vrai à l'aide de Google. Donnez aux étudiants l'ensemble de validation, c'est-à-dire trois documents, chacun avec quelques mots-clés, comme s'il s'agissait d'une requête pour un moteur de recherche ; Exécutez les fonctions de similarité pour fournir un classement des documents du corpus et donner aux étudiants le meilleur classement. Calculez le gagnant à l'aide de la mesure F1. Expliquez aux élèves ce qu'est la F1, la mémorisation et la précision. [Séance d'exposition de 60 minutes] 8.2. Débat, réflexion et conclusions. [Séance d'exposition de 60 minutes]
Évaluation - Évaluation	**Évaluation et réflexion** 1. Évaluez la compréhension et l'application par les élèves des opérateurs et des concepts de l'algèbre ensembliste et vectorielle au moyen d'évaluations basées sur des projets, de présentations et de réflexions écrites. 2. Encouragez les élèves à réfléchir à leurs expériences d'apprentissage, en soulignant la relation entre les mathématiques et l'exploration de textes. 3. Demandez aux élèves de concevoir et de présenter une méthodologie de moteur de recherche et un prototype non fonctionnel à l'aide de ce qu'ils ont appris.

Présentation - Reporting - Partage *Prolongations - Autres informations*	1. Fonction qui calcule la similitude entre deux documents texte. 2. Une présentation en PowerPoint décrivant une méthodologie et un prototype non fonctionnel d'un nouveau moteur de recherche proposé par les étudiants à partir de ce que nous avons développé dans l'atelier (la fonction de similarité de document).

STEAME ACADEMY Prototype/Guide pour l'Approche de l'Apprentissage et de la Créativité
Formulation du plan d'action

Grandes étapes de l'approche d'apprentissage STEAME :

ÉTAPE I : Préparation par un ou plusieurs enseignants

1. Formuler des premières réflexions sur les secteurs/domaines thématiques à couvrir
2. S'engager dans le monde de l'environnement au sens large / travail / affaires / parents / société / environnement / éthique
3. Groupe d'âge cible des élèves - S'associer au programme officiel - Fixer des buts et des objectifs
4. Organisation des tâches des parties concernées - Désignation du coordinateur - Lieux de travail, etc.

ÉTAPE II : Formulation du plan d'action (étapes 1 à 18)

Préparation (par les enseignants)

1. Relation avec le monde réel – Réflexion
2. Incitation – Motivation
3. Formulation d'un problème (éventuellement par étapes ou phases) résultant de ce qui précède

Développement (par les élèves) – Orientation et évaluation (dans le 9-11, par les enseignants)

4. Création d'arrière-plan - Recherche / Collecte d'informations
5. Simplifiez le problème : configurez le problème avec un nombre limité d'exigences
6. Case Making - Designing - Identification des matériaux pour la construction / l'aménagement / la création
7. Construction - Flux de travail - Mise en œuvre des projets
8. Observation-Expérimentation - Conclusions initiales
9. Documentation - Recherche de domaines thématiques (domaines d'IA) liés au sujet étudié - Explication basée sur des théories existantes et/ou des résultats empiriques
10. Collecte des résultats / informations sur la base des points 7, 8, 9
11. Première présentation de groupe par les étudiants

Configuration et résultats (par les étudiants) – Orientation et évaluation (par les enseignants)

12. Configurer les modèles STEAME pour décrire / représenter / illustrer les résultats
13. Étudier les résultats en 9 et tirer des conclusions, en utilisant 12
14. Applications dans la vie quotidienne - Suggestions pour développer 9 (Entrepreneuriat - SIL days)

Évaluation (par les enseignants)

15. Examinez le problème et examinez-le dans des conditions plus exigeantes

Réalisation de projet (par les étudiants) – Orientation et évaluation (par les enseignants)

16. Répéter les étapes 5 à 11 avec les exigences supplémentaires ou nouvelles formulées à l'article 15
17. Investigation - Etudes de cas - Expansion - Nouvelles théories - Mise à l'épreuve de nouvelles conclusions
18. Présentation des conclusions - Tactiques de communication.

ÉTAPE III : STEAME ACADEMY Actions et coopération dans des projets créatifs pour les élèves

Titre du projet : _____
Brève description/aperçu des dispositions organisationnelles / responsabilités d'action

ÉTAPE	**Activités/Étapes** Enseignant 1(T1) Coopération avec T2 et l'orientation des étudiants	**Activités / Étapes** **Par les étudiants** Groupe:____	**Activités / Étapes** Enseignant 2 (T2) Coopération avec T1 et Orientation des étudiants
Un	Préparation des étapes 1,2,3		Coopération à l'étape 3
B	Orientation à l'étape 9	4,5,6,7,8,9,10	Accompagnement du support à l'étape 9
C	Évaluation créative	11	Évaluation créative
D	Direction	12	Direction
E	Direction	13 (9+12)	Direction
F	Organisation (SIL) STEAME dans la vie	14 Rencontre avec des représentants d'entreprises	Organisation (SIL) STEAME dans la vie
G	Préparation de l'étape 15		Coopération à l'étape 15
H	Direction	16 (répétitions 5-11)	Conseils d'assistance
Je	Direction	17	Conseils d'assistance
K	Évaluation créative	18	Évaluation créative