



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 STEAME FACILITATION PÉDAGOGIQUE PLAN D'APPRENTISSAGE ET DE CRÉATIVITÉ (PLAN L&C) - NIVEAU 1 ÉLÈVES ENSEIGNANTS : Text Mining : ces documents sont-ils les mêmes ?

S **T** **Eng** **A** **M** **Ent**



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	132 heures		17 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	<p>Enseignant 1 (Sciences)</p> <ul style="list-style-type: none">• 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 <p>Enseignant 2 (Ingénierie)</p> <ul style="list-style-type: none">• Programmation Python pour l'exploration de texte• Algèbre vectorielle dans Excel <p>Enseignant 3 (Mathématiques)</p> <ul style="list-style-type: none">• Fonction cosinus, algèbre vectorielle• Opérateurs de plateau, intersection
Organisation STEAME in Life (SiL)	<p>L'enseignant 1 coopère avec l'enseignant 2 et l'enseignant 3 pour :</p> <ul style="list-style-type: none">- 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 <p>L'enseignant 1 coopère avec l'enseignant 2 pour :</p> <ul style="list-style-type: none">- rassembler les corpus nécessaires aux activités pratiques- Annoter les corpus nécessaires pour chaque exercice <p>L'enseignant 1 coopère avec l'enseignant 3 pour :</p> <ul style="list-style-type: none">- Introduire des opérateurs d'ensemble dans un environnement de traitement de texte- Introduire l'algèbre vectorielle dans un environnement de traitement de texte. <p>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.</p>

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, modélisation de données non structurées telles que des images et du texte).
2. Modélisation de texte
 - 2.1. Rassemblez les corpus qui seront utilisés pour les exercices et le travail de projet (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. Modèles pour représenter du texte, exemples : Modèle booléen, TF, TFxIDF, autre. Reportez-vous à Sac de mots.
 - 2.3. Opérateurs de plateau (union, intersection, etc.).
 - 2.4. Implémenter plusieurs versions d'un outil dans Excel pour calculer la similarité entre deux documents texte basés sur des opérateurs d'ensemble et/ou l'algèbre vectorielle (produit interne, cosinus, intersection et union d'ensembles, ...).
 - 2.5. Introduisez l'algèbre vectorielle : produit interne et cosinus.
 - 2.6. Montrez l'impact/la pertinence de la fréquence des termes par rapport à la valeur booléenne. Indiquez clairement que le cosinus fonctionne correctement.
 - 2.7. Contestations de similitude fondées uniquement sur l'ATF (les termes présents dans tous les documents n'ont aucun pouvoir discriminatoire) ; solutions alternatives.
 - 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. Recherchez R, Python ou d'autres bibliothèques pour l'exploration de texte. Préparez un référentiel et un guide d'installation pour que les étudiants puissent installer cet environnement de développement.
 - 3.2. Implémentez une fonction pour calculer la similarité entre deux documents à l'aide de R, Python ou autre. Fournir une mise en œuvre

pour tous les élèves si nécessaire (si les élèves ne peuvent pas développer la leur en temps voulu).

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. Concevez le projet (meilleure fonction de similarité pour les documents texte) et rédigez le guide.
- 5.2. Rassembler et annoter un corpus et un ensemble de requêtes pour évaluer la précision et le rappel produits par les implémentations des fonctions de similarité de texte par les étudiants. La précision et le rappel seront évalués à partir des requêtes/documents de test statiques (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 texte, chacun avec quelques termes comme s'il s'agissait d'une requête pour un moteur de recherche ; pour chacune de ces « requêtes », indiquez le classement des documents dans le corpus des étudiants, celui-ci sera utilisé pour calculer F1 à la fin et annoncer le gagnant).
- 5.3. Fournir des informations sur les techniques de prétraitement (suppression des mots vides, stemming, etc.) et leur mise en œuvre à l'aide des bibliothèques de fouille de texte sélectionnées pour le projet.

6. Lien vers les moteurs de recherche

- 6.1. Préparez le scénario de validation. 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 itératif.
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, <i>conseils de dépannage</i>	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.
<i>Santé et sécurité</i>	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.

5. Mise en œuvre

Activités pédagogiques, procédures, réflexions	<p>Structure de l'atelier</p> <ol style="list-style-type: none"> 1. Introduction [1 activité, 16 heures] <ol style="list-style-type: none"> 1.1. Vue d'ensemble du Data Mining et du Machine Learning : perspective historique, tâches/problèmes, applications, modélisation de données non structurées telles que des images et du texte). [16 heures] 2. Modélisation de texte [8 activités, 52 heures] <ol style="list-style-type: none"> 2.1. Rassemblez les corpus qui seront utilisés pour les exercices et le travail de projet (doit avoir de petits documents avec un lexique très réduit ; inclure des documents où TF peut faire une différence par rapport au modèle booléen). [16 heures] 2.2. Modèles pour représenter du texte, exemples : Modèle booléen, TF, TFxIDF, autre. Reportez-vous à Sac de mots. [8 heures] 2.3. Décrire les opérateurs d'ensemble pertinents (union, intersection, etc.). [4 heures] 2.4. Implémenter plusieurs versions d'un outil dans Excel pour calculer la similarité entre deux documents texte basés sur des opérateurs d'ensemble et/ou l'algèbre vectorielle (produit interne, cosinus, intersection et union d'ensembles, ...). [4 heures] 2.5. Introduisez l'algèbre vectorielle : produit interne et cosinus. [4 heures] 2.6. Montrez l'impact/la pertinence de la fréquence des termes par rapport à la valeur booléenne. Démontrez à l'aide d'exemples. Indiquez clairement que le cosinus fonctionne correctement. [4 heures] 2.7. Décrire les défis de la similitude lorsqu'elle est fondée uniquement sur la TF (les termes présents dans tous les documents n'ont pas de pouvoir discriminatoire) ; solutions alternatives. Démontrez à l'aide d'exemples. [4 heures] 2.8. Introduire IDF et TFxIDF ; discuter, concevoir et implémenter dans Excel une mesure de similarité basée sur le modèle TFxIDF à utiliser pour les documents jouets avec un lexique réduit. [8 heures]
--	--

3. Mise en œuvre [2 activités, 32 heures]
 - 3.1. Recherchez R, Python ou d'autres bibliothèques pour l'exploration de texte. Préparez un référentiel et un guide d'installation pour que les étudiants puissent installer cet environnement de développement. [16 heures]
 - 3.2. Implémentez une fonction pour calculer la similarité entre deux documents à l'aide de R, Python ou autre. Fournir une mise en œuvre pour tous les élèves si nécessaire (si les élèves ne peuvent pas développer la leur en temps voulu). [16 heures]
4. Exploration mathématique [2 activités, 4 heures]
 - 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. [2 heures]
 - 4.2. Animer la discussion sur les principes mathématiques qui sous-tendent le traitement de texte. [2 heures]
5. Projets culminants [3 activités, 20 heures]
 - 5.1. Concevez le projet (meilleure fonction de similarité pour les documents texte) et rédigez le guide. [8 heures]
 - 5.2. Rassembler et annoter un corpus et un ensemble de requêtes pour évaluer la précision et le rappel produits par les implémentations des fonctions de similarité de texte par les étudiants. La précision et le rappel seront évalués à partir des requêtes/documents de test statiques (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 texte, chacun avec quelques termes comme s'il s'agissait d'une requête pour un moteur de recherche ; pour chacune de ces « requêtes », indiquez le classement des documents dans le corpus des étudiants, celui-ci sera utilisé pour calculer F1 à la fin et annoncer le gagnant). [8 heures]
 - 5.3. Fournir des informations sur les techniques de prétraitement (suppression des mots vides, stemming, etc.) et leur mise en œuvre à l'aide des bibliothèques de fouille de texte sélectionnées pour le projet. [4 heures]
6. Lien vers les moteurs de recherche [1 activité, 8 heures]
 - 6.1. Préparez le scénario de validation. 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. [8 heures]

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

1. Fonction qui calcule la similitude entre deux documents texte.

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).

Prolongations - Autres informations

Ressources pour l'élaboration du modèle de plan d'apprentissage et de créativité de STEAME ACADEMY

Dans le cas de l'apprentissage par le biais d'une activité basée sur un projet

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