**Rapport Professionnel :**

**Table des matières**

1. **Introduction générale**
2. **C6 : Veille technique et réglementaire**
   * **2.1 Objectifs de la veille**
   * **2.2 Planification et organisation de la veille**
   * **2.3 Outils et sources d’informations**
   * **2.4 Résultats de la veille : Analyse des tendances récentes**
3. **C7 : Benchmark des services IA**
   * **3.1 Expression du besoin en intelligence artificielle**
   * **3.2 Critères d’évaluation des services IA**
   * **3.3 Analyse détaillée des services IA**
     + **3.3.1 SentenceTransformer**
     + **3.3.2 Amazon Personalize**
     + **3.3.3 Pytorch**
   * **3.4 Comparaison des solutions et recommandations**
4. **C8 : Paramétrage du service IA**
   * **4.1 Installation et configuration**
   * **4.2 Gestion des accès et authentification**
   * **4.3 Tests fonctionnels**
   * **4.4 Monitorage et documentation technique**
5. **Conclusion générale**

**Introduction générale**

L'intelligence artificielle (IA) a transformé la manière dont les utilisateurs interagissent avec les plateformes numériques, en particulier dans les domaines de la recommandation personnalisée.

Ce rapport présente un projet visant à intégrer un service d'intelligence artificielle dans une plateforme de recommandations de livres.

Dans le cadre de ce projet, trois compétences spécifiques ont été mises en œuvre :

1. Veille technique et réglementaire (C6) : Suivi des tendances récentes et des évolutions dans le domaine de l'intelligence artificielle appliquée aux systèmes de recommandation
2. Benchmark des services IA (C7) : Analyse comparative des services d'IA disponibles sur le marché, évaluation de leur adéquation avec les besoins du projet, et recommandation d'une solution adaptée.
3. Paramétrage et intégration du service IA (C8) : Installation, configuration et intégration du service IA choisi dans l'infrastructure existante, en respectant les spécifications techniques et fonctionnelles du projet.

**C6 : Veille technique et réglementaire**

**2.1 Objectifs de la veille**

L'objectif principal de cette veille est d’identifier, d’analyser et de suivre les tendances récentes et les évolutions dans le domaine de l'intelligence artificielle appliquée aux systèmes de recommandation.

Cette veille est essentielle pour garantir que les technologies et les outils utilisés dans le cadre du projet soient à jour, efficaces, et conformes aux réglementations en vigueur, notamment en matière de protection des données personnelles.

Les objectifs spécifiques sont les suivants :

* Identifier les nouvelles technologies d'IA pour les systèmes de recommandation : exploration des algorithmes et modèles récents, des frameworks utilisés, et des avancées en matière de personnalisation.
* Veiller à l'efficience et à la performance des recommandations : garantir une expérience utilisateur fluide en étudiant les meilleures pratiques d’implémentation de l’IA dans les systèmes de recommandation.

**2.2 Planification et organisation de la veille**

La veille a été organisée sur une base hebdomadaire pour assurer un suivi régulier des évolutions du domaine. Chaque semaine a été dédiée à un thème spécifique en fonction des besoins du projet et des évolutions attendues dans le domaine de l'IA.

Voici la planification établie :

* Semaines 1 à 4 : Revue des nouvelles solutions IA et des modèles de recommandation (tels que les réseaux neuronaux, le filtrage collaboratif, et les modèles hybrides).
* Semaines 5 à 8 : Focus sur la réglementation des données personnelles et son impact sur les systèmes de recommandation utilisant l’IA, avec un accent particulier sur la conformité au RGPD.
* Semaines 9 à 12 : Exploration des solutions de déploiement et d’intégration MLOps (Machine Learning Operations), en lien avec l'automatisation du monitoring et l’optimisation des performances.

Ces phases ont permis une collecte d’informations diversifiée et enrichie, garantissant que tous les aspects critiques du projet aient été couverts avant son implémentation.

**2.3 Outils et sources d’informations**

Pour assurer une veille efficace, des outils d'agrégation ont été utilisés pour centraliser et organiser les informations. Voici les outils principaux choisis en fonction de leurs fonctionnalités et de leur compatibilité avec les besoins du projet :

* Google Alerts : Utilisé pour recevoir des alertes sur les mots-clés spécifiques tels que ‘systèmes de recommandation’, ‘IA’ ou autresUne image contenant texte, logiciel, Page web, Icône d’ordinateur

  Description générée automatiquement

LinkedIn et Twitter : Suivi des publications d'experts dans le domaine de l'intelligence artificielle et des systèmes de recommandation :

* Yann LeCun : Un pionnier de l'apprentissage profond, très actif dans le partage d’avancées en IA.
* Xavier Amatriain : Expert en systèmes de recommandation, ancien responsable chez Netflix.

Une image contenant capture d’écran, Espace lointain, Visage humain, Univers

Description générée automatiquement

Les sources utilisées ont été sélectionnées pour garantir leur fiabilité et leur pertinence mais aussi variées pour permettre d’avoir un nombre de sources différentes importantes :

* LarevueIA : Site francais accès sur l’univers de l’IA passant du NLP, de la Vision au Machine Learning. L’idée était de commencer par bien comprendre comment fonctionne les systèmes de recommandations, quels modèles sont les plus adaptées à chaque type de données (textuelle, vocale …). Le site propose également une ‘newsletter AI news’.
* towardsdatascience : plateforme en anglais qui propose des articles détaillés sur l'intelligence artificielle, le machine learning et les algorithmes de recommandation. Les experts y partagent des cas d'utilisation, des tutoriels et des analyses de performance. Ce site est une référence pour comprendre les concepts avancés dans les systèmes de recommandation, notamment les modèles hybrides, les réseaux de neurones, et le Deep learning.
* KDnuggets : une excellente source pour obtenir des insights sur les algorithmes de recommandation, les technologies utilisées par des grandes entreprises. Ce site est une référence dans le domaine de la science des données, du big data et de l'IA

**2.4 Résultats de la veille : Analyse des tendances récentes**

**1. L'essor des modèles hybrides pour maximiser la précision des recommandations**

L’une des principales tendances identifiées est l’adoption croissante de modèles hybrides dans les systèmes de recommandation. Les modèles hybrides, qui combinent différentes approches telles que le filtrage collaboratif et le filtrage basé sur le contenu, sont capables d'améliorer la précision des recommandations. Selon un article de Vandana Singh sur Towards Data Science (2021), cette hybridation permet de pallier les limites des modèles purement collaboratifs (qui nécessitent un historique d'interactions important) et ceux basés sur le contenu (qui dépendent des caractéristiques explicites des items recommandés).

Ces modèles hybrides peuvent aussi intégrer des données contextuelles comme l’heure, l’appareil utilisé ou encore la localisation pour affiner la pertinence des recommandations. Cette approche est particulièrement intéressante pour les plateformes qui souhaitent proposer des recommandations plus adaptées aux préférences des utilisateurs.

**2. L'importance croissante des Graph Neural Networks (GNN)**

Les Graph Neural Networks (GNN) ont vu leur adoption augmenter dans le domaine des systèmes de recommandation. Ces réseaux de neurones permettent de modéliser des interactions complexes sous forme de graphes, ce qui est particulièrement adapté pour des systèmes où les relations entre utilisateurs et produits sont denses et multidimensionnelles. Mounica Maddikunta, dans un article sur KDnuggets (2020), met en avant que les GNN permettent d'améliorer la précision en capturant les relations entre utilisateurs et éléments en exploitant les connexions de manière plus profonde qu’avec les approches traditionnelles.

Par exemple, un utilisateur qui partage des goûts similaires avec un groupe de personnes pourra recevoir des recommandations basées sur les préférences de ce groupe, même sans interaction directe. Cette capacité à comprendre des relations non linéaires renforce la capacité des systèmes de recommandation à proposer des contenus plus diversifiés et adaptés.

**3. Développement des modèles de recommandation en temps réel**

L’évolution des attentes des utilisateurs pousse les entreprises à développer des modèles de recommandation en temps réel. Ces systèmes doivent être capables de s'adapter rapidement aux nouvelles actions des utilisateurs pour offrir une expérience fluide et personnalisée. Dans un article publié par Julien Simon, (expert en IA chez AWS), sur LarevueIA (2022), il est expliqué que des solutions comme Amazon Personalize utilisent des algorithmes de deep learning optimisés pour ajuster les recommandations en temps réel.

Cette approche est particulièrement bénéfique pour des plateformes telles qu’Amazon ou Netflix, où les données en temps réel (par exemple, les derniers achats ou les films récemment regardés) influencent directement les suggestions futures. Cela permet de maintenir un haut niveau d’engagement et de pertinence dans les recommandations.

**4. L’utilisation des modèles auto-supervisés pour le problème du démarrage à froid**

Le démarrage à froid, c'est-à-dire l'absence de données historiques pour un nouvel utilisateur ou un nouveau produit, constitue un défi majeur dans les systèmes de recommandation. L’adoption croissante de modèles auto-supervisés basés sur des architectures comme BERT (Bidirectional Encoder Representations from Transformers) offre une solution innovante à ce problème. D’après Thomas Wolf, chercheur chez Hugging Face, ces modèles sont capables de générer des recommandations en analysant directement les métadonnées et les descriptions des items, même en l'absence d’historique utilisateur.

L'article de Wolf publié sur Medium (2020) montre que ces modèles peuvent comprendre les relations sémantiques entre des descriptions textuelles et ainsi recommander des livres, par exemple, en fonction de leur contenu plutôt que des interactions passées, rendant le système plus agile dès les premières interactions.

**5. Accent sur l’éthique et la transparence dans les recommandations**

L’éthique et la transparence sont devenues des enjeux centraux dans le développement des systèmes de recommandation. Les utilisateurs souhaitent désormais comprendre pourquoi certains contenus leur sont suggérés. Dans un rapport publié par Clara Tisserant sur LarevueIA (2022), il est mis en avant que les entreprises doivent intégrer des mécanismes de transparence algorithmique pour expliquer les choix de recommandations. Cela passe par des outils permettant aux utilisateurs d’avoir un aperçu des raisons derrière une recommandation, et par des efforts pour minimiser les biais présents dans les modèles. Les systèmes doivent veiller à ne pas favoriser certains types de contenus au détriment d’autres, pour offrir des suggestions plus diversifiées et justes.

Conclusion

Ces tendances montrent que les systèmes de recommandation basés sur l’intelligence artificielle évoluent rapidement pour répondre aux attentes des utilisateurs et aux avancées technologiques. Les modèles hybrides, les Graph Neural Networks et les modèles auto-supervisés figurent parmi les approches les plus prometteuses pour offrir des recommandations plus précises, personnalisées et éthiques. L’optimisation en temps réel et la prise en compte des préoccupations éthiques sont également des éléments essentiels pour améliorer l’expérience utilisateur sur les plateformes de recommandation.

**C7 : Benchmark des services IA**

**3.1 Expression du besoin en intelligence artificielle**

L'objectif de ce projet est d'intégrer une solution d'intelligence artificielle capable de fournir des recommandations personnalisées de livres en fonction de leurs métadonnées, car les données utilisateurs ne sont pas disponibles dans ce projet (**pour l’instant en tout cas**).

Le système devra analyser les descriptions, les catégories, et d'autres informations issues des livres pour générer des suggestions pertinentes basées uniquement sur les données disponibles. Le besoin spécifique en IA est donc axé sur un modèle de recommandation basé sur le contenu, sans données utilisateur.

Les contraintes techniques sont les suivantes :

* Données d'entrée : Les données de base sont sous forme de fichier CSV contenant les colonnes suivantes : title, authors, description, published\_year, average\_rating, categories.
* Base de données : PostgreSQL pour le stockage et la gestion des données.
* Budget : Une solution gratuite est préférée afin de minimiser les coûts.
* Infrastructure : Utilisation d'un environnement local ou basé sur des fichiers CSV, sans besoin d'infrastructure complexe ou gourmande en ressources.

**3.2 Critères d’évaluation des services IA**

Les services d'intelligence artificielle évalués dans le cadre de ce benchmark sont analysés selon les critères suivants :

* Adéquation fonctionnelle : Capacité du service à fournir des recommandations basées sur le contenu, à partir des métadonnées des livres (descriptions, catégories).
* Facilité d'intégration : Compatibilité avec un environnement local utilisant PostgreSQL et des fichiers CSV, ainsi qu'une implémentation via des scripts Python.
* Coût : Solutions gratuites ou à faible coût.
* Personnalisation : Possibilité de construire des modèles sur mesure et d'ajuster les paramètres pour des besoins spécifiques.

**3.3 Analyse détaillée des services IA**

**3.3.1 SentenceTransformer (Solution IA utilisée actuellement)**

* Description : SentenceTransformer est un modèle pré-entraîné pour générer des **embeddings** à partir de textes. Il est utilisé pour comparer les descriptions des livres en calculant leur similarité cosinus.
* Adéquation fonctionnelle : Le modèle permet de générer des recommandations pertinentes sans nécessiter de données utilisateurs, et les résultats sont déjà satisfaisants sans fine-tuning.
* Facilité d'intégration : Compatible avec l'infrastructure locale et les bibliothèques Python (pandas, scikit-learn), il s'intègre facilement avec PostgreSQL et les fichiers CSV.
* Coût : SentenceTransformer est gratuit.
* Personnalisation : Bien que le modèle fonctionne bien sans fine-tuning, un fine-tuning avec PyTorch est prévu pour affiner davantage les recommandations.

**3.3.2 Amazon Personalize**

* Description : Amazon Personalize est un service basé sur le cloud proposé par Amazon Web Services (AWS) qui utilise des algorithmes de machine learning pour générer des recommandations en temps réel. Il s'agit d'une solution clé en main pour les systèmes de recommandation, avec une intégration aisée dans des environnements complexes.
* Adéquation fonctionnelle : Amazon Personalize est principalement conçu pour des systèmes de recommandation basés sur le comportement des utilisateurs. Il peut, toutefois, utiliser des métadonnées pour générer des recommandations, mais il est particulièrement efficace pour les systèmes où les interactions utilisateur sont un facteur majeur.
* une infrastructure cloud avec des coûts d'intégration et d'hébergement. Pour des projets comme celui-ci, basés uniquement sur des fichiers CSV et des données locales, cette solution pourrait être surdimensionnée.
* Coût : Le modèle de tarification à l'usage d'Amazon Personalize peut devenir onéreux si le projet évolue vers une utilisation intensive.

**3.3.3 PyTorch (Solution IA pour fine-tuning futur)**

* Description : PyTorch est une bibliothèque open-source permettant de développer des modèles sur mesure, comme des réseaux de neurones pour des tâches complexes de traitement du langage.
* Adéquation fonctionnelle : PyTorch sera utilisé pour le fine-tuning de SentenceTransformer, ce qui permettra d'améliorer la pertinence des recommandations en ajustant les paramètres du modèle.
* Facilité d'intégration : PyTorch s'intègre parfaitement dans notre environnement local (PostgreSQL, fichiers CSV) via des scripts Python.
* Coût : PyTorch est gratuit.
* Personnalisation : Le fine-tuning avec PyTorch offre une flexibilité pour optimiser les performances du modèle et répondre aux besoins spécifiques du projet.

**3.4 Comparaison des solutions et recommandations**

Pour le moment, SentenceTransformer est utilisé efficacement sans fine-tuning, fournissant des recommandations basées sur la similarité des descriptions. Cependant, dans une phase future, un fine-tuning avec PyTorch permettra de mieux personnaliser le modèle et d'améliorer encore plus la pertinence des recommandations.

**C8 : Paramétrage du service IA**

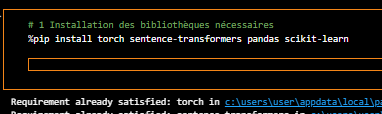
Dans cette section, nous décrivons le processus de configuration du service IA basé sur SentenceTransformer et PyTorch, utilisé pour générer des recommandations de livres en fonction des similarités entre les descriptions des livres.

Ce service a été sélectionné pour ses avantages en matière de flexibilité, de simplicité d’intégration et de gratuité, en adéquation avec l'infrastructure du projet.

**4.1 Installation et configuration**

L'installation sur un environnement local a été réalisée en suivant ces étapes :

1. Pré-requis :
   * Python version 3.8+
   * PostgreSQL pour le stockage des données.
   * Fichier CSV contenant les données des livres à recommander
2. Étapes d’installation de Pytorch :
   * Utilisation de pip pour installer Pytorch et ses dépendances, telles que pandas pour manipuler les fichiers CSV mais aussi les bibliothèques de traitement de données comme numpy et scikit-learn :

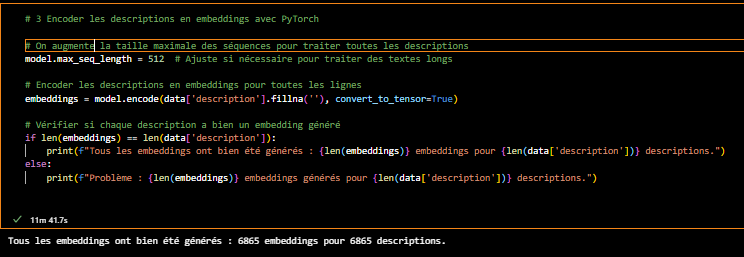


1. Configuration de l’environnement local :
   * Chargement des données du fichier CSV dans un DataFrame avec pandas :

Une image contenant texte, capture d’écran, logiciel, Logiciel multimédia

Description générée automatiquement

* + Utilisation du modèle SentenceTransformer ('all-MiniLM-L6-v2') pour encoder les descriptions en embeddings et calculer la similarité cosinus.



**4.2 Gestion des accès et authentification**

1. Système d'authentification avec JWT :

Pour sécuriser l'API REST qui fournit les recommandations, nous avons mis en place un système d'authentification utilisant JWT (JSON Web Tokens).

* Lorsqu'un utilisateur s'authentifie, un token JWT est généré, contenant des informations sur l'utilisateur, et est ensuite utilisé pour chaque requête suivante dans l'en-tête Authorization : Bearer <token>.
* Chaque requête vers une route protégée nécessite ce token, qui est validé par le serveur avant traitement.

2. Gestion des rôles utilisateurs :

* Les utilisateurs peuvent avoir des rôles différents (administrateur, utilisateur standard), définissant les permissions pour certaines fonctionnalités, telles que :
  + Accès aux recommandations de livres pour les utilisateurs standards.
  + Accès à des fonctionnalités avancées pour les administrateurs, comme la gestion des livres.

**4.3 Tests fonctionnels**

J’ai mis en place des tests pour valider les différentes fonctionnalités du système, en particulier l'authentification et la récupération de données.

1. Tests de connexion et génération de token :

* Test de connexion réussie : Validation de la génération correcte d'un token JWT lors d'une connexion avec des informations valides (administrateur).
* Test de connexion échouée : Validation que les informations incorrectes retournent un code de statut 401 (non autorisé).

2. Tests d'accès sécurisé :

* Test des routes protégées : Seules les requêtes contenant un token JWT valide peuvent accéder aux routes protégées comme /books pour récupérer les recommandations de livres.

3. Tests de récupération de données :

* Test de récupération des livres : Validation de la capacité à récupérer la liste des livres disponibles via un token JWT après authentification.
* Test de récupération d'un livre spécifique : Validation que le système peut récupérer un livre par titre et afficher des informations détaillées (titre, année, note).
* Test de Récupéation de Recommendations : Validation de la capacité à utiliser la fonction créée pour recommander 3 livres en fonction d’un envoyé.

4. Tests de révocation de token :

* Simulation de la déconnexion d'un utilisateur pour révoquer son token, garantissant que les requêtes futures avec ce token ne sont plus autorisées.

**4.4 Monitorage et documentation technique ( A FINIR)**

1. Monitorage des requêtes API :
   * Un système de journalisation a été mis en place pour suivre les tentatives d'authentification et l'utilisation des tokens JWT. Des outils comme Flask-Logging ou Prometheus sont utilisés pour suivre les requêtes réussies ou échouées.
2. Documentation technique :
   * La documentation détaille les procédures d'installation, la configuration des clés secrètes pour les JWT, et l'intégration de l'API REST. Elle inclut également les dépendances nécessaires, comme Flask-JWT-Extended dans un fichier requirements.txt.

**5.Conclusion générale**

Le service IA basé sur SentenceTransformer fonctionne efficacement, même sans fine-tuning, et satisfait les besoins actuels du projet. Une étape de fine-tuning avec PyTorch est envisagée pour améliorer la précision des recommandations.