Sujet de Projet 8INF919 : Plateforme de recommandation des produits Chanel

---

Partie 1 : Analyse approfondie du jeu de données (20 points)

1. **Exploration initiale :**
   - Analyse des colonnes du dataset, notamment category2_code, title, image, et price.
   - Distribution des catégories (category2_code), des prix, et longueur des descriptions (title).
   - Analyse de la qualité et de la diversité des images (dimensions, couleurs, formats).
   - Identification des biais éventuels dans le dataset, comme la sur-représentation de certaines catégories.
2. **Préparation des données :**
   - Prétraitement des images : redimensionnement, normalisation, et augmentation.
   - Nettoyage des descriptions textuelles : suppression des doublons, uniformisation des formats.
   - Vérification des correspondances entre les images, les descriptions, et les catégories.
3. **Visualisation :**
   - Graphiques pour la distribution des catégories et des prix.
   - Exemples visuels des produits pour chaque catégorie majeure.

Partie 2 : Comparaison des embeddings visuels (30 points)

1. **Méthodes pour extraire les embeddings visuels :**
   - **Méthode 1 : Modèle entraîné pour la classification.**
     - Entraîner un modèle CNN (ou Vision Transformer) pour classifier les images selon category2_code.
     - Extraire les embeddings d'une couche intermédiaire.
   - **Méthode 2 : Utilisation d'un modèle pré-entraîné.**
     - Appliquer un modèle pré-entraîné comme CLIP, ResNet, ou Vision Transformer.
     - Extraire les embeddings des images directement à partir du modèle.
   - **Méthode 3 : Self-supervised learning. (BONUS – 5 points)**
     - Implémenter un encodeur non-contrastif (par exemple, SimSiam ou VICReg).
     - Former l'encodeur en mode self-supervised pour générer des embeddings adaptés.
2. **Comparaison des trois méthodes :**
   - Analyser la qualité des embeddings sur des critères comme la cohérence intra-classe et les distances inter-classe.

- Sélectionner des exemples représentatifs (par exemple, un sac, un parfum, un produit cosmétique) et comparer les distances entre leurs embeddings générés par les trois méthodes.
- Visualisation des embeddings (via t-SNE ou UMAP) pour illustrer les regroupements.

Partie 3 : Analyse et comparaison des embeddings textuels (30 points)

1. **Traduction des descriptions :**
   - Utiliser un outil de traduction automatique (comme DeepL ou Hugging Face Transformers) pour convertir les descriptions (title) en anglais.
2. **Génération des embeddings textuels :**
   - Utiliser des modèles NLP pré-entraînés comme BERT, DistilBERT ou Sentence-BERT pour extraire les embeddings des descriptions.
   - Explorer différentes options comme la moyenne des tokens ou l'utilisation directe de la sortie CLS.
3. **Analyse des embeddings :**
   - Calculer les distances entre les embeddings des descriptions pour des exemples parlants (par exemple, des produits très similaires ou totalement différents).
   - Comparer les regroupements textuels avec ceux observés dans les embeddings visuels.
4. **Visualisation :**
   - Réduction dimensionnelle et visualisation des embeddings textuels pour détecter des clusters potentiels.

Partie 4 : Création d'une plateforme de système de recommandation (BONUS) (20 points)

1. **Développement de l'interface :**
   - Utiliser une librairie comme **Streamlit** ou **Dash** pour concevoir une plateforme interactive.
2. **Fonctionnalités principales :**
   - **Option 1 : Recherche par image.**
     - L'utilisateur charge une image, et la plateforme propose les 10 articles les plus similaires visuellement.
     - Affichage des images des articles recommandés.
   - **Option 2 : Recherche par texte.**
     - L'utilisateur saisit une description textuelle, et la plateforme propose les 10 articles les plus similaires en termes de description.
     - Affichage des descriptions et images des articles recommandés.
   - **Option 3 : Recherche combinée.**
     - L'utilisateur fournit une image et un texte.
     - La plateforme combine les similarités visuelles et textuelles pour proposer 10 articles pertinents.
3. **Approche technique :**
   - Calcul de la similarité entre les embeddings avec des métriques comme la cosine similarity.

    - Pondération des similarités pour l'option combinée.
4. **Validation :**
    - Tester la plateforme sur des cas pratiques pour vérifier la pertinence des recommandations.

Livrables :

- Rapport détaillé avec analyses, méthodologies, et résultats.
- Code source documenté pour l'extraction des embeddings, les comparaisons et la plateforme.
- Prototype fonctionnel du système de recommandation.

Ressources :

Jeu de données :
https://huggingface.co/datasets/DBQ/Chanel.Product.prices.Germany?row=0

Instructions :

Le projet est à rendre avant le **11 décembre à 23h59**. Les étudiants devront indiquer les noms de chaque membre du groupe dans le document final.

Le projet est à déposer sur Moodle dans la section spécifique du projet.