**Une image contenant texte, Appareils électroniques, capture d’écran, ordinateur

Description générée automatiquement**

**Sujet 2025 :**



[https://www.imdb.com](https://www.imdb.com/)

L’Internet Movie Database (littéralement « Base de données cinématographiques d'Internet »), abrégé en IMDb, est une base de données en ligne sur le cinéma mondial, sur la télévision, et plus secondairement les jeux vidéo. IMDb restitue un grand nombre d’informations concernant les films, les acteurs, les réalisateurs, les scénaristes et toutes personnes et entreprises intervenant dans l’élaboration d’un film, d’un téléfilm, d’une série télévisée ou d’un jeu vidéo. L’accès aux informations publiques est gratuit. Il est possible de les télécharger ici :

<https://developer.imdb.com/non-commercial-datasets/>

Un service payant, IMDbPro, donne accès aux informations supplémentaires susceptibles d’intéresser les professionnels. Créé le 17 octobre 1990 par l'Anglais Col Needham, c’est un site visité, en 2010, par plus de 57 millions d’usagers uniques chaque mois, ce qui le plaçait au 39e rang des sites les plus visités au monde. Il appartient depuis 1998 à Amazon.

Le Stanford Artificial Intelligence Laboratory (SAIL) a mis à disposition gratuitement le jeu de données IMDb utilisé pour l’article *Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).*

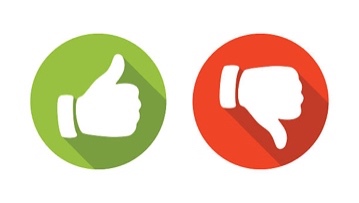
ll s'agit d'un ensemble de données pour la classification binaire de sentiments (positifs ou négatifs) contenant 25 000 critiques de films pour l'entraînement et 25 000 autres critiques pour le test. Des données non étiquetées supplémentaires peuvent également être utilisées. Ce jeu est disponible à l’adresse <https://ai.stanford.edu/~amaas/data/sentiment/>

**Vous réaliserez une base de données SQL Server de façon à stocker les données issues d’IMDB.** Vous devrez tout d’abord constituer à partir des fichiers plats un modèle relationnel OLTP afin de limiter la redondance et le volume des données. Vous appliquerez toutes les techniques vues en optimisation de bases de données (sur Oracle) - index, tables partitionnées, etc. - afin d’optimiser cette grosse base de données. Vous importerez les données en Python, avec Knime ou SSIS, au choix.

Vous créerez des rapports PowerBI pour visualiser/naviguer dans les données de la base OLTP.

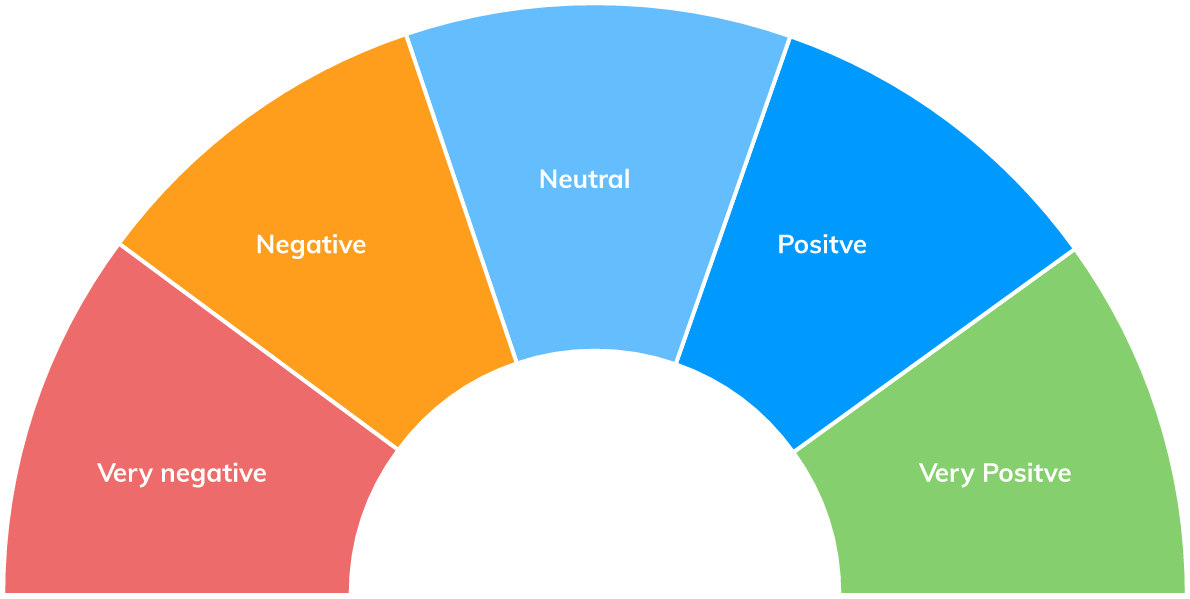
Si nécessaire, vous mettrez en place une base de données OLAP qui permettra d’alimenter des rapports analytiques PowerBI.

Une fois ceci fait, IMDb souhaite également que **vous mettiez en place un modèle IA en Python qui calcule automatiquement la polarité d’une critique (négative ou positive) à partir du seul texte de cette critique**. Ce modèle permettra ainsi de proposer automatiquement à chaque auteur d’une critique de qualifier sa critique de positive ou négative à l’aide de logos de type



**Un deuxième modèle IA permettra de proposer automatiquement une note sur 10 à l’auteur d’une critique.** Ces deux modèles IA ont pour objectif final de renforcer la cohérence entre les critiques, leur qualification et les notes associées.

Un troisième besoin, pour lequel les techniques d’IA seront mobilisées, est la **création d’un score sous forme de grades** **(nombre à définir)** qui permette de sortir d’une évaluation binaire (négatif ou positif) pour proposer **une évaluation** qui ne soit pas aussi exigeante qu’une notation sur échelle à 10 niveaux. Ce score pour par exemple être du type



**Un dernier modèle IA devra enfin être construit de façon à proposer automatiquement ce score à chaque saisie d’une critique.** Une attention particulière sera portée au fait que les modèles IA devront être développés en maîtrisant la consommation de ressources : seule l’utilisation d’un PC standard, et non l’usage d’une station de calcul intensif, sera autorisée. Par ailleurs, vous veillerez à adopter des plans d’expérience qui soient les plus frugaux possible.

**4 Groupes de** 4 à 5 étudiants

**Planning :**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **semaine** | **base de données** | **rapports Power BI** | **IA** | **autonomie** | **Soutenance** |
| **S2** | V. Couturier (8h) |  |  | 12h |  |
| **S3** |  | V. Couturier (4h) | A. Bhanot (8h) | 16h |  |
| **S7** |  |  | A. Bhanot (4h) | 12h | A. Bhanot / V. Couturier (4h) |

* **soutenance finale en S7, 4 heures : V. Couturier, A. Bhanot**

**Date limite de remise : 14/02 (à la fin de la séance d’autonomie)**

**Livrables :**

– Documents de suivi du projet (GANTT final)

– Document technique :

– modèle et script optimisé de la base OLTP en expliquant les optimisations,

– modèle en étoile expliqué et choix d’architecture,

– modèles IA produits en expliquant la démarche de développement et d’optimisation de ces modèles (données d’apprentissage, types de modèles, paramétrages, performances atteintes).

– Livrables techniques : script des bases, rapports power BI, scripts Python

– Soutenance finale du projet dont une conclusion en anglais

– Abstract d’1/2 page présentant le sujet (thème et résultats)

- Répartition du travail

**SQL Server :**

* Voir les procédures de connexion à SQL Server
* Voir la procédure de connexion Power BI à SQL Server

**Soutenance (14 février matin) :** 4h.

Contenu de la soutenance :

* Contexte
* Besoins :
  + Besoins liés aux rapports et à l’analyse de données
* Gestion de projets (ou à la fin) : GANTT (combien de séances sur la partie OLTP, BI, Analyse de données).
* Base de données OLTP :
  + Conception : MCD, MLD, MPD. Présenter les choix de conception pris.
  + Optimisations : optimisations de modèle (MPD), index (sur quels champs ?), tables partitionnées et/ou clusters (quelles tables et pourquoi), etc. Vous devrez justifier vos choix. Pour index, montrer par exemple un/des plans d’exécution.
  + Import des données : quelle source ? comment ? volumétrie ?
* Base de données OLAP (si nécessaire) (DW) :
  + Besoins remplis par la base OLAP / rapports (retour sur les besoins)
  + Conception (schéma en étoile ou flocon). Présenter les choix de conception pris.
  + Développement :
    - Alimentation OLTP -> OLAP : Expliquer un ou qq exemples de transformation et leur type
    - Rapports PowerBI : liste, exemples, démo.
* Architecture OLTP / OLAP (flux entre les 2 BDs, positionnement des rapports : sur quelle base ?)
* Analyse de données/IA :
  + Modèles IA développés (description des classifieurs et clusters obtenus, performances atteintes).
  + Démarche de développement et d’optimisation des modèles (données d’apprentissage, types de modèles, paramétrages)
  + Un test en live en prenant un commentaire sur IMDb serait souhaitable.
* Conclusion en anglais

25 minutes y compris démo + 15 minutes de questions.

**Finalisation des documents le 14/02 PM.**