

CDC N°220318

Rotten Tomatoes

Cahier des Charge (CDC) Analyse de Données et Perspectives IA

Auteurs : Maïna LE DEM
Lucas LUSTOSA
Romain HELLER

Indice : A

Date d'émission : 18/03/2022



SOMMAIRE

1) Introduction	3
1.1 - Motivations	3
1.2 - Contexte	3
2) Documentation	3
2.1 - Terminologie métier.....	3
2.2 - Bibliographie.....	4
3) Profil des utilisateurs finaux	5
3.1 - Analyse.....	5
3.2 - Base de données	5
3.3 - Exploitation des données.....	5
4) Fonctionnalités à réaliser	5
4.1 – Portée	5
4.2 - Exclusions	5
5) Contraintes du livrable	6
5.1 - Contraintes matérielles	6
5.2 - Contraintes logicielles	6
5.3 - Contraintes fonctionnelles.....	6
5.4 - Contraintes d'ergonomie	6
5.5 - Autres facteurs de qualité exigés par le client.....	6
6) Aspects contractuels	6
6.1 - Délais et protocole de livraison	6
6.2 - Aspect juridique et commerciaux.....	6



1) Introduction

La société Rotten Tomatoes est un site Web, créé en 1998, dévolu aux critiques et aux informations sur les films.

Rotten Tomatoes fait partie des 300 sites les plus visités aux États-Unis.

1.1 - Motivations

- **Le client :** Fandango

Fandango est une société américaine de billetterie qui vend des billets de cinéma par le biais de leur site internet ainsi que par leur application mobile. Depuis le rachat de Flixter à Warner Bros en 2016, c'est Fandango qui opère sa filiale Rotten Tomatoes et sa base de données répertoriant les critiques de films ou de DVD.

- **Le problème :**

Fandango aimerait pouvoir évaluer les possibilités d'exploitation de ces données afin d'en tirer d'autres sources de revenus complémentaires.

- **L'existant :**

La base de données date de 1998, et sa structure n'a pas été mise à jour depuis. Les données y sont classées de manière tabulaire, rendant les recherches compliquées.

- **Le besoin non satisfait :**

La structure actuelle de la base de données ne permet pas d'avoir une vue précise de la qualité des données contenue, ni d'évaluer les possibilités d'exploitation de celles-ci.

- **Les objectifs :**

Une analyse de la qualité des données et des relations entre les différents champs de la base.
Une meilleure structure des données pour une exploitation rationnelle.
Des préconisations sur les possibilités d'exploitation de ces données.

1.2 - Contexte

- **Précisions sur le client :**

Le site Web Rotten Tomatoes permet de comparer les notes données par les utilisateurs réguliers (score d'audience) et les notes données/avis fournis par les critiques (tomatomètre) qui sont membres certifiés de diverses guildes d'écrivains ou associations de critiques de cinéma.

- **Marché :**

Pour l'activité « ventes de billets », Fandango reste le premier acteur sur le marché, devant la FNAC et France-Billets.

Les concurrents directs de Rotten Tomatoes sont *Internet Movie Database (imdb)* et *Allociné* pour l'information sur les films et les acteurs.

Un concurrent de taille a fait son entrée dans le domaine de l'agrégation de contenu en général, et donc de la présentation de films : Google propose des informations pertinentes sur les films dans ses résultats de recherche, mais ne traite pas (encore ?) la qualité et la pertinence des critiques.

2) Documentation

2.1 - Terminologie métier

Tomatometer :



Le « Tomatomètre » est l'instrument de mesure principal du site : il permet au lecteur, en un seul coup d'œil, d'évaluer l'avis donné sur le film. Il fonctionne sur le pourcentage de critiques positives faites sur le film :

- À plus de 75 % : le film reçoit le label « fraîcheur certifiée » (Certified fresh) et est marqué par une tomate rouge
- Entre 70 % et 75 % : si le film a fait l'objet de beaucoup de critiques, il peut recevoir le label de « fraîcheur certifiée ». Dans tous les cas, il est marqué d'une tomate rouge
- Entre 60 % et 70 % : le film est considéré comme « frais » (fresh) et il est marqué d'une tomate rouge
- En dessous de 60 % : le film est considéré comme « pourri » (rotten) et il est marqué d'une tomate éclatée verte.
- Certains films font l'objet d'un taux d'approbation de 100 % mais ne reçoivent pas le label de fraîcheur certifiée car le nombre de critiques émises est trop faible.

2.2 - Bibliographie

Base de Données fournie par le Client :

- 1) **rotten_tomatoes_link** : lien à partir duquel les données des films ont été extraites
- 2) **movie_title** : titre du film tel qu'il est affiché sur le site Web de Rotten Tomatoes
- 3) **movie_info** : petite description du film
- 4) **critics_consensus** : commentaire de Rotten Tomatoes
- 5) **content_rating** : catégorie basée sur l'adéquation du film au public
- 6) **genres** : genres de films
- 7) **directors** : nom du ou des réalisateurs
- 8) **authors** : nom de l'auteur ou des auteurs
- 9) **actors** : nom des acteurs principaux
- 10) **original_release_date** : date de sortie du film
- 11) **streaming_release_date** : date à laquelle le film est sorti en streaming
- 12) **runtime** : durée du film (en minutes)
- 13) **production_company** : nom de la société de production
- 14) **tomatometer_status** : valeur tomatométrique
 - "Rotten" Pourri (moins de 60 % d'avis positifs),
 - "Fresh" Frais (au moins 60 % d'avis positifs)
 - "Certified Fresh" Frais certifiés (au moins 75 % d'avis positifs, au moins 80 avis dont au moins 5 des meilleurs critiques)
- 15) **tomatometer_rating** : pourcentage de critiques positives
- 16) **tomatometer_count** : notes critiques comptées pour le calcul du statut de tomatomètre
- 17) **audience_status** : valeur d'audience
 - "Spilled" déversé (moins de 60 % des utilisateurs ont donné une note d'au moins 3,5)
 - "Upright" droit (au moins 60 % des utilisateurs ont donné une note d'au moins 3,5)
- 18) **audience_rating** : pourcentage d'évaluations positives des utilisateurs
- 19) **audience_count** : les notes des utilisateurs comptées pour le calcul du statut d'audience
- 20) **tomatometer_top_critics_count** : nombre de notes des meilleurs critiques
- 21) **tomatometer_fresh_critics_count** : nombre de notes de critiques fraîches
- 22) **tomatometer_rotten_critics_count** : nombre de critiques pourries



3) Profil des utilisateurs finaux

3.1 - Analyse

L'analyse stratégique est destinée à des décideurs. Elle doit être compréhensible et illustrée pour pouvoir servir d'outils d'aide à la décision.

3.2 - Base de données

La base de données sera opérée par le service Infrastructure et Réseaux, qui maîtrise l'exploitation de serveur MySQL et de scripts Python.

3.3 - Exploitation des données

Les résultats exploratoires seront à destination d'équipes d'analystes qui sont capable d'utiliser des note-books Jupyter.

4) Fonctionnalités à réaliser

4.1 - Portée

Priorité	Fonction	Description	Critère de performance
1	Analyse des données existantes		Analyse complète : présentation du dataset et graphiques obtenus, sélection des variables d'intérêt.
2	Création d'une nouvelle base de données	Un schéma de base MySQL permettant l'exploitation des données	Modèle Conceptuel de données cohérent, respect des bonnes pratiques en la matière (ACID, SOLID).
2	Retraitement et intégration des données dans la base	Stratégie de préparation des données (manquantes et aberrantes) et scripts d'import des données	Scripts d'importation Données pertinentes en base.
3	Analyse stratégique	Préconisations sur les possibilités d'exploitation de ces données	Pertinence des préconisations par rapport aux contextes définis par le Client.
Option	Chiffrage des préconisations	Propositions de MVP réalisables pour les préconisations les plus pertinentes	En fonction des choix du client, une première version peut être imaginée.

4.2 - Exclusions

Cette prestation s'arrête à l'exploration préliminaire des données. Elle n'inclue aucun développement supplémentaire, comme des modèles d'intelligence artificielle où leurs entraînements, car l'exploration des données n'est faite que sur une partie réduite de la base de données originale.



5) Contraintes du livrable

5.1 - Contraintes matérielles

Les caractéristiques du serveur permettant la migration des données seront évaluées en fonction de la volumétrie des données à traiter.

5.2 - Contraintes logicielles

Les versions des outils permettant l'exploitation du livrable seront les versions stable et/ou LTS à la date de livraison.

5.3 - Contraintes fonctionnelles

Les dépendances seront indiquées aux équipes du Client dans les règles de l'art.

5.4 - Contraintes d'ergonomie

Le dépôt contiendra un readme permettant aux équipes du Client de trouver rapidement les parties les concernant.

5.5 - Autres facteurs de qualité exigés par le client

Sans objet

6) Aspects contractuels

6.1 - Délais et protocole de livraison

- **Echéances prévues** : restitution de produit finalisé
- **Date de livraison** : 23/03/2022
- **Restitution** : sous forme d'un dépôt github contenant les livrables
- **Documentation à fournir** : Inclue dans les livrables
- **Délais de maintenance** : sans objet
- **Formations prévues** : Une présentation des éléments au Client

6.2 - Aspect juridique et commerciaux

L'entreprise Central Statistical Office of Breizhmeiz de réserve les droits d'auteur et les licences d'exploitation sur l'intégralité des éléments livrés, à l'exception des données fournies par le Client.

Fait le 23/03/2022, A Breizhmême,

Le représentant du CSOB*

Signature

(Précédé de la mention « lu et approuvé »)

Le représentant de Fandango*

Signature

(Précédé de la mention « lu et approuvé »)



**La signature des deux parties vaut pour acte d'engagement*