Web Mining

Cahier des charges : Collecte et analyse de données sur le site JVC (jeuxvideo.com)

1 CONTENTS

2 BUT ET OBJECTIFS

Le but de ce projet est de collecter les tests et commentaires contenus sur le site jeuxvideo.com (JVC) à la section tous les jeux (https://www.jeuxvideo.com/tous-les-jeux), puis d'analyser les données collectées afin de visualiser quels sont les genres de jeux les plus tendance, évaluer si les jeux sont notés et commentés selon les mêmes critères (analyse de sentiment), permettre de rédiger un commentaire et d'évaluer automatiquement la note du commentaire.

Must have :

- Récolte des données à l'aide d'un ou des spiders.
- Indexation des données récoltées
- Visualisation des données sous formes de graphique.
- Analyse de sentiments pour rechercher une corrélation entre certains mots et la note.

Should have :

- Meilleures analyses de sentiments :
  - Différence en fonction du genre de jeux
  - Différence de langage dans les commentaires si le jeu est bien ou mal noté par le site (review bombing)
- Meilleures visualisations des données
  - Comparatif des notes des joueurs selon le genre.
  - Cycle de vie d'un jeu en fonction des commentaires (note, nombre de commentaires)

Could have :

- Rédiger un commentaire et proposer une note
- Donner la possibilité de choisir plusieurs modèles
- Affichage de l'activité d'un utilisateur (profilage, nombre d'avis, notes, etc…)

Au 24.04.2022, 43'397 jeux sont recensés sur lesquels se trouvent des informations propres à chaque jeu comme le titre, le synopsis, des notes et commentaires d'évaluation, la date de sortie ou encore le genre de jeu. Toutes ces informations peuvent être collectées avec la bibliothèque scrapy-splash, la démonstration de collecte des données sur ce site a été réussie durant le labo1 de WEM2022.

La première étape consistera à collecter les données et les indexer afin de créer un dataset avec les features qui seront intéressantes au projet.

3 PRÉSENTATION DES DONNÉES

Une première table est créée pour obtenir des informations sur les jeux. La seconde table contient les commentaires et tests des lecteurs associé à un jeu.

Table `games` :

- `id_jvc` *(ID)* : ID associé au jeu par le site JVC.
- `title` *(string)* : Nom complet du jeu.
- `synopsis` *(string)* : Résumé du jeu.
- `editorial_grade` *(int)* : Valeur de la note attribuée par les journalistes de JVC.
- `users_grade` *(float)* : Valeur moyenne des notes attribués par les lecteurs de JVC.
- `release_date` *(date)* : Date de sortie du jeu.
- `genres` *(string/list)* : Genres du jeu.
- `support` *(string)* : Plateforme du jeu.

Table `comments` :

- `id_game` *(Foreign Key)* : ID du jeu associé au commentaire.
- `grade` *(int)* : Note attribuée par le lecteur.
- `comment` *(string)* : Commentaire du lecteur (avis).
- `date` *(date)* : Date de parution du commentaire.
- `username` *(string)* : Nom d'utilisateur du commentaire.

Concernant le droit d'utilisation des données, le site JeuxVideo.com n'émet aucune opposition à la collectes de données pour une utilisation privée, selon l'article 9.1 des conditions générales d'utilisation du site;

« La société Webedia ne confère à l'utilisateur qu'un droit non exclusif et incessible d'utilisation (l'utilisation s'entend d'un usage non commercial, caractérisé par la navigation, la participation et le choix de la souscription aux différents services) de son site et de ses services, et se réserve par conséquent les droits d'exploitation de diffusion, cession, ainsi que tout autre droit sur les éléments qui constituent son site et ses services[1] *».*

De plus, il est strictement interdit de reproduire, totalement ou partiellement, le contenu de cette œuvre, sans l'accord écrit et préalable de la société Webedia. Il est aussi interdit d'employer les éléments, le contenu du site et de ces services à des fins commerciales.

[1] https://www.jeuxvideo.com/cgu.htm

Le site est bienveillant concernant la bonne entente et le respect de chaque utilisateur l'un envers l'autre. Et ne tolère pas les comportements diffamatoires, injurieux, discriminatoires, dénigrants ou contrevenants.

4 TECHNOLOGIES ET MÉTHODES

L'application se fait en python et utilise divers packages pour extraire les données, les analyser et les présenter. L'extraction des données se fait grâce à scrapy-splash, permettant ainsi la lecture d'une page web. Certaines données peuvent être extraites via les standards (schema.org, …) et d'autres nécessitent l'accès via des *IDs* ou *Class* depuis les balises HTML.

Une fois les données récoltées, l'objectif est de les sauvegarder sur une base de données via le moteur *Elastic Search* et son package python associé permettant ainsi de les indexer. La phase d'analyse se suit pour obtenir et extraire des informations statistiques sur les jeux. Puis effectuer l'analyse de sentiment via des outils d'analyse de texte sur les commentaires/avis des lecteurs.

Analyse de texte :

- Utilise des outils machine learning de classification (Bayes, Logistic Regression, …)
- Sélection des mots à connotation positive ou négative (SentiwordNet).
- Embedding des mots (TFIDF, Word2Vec, BERT, etc…)
- Pointwise Mutual Information (PMI) et Semantic Orientation.
- Supervised Regression (linear, random forest regressors, …)

La dernière étape consiste à proposer une visualisation des résultats via une page web effectuée avec *Dash* et *Plotly*.

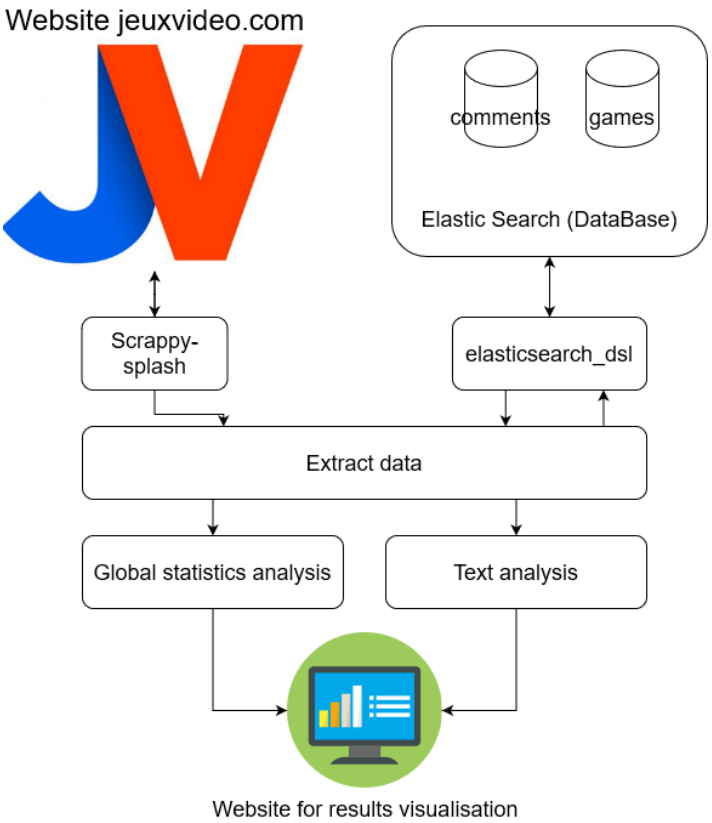

Figure 1: Diagramme du fonctionnement global du projet.

5 RÉSULTATS ATTENDUS

Un site web local permettant d'afficher les statistiques et l'analyse de sentiment obtenu à l'aide des informations extraites par le crawling du site.

6 RISQUES, POINTS CRITIQUES OU PROBLÈMES RENCONTRÉS JUSQU'À PRÉSENT

Le site JVC fonctionne avec une mise à jour automatique des données sans réactualisation des pages dans un grand nombre d'endroits. Cette technique utilise la fetch API et rend l'utilisation d'un site moins « lourde » et plus rapide car elle ne recharge pas toutes les données.

Une telle fonctionnalité cause des problèmes à notre crawler. En effet, la page est mise à jour après la première réponse GET fournie. C'est seulement par la suite que d'autres requêtes sont faîtes sur une API externe du site. Un tel problème peut être résolu avec l'utilisation du package `scrapy-splash`.

L'analyse textuelle comporte le plus de risques pour obtenir des résultats concluants. En effet, certains commentaires pourraient ne pas comporter assez d'informations pour définir son sentiment global. Le « Review Bombing » par exemple, ne sert qu'à baisser la moyenne globale des lecteurs mais pourrait ne pas expliquer la raison de la mauvaise/bonne note avec les mots décrits sur l'avis.

7 PLANIFICATION DES PROCHAINES ÉTAPES DU PROJET

La planification du projet s'accordera selon le calendrier publié sur la page moodle de MA-WEM. Les milestones sont au 13.05, une présentation des méthodes d'analyse utilisées dans le projet, puis le rendu du projet au 19.06.2022 et enfin la présentation du projet le 24.06.2022.