ANALYSE DES DONNÉES JVC

Cahier des charges WEM 2022 Gr.1

Abstract

Le site internet JeuxVideo.Com (JVC) contient des descriptions et des commentaires de joueurs sur tous les jeux vidéo. L'intérêt de ce projet est de collecter et analyser les données contenues sur ce site afin de mieux comprendre l'avis des joueurs de jeux vidéo.

Campos Carvalho Cédric, Feuillade Florian, Ramosaj Nicolas

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. Ces fonctionnalités ciblent particulièrement les éditeurs de jeux-vidéos. Ainsi, permettant à l'entreprise d'obtenir des statistiques globales sur leurs jeux ou ceux des concurrents se positionnant sur un même marché.

Must have :

- Récolter en une seule fois les données à l'aide d'un spider.
- Indexer les données récoltées.
- Analyser la corrélation entre certains mots et la note avec une analyse de sentiments.
- Visualiser les données sous formes de graphique.

Should have :

- Améliorer les analyses de sentiments :
  - Classifier les termes récurrents en fonction du genre de jeux.
  - Identifier les commentaires classifiés comme « review bombing ».
- Améliorer les visualisations des données
  - Comparer les notes des joueurs selon le genre.
  - Evaluer le cycle de vie d'un jeu en fonction des commentaires (note, nombre de commentaires).

Could have :

- Rédiger un commentaire et proposer une note.
- Obtenir un suivi sur les commentaires et notes des utilisateurs.

3 PRÉSENTATION DES DONNÉES

Le nombre de données contenues sur le site de JVC est suffisante pour permettre le développement de ce projet. Au 24.04.2022, 43'397 jeux étaient 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. Pour ce travail, l'équipe a décidé d'utiliser 1000 jeux en prenant compte tous les commentaires de chaque jeu. La répartition des commentaires peut être plus au moins aléatoire selon sa popularité.

Ci-dessous nous avons réunis les données, avec leur description, que nous allons récolter avec le spider :

- `Id JVC` : ID associé au jeu par le site JVC.
- `Title` : Nom complet du jeu.
- `Synopsis` : Résumé du jeu.
- `Editorial grade` : Valeur de la note attribuée par les journalistes de JVC.
- `Users grade` : Valeur moyenne des notes attribués par les lecteurs de JVC.
- `Release date` : Date de sortie du jeu.
- `Genres` : Genres du jeu.
- `Support` : Plateforme du jeu.
- `Grade` : Note attribuée par le lecteur.
- `Comment` : Commentaire du lecteur (avis).
- `Date` : Date de parution du commentaire.
- `Username` : 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.

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.

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

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 on effectue l'analyse de sentiments via des outils d'analyse de texte sur les commentaires/avis des lecteurs.

Analyse de texte :

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

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

Le fonctionnement de notre architecture est illustré par la Figure 1.

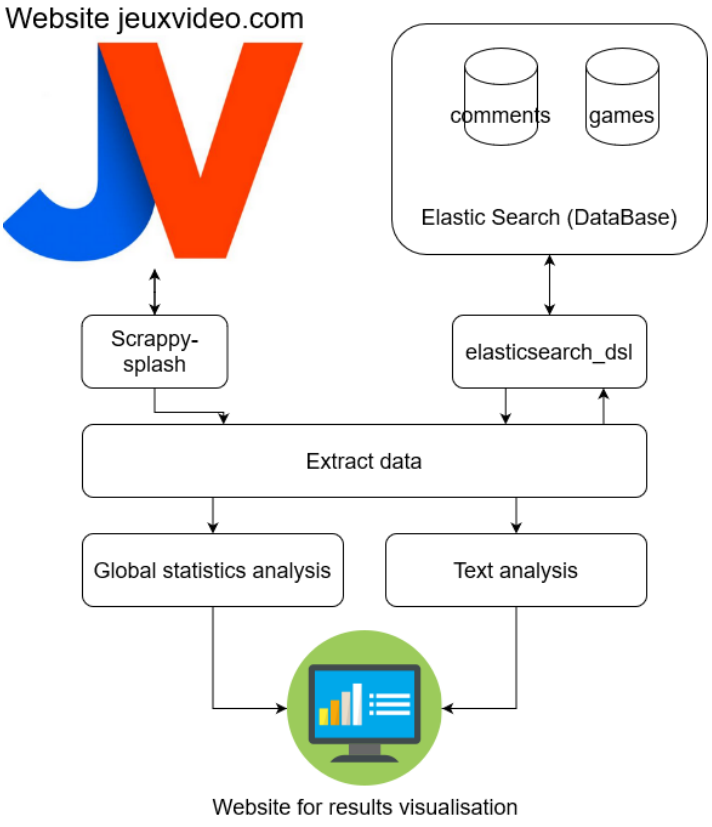


Figure 1: Diagramme du fonctionnement global du projet.

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. La Figure 1 montre le résultat attendu en termes d'architecture du logiciel. En annexes une maquette préliminaire du site est disponible. L'objectif du site web est de proposer une visualisation détaillée des statistiques au niveau des jeux et genres. La maquette sur la Figure 2 montre une première idée sur les blocs principaux.

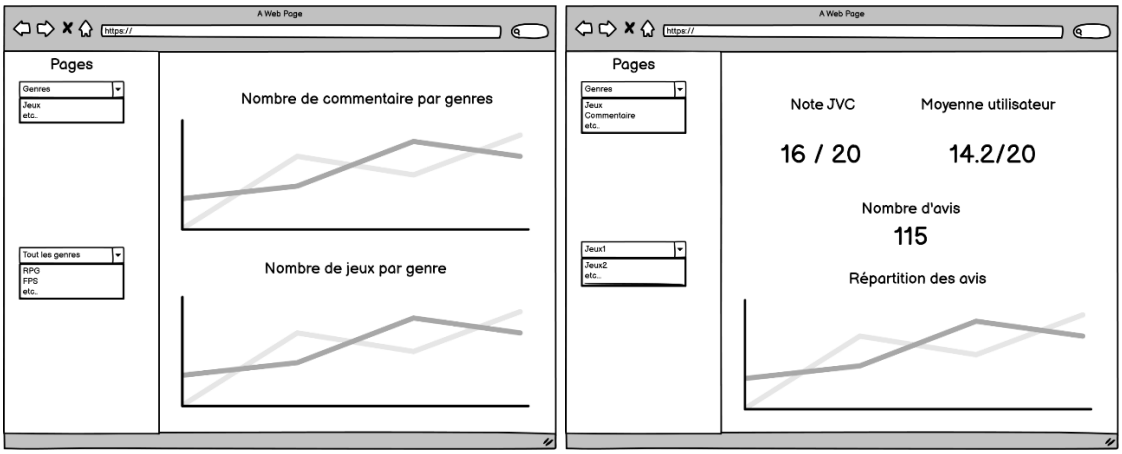


Figure 2 : Maquette site web

Par notre projet, nous avons comme attente un logiciel qui puisse permettre des analyses comme si nous étions développeurs de jeux vidéo pour une société et pouvoir situer et mesurer le succès de notre/nos produits par rapport à la concurrence et dans la communauté des joueurs.

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.

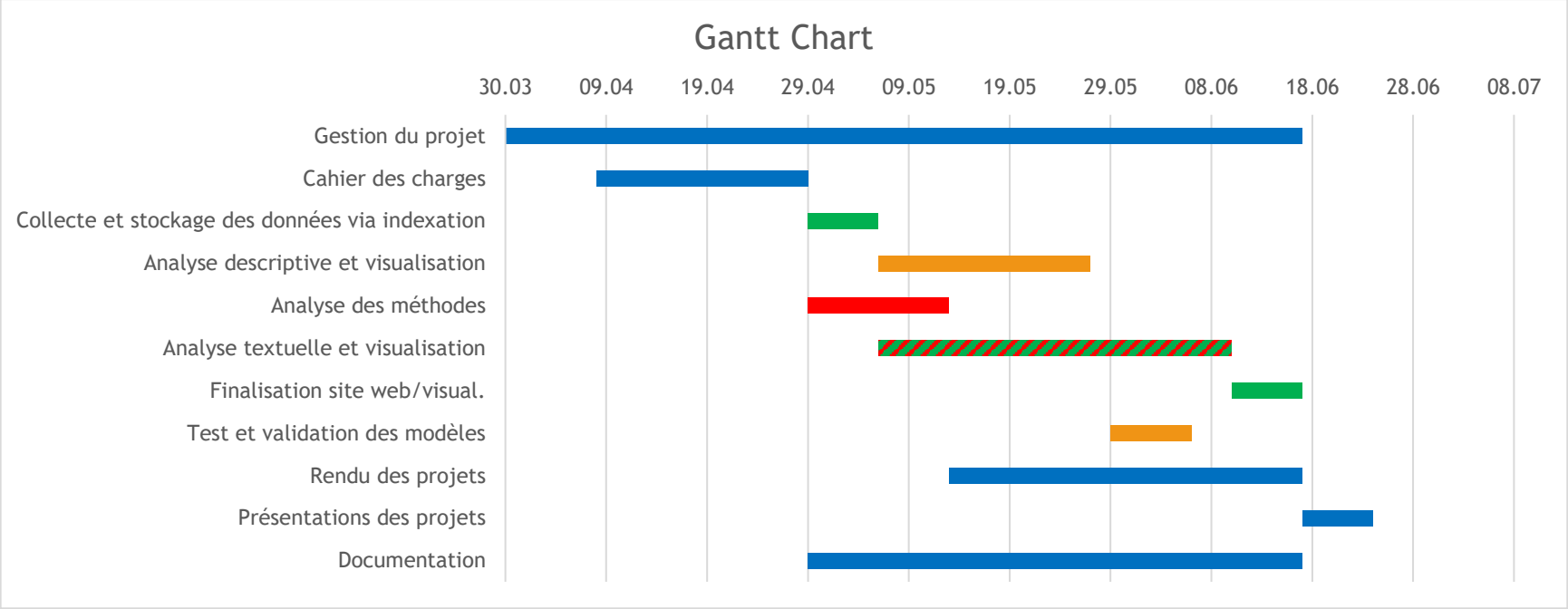
Une telle fonctionnalité cause des problèmes à notre crawler. En effet, la page est mise à jour après la première réponse GET fournie. C'est seulement par la suite que d'autres requêtes sont faîtes sur une API externe du site. Un tel problème peut être résolu avec l'utilisation du package `scrapy-splash`.

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.

De plus, le site JVC est principalement en français et les commentaires des joueurs est en français, un challenge sera de trouver les librairies performantes permettant l'analyse de texte en français.

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.



Membres	Couleur
Tout le monde	
Campos Carvalho Cédric	
Feuillade Florian	
Ramosaj Nicolas	