Gabriel Laffitte – Briac Marchandise

October 12th Deliverable

Data Camp

Professor : Thierno Diallo

Table des matières

Contexte et enjeux

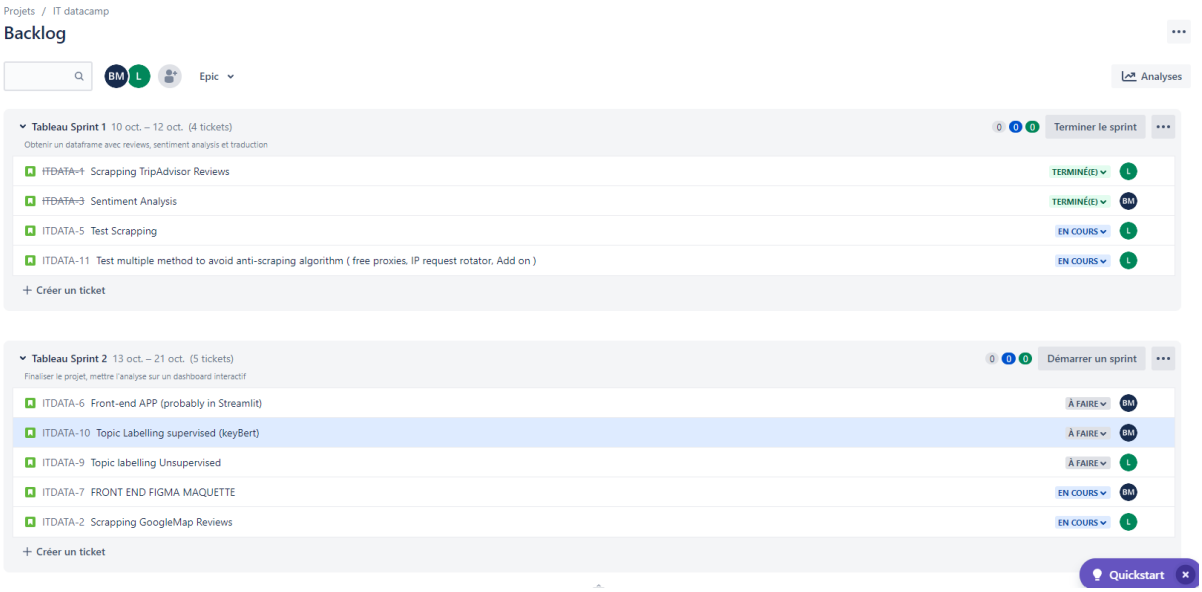
Une entreprise sur Internet cherche à connaitre précisément sa réputation sur Internet et ainsi connaitre les axes sur lesquels s'améliorer. Nous prendrons comme entreprise le restaurant Portobello pour lequel nous collecterons l'ensemble des commentaires sur les plateformes comme TripAdvisor ou Google Reviews.

Le but de ce projet est de nous faire découvrir les algorithmes de Webscrapping classifier les différents commentaires collectés selon des mots clés, les analyser afin d'obtenir les axes de progression pour l'entreprise et enfin , réaliser un dashboard pour permettre au client de visualiser ces analyses.

Nous allons vous présenter les solutions auxquelles nous avons réfléchies pour mener à bien ce projet.

Méthode de travail

Nous travaillons en sprint d'une semaine en méthode Agile définis en avance par des backlogs. Comme vous pouvez le voir sur le Backlog, il nous reste deux étapes à finaliser pour terminer le sprint courant du 10 au 12 octobre. Nous entamerons ensuite un dernier sprint jusqu'au 21 octobre, date du rendu. Afin de planifier au mieux notre travail, nous avons décidé d'utiliser l'outil Jira. Nous pouvons nous rendre compte des tâches à effectuer, des tâches qui sont en cours de production et les tâches déjà finalisée. Les tâches sont assignées selon les envies et les volontés de progression dans un domaine de chacun. Nous profitons des séances de cours régulières pour faire le point sur l'avancement du projet.



Approche technique

Collecte	Technologie utilisée	Taches	Pourquoi ?
WebScraping	Selenium	Ouvre un webdriver, fais les cliques nécessaires et récupere le code html après clique + utilisation de proxies pour éviter les algorithmes anti-scrapping.	Librairie python, intégrable directement sur un notebook, performant
	BeautifulSoup	Parser le code html et l'ordonner en dataframe.	Utiliser en binôme avec selenium et performant
Exploration			
	Pandas	Permet de manipuler des dataframes et de se rendre compte des taches de preprocessing à effectuer.	Libraire performante pour manipuler des dataframes
	Plotly matplotlib	Visualiser la donnée, on a par exemple regardé l'évolution du score des reviews dans le temps.	Libraire performante pour visualiser la donnée directement dans un jupyter notebook
Exploitation			
Sentiment Analysis	Deep_translator Google translator	Permet de traduire en anglais directement en notebook les texts et de reconnaitre la langue utilisée.	Librairie rapide, fiable avec auto détection de la langue. Intégration rapide
	TensorFlow, Keras	Permet de mettre en place un réseau de neurones capable de labelliser les commentaires en postif ou negatif.	Libraire basique de deep learning que l'on a utilisé dans le cours de machine learning. On test cette méthode afin de la comparer à d'autres modeles qui semblent plus performants.

	Transfomers pipeline Distilbert-base-uncased-finetuned-sst-2-englisg	Modele provenant de la plateform Huggingface, permettant d'avoir une labellisation avec une meilleur précision et un meilleur score f1. En plus de la labellisation on obtient une probabilité de label.	Reference dans le domaine plus performant que notre modele tensorflow. Très rapide 55sec 700 commentaires.
Topic labelling	T0pp	Modele de zeroshortlearning qui nous permet de donner un score entre 0 et 5 à chaque review et d'obtenir une autre labellisation.	T0pp est meilleur que gpt3 dans eles taches demandés, et il est 16 fois plus petit. Seulement il prend 45go de ram et son utilisation requiert plusieurs heures mais les résultats sont plus précis et ont un meilleur f1 score que sst2.
	Gensim	Gensim nous permet de créer un word embedding et de représenter chaque text comme un vecteur. On l'utilise pour faire des rapprochements de topics.	La librairie la plus rapide de vector embedding, beaucoup de documentation
	KeyBert	KeyBert nous permet d'extraire les mots clés de chaque commentaire.	KeyBert comprend en première partie un document embedding avec BERT puis un N-gram word et finalement, une mesure de similarité pour trouver les mots qui sont le plus similaire avec la review ie les mots clés. Le preprocessing est déjà pris en charge par BERT. C'est très rapide 40 secondes pour 700 commentaires et performant.
Mise en production	KNN sklearn	Permet de faire un clustering nous permettant de déterminer les groupes et de définir des topics.	La référence en terme d'algorithme de clustering
	Streamlit	Permet de faire une web application en	La solution la plus simple et rapide pour

		utilisant uniquement du code python	avoir une web application.
	Instance azur	Cloud service	On a préféré utiliser azur à la place de AWS pour bénéficier des crédits.
	Docker (si on a le temps)	Découpage du projet en microservices. Permet d'automatiser le déploiement dans le conteneur, contrôle les versions et permet un contrôle des couts de l'instance	Technologie importante pour la mise en production.
	MySQL	Permet de stocker la donnée.	Requêtable directement via notebook.

Fonctionnalités de notre rendu final

Voici le diagramme Use case de notre dashboard qui permettra à notre client de visualiser l'ensemble de ses données d'e-réputation.

Use case diagramm : Dashboard Portobello e-Reputation

Briac Marchandise | October 12, 2022