Data collection and Storage
Web scraping project

GPT3: does GPT3 represent a breaking point in the evolution of artificial intelligence?

FORAY Léo-Paul

HENRY Steven

MAUGER Mika

MU Maxime

9 mars 2023

Sommaire

Introduction

Contexte

L'intelligence artificielle (IA) est une discipline en constante évolution qui a connu des avancées significatives ces dernières années. Avec l'émergence de nouvelles technologies, l'IA est devenue un domaine clé de la recherche scientifique. GPT-3 (Generative Pre-trained Transformer 3) est une nouvelle technologie d'IA qui a attiré l'attention de nombreuses personnes allant des experts du domaine jusqu'aux utilisateurs ordinaires.

GPT-3 est un modèle de langage de grande envergure, qui a été développé par la société OpenAI. Il utilise des réseaux de neurones dits « transformer » qui ont été pré-entraînés pour prédire la probabilité des mots suivants dans une séquence de texte. GPT-3 a la capacité de générer du texte, de répondre à des questions et de traduire des langues.

Le développement de GPT-3 représente pour certains le point culminant dans l'évolution de l'IA, car il a amélioré les performances des modèles de langage existants. Il est devenu un sujet d'étude intéressant pour les chercheurs et les experts en IA, qui cherchent à comprendre comment cette technologie peut être utilisée pour résoudre des problèmes complexes.

De cette technologie est notamment né l'agent conversationnel ChatGPT, aussi développé par OpenAI. Ce qui rend ChatGPT si populaire est sa pertinence. En effet, dans de nombreux domaines, ChatGPT fournit des informations adéquates en un temps record. De plus son accès grand public le place au sommet des agents conversationnels en vogue.

Problématique

Dans le cadre de notre projet de web scraping, nous avons dans un premier temps élaboré un plan d'expérience (DOE en anglais) pour encadrer notre étude et les expériences qui la composent. Puis nous avons dans un second temps collecté des données à partir de différentes sources que nous estimions pertinentes dans ce contexte. L'objectif principal de notre étude est de répondre à la question : "Does GPT-3 represent a breaking point in the evolution of artificial intelligence ?" que nous avons reformulé en « La technologie GPT3 est-elle une révolution dans le champ de l'intelligence artificielle ? ». Pour cela nous utilisons les données collectées pour déterminer l'impact de GPT-3 sur l'évolution de l'IA en étudiant l'opinion publique.

En utilisant le DOE, nous avons pu planifier les expériences de manière à maximiser la précision et la fiabilité des résultats. Nous avons également pu identifier les facteurs les plus importants à considérer lors de l'analyse des données, tels que la source de données, la période de collecte des données et les mots clés utilisés dans la recherche. Grâce à cette méthodologie, nous avons pu obtenir des résultats précis et significatifs, qui ont contribué à notre compréhension de l'impact de GPT-3 sur l'évolution de l'IA.

Design of Experiment (DoE)

Variables

Nous avons choisi de nous intéresser à certaines variables pour répondre à cette question, qui sont les suivantes :

- Le ressenti : Le commentaire/l'article émet-il un jugement positif ou négatif sur GPT-3.
- Le nombre de likes : Combien de personne partage l'avis de ce commentaire/tweet.

Population et échantillon

Nous avons ciblé trois populations différentes pour notre étude : les utilisateurs de Twitter, les utilisateurs de YouTube et les journaux présents sur Mediastack. Nous avons prévu de collecter un échantillon minimum de 10000 données de pour les deux premières sources, car leur population est de plusieurs centaines de millions d'utilisateurs. Pour Mediastack, qui a une population constituée de 7800 sources, nous avons récolté 200 résumés d'articles.

Hypothèses

(H_0) est que l'impact de GPT-3 représente un point de rupture dans l'évolution de l'IA,

Tandis que notre hypothèse alternative :

(H_1) est que l'impact de GPT-3 ne représente pas un point de rupture dans l'évolution de l'IA.

Récolte des données

Nos sources

Nous avons soigneusement sélectionné nos sources de données en ligne pour notre projet de collecte de données. Nous avons choisi de consulter Twitter, YouTube et des articles diversifiés provenant de différentes sources professionnelles, car ce sont des sources de données largement utilisées et accessibles pour une grande partie de la population.

- Twitter est une plateforme de médias sociaux très populaire où les utilisateurs peuvent s'exprimer librement sur divers sujets, y compris les avancées technologiques. Cette plateforme nous a permis de collecter des tweets sur GPT-3 qui nous ont donné une idée de l'opinion publique sur cette technologie.

- YouTube est également une plateforme de médias sociaux très populaire où les utilisateurs peuvent partager des vidéos et des commentaires sur divers sujets. Nous avons utilisé YouTube pour collecter les commentaires des vidéos s'exprimant au sujet de GPT-3, afin de mieux comprendre les discussions en ligne sur cette avancée technologique.

- Les journaux sont généralement une source d'informations fiable et équilibrée qui peuvent fournir des analyses approfondies sur les avancées technologiques. Nous avons choisi Mediastack comme source de données de journal car il fournit un accès facile et pratique à une grande variété de sources d'actualités. En utilisant cette API, nous avons pu récupérer des articles pertinents sur GPT-3 provenant de sources fiables et variées, ce qui nous a permis de disposer d'une analyse complète des différentes perspectives sur cette avancée technologique.

En combinant ces sources de données, nous avons pu obtenir une vue d'ensemble de l'opinion publique sur GPT-3 et sa place dans l'évolution de l'intelligence artificielle. Ce choix nous a permis de collecter des données variées et pertinentes pour notre analyse.

La méthode de collecte

Les données collectées sur la plateforme YouTube ont été obtenues à l'aide d'une bibliothèque Python officielle de Google, qui permet l'accès à l'API de YouTube. Cette API permet aux utilisateurs d'effectuer jusqu'à 10 000 requêtes par jour, ce qui est largement suffisant pour notre projet. Pour collecter les données sur la plateforme Twitter, nous avons opté pour la bibliothèque Python Tweepy qui utilise l'API de Twitter. Nous avons préféré cette bibliothèque en raison de sa grande intuitivité et de son efficacité pour récupérer les données. Il est important de souligner que l'API de Twitter autorise jusqu'à 1 million de requêtes par mois, ce qui représente une quantité plus que suffisante pour notre projet.

Pour la collecte de données à partir de Mediastack, nous avons utilisé leur API pour récupérer des articles de presse pertinents pour notre étude. Nous avons utilisé les paramètres permettant d'inclure des mots-clés (ChatGPT, GPT3, OpenAI), des langues (français et anglais) pour récupérer des articles permettant de répondre à notre problématique. La principale problématique a été la limitation en nombre d'articles renvoyés (100 par requête), mais aussi le formatage des caractères spéciaux qui apparaissait en sous format Unicode.

Concernant les données de YouTube, nous avons dû faire appels aux services de Google. Google proposait son propre client API qui nous a servis à requêter Youtube. Dans un premier temps il était question de rechercher les vidéos sous le mot-clé « ChatGPT » puis de récupérer tous les commentaires des vidéos associées. Comme ses plateformes consœurs, l'API de Youtube était limitée en taille de donnée par requête et en nombre de requêtes par jour (10000). Une fois de plus, le multilinguisme de la plateforme nous imposait une attention particulière à l'encodage de nos données.

Une fois que nous avons récupéré les données à partir de ces sources, nous avons utilisé des outils de traitement de données, tels que pandas, pour nettoyer et préparer les données pour l'analyse. Nous avons également pris en compte les différentes limitations et contraintes imposées par chaque source, telles que les limites de l'API de Twitter en termes de volume de données collectées et la nécessité d'extraire les commentaires à partir des vidéos YouTube.

Enfin, il est important de souligner que la collecte de données est une étape qui nécessite beaucoup de temps et d'effort. Il est également important de prendre en compte les différentes contraintes et limites imposées par chaque source de données, afin de garantir la qualité et la pertinence des données collectées pour l'analyse.

Problèmes rencontrés

LinkedIn

Nous avons rencontré des difficultés pour collecter des données sur LinkedIn en raison de sa politique d'accès restreint à son API. Contrairement à d'autres plateformes, LinkedIn demande aux utilisateurs de créer une application sur le portail de LinkedIn Developers et de compléter plusieurs étapes pour accéder à son API.

Ces obstacles nous ont conduit à prendre la décision de ne pas inclure LinkedIn dans notre collecte de données. Il est également important de souligner que LinkedIn est une plateforme de médias sociaux professionnels, où les utilisateurs partagent principalement des informations liées à leur carrière et à leur domaine d'expertise. Bien que LinkedIn puisse contenir des informations pertinentes sur l'évolution de l'intelligence artificielle, obtenir des données de manière fiable et complète serait trop difficile et fastidieux. Nous avons donc choisi de nous concentrer sur des sources de données plus accessibles qui nous permettent d'obtenir des données pertinentes de manière plus efficace.

Collecte de données dynamiques sur Instagram à l'aide de Selenium

Dans notre processus de collecte de données, nous avons choisi d'utiliser Selenium en tant qu'outil pour la récupération de données sur la plateforme d'Instagram. Plus précisément, nous avons utilisé Selenium pour extraire des données à partir de pages Web dynamiques, qui ne sont pas facilement accessibles via une API ou une méthode de collecte de données automatisée plus traditionnelle.

Selenium est un outil open source qui permet de contrôler un navigateur Web et de simuler des actions d'utilisateur pré-enregistrées, telles que la saisie de formulaires, le clic sur des boutons et la navigation sur des pages Web. Il est souvent utilisé dans le cadre de tests automatisés pour les applications Web, mais il peut également être utilisé pour extraire des données à partir de pages Web.

Cependant, dans le cadre de notre étude nous avons rencontré des problèmes lorsque notre compte Instagram a été suspendu. Nous avons identifié que cette suspension était due à l'utilisation de Selenium, qui a été considéré comme un comportement de bot par Instagram. Nous avons donc décidé de ne plus utiliser Selenium en raison de sa complexité avec la plateforme Instagram.

Nettoyage et stockage des données

Nous avons décidé d'utiliser MongoDB pour stocker les données collectées lors de notre web scraping. MongoDB est une base de données NoSQL (type document) qui offre une grande flexibilité pour stocker des données. C'est une solution populaire pour stocker des données volumineuses et variées telles que des tweets, des commentaires de vidéos YouTube et des articles de presse.

La flexibilité de MongoDB est particulièrement utile pour stocker les données collectées à partir de sources différentes. Par exemple, les tweets collectés à partir de Twitter peuvent varier considérablement en termes de longueur et de contenu, tandis que les commentaires YouTube et les articles de presse peuvent contenir des informations structurées telles que des titres, des dates et des auteurs. MongoDB permet de stocker ces données hétérogènes de manière flexible et de les interroger facilement.

MongoDB est également conçu pour être scalable, ce qui signifie qu'il peut gérer des volumes de données importants et peut facilement être étendu pour gérer des données supplémentaires si nécessaire. Avec MongoDB, nous pouvons facilement ajouter ou supprimer des données et optimiser les requêtes pour récupérer les informations dont nous avons besoin. Cette évolutivité est particulièrement importante pour nos besoins car nous avons collecté des données provenant de sources multiples et variées, et nous avons besoin de stocker ces données dans un format facilement accessible pour les analyses ultérieures.

Dans l'ensemble, l'utilisation de MongoDB pour stocker les données collectées lors de notre web scraping est pertinente en raison de sa flexibilité, de son évolutivité et de sa communauté d'utilisateurs active. Avec cette base de données, nous pouvons stocker et interroger facilement les données provenant de multiples sources et les utiliser pour des analyses approfondies sur l'opinion publique concernant GPT-3 et son impact sur l'évolution de l'intelligence artificielle.

Analyse

Les données récoltées donnent accès à de nombreuses informations, mais nous nous sommes efforcés de sélectionner seulement celles qui sont pertinentes, car certaines API telles celle de YouTube renvoient bien trop d'informations qui seraient inutiles pour répondre à notre problématique.

Nous avons ainsi récupérés les commentaires des vidéos YouTube utilisant le mot-clé « GPT-3 », ainsi que les potentielles réponses à ces commentaires. Pour chacun de ces commentaires, nous avons conservé : l'identifiant de référence vidéo, le texte original, l'identifiant de l'auteur, la date de publication, le nombre de

mentions « j'aime » ainsi que l'identifiant de l'éventuel commentaire auquel il répond, ce qui permet d'analyser l'évolution des avis au cours du temps, ou même d'analyser des discussions complètes autour de ce sujet.

Les « Tweets » que nous avons récupéré sur l'API Twitter sont accompagnés de leur nombre de « retweets » et de mentions « j'aime » chacun de ceux-ci a eu. Le gros avantage de cette source est aussi la quantité disponible, avec pas moins de 10 000 tweets exportés, tout comme les commentaires YouTube qui étaient au nombre de 60 000. Ainsi, nous avons pu récolter pour chaque tweet ces différentes informations : le nom d'utilisateur, le texte, le nombre de réactions. Pour compléter les informations fournies par ces Tweets, nous avons pris la liberté d'ajouter une donnée sentiment qui fournit une information sur l'avis du tweet (positif=1, neutre=0, négatif=-1). Avec ceci, nous avons généré un histogramme des tweets pour se rendre visuellement compte des différentes périodes d'expression sur le sujet (Voir annexe 1).

Les articles Mediastack sont au nombre de 200, mais la qualité de ces sources par rapport à notre problématique en fait une donnée très intéressante. En effet, ils sont écrits par des journalistes professionnels, qui donnent directement leur avis dessus. Étant trop longs pour être tous récupérés, seul le début de l'article et un lien vers sa version complète ont été récupérés, mais la présence du nom du média, de l'auteur ainsi que de la date de parution en fait une source fiable et pertinente pour notre étude.

Conclusion

Dans l'ensemble, notre projet de web scraping nous a permis de collecter des données à partir de différentes sources pour répondre à notre question de recherche sur l'impact de GPT-3 sur l'évolution de l'IA. Grâce à l'utilisation du DOE et à l'analyse de données, nous avons pu obtenir des résultats significatifs qui nous ont permis de tirer des conclusions sur notre hypothèse de recherche.

Annexes

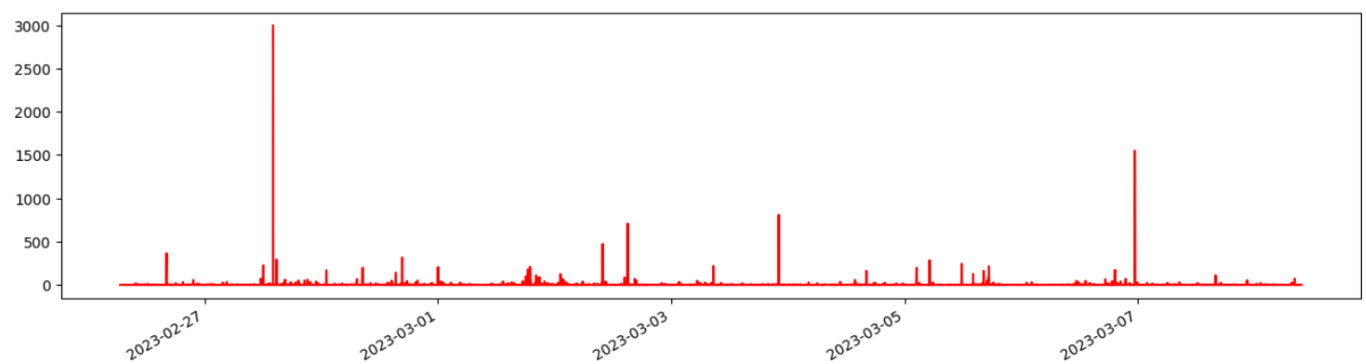

Figure 1 - Histogramme des tweets exportés