Introduction

L'approche MLOps vise à automatiser et à optimiser le déploiement des modèles de machine learning en production. Elle comprend un pipeline de développement, de test et de déploiement, ainsi qu'un suivi continu de la performance du modèle une fois en production. Cette approche est cruciale pour assurer la scalabilité, la fiabilité et la maintenabilité des systèmes d'apprentissage automatique dans des environnements réels.

Pour récupérer les questions pertinentes de Stack Overflow, nous avons utilisé l'API Stack Exchange avec des filtres spécifiques. Nous avons filtré les questions avec certaines contraintes que nous avons jugées nécessaires pour la qualité des données. C'est entre autres le nombre de vues, le score et de leur nombre de réponses. Nous avons extrait en tout 50000 questions pour notre projet.

Une fois les questions extraites, nous avons effectué un text processing pour rendre les données utilisablese. Cela comprenait la suppression de la ponctuation, des stopwords et la lemmatisation des mots. Nous avons également utilisé BeautifulSoup pour éliminer les balises HTML dans le corps des questions. Enfin, nous avons séparé les questions en ensembles de formation et de test, et nous avons créé des features à l'aide de CountVectorizer.

Pour proposer des mots clés de manière non supervisée, nous avons utilisé une combinaison de countVectorizer et de LDA. Afin de determiner le nombre optimal de topics nous avons combiner la méthode de perplexité et la cohérence score. Cela nous a permis de porter notre choix sur 24 topics que de nous avons pu visualiser avec LDAvis. Ensuite, nous avons attribué 20 mots clés aux questions de test en utilisant les topics générés à partir des questions de formation.

Quant à l’approche supervisée nous avons d’abord traiter la colonne tags afin de ne garder que les tags les plus fréquents en utilisant également le multilabelencoder. Compte tenu des ressources limitées de notre ordinateur, nous avons retenu que les 10 tags les plus fréquents. Par la suite, nous avons entraîné 7 7 algorithmes dont 3 modèles embedding, en utilisant OneVsRestClassifier. Pour l’évaluation comprative des algorithmes, 5 métriques ont été utilisés en plus du temps d’entraînement. Dans cette étape, nous avons exploré des techniques d'embedding telles que Word2Vec, BERT et USE pour créer des features à partir des questions. Nous avons utilisé ces features dans une approche supervisée similaire à celle de l'étape précédente et avons comparé les performances avec les autres méthodes.

Enfin, nous avons déployé notre modèle de prédiction des mots clés en utilisant un pipeline de mise en production. Nous avons utilisé Git pour gérer le code et avons déployé l'API sur streamlit cloud. Nous avons également créé une interface locale permettant aux utilisateurs de saisir une question et de recevoir les mots clés proposés par l'API.

Conclusion

En conclusion, nous avons réussi à mettre en place un pipeline complet pour la proposition de mots clés à partir des questions de Stack Overflow. Nous avons exploré des approches non supervisées et supervisées, ainsi que des techniques avancées d'embedding. Notre modèle est maintenant prêt à être utilisé en production, avec une interface conviviale pour les utilisateurs finaux.