Nam Nguyen - 22.11.2025

Interview 3 avec un Datascientist en extraction et génération de données en personne:

On a un problème avec le chatbot, c'est à cause de la mauvaise qualité de données ?
Oui, vous créez un pipeline qui avale des erreurs : mots coupés, chiffres mal lus, tableaux désalignés… Ensuite, lorsque vous entraînez un modèle ou générez des visualisations, ces erreurs se propagent. C'est pour cela qu'un prétraitement rigoureux est indispensable : normalisation, correction des symboles, vérification des données numériques…

Quelle approche vous nous conseillez d'adopter alors dans ce cas-là ?

Si vous avez essayé l'extraction automatique mais que cela ne marche pas, il vaut mieux réorienter son approche. Le document a été peut être scanné de travers, compressé ou annoté, La méthode OCR permet de récupérer un "premier jet", puis il faut reconstruire le texte à la main ou semi-automatiquement. Le scraping n'est jamais entièrement automatique. Il faut accepter cette zone d'intervention humaine.

Qu'est que vous nous conseillez avec le temps restant (1h) avant le rendu final ?

Je vous recommande de générer les données synthétiques via l'IA parce que cela permettra de contourner le chaos des documents réels lorsque le but est de tester un concept. Dans votre hackathon, vous deviez démontrer la faisabilité d'un pipeline, pas encore produire une base parfaite. La génération IA permettra de produire des tables propres, cohérentes et structurées qui imitent fidèlement la forme des vraies données. Cela aide énormément lors des phases de prototypage.

Quels types de jeux de données peuvent être générés par l'IA dans ce contexte ?
Vous pouvez générer des tableaux statistiques, des blocs textuels résumant les résultats, mais aussi des données structurées alignées sur la syntaxe d'un fichier réel, par exemple, un tableau de résultats cliniques, ou un tableau "fact box" utilisé en risk literacy. Tant que la structure est respectée, ces données servent parfaitement de terrain d'essai.

À plus long terme, comment voyez-vous l'évolution de l'extraction scientifique par IA ?
On va vers des modèles capables de lire un PDF comme un humain, en combinant vision, langage et logique. Mais paradoxalement, la standardisation des formats sera tout aussi importante. Tant que les articles restent hétérogènes, les IA devront continuer à improviser. C'est pour ça que l'open science et les formats ouverts vont devenir essentiels dans les années à venir.