eOnsight

Stage 2024 - Extraction de données via OCR

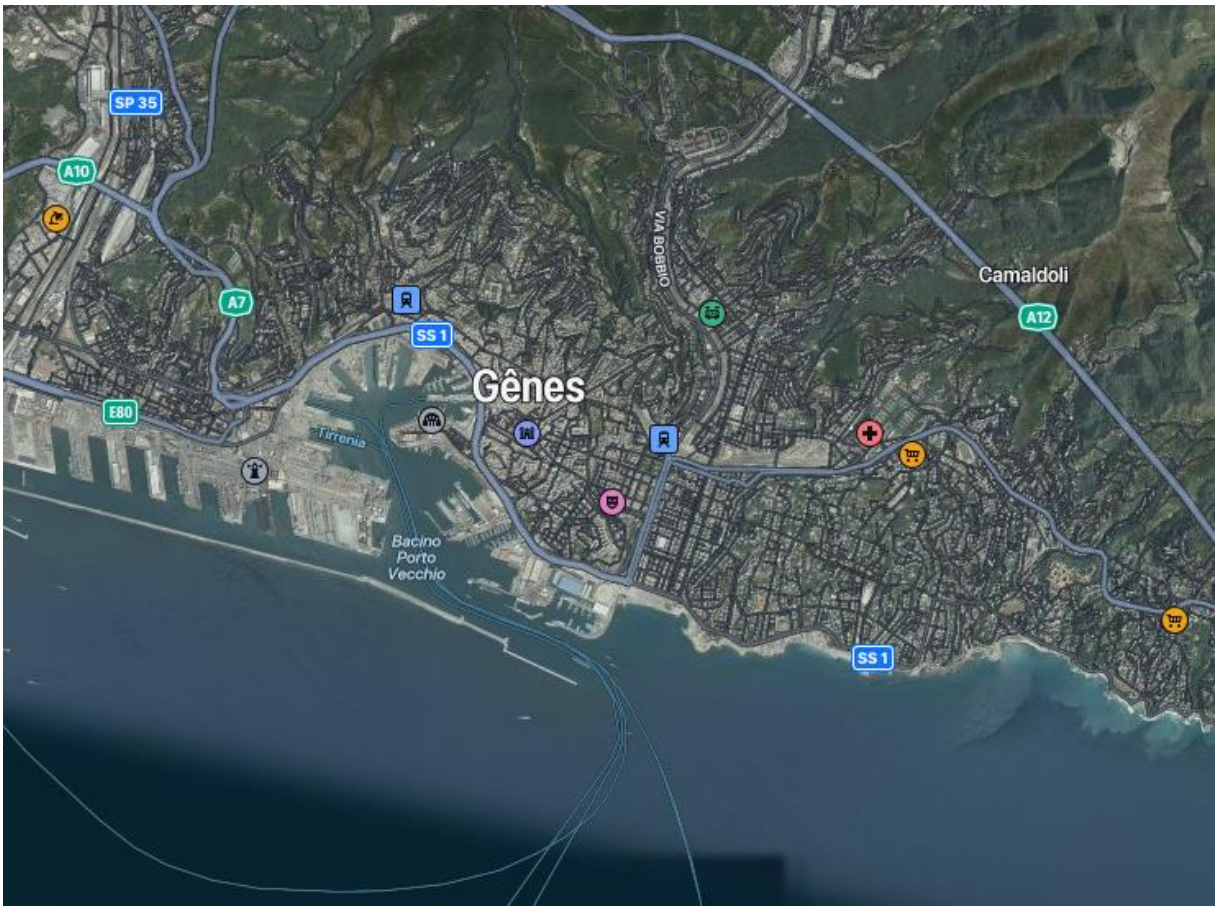
Contexte

Dans le cadre d'un projet d'ingestion de rapport d'inspection d'infrastructures et plus généralement de numérisation de documents sous formats hétéroclites, maîtriser les outils et techniques d'OCR est essentiel. C'est la première brique nécessaire à l'élaboration de pipeline d'analyse IA comme dans le cas d'intégration via LLM.

L'objectif d'eOnsight est de croiser les données pouvant être ingérées de ces rapports d'inspection avec des données satellites ou autres données disponibles sur Internet.

Préparation

Dans cet exercice, nous nous intéresserons uniquement à l'extraction des textes de l'image suivante :



Genova.png

Vous trouverez également l'image au format PNG en pièce jointe.

Résultat attendu

Nous vous proposons un rendu via un repository Gitlab ou Github incluant l'implémentation ou l'intégration d'un outil OCR ayant l'image en entrée et extrayant les **textes de l'image** (avec au moins les noms des villes de **Gênes** et de **Camaldoli**), la **localisation de ces champs de textes sur l'image** et **le taux de confiance** associés au processus d'extraction/reconnaissance de texte.

L'output devra être au **format JSON**.

Cette implémentation ou intégration peut être faite dans le langage de programmation de votre choix.

On tiendra fortement compte de la forme, de la simplicité d'usage et de reproduction des résultats. On attend donc un README.md et, par exemple, si vous comptez développer en Python avec des dépendances pip, n'oubliez pas un fichier requirements.txt standard. Faites comme si vous deviez transmettre votre projet à une autre personne qui n'a aucune autre information que ce que vous lui livrez.

Suggestion

Vous pouvez ne pas réinventer l'existant et intégrer n'importe quel outil qui vous semblera pertinent. N'hésitez pas à rechercher des benchmarks et essayer de reproduire des résultats existants. Cela dit, toute autre façon de faire sera tout aussi bienvenue.

Pour aller plus loin

Si le sujet vous inspire, vous pouvez essayer différentes choses en supplément comme développer une approche RAG par exemple.

Le sujet est trop simple pour vous ? Dans ce cas, nous vous proposons d'extraire le tableau suivant en gardant la structuration de celui-ci.

EQUIPEMENTS				
SUR OUVRAGE	SUBDI		CDOA	
	classe	S	classe	S
. Chaussée	2			
. Trottoirs et bordures	1			
. Dispositifs de retenue	3			
. Corniches	4			
. Dispositifs d'évacuation des eaux	2			
. Joints de chaussée et de trottoirs	S.O.			
. Autres équipements sur ouvrage	S.O.			

Extrait_IQOA_data.png