Module IaaS / Cloud : inclusion projet fil rouge P2022v2

Introduction

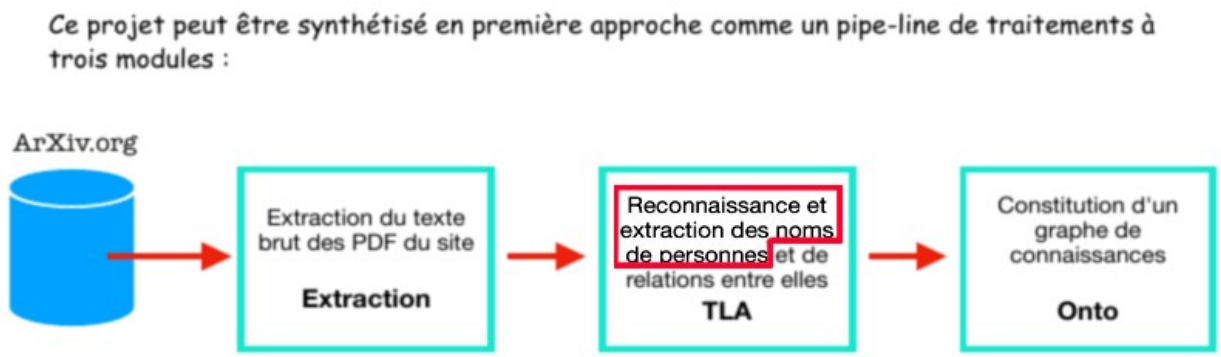
Le projet fil rouge est augmenté par ce document.

La notation du module IaaS / Cloud se basera intégralement sur ce projet.

Nous avons vu en cours les principes de base du Cloud Public, avec l'exemple concret de Amazon Web Services. Ce projet va mettre en pratique les notions que nous avons vues.

Le projet fil rouge vous demande de développer un pipeline de génération d'un graphe de connaissance à partir de l'extraction de PDF depuis un site web.

Cette augmentation du projet vous demande d'**implémenter les étapes d'OCR et d'extraction des noms avec les services AWS**, soit l'étape encadrée en rouge ci-dessous :



Besoin exprimé

Fournissez-moi une solution codée avec Python accompagnée de fichiers PDF. Ces fichiers PDF seront analysés par la solution afin d'en faire les extractions de noms de personnes.

Vos fichiers PDF serviront à valider que votre code fonctionne dans des situations idéales ; bien sûr je testerai votre code avec d'autres PDFs issus ou non de ArXiv.org ; votre code devra savoir les analyser également.

La solution devra générer une sortie dans un format json.

Au moins un service AWS devra être utilisé (par exemple, Comprehend, Lambda, Textract, S3, EC2 …).

Il doit être facile pour le correcteur de lancer votre code pour le vérifier. Vous devez m'inclure une procédure de test qui commence par l'étape « faire un git clone » ou « décompresser l'archive .zip » et termine par « on peut observer que le nom de l'auteur du pdf xxxx.pdf est affiché à tel endroit ».

Je vous recommande très fortement de tester votre procédure de A à Z dans une machine virtuelle neuve sous Linux ou sur un autre ordinateur. Rappelez-vous que je corrige sous Linux.

Notation

Le barème est le suivant :

Critère	points
La solution fournie répond **à l'ensemble du besoin exprimé.** Relisez ce paragraphe autant de fois que nécessaire avant de m'envoyer votre projet.	8
La solution n'explose pas en vol sur d'autres PDFs	2
La solution échoue élégamment lorsqu'elle se voit présenter des PDF non conformes	1
Votre projet comporte un requirements.txt, ou autre gestion des dépendances élégante.	1
La procédure d'installation est exhaustive	2
Vos instructions ne nécessitent pas sudo (en dehors de l'installation d'outils systèmes)	2
Si vous m'avez contacté par email au sujet du projet : • Ajout du point pour le respect du protocole ci-dessous • Soustraction du point en cas de non respect du protocole	±1
Vos instructions ne nécessitent pas de modifier le code pour y mettre les identifiants AWS ni mots de passe	2
Votre rapport fait un comparatif quantitatif et qualitatif des performances du service AWS par rapport à vos implémentations locales	2
Votre solution utilise Lambda & API Gateway	2
Vous me livrez de quoi déployer la grande majorité de votre solution en Infrastructure as Code (IaC) (Terraform, serverless framework, ...)	2

La note maximale ne pourra pas excéder 20/20.

Protocole de contact

En cas de difficultés, demandez d'abord à vos collègues ce qu'ils pensent de votre problème. Si un point du projet est flou pour plusieurs personnes, alors vous pouvez m'envoyer un e-mail en mettant en copie les collègues que vous avez sollicités à ce sujet **ET** la liste de distribution de la promo ms.sio.eleves.p2022@listes.centralesupelec.fr.
Je vous adresserai ma réponse en faisant « Répondre à tous », ce qui permettra à toute la promotion d'être équitablement informé.

Restitution

- Vous devez me livrer un document dédié à ce sous-projet au projet de deux pages maximum (hors schémas et annexes). Vous êtes encouragés à me renvoyer vers les chapitres pertinents de votre rapport principal lorsque nécessaire.
- L'ensemble du projet est à remettre sur Edunao (des précisions suivront) au plus tard le vendredi 8 avril 2022 à 23h59.

Annexe 1 : tests accès CLI AWS

```
$ vim ~/.aws/credentials
$ aws sts get-caller-identity
```