Rapport AWS – Projet Fil Rouge – Simon ADDA

Vous trouverez l'ensemble du projet dur Edunao ou directement sur le lien GitHub : https://github.com/SimonADDA/PFR . Le service se trouve dans le répertoire **/AWS**

Vous trouverez dans **/AWS/Readme.md** les informations d'installations de ce service.

Pour ce projet, j'ai décidé d'utiliser le service comprehend de AWS. Amazon Comprehend est un service de traitement du langage naturel (NLP) qui utilise le machine Learning (ML) pour découvrir des informations et des relations utiles dans un texte.

Pour implémenter ce service, j'ai donc utiliser le module boto3 sur python.

L'objectif est donc d'extraire à partir de pdf fourni par l'utilisateurs, les entités nommées de ce dernier. Pour ce faire j'ai réutilisé les modules pythons permettant d'extraire les textes d'un pdf avec **pdfminer**. Ce package contrairement a **pdftotexte** est bien plus efficace pour la lecture de PDF comportant des colonnes.

Il est important de faire un pré traitement sur les textes car ces derniers sont trop lourds pour le service Comprehend. J'ai donc, comme pour le projet Fil rouge, extrait seulement les références de ce texte en ne récupérant que les mots après le mot « Références » dans le texte.

Une fois les textes extraits et pré traitement effectué, nous faisons donc appel au service comprehend de AWS. Ce dernier fourni un résultat sous format de json qui comporte un score sur chaque mot du texte avec sa classe grammaticale. J'ai donc décidé pour avoir un résultat pertinent de ne récupérer que le mot donc la classe grammaticale est PERSON et qui a un score supérieur a 80%.

Une fois extrait, nous avons donc les résultats encore une fois sous format json et ce dernier est accessible dans les répertoires /Download et /Professeur.

Détails dossier /AWS :

/AWS : *Dans ce répertoire l'utilisation du service Comprehend de AWS. Ce dernier permet d'extraire les entités nommées d'un texte.*

On y trouve deux scripts python et un notebook

- AWS_Exemple.py
- AWS_Teacher.py
- AWS_service.ipynb

Pour l'exemple, les résultats se trouvent dans le fichiers json : **/Download/EntitiesPDF.json**

Pour votre propre pdf, les résultats se trouvent dans le fichiers json : **/Professeur/AWS_json_teacher.json**