Adel Moumen CMI Informatique 6 mai 2020

Fiche mission – Stage d’été 2020

Intelligence artificielle, Big Data, Système multi-agents, Traitement automatique du langage naturel

Table des matières

Contexte professionnel de la mission

Le projet s'inscrit dans le cadre du développement d'un annuaire présentant l'ensemble des clubs d'arts martiaux et de sports de combat en France. La mission se divisera en plusieurs axes que le stagiaire pourra emprunter. L'objectif du projet et de produire de nouveaux textes à partir des objets des associations afin de proposer un contenu à la fois intéressant et pertinent pour les utilisateurs de l'annuaire, mais aussi pour les moteurs de recherche. L'accent sera mis sur la résolution du problème du duplicate content.

Création d'un corpus de données

Afin de disposer des données en ce qui concerne l'ensemble des clubs, l'étudiant s'équipera de Python 3 ainsi que de Selenium pour scrapper la base de données du Journal officiel des Associations. Il devra fournir dans un fichier CSV l'ensemble des données récoltées dans un format qui lui sera proposé.

Analyse du corpus

Une analyse de l'ensemble des documents écrits sera réalisée. L'objectif est d'obtenir une représentation de la manière dont la langue française est utilisée par l'humain afin de proposer un modèle permettant de représenter et délimiter la thématique des arts martiaux et des sports de combat. Le stagiaire devra par exemple développer un outil capable de collecterle nombre de mots utilisés et leur fréquence d'apparition. Il pourra également collecter le nombre de syntagmeset de paradigmes utilisés. Toutes les fonctions développées seront utilisées dans les développements qui vont suivre. L'étudiant utilisera Python 3 ainsi que la librairie NLTK.

Extraction de connaissances

Après avoir analysé le corpus, le stagiaire devra reconnaître et lister les types de connaissances qu'il est possible d'extraire de chaque texte afin de les modéliser sous la forme d'un axe sémantique.Chaque texte pourra donc prendre la forme d'un graphe sémantique. Ces graphes pourront être enrichis par la suite.

Enrichissement des connaissances

À partir des connaissances acquises par le biais des textes et la construction d'un modèle de connaissances représentant les différents concepts et propriétés présentes sur des bases de données comme DBPedia, le stagiaire enrichira les connaissances acquises.Cela permettra de former des axes secondaires afin d'étoffer le contenu des documents à produire qui risquent d'être relativement courts. On distinguera alors dans notre modèle la notion de connaissances acquises et de connaissances apprises.

Génération automatique d'un texte à partir d'un graphe de connaissances

Maintenant que chaque texte dispose de suffisamment de connaissances, en utilisant une technique comme le text spinning, l'étudiant pourra générer de nouveaux textes plus pertinents que les documents originaux. Il devra également faire évoluer la méthode de spinning classique afin de résoudre en amont un maximum de problématiques en lien avec le duplicate content.

Vers le développement d'un système multi-agents

Considérant un Agent comme un porteur d'un message, le stagiaire développera la base d'un système capable de représenter l'ensemble des documents du corpus. Il développera des interactions entre les agents pour entamer la résolution de problèmes en lien avec le duplicate contentet l'extraction de connaissances. Un cahier des charges plus fourni sera rédigé pendant la période de stage en fonction des résultats obtenus lors des premières phases du travail.