Bute : réaliser une IA capable de modéliser avec le langage UML et une documentation.

Historique / contexte générale du projet :

En tant qu'étudiant en Master Réseaux et Systèmes Informatiques à la faculté des sciences et techniques à Settat, ma formation se concentre sur le développement web et les techniques de génie logiciel qui vise à créer des applications web et de les documenter. Cette documentation inclut la conception et les explications utilisant le langage UML (Unified Modeling Language ou Langage de Modélisation Unifié).

La réalisation de la conception avec UML repose sur des règles et des conventions qui établissent un cadre commun pour la communication et la représentation des systèmes logiciels. Cependant, cette démarche peut prendre du temps, car il est essentiel de respecter ces règles et de maintenir la cohérence tout au long du processus de conception.

Mon objectif est de créer une application qui, lorsqu'on lui fournit un cahier des charges, effectue une analyse linguistique. Ensuite, elle génère les principaux résultats de cette analyse, notamment un diagramme de cas d'utilisation et un diagramme de classes, accompagnés d'explications détaillées.

Introduction NLP

Le traitement automatique du langage naturel (NLP) représente une branche cruciale de l'intelligence artificielle qui se concentre sur la capacité des machines à comprendre, interpréter et générer le langage humain de manière similaire à celle des êtres humains. En combinant les principes de la linguistique, de l'informatique et de l'apprentissage automatique, le NLP permet aux ordinateurs de traiter et d'analyser un large éventail de données textuelles, ouvrant ainsi la voie à un certain nombre d'applications pratiques dans divers domaines.

L'objectif principal du NLP est de permettre aux machines de comprendre le langage humain dans toute sa complexité, y compris ses nuances, ses ambiguïtés et ses variations. Pour ce faire, le NLP englobe un ensemble de techniques et de tâches, allant de la tokenization de texte à l'analyse syntaxique en passant par la reconnaissance d'entités nommées, l'analyse de sentiment, la traduction automatique, la génération de texte et bien plus encore.

Au cours des dernières années, le domaine du NLP a connu des avancées significatives grâce à l'essor des données massives, à l'amélioration des modèles et des techniques, ainsi qu'à l'augmentation de la puissance de calcul. Ces progrès ont ouvert de nouvelles perspectives dans des domaines tels que la recherche d'informations, les assistants virtuels, la traduction automatique, l'analyse des médias sociaux, la médecine et bien d'autres encore.

Dans ce rapport, nous explorerons

Le langage Humaine - - Les différents niveaux du langage - -

L'analyse du langage nécessite une compréhension approfondie de sa structure à de multiples niveaux : quels sont les mots ? Que signifient-ils ? Comment se combinent-ils pour former des phrases ? Comment contribuent-ils au sens global de la phrase ? En outre, comment fonctionne le monde et la pensée humaine dans ce monde ? Par exemple, pour participer à une conversation, un individu doit non seulement maîtriser le langage utilisé, mais aussi les règles du monde dans lequel lui et son interlocuteur évoluent, ainsi que les règles élémentaires de la conversation.

Les linguistes distinguent plusieurs niveaux permettant l'analyse ou la génération d'un énoncé en langage naturel, depuis la réception des sons (ou leur prononciation) jusqu'à la compréhension profonde des mots dans leur environnement. Ces niveaux de connaissance demeurent pertinents lorsqu'on aborde l'analyse automatique du langage. Les principales spécialités sont :

* La phonétique et la phonologie, qui étudient les liens entre les mots et les phrases et les sons qui les réalisent à l'oral.
* La morphologie, qui examine la construction des mots et leur rôle dans la phrase.
* La syntaxe, qui étudie la manière dont les mots se combinent pour former des syntagmes, des propositions et des phrases correctes.
* La sémantique, qui explore la signification des mots lorsqu'ils sont insérés dans une phrase, indépendamment du contexte.
* La pragmatique, qui analyse comment les phrases peuvent être interprétées en fonction de leur contexte d'énonciation, comprenant les interlocuteurs, les phrases précédentes et la connaissance commune du monde.

Pour illustrer la distinction entre syntaxe, sémantique et pragmatique, Allen propose les exemples suivants. Considérons ces phrases en tant que candidats pour le titre d'un mémoire, énoncées sans contexte :

* "Le domaine du langage naturel et de son traitement automatique se trouve au cœur de la problématique de l’extraction et de la recherche d’information."
* "Les grenouilles vertes ont des gros nez."
* "Les idées vertes ont des gros nez."
* "Vertes des ont les idées nez gros."

La première phrase semble constituer un début raisonnable pour un tel rapport, couvrant les aspects syntaxiques, sémantiques et pragmatiques. La deuxième phrase est correcte sur le plan syntaxique et sémantique, mais pas pragmatique. La troisième phrase est à la fois pragmatiquement et sémantiquement incorrecte. Enfin, la quatrième phrase est tout simplement inintelligible, ne respectant aucune structure grammaticale admise en français.

Qu'est ce qu'une phrase ?

Une phrase comprend toujours au moins deux éléments essentiels : le **sujet** et le **verbe**. Le sujet est la personne, l'animal, la chose ou l'idée dont on parle dans la phrase. Le verbe exprime l'action que le sujet fait ou l'état dans lequel il se trouve.

ou trois éléments : Sujet + verbe + complément

* **Le chat** **attrape** **une souris**. (Sujet : le chat ; verbe : attrape ; complément : une souris)
* **Marie** **chante** **une chanson**. (Sujet : Marie ; verbe : chante ; complément : une chanson)

Dans notre travail, nous nous concentrons exclusivement sur la langue écrite, où les mots sont les entités les plus étudiées. Cela dit, la structure interne du mot, y compris ses différentes flexions, est également prise en compte. Toutefois, la distinction entre les mots est relativement claire dans un texte écrit, ce qui n'est pas toujours le cas dans la langue parlée.

La relation entre Analyse linguistique et la conception :

Dans l'ensemble, cette analyse fournit un cadre pour comprendre comment les éléments linguistiques peuvent être utilisés pour représenter un système en UML .

L'idée sous-jacente à cette analyse linguistique est que le langage naturel est bien plus qu'une simple collection de mots et de phrases. Il contient des informations riches et subtiles sur la manière dont les concepts sont liés les uns aux autres dans notre monde, ainsi que sur la structure et la nature de ces concepts eux-mêmes.

L'analyse linguistique fournit des indications sur la manière dont les éléments linguistiques tels que les noms et les groupes nominaux peuvent être interprétés dans le cadre de la modélisation conceptuelle.

En explorant ces nuances linguistiques, nous pouvons déduire des informations importantes sur la manière dont les éléments linguistiques peuvent être utilisés pour représenter des concepts et des relations dans un système d'information. Par exemple, les noms et les groupes nominaux peuvent être interprétés comme représentant des classes ou des entités dans un système, tandis que les verbes indiquent des associations entre ces concepts.

Les articles définis et indéfinis, tels que "le", "la", "un" ou "une", indiquent la multiplicité des entités représentées. De même, les pronoms démonstratifs comme "ce", "ces" et "cette", ainsi que les synonymes, peuvent être interprétés comme faisant référence à des classes identiques.

Les possessifs, tels que "son", "sa" ou "ses", peuvent être attribués à des attributs ou à des classes, selon que la possession représente une caractéristique simple du possesseur ou une relation structurelle plus complexe entre le possesseur et la possession.

Les verbes peuvent indiquer des associations entre concepts, tandis que les participes présents peuvent représenter des associations continues.

Cette analyse permet donc de construire un cadre conceptuel solide pour la modélisation et la représentation des connaissances dans un système d'information. En comprenant comment les éléments linguistiques sont utilisés pour exprimer des concepts et des relations, nous pouvons concevoir des modèles plus précis et plus riches, capables de capturer la complexité du monde réel de manière plus fidèle. Cela peut être particulièrement utile dans des domaines tels que la recherche d'information, la gestion des connaissances et la modélisation des processus métier, où la précision et la richesse sémantique sont essentielles pour une compréhension et une manipulation efficace des données.

NLP vs NLU

Outils et technologie

Python :

VsCode :

Les bibliothèques : SpaCy, RE

les expressions regilieres.

## **Les bibliothèques SpaCy et RE**

**SpaCy** et **RE** sont deux bibliothèques open-source populaires pour le traitement du langage naturel (NLP) en Python. Elles permettent d'effectuer diverses tâches, telles que la tokenisation, l'étiquetage morphosyntaxique, l'analyse de dépendances, la reconnaissance d'entités nommées et l'extraction d'informations.

**SpaCy** est une bibliothèque plus complète et plus facile à utiliser que RE. Elle offre une large gamme de fonctionnalités et de modèles pré-entraînés pour différents langages, dont le français. SpaCy est particulièrement utile pour les tâches suivantes :

* **Tokenisation:** Découper un texte en jetons (mots, ponctuation, etc.).
* **Étiquetage morphosyntaxique:** Attribuer à chaque jeton sa classe grammaticale (nom, verbe, adjectif, etc.) et ses caractéristiques morphologiques (genre, nombre, temps, etc.).
* **Analyse de dépendances:** Déterminer les relations grammaticales entre les jetons d'une phrase.
* **Reconnaissance d'entités nommées:** Identifier les noms propres dans un texte (personnes, lieux, organisations, etc.).
* **Extraction d'informations:** Extraire des informations spécifiques d'un texte, comme les dates, les noms de produits ou les coordonnées.

**RE** est une bibliothèque plus spécialisée que SpaCy. Elle est conçue pour la reconnaissance d'expressions régulières dans le texte. RE est particulièrement utile pour les tâches suivantes :

* **Recherche de motifs:** Trouver des occurrences d'un motif donné dans un texte.
* **Extraction de données:** Extraire des données spécifiques d'un texte en utilisant des expressions régulières.
* **Classification de texte:** Classer un texte en fonction de son contenu en utilisant des expressions régulières.

**Voici un tableau récapitulatif des principales différences entre SpaCy et RE:**

| **Fonctionnalité** | **SpaCy** | **RE** |
| --- | --- | --- |
| **Facilité d'utilisation** | Plus facile à utiliser | Plus difficile à utiliser |
| **Fonctionnalités** | Plus complète | Moins complète |
| **Modèles pré-entraînés** | Oui | Non |
| **Langages pris en charge** | Plus de langages | Moins de langages |
| **Tâches typiques** | Tokenisation, étiquetage morphosyntaxique, analyse de dépendances, reconnaissance d'entités nommées, extraction d'informations | Recherche de motifs, extraction de données, classification de texte |

**En résumé, SpaCy est un bon choix pour les débutants en NLP et pour les tâches générales de traitement du langage naturel. RE est un bon choix pour les utilisateurs plus expérimentés qui ont besoin de fonctionnalités avancées de reconnaissance d'expressions régulières.**

**Informations complémentaires:**

* **Site web de SpaCy:**<https://spacy.io/>
* **Documentation de SpaCy:**<https://spacy.io/api/doc>
* **Site web de RE:**<https://www.merriam-webster.com/dictionary/re>
* **Documentation de RE:**<https://docs.getre.io/latest/docs/start_here>

Diagramme de cas d’utilisation :

Définition :

Est un description détaillée d'une fonctionnalité ou d'un ensemble de fonctionnalités d'un système logiciel du point de vue des utilisateurs.

Théorique :

commençant par définir “chercher” c'est quoi le strict minimum pour dire ce que “X” c'est une cas d'utilisation ?

On va prendre des exemple pour extraire la théorie générale:

les phrases sont extraits à partir d'un cahier des charges d’un **Gestion des projets de recherche**

***“Vous le trouverez dans le projet ”***

***Les verbes “ER”:***

* La partie web du dépôt de documents sert à générer de manière automatique un site Web dont l’accès en lecture est toujours public.
* peut valider ou rejeter une création de projet.
* qui a la possibilité d’ajouter et de retirer des membres au projet.
* peut ajouter un des membres du projet au groupe d’administrateurs du projet.
* peut créer une liste de diffusion associée au projet en plus des listes prédéfinies vues précédemment.
* peut demander la clôture du projet.
* peut refuser de clôre un projet.
* permettra de réaliser ensemble des productions de documents électroniques.
* permet de gérer des projets, des personnes et des groupes.
* permet de créer et de supprimer les informations sur les personnes, les projets et les groupes.
* permet de réaliser ensemble des productions de documents électroniques.
* Pour obtenir un compte sur le portail, un nouvel utilisateur doit passer par une procédure d’inscription qui lui demande les informations suivantes : son nom et son prénom, son adresse Email et l’identifiant qu’il désire utiliser sur le portail.

**ANALYSE :**

Les cas d'utilisation se situent entre le verbe jusqu'à la fin ”.”, des fois il atteint pas la fin de la phrase “.”

Une phrase peut contenir un ou plusieurs cas d'utilisations.

peut être que ce n'est pas une cas d'utilisation (Phrase 1) , car il parle d'une fonctionnalité que l'app doit faire.

il faut que le **sujet soit un humain “vivant”.**

***Les verbes “RE”:***

* l'application web permet le travail collaboratif d’un ensemble de personnes qui participent à un projet pour réaliser ensemble des productions de documents électroniques.
* Il offre des ressources à chaque groupe. “il = Le portail ”
* Le système doit permettre de créer et de supprimer les informations sur les personnes, les projets et les groupes.
* Le système est public,C’est-à-dire qu’il n’y a pas de contrôle lors de l’inscription des utilisateurs qui sont associés dès leur inscription au groupe Default.
* qu’il désire utiliser sur le portail.
* L’identifiant de connexion doit être unique.
* Le portail gère l’accès à un ensemble de ressources qui peuvent être associées à un utilisateur ou à un projet.
* Le projet peut être public ou privé.
* Un administrateur de la plateforme peut refuser de clôre un projet.
* Les projets clos peuvent être détruits après deux ans.

**ANALYSE :**

**le sujet généralement c'est le portail.**

Le cas d'utilisation est loin des verbes qui se terminent avec “re”.

Les verbes “RE” permettent de décrire un scénario.

la 3eme phrase contient une cas d'utilisation , mais déjà il existe un verbe “ER”

***Les verbes “RI”:***

* Le système peut offrir de plus un compte public appelé guest qui est associé au groupe Guests.

**ANALYSE :**

Pas de cas d'utilisation.

***Conclusion*** : Les cas d'utilisation sont la plupart des cas , sont situés après les verbes qui se terminant avec "ER" ; qui décrivent les fonctionnalités principales de l'application web.

Problème 😡: Comment extraire exactement le cas d'utilisation ?

Malheureusement la vie n'est pas belle , faut régler les problèmes ; Les cas d'utilisation se situent entre le verbe jusqu'à la fin, des fois il atteint pas la fin de la phrase “.” et une phrase peut contenir un ou plusieurs cas d'utilisations.

Solution 🙂:

1- on commence par la phrase qui peut contenir un ou plusieurs cas d'utilisations . Alors cela est possible que si seulement si la phrase contient “, et ” un ou plusieurs de ces éléments.

exemples : doit permettre de créer et de supprimer les informations sur les personnes, les projets et les groupes.

à un niveau d'abstraction plus haut la phrase contient 1 seul cas d'utilisation :

* créer et de supprimer les informations sur les personnes, les projets et les groupes.

on peut la découper en 4 :

* créer et de supprimer les informations sur les personnes.
* créer et de supprimer les projets.
* créer et de supprimer les groupes.

même en 6 :

* créer les informations sur les personnes
* créer les informations sur les projets
* créer les les groupes
* supprimer les informations sur les personnes
* supprimer les informations sur les projets
* supprimer les les groupes

! Je vais travailler sur un niveau d'abstraction plus élevé dans les cas d'utilisations.

2- Les cas d'utilisation se situent entre le verbe jusqu'à la fin, des fois il atteint pas la fin de la phrase :

**conclusion générale:**

Jusqu'en moment nous avons pas bcp d'information pour passer une mais on peut dire que CAS D'UTILISATION = Verbe + GROUPE NOMINALE. Et il se trouve dans une phrase entre le Sujet (un être humain) + verbe . . . . +GROUPE NOMINALE.

Procédure de code :

Extraire les phrases qui contiennent des sujets Humaine

Extraire les phrases qui contiennent les verbes à l’infinitif.

import spacy

nlp = spacy.load("fr\_core\_news\_sm")

text = """

Je veux manger une pomme.

Tu dois aller à l'école.

Il aime courir dans le parc.

"""

for token in nlp(text):

if token.pos\_ == "VERB":

if token.infinitive\_form:

print(token.sent)

vérifier si le sujet de la phrase est un être humain “Point rouge ”

Extraire de 1 du verbe jusqu'à la fin de la phrase .

Extraire de 2 les groupes nominaux et les ajouter à chaque fois au verbe.

Généraliser le module