TP4_TAL_rapport

Vanessa Guerrier

Avril 2025

1 Introduction

Ce rapport a pour but d'expliquer ce que nous avons réalisé durant le projet final de traitement automatique du langage naturel. Nous commencerons par une présentation de l'objectif du projet, puis nous détaillerons les étapes de mise en œuvre, les modèles utilisés et les résultats obtenus lors des tests.

2 Objectif du projet

Le but de ce projet est de développer une interface en langage naturel permettant d'interroger une base de données en SQL. Contrairement à l'approche basée sur une grammaire, où l'on construisait les requêtes à l'aide de notions de compilation, la formulation des requêtes dans ce projet est plus libre. Pour cela, nous utilisons le modèle **Intention-Concepts-Valeurs (ICV)**.

3 Extraction des intentions et des concepts

Nous commençons par identifier, dans chaque phrase, les intentions de l'utilisateur. Pour cela, nous avons défini une fonction `convert` qui prend en entrée une phrase, et renvoie cette phrase où les mots issus de la base sont remplacés par l'un des cinq concepts suivants :

- `name` : pour les titres de films,
- `producteur` : pour les noms de réalisateurs,
- `acteur` : pour les noms d'acteurs,
- `year` : pour les années de sortie des films,
- `type` : pour les genres de films.

et un tableau contenant les concepts et leur valeurs.

4 Création des jeux de données et étiquetage

À partir d'un corpus contenant différentes phrases et leur requête SQL associée, nous avons définie la fonction find_label qui crée deux étiquettes (labels) correspondant aux intentions de l'utilisateur :

- Une étiquette pour la clause `SELECT`,
- Une étiquette pour la clause `WHERE`.

Ces étiquettes nous permettent de superviser l'apprentissage de deux modèles, chacun dédié à prédire l'une des deux clauses SQL.

5 Entraînement des modèles

Pour entraîner nos modèles, nous avons utilisé le composant `Pipeline` de `scikit-learn`, qui permet de chaîner un vecteuriseur (`CountVectorizer` ou `TfidfVectorizer`) avec un perceptron. Nous avons séparé notre corpus en deux parties :

- 70% des données pour l'apprentissage,
- 30% des données pour le test.

L'évaluation sur les données de test nous a permis de comparer différentes combinaisons de modèles. Nous avons constaté que l'utilisation de `TfidfVectorizer` avec un perceptron offrait les meilleures performances pour la prédiction de la clause `WHERE`, avec un écart de 0,02% en précision par rapport aux autres configurations. Nous avons donc privilégié cette combinaison dans notre solution finale.

6 Création de la requête SQL

Une fois notre modèle entraîné, nous l'avons utilisé pour prédire les différentes parties de la requête SQL à partir de la phrase en langage naturel. La fonction `produire_requete` nous permet alors de construire automatiquement la requête correspondante.

7 Évaluation de notre modèle

7.1 Évaluation quantitative

Nous avons évalué notre système selon trois axes principaux. La métrique retenue est la **précision**.

- **Capacité à extraire correctement les concepts et valeurs des requêtes SQL à partir des questions en langage naturel :**

  - **Précision :** 95%.
  - **Causes des erreurs :** Certaines valeurs ne sont pas présentes dans la base de données, ou bien elles n'apparaissent pas telles quelles dans la phrase.
  - **Solution :** Pour corriger ce problème, on pourrait introduire des concepts spécifiques représentant directement les attributs de la base de données.

- **Performances du système sur les deux types d'intentions : SELECT et WHERE**
  - **Précision :** `SELECT = 99%`, `WHERE = 92%`.

- **Capacité du système à produire la bonne réponse (résultat de la requête SQL sur la base de films) :**
  - **Précision :** 99,4%.
  - **Causes des erreurs :** Elles proviennent principalement de fautes d'orthographe dans la requête ou la phrase.
  - **Solution :** Dans la fonction `convert`, il serait possible d'ignorer la casse et les accents lors de la comparaison des mots avec ceux de la base de données. Ensuite, dans la requête finale, nous pourrions réattribuer au concept la forme exacte du mot tel qu'il apparaît dans la base.

7.2 Évaluation qualitative

Ici, en ce qui concerne la qualité, notre système n'est pas très performant au vu des requêtes SQL générées : seulement 15 % des requêtes sont correctes, 40 % sont incorrectes et 45 % sont partiellement correctes.

8 Comparaison entre les types d'évaluation

Nous constatons que notre système est plus performant lors des évaluations quantitatives que lors des évaluations qualitatives. Cela peut s'expliquer par la manière dont les questions ont été formulées lors de l'évaluation qualitative, par des fautes d'orthographe, ou encore par le fait que certaines informations demandées ne figurent pas dans notre base de données.

On peut ainsi conclure qu'un bon score lors d'une évaluation quantitative ne garantit pas nécessairement qu'un système est de bonne qualité.

Question	Requête SQL générée	Score
Quels sont les films réalisés par Christopher Nolan ?	SELECT titre FROM films WHERE realisateur = 'Christopher Nolan';	2
Quels films sont sortis après 2010 ?	SELECT titre FROM films WHERE annee BETWEEN 2010 AND 2010;	1
Quels films appartiennent au genre science-fiction ?	SELECT titre FROM films WHERE (acteur1 = 'player' OR acteur2 = 'player' OR acteur3 = 'player') AND genre = 'type';	0
Quels films mettent en vedette Leonardo DiCaprio ?	SELECT titre FROM films WHERE ((acteur1 = 'Leonardo DiCaprio' OR acteur2 = 'Leonardo DiCaprio' OR acteur3 = 'Leonardo DiCaprio')) AND ((acteur1 = 'Leonardo DiCaprio' OR acteur2 = 'Leonardo DiCaprio' OR acteur3 = 'Leonardo DiCaprio'));	1
Quels sont les films réalisés par Quentin Tarantino entre 1990 et 2000 ?	SELECT titre FROM films WHERE annee BETWEEN 1990 AND 1990;	1
Quels sont les films où jouent à la fois Tom Hanks et Meg Ryan ?	SELECT titre FROM films WHERE ((acteur1 = 'Tom Hanks' OR acteur2 = 'Tom Hanks' OR acteur3 = 'Tom Hanks')) AND ((acteur1 = 'Tom Hanks' OR acteur2 = 'Tom Hanks' OR acteur3 = 'Tom Hanks'));	1
Combien de films appartiennent au genre comédie ?	SELECT titre FROM films WHERE (acteur1 = 'player' OR acteur2 = 'player' OR acteur3 = 'player') AND genre = 'type';	0
Quels sont les films dans lesquels joue Robert De Niro sortis avant 2000 ?	SELECT titre FROM films WHERE (acteur1 = 'Robert De Niro' OR acteur2 = 'Robert De Niro' OR acteur3 = 'Robert De Niro') AND genre = 'type';	0
Quels films ont pour genre Drame et sont sortis entre 2005 et 2015 ?	SELECT titre FROM films WHERE annee BETWEEN 2005 AND 2005;	1
Quels réalisateurs ont travaillé sur plus de 3 films ?	SELECT titre FROM films WHERE titre LIKE 'Burning%' ORDER BY annee DESC;	0
Quels sont les réalisateurs qui ont travaillé avec Tom Cruise ?	SELECT titre FROM films WHERE (acteur1 = 'Tom Cruise' OR acteur2 = 'Tom Cruise' OR acteur3 = 'Tom Cruise') AND genre = 'type';	0
Quels sont les films qui ont pour genre Thriller, réalisés entre 1995 et 2005 ?	SELECT titre FROM films WHERE annee BETWEEN 2005 AND 2005;	1
Quels films ont le même réalisateur que Inception ?	SELECT titre FROM films WHERE annee BETWEEN 1954 AND 1966;	0
Quels sont les films sortis entre 1980 et 1989 ?	SELECT titre FROM films WHERE annee BETWEEN 1980 AND 1989;	2
Quels films de Science-fiction sont sortis après 2015 ?	SELECT titre FROM films WHERE (acteur1 = 'player' OR acteur2 = 'player' OR acteur3 = 'player') AND genre = 'Science-fiction';	0
Quels films sont sortis la même année que "Titanic" ?	SELECT titre FROM films WHERE titre = 'name';	0
Quels films sont sortis exactement en l'an 2000 ?	SELECT titre FROM films WHERE annee BETWEEN 2000 AND 2000;	1
Quels sont les films sortis pendant les années 1970 à 1979 ?	SELECT titre FROM films WHERE annee BETWEEN 1970 AND 1979;	2
Je Voudrais les films sorti en en 2024 ?	SELECT titre FROM films WHERE annee BETWEEN 2024 AND 2024;	1
Quels sont les films dans lesquels joue Brad Pitt mais pas Angelina Jolie ?	SELECT titre FROM films WHERE (acteur1 = 'Brad Pitt' OR acteur2 = 'Brad Pitt' OR acteur3 = 'Brad Pitt') AND genre = 'type';	1

Table 1: Résultats de l'évaluation qualitative