Classification automatique de mesures de similarité

Cahier des charges

Année universitaire 2023 - 2024

Étudiants : SEN Abdurrahman
VADUREL Benjamin

Encadrants : DUCHATEAU Fabien
FAVETTA Franck

I. Préambule

Ce projet prend place dans le cadre de l'UE Ouverture à la recherche pendant la première année du Master Informatique à l'Université Lyon 1 Claude Bernard. Merci à DUCHATEAU Fabien et FAVETTA Franck de nous encadrer et merci à BRANDEL Sylvain d'orchestrer cette UE.

II. Contexte

L'appariement d'entités consiste à **identifier des instances équivalentes dans des jeux de données** [1].
Par exemple, dans les tableaux ci-dessous, les instances *123* (à gauche) et *P1* (à droite) correspondent au même objet du monde réel (i.e., le roman *les deux tours*).

id	nom	type
123	Les deux tours	livre
345	Holy grail	DVD
567	Les deux tours	DVD

id	titre	catégorie
P1	Le seigneur des anneaux - Les 2 tours	roman
P2	Holy Grail - Dark Passenger	musique

Pour détecter ces entités correspondantes, on utilise des **mesures de similarité** qui calculent un score entre les propriétés des entités à comparer. Par exemple, le ***coefficient de Dice*** calcule un score de 0,51 entre les chaînes *Holy Grail* et *Holy Grail - Dark Passenger*, tandis que la ***distance de Jaro*** (et par extension celle de ***Jaro-Winkler*** également) nous donne un score de 1 car le long préfixe identique a un grand poids dans le calcul de ces distances.

Il existe de nombreuses mesures de similarité [4, 5, 6], qui sont généralement catégorisées selon leur fonctionnement : certaines considèrent des ***tokens*** (ou jetons représentant des mots dans une phrase par exemple) pour le calcul du score, d'autres exploitent les ***manipulations*** de caractère nécessaires pour arriver d'une chaîne à une autre, d'autres encore se basent sur la ***phonétique*** des chaînes de caractères pour les comparer.

Dans ce projet, nous voulons **classer ces mesures** selon leurs résultats sur un jeu de données, afin de pouvoir les étudier et les comparer (robustesse sur le volume du jeu de données par exemple).

III. Objectifs

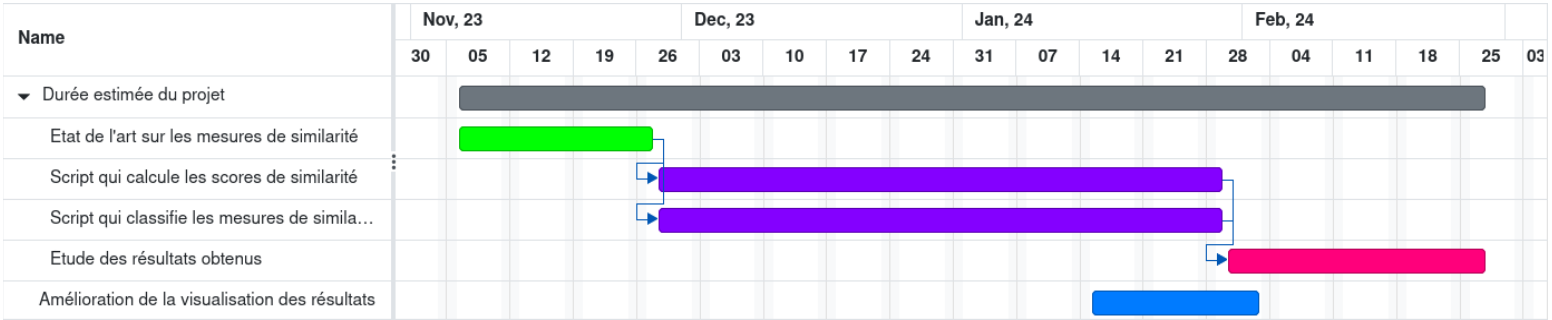
L'objectif global de ce projet est de proposer une nouvelle classification de mesures de similarité entre deux entités afin de pouvoir, à terme, vérifier la qualité d'un appariement sans utiliser un jeu de données expertisé.

La liste ci-dessous indique toutes les étapes par lesquelles passera le projet :

- Rédiger un court **état de l'art sur les mesures de similarité et leur classification**;
- Améliorer un script Python existant qui **calcule des scores de similarité, pour un ensemble de mesures et un jeu de données**. Le calcul des scores de similarité est réalisé à partir des deux bibliothèques Python : *Talisman* et *Jellyfish*
    - Jellyfish [7] permet d'apparier et/ou de mesurer la distance entre deux chaînes de caractères de façon approximative ou phonétique ;
    - Talisman [8] propose d'autres méthodes pour mesure les similarités entre deux chaînes de caractères;
- Développer un script Python pour **classifier les mesures de similarité. On pourra utiliser des algorithmes de clustering** (bibliothèque *scikit-learn*);
- **Étudier les résultats obtenus** (comparaison avec la classification traditionnelle, robustesse de la classification, configuration des paramètres comme le nombre de catégories, etc.).
- **Proposer une meilleure visualisation des clusters** qui n'est pas pratique actuellement. On obtient simplement un affichage en console où chaque donnée est associée à un cluster (par exemple : *("dice", 0), ("jaro", 1), ("jaro-winkler",0)*).

IV. Calendrier prévisionnel

Le projet devrait s'étendre **de début novembre 2023 à fin février 2024**.
Le diagramme de Gantt ci-dessous présente pour chaque étapes une période de temps qui lui sera attribuée :



V. Références

[1] Wei Shen, Jianyong Wang, and Jiawei Han.
Entity linking with a knowledge base: Issues, techniques, and solutions.
Knowledge and Data Engineering, IEEE Transactions on, 27(2):443–460, 2015. Lien TKDE14

[2] Indice de Sørensen-Dice — Wikipédia

[3] Distance de Jaro-Winkler — Wikipédia

[4] Seung seok Choi and Sung hyuk Cha. **A survey of binary similarity *and distance measures***.
Journal of Systemics, Cybernetics and Informatics, pages 43–48, 2010. Lien JSCI2010

[5] Nick Koudas, Sunita Sarawagi, and Divesh Srivastava.
Record linkage: similarity measures and algorithms.
In *SIGMOD*, pages 802–803. ACM, 2006. Lien SM2006

[6] Najlah Gali, Radu Mariescu-Istodor, and Pasi Fränti.
Similarity measures for title matching.
In *2016 23rd International Conference on Pattern Recognition*, pages 1548–1553. IEEE, 2016. Lien ICPR2016

[7] https://yomguithereal.github.io/talisman/metrics/

[8] https://pypi.org/project/jellyfish/