UNIVERSITÉ DE LYON

Université Claude Bernard Lyon 1

Classification automatique de mesures de similarité

Cahier des charges

Etudiant: Ghiles Boussahla (Master 1 informatique)

Encadrants: Fabien Duchateau (Maître de conférence, LIRIS)
Franck Favetta (Maître de conférence, LIRIS)

Projet d'ouverture à la recherche 2022-2023

1- Contexte :

La quantité de données numériques crées ou répliquées a augmenté d'une manière exponentielle et cela est connu sous le nom de « big data ». Ce phénomène de données massives crée le besoin en analyse de données afin d'extraire des informations intéressantes. Même si le langage naturel est celui utilisé dans la majorité des données numérique, celui-ci reste très ambigu, en particulier les entités désignées, car une entité peut avoir plusieurs noms et un nom peut désigner plusieurs différentes entités.

L'objectif sera de classifier de manière automatique des mesures de similarité, ces dernières consistent à identifier des objets équivalents dans des jeux de données. Par exemple dans les tableaux ci-dessous, l'objet 123 du premier jeu de données correspond à l'objet 1 dans le second. Pour détecter ces objets correspondants, on utilise des mesures de similarité qui calculent un score entre les propriétés des objets à comparer, ces mesures utilisent des distances : euclidienne, de Manhattan, le coefficient de Dice (qui donne un score de 0.51 entre les chaînes 'Holy Grail' et 'Holy Grail – Dark Passenger' sur les exemples ci-dessous. L'objectif est de classer ces différentes mesures de similarité selon leurs résultats sur des jeux de données.

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

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

2- Présentation du projet

2.1- Objectifs

L'objectif de ce projet sera de calculer un indice de similarité entre des jeux de données, en utilisant des mesures différentes et de les classifier selon leurs résultats.

On pourra donc s'intéresser à plusieurs problématiques telles que la configuration des mesures et des algorithmes de calcul de similarités, le changement du nombre de mots des jeux de données (100, 500, 10 000 mots) ou l'utilisation de différents algorithmes de classification.

- État de l'art sur la classification des mesures de similarité.
- Le développement d'un script qui calcule des scores de similarité, pour un ensemble de mesures et un jeu de données.
- Application des algorithmes de classification (clustering) pour catégoriser ces mesures.
- Étude sur les résultats obtenus en comparaison avec la classification traditionnelle, configuration des paramètres comme le nombre de catégories, etc.

2.2- Travail à faire

Tâche 1 : Définition de la taxonomie et des mesures de similarité

a. Etat de l'art sur les taxonomies utilisées en recherche et l'ensemble des mesures de similarité.[1]

[1] Article http://dbgroup.cs.tsinghua.edu.cn/wangjy/papers/TKDE14-entitylinking.pdf

b. Choix et définition des mesures qui seront utilisées (coefficient de Dice, distance de Manhattan, l'index de Jaccard, etc.).[2]

Tâche 2 : Analyse de similarité entre jeux de données

a. Développement de l'algorithme pour calculer l'indice de similarité entre des jeux de données.
b. Expérimentations avec des listes de mots de taille ou d'hétérogénéité variable.
c. Analyse manuelle des résultats.

Tâche 3 : Classification automatique des mesures de similarité

a. Sélection et utilisation de différents algorithmes d'apprentissage ('clustering') pour classifier les mesures de similarité utilisées en fonction de leurs résultats.
b. Comparaison des résultats avec la classification traditionnelle et le changement de paramètres tel que le nombre de catégories (cluster).

3- Contraintes

Langage : Python (ainsi que des bibliothèques tel que scikit-learn[3], pandas[4], etc.), Anaconda, Jupyter Notebook.

Versionning : Gitlab.

Organisations : 2 réunions /mois.

4- Livrables

- Cahier des charges
- Algorithme de classification automatique des mesures de similarité.
- Rendu intermédiaire : en cours d'organisation.
- Rapport final et vidéo de vulgarisation

[2] Article Talisman - Metrics (yomguithereal.github.io)
[3] https://scikit-learn.org/stable/
[4] https://pandas.pydata.org/

5- Diagramme de Gantt