Master 1 Informatique - Web Sémantique

TP2 - Construction d'un outil d'intégration de données

Hugo Maitre
Adrien Linares
Mitra Aelami
Amirhossein Nasri

Enseignant :
Pr. Konstantin Todorov

19 mars 2022

Table des matières

1 Introduction

L'intégration des données sémantiques est un composant fondamental dans le domaine du web sémantique. Pour ce faire nous avons besoin d'ontologies, elles décrivent des concepts et des relations d'un domaine particulier.

Pour réaliser ce projet, nous a été attribuées deux ontologies sous la forme de graphes rdf. Le premier graphe, dit graphe source rassemblait des oeuvres musicales de la Bibliothèque Nationale de France (BnF), le second, dit target, était quant à lui de la Philarmonie de Paris. Enfin un benchmark (fichier d'alignements [1] de références) nous a été donné pour évaluer notre système.

Nous avons travaillé avec Python qui était le langage avec lequel le groupe se sentait le plus à l'aise pour ce type de travail. Nous avons de plus tout fait sur Google Colab, en utilisant parfois − avec parcimonie bien-sûr − les GPU de Google, certains calculs étant très longs sur une architecture lambda. Notre travail, rendu sous le format d'un notebook, se trouve sur le dépot git ici.

1. ensemble de correspondances entre les éléments de deux ontologies hétérogènes

2 Développement

2.1 Preprocessing

En ce qui concerne les preprocessing, nous nous sommes d'abord débarassés des noeuds et des urls des objets en filtrant sur les graphes originaux. Pour cela on a manipulé des arrays avec pour colonnes sujet, prédicat, objet et comme lignes les triplets résultants du graphe.

Nous avions également voulu traduire les données qui étaient en anglais vers le français mais lorsque nous avions vu que cet opération était beaucoup trop couteuse en temps cpu nous l'avons par la suite écarté. Nous avons par la suite sauvegarder les fichiers résultants sous un format csv pour essayer de perdre moins de temps à l'éxécution.

Nous avons également utilisé certains des prétraitements effectués pour l'UE Machine Learning 1, notamment en se servant de la librairie NLTK. Certains des preprocessing utilisés incluent la tokénisation de chaque mot, la suppression de la ponctuation et des stopwords, le passage des majuscules en minuscules, la lemmatisation de chaque mot, la suppression des pronoms, déterminants verbes à l'infinitif et d'autres. Ces prétraitements ont été effectué sur les deux graphes.

2.2 Mesures de similarité utilisées

La similarité agit comme une valeur de confidence entre une paire de concepts appartenant à deux ontologies différentes. Nous avons fait le choix dans le cadre de ce projet d'en privilégier certaines au détriment d'autres. Parmi celles retenues citons :

1. Identity similarity
2. Normalized Levenshtein similarity
3. JaroWinkler similarity
4. Jaccard similarity
5. Cosine similarity
6. Overlap coefficient similarity
7. Sørensen–Dice coefficient similarity
8. Sift4 similarity

En ce qui concerne la dernière, elle a été inspirée de JaroWinkler et d'une mesure de plus longue sous séquence. On peut retrouver plus de détails ici, celle-ci n'étant que peu documentée sur le web.

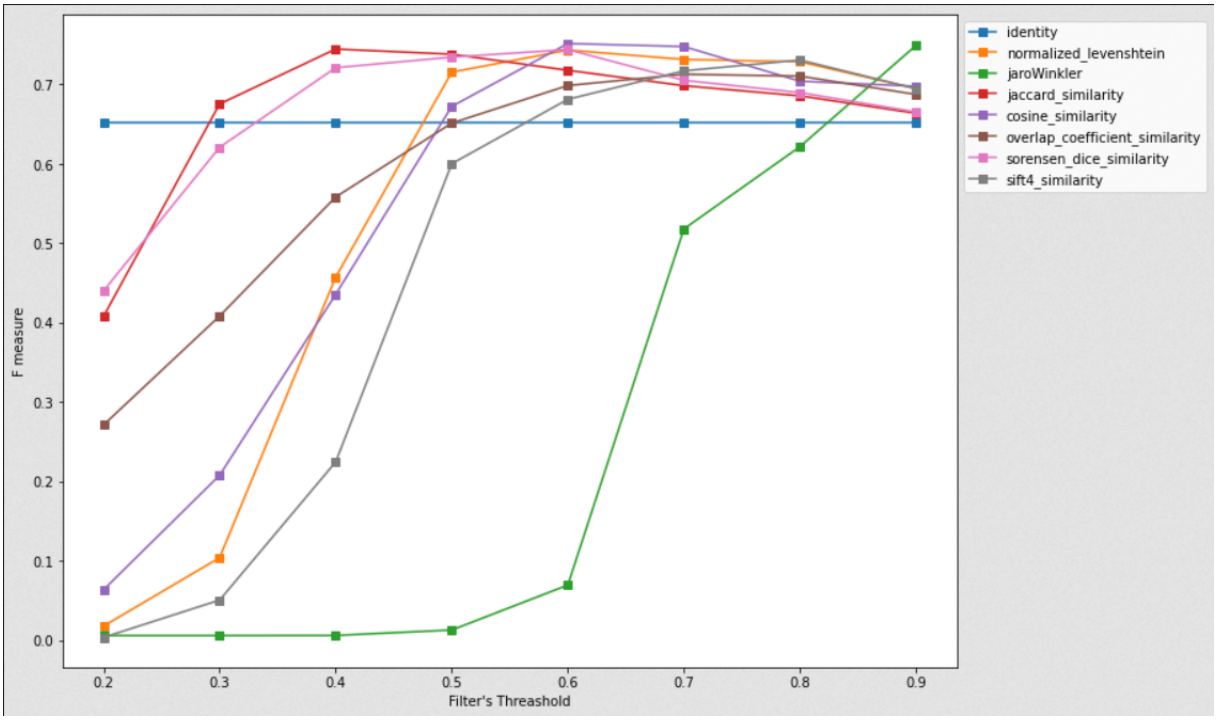
2.3 Comparaison et matching strategy

Nous avons premièrement récupérer pour chaque graphe les propriétés avec leurs occurences respectives ainsi qu'une fonction permettant d'avoir les propriétés communes à nos deux graphes par ordre croissant d'occurence. Nous avons également défini une première fonction pour permettre à l'utilisateur d'intéragir avec notre système, celle-ci attend un entier en entrée entre 0 et 4, 0 étant la propriété ayant l'occurence la plus élevée et 4 la moins élevée. On veut un nombre entre 0 et 4 car c'est le nombre de propriétés restantes après filtrage (celles apparaissant dans le masque d'avant). Une autre fonction permet de choisir une propriété donnée.

2.4 Évaluation

2.5 Résultats

Différents seuils amènent à des résultats différents, le graphique ci-dessous présente ces différences. Les seuils sont représentés en fonction du f-measure.



3 Conclusion

3.1 Principales difficultés

Nos principales difficultés se sont portées principalement sur

3.2 Perspectives d'amélioration

On pourrait imaginer une interface graphique pour l'utilisateur qui sera au final bien plus intuitive, qu'elle soit mobile non.

3.3 Bilan

Nous avons apprécié travailler sur ce projet qui a été un complément au premier et nous a enseigné beaucoup plus sur le web sémantique. Le contexte était d'autant plus intéressant pour certains de nous qu'il portait sur le monde des oeuvres musicales.