Module GMD : Projet 2025

Réalisation d'un système d'intégration de données biomédicales

10 mars 2025

Contexte du projet et définitions

Vous développerez un système qui permet à un utilisateur de proposer un ou plusieurs symptômes pour savoir 1) quelles maladies pourraient causer ce(s) symptôme(s) ou 2) quels médicaments pourraient en être la cause. Pour cela votre système devra permettre d'intégrer des données (réelles) diverses sur les symptômes, maladies, et médicaments qui sont actuellement réparties dans plusieurs sources de données.

Pour le projet, nous considérons ainsi trois types d'entités : des signes et symptômes, des maladies, et des médicaments.

Définitions :

Signes et symptômes : observations de modifications relativement ponctuelles de l'état d'un individu. Ces observations peuvent être la conséquence de dysfonctionnements plus globales que sont les maladies. Les signes et symptômes peuvent être observés de façon subjective ou objective par l'individu lui-même ou par un clinicien. Il arrive également que la survenue de signes et symptômes soit causée par un médicament. On parle alors des effets secondaires du médicament

Maladies : altération des fonctions ou de la santé d'un organisme vivant, partielle ou totale, permanente ou passagère. Elle se manifeste par un ensemble de signes et symptômes spécifiques, reflétant une déviation par rapport à l'état physiologique normal.

Médicament : substance ou une composition de substances prescrite pour diagnostiquer, prévenir, soulager, traiter ou guérir une maladie ou ses symptômes. On parle dans ce cas de l'indication du médicament. Il arrive que 2 médicaments soient prescrits aux patients, le premier pour traiter une maladie, et le second pour limiter un effet secondaire du premier.

Nous nous intéressons aux types de relations suivantes entre nos 3 entités :

— Un signe ou symptôme peut-être soit la manifestation d'une maladie.
— Un signe ou symptôme peut également être l'effet secondaire d'un médicament.
— Une maladie ou un symptôme peut être l'indication d'un médicament.

Objectif du projet

Chaque groupe développera un système d'intégration de données de type médiateur qui permet de retrouver :

— à partir d'un signe ou symptôme, l'ensemble des maladies qui pourraient le causer mais également les médicaments qui pourraient de façon indésirable en être la cause.
— Si le signe ou symptôme peut-être causé par une maladie, la liste des médicaments qui pourraient la traiter ; et si c'est un effet indésirable, la liste des médicaments qui pourraient traiter l'effet indésirable.

Exigences du projet

1. La langue du projet sera l'anglais ou le français (pour les commentaires du code, la documentation et les éventuelles interfaces), mais pas un mélange des deux.
2. Le système doit considérer simultanément les différentes sources de données suivantes : (cf. Section sur les sources de données).
3. Le système d'intégration doit suivre l'architecture de type médiateur.
4. L'utilisateur doit pouvoir faire une requête par nom de signes et symptômes pour retrouver maladies et médicaments associés.
5. L'utilisateur doit pouvoir écrire une requête avec l'opérateur logique ET. La liste de résultats correspond à l'intersection des éléments associés.
6. La présentation des résultats à l'utilisateur doit lui permettre de distinguer clairement à quoi correspondent les résultats.

Les sources de données

Source 1 : DrugBank Format : XML. Cette source contient, entre autre, des données sur l'indication du médicament (attribut Indication), ses effets secondaires (attribut Toxicity). Attention, car ces données sont présentes sous la forme de phrases qui contiennent des noms de maladies. Pour simplifier, nous considérerons qu'un nom de maladie présent dans l'attribut Indication (respectivement Toxicity) est une indication (respectivement un effet secondaire).

Source 2 : OMIM Format : Texte + CSV. Cette source contient des données sur les maladies génétiques, notamment leur signes et symptômes, dans les sections marquées par la balise *FIELD* CS. Vous trouverez également le fichier omim_onto.csv qui permet d'associer des CUI à certains éléments d'OMIM.

Source 3 : Sider 4.1 Format : MySQL/TSV. Cette source est composée de quatre tables (meddra, meddra_all_indications, meddra_all_se* [side effects], meddra_freq) qui contiennent les indications et les effets secondaires données par les notices d'utilisations de médicaments. Sider propose des identifiants appelés CUI que l'on retrouve dans de nombreuses bases de données.

Source 4 : HPO et HPO Annotations Format : OBO + SQLite. HPO (Human Phenotype Ontology) est un vocabulaire contrôlé de référence pour les signes et symptômes. Il contient notamment une liste de synonymes pour les signes et symptômes. HPO Annotations contient des associations entre les identifiants des signes et symptômes de HPO et les maladies de OrphaData et OMIM.

Source 5 : STITCH et ATC Format : TSV + texte. Localisation Arche ET `http://stitch.embl.de/download/chemical.sources.v5.0.tsv.gz`. Ces sources sur les médicaments vont vous permettre d'associer un label aux médicaments par le cheminement suivant : SIDER : stitch_compound_id -> STITCH :compound_id -> Code ATC -> Label.

Conseils

Mediateur : les données doivent rester dans leurs sources d'origine. Lorsque l'utilisateur pose une requête, celle ci est traduite pour être posée aux différentes sources de données, les résultats de chaque source de données sont ensuite regroupés de façon cohérente avant d'être présentés à l'utilisateur. **Cependant**, si vous estimez qu'un fichier pourrait

être utilisé dans un certain type de base de données, vous pouvez changer son type et **justifierez** les raisons de votre choix.

Utilisation d'indexes : il est fortement recommandé de faire des indexes full text pour les fichiers texte.

Schémas : En premier lieu vous devez établir le de chacune des sources le plus en détail possible ainsi qu'un schéma global qui sera utilisé dans votre application. Le travail pourra ensuite être répartit sur les mappers/wrapers pour les différentes sources.

Gestion de code : Utilisez un Git pour mettre en commun le code ainsi que les documents de conception. Attention cependant à ne pas déposer les données dessus. Déposez seulement du codes ou de la doc.

La soutenance : 25 Mars 14h-16h

Durée : environ 20-25 minutes / trinômes. 5-10 minutes de présentation des mappings, des choix technologiques et théoriques. 15 minutes de démonstration du système + questions.

Barème

1. La base : La création d'un système qui répond aux exigences précédentes assurera une note de 10/20 au groupe.
2. Les fonctionnalités supplémentaires ajouteront des points au dessus de cette note de 10.

Les fonctionnalités supplémentaires

Des points supplémentaires peuvent être gagnés en développant les fonctionnalités suivantes (le chiffre annoncé est un nombre de points bonus maximum par fonctionnalité) :

— **Quantification de la qualité des mappings :** +2 points. Vous comptez et présentez le nombre de correspondances entre les sources.
— **Statistiques sur les sources de données :** +2 points. Vous présentez vos données, leur compositions et répartitions. (Un minimum de rigueur mathématique est attendu si vous trouvez des lois de répartitions)
— **Une interface graphique :** +1 point. Le système propose une interface graphique complète et intuitive.
— **Requête avec des OU et des ET :** +1 point. Il est possible de faire une requête avec des signes et symptômes articulés avec des OU et des ET.
— **Utilisation des synonymes :** +1 point. L'utilisateur doit pouvoir faire une requête avec des synonymes de noms de signes et symptômes.
— **Utilisation de jokers dans la requête :** +2 points. L'utilisateur doit pouvoir écrire une requête partielle en utilisant des caractères joker.
— **Tri des résultats :** +1 point. Les résultats sont triés suivant un score.
— **Fournir la provenance des données :** +1 point. L'utilisateur doit pouvoir savoir de quelle source proviennent les résultats d'une requête.
— **Visualisation des résultats :** +1 point. Un point pourra être accordé si le groupe propose une visualisation intéressante des résultats.

Dates importantes

Les accès vers le code doivent être donnés au plus tard 24h avant votre soutenance.