Zadanie zaliczeniowe - wersja 1.0 - 24.XII.2024

Baza danych DrugBank to ogólnodostępna i bezpłatna baza informacji o lekach (substancjach leczniczych). Została utworzona w 2006 roku przez zespół Craiga Knoxa i Davida Wisharta z Wydziału Informatyki i Nauk Biologicznych Uniwersytetu Alberty w Kanadzie. Łączy dane z dziedziny chemii, biochemii, genetyki, farmakologii i farmakokinetyki.

Ponieważ dostęp do pełnej bazy danych wymaga utworzenia konta, wypełnienia formularza z podaniem uzasadnienia prośby o dostęp i uzyskanie akceptacji, na potrzeby niniejszego projektu zaliczeniowego udostępniony zostanie Państwu plik drugbank_partial.xml z okrojoną wersją bazy. Plik ten zawiera dane dla 100 leków (pełna baza opublikowana 2024-03-14 posiada informacje o ponad 16000 leków).

Projekt zaliczeniowy polega na przeanalizowaniu zawartości okrojonej wersji bazy i utworzeniu różnego rodzaju tabel i wykresów podsumowujących zawartość bazy leków.

1) Utworzyć ramkę danych, która dla każdego leku zawiera następujące informacje: unikalny identyfikator leku w bazie DrugBank, nazwę leku, jego typ, opis, postać w jakiej dany lek występuje, wskazania, mechanizm działania oraz informacje z jakimi pokarmami dany lek wchodzi w interakcje. (**4 pkt**)

2) Utworzyć ramkę danych pozwalającą na wyszukiwanie po DrugBank ID informacji o wszystkich synonimach pod jakimi dany lek występuje. Napisać funkcję, która dla podanego DrugBank ID utworzy i wyrysuje graf synonimów za pomocą biblioteki NetworkX. Należy zadbać o czytelność generowanego rysunku. (**4 pkt**)

3) Utworzyć ramkę danych o produktach farmaceutycznych zawierających dany lek (substancję leczniczą). Ramka powinna zawierać informacje o ID leku, nazwie produktu, producencie, kod w narodowym rejestrze USA (ang. *National Drug Code*), postać w jakiej produkt występuje, sposób aplikacji, informacje o dawce, kraju i agencji rejestrującej produkt. (**4 pkt**)

4) Utworzyć ramkę danych zawierającą informacje o wszystkich szlakach (sygnałowych, metabolicznych) z jakimi jakikolwiek lek wchodzi w interakcje. Podać całkowitą liczbę tych szlaków. (**4 pkt**)

5) Dla każdego szlaku sygnałowego/metabolicznego w bazie danych podać leki, które wchodzą z nim w interakcje. Wyniki należy przedstawić w postaci ramki danych jak i w opracowanej przez siebie formie graficznej. Przykładem takiej grafiki może być graf dwudzielny, gdzie dwa rodzaje wierzchołków to szlaki sygnałowe i leki, a poszczególne krawędzie reprezentują interakcję danego leku z danym szlakiem sygnałowym. Należy zadbać o czytelność i atrakcyjność prezentacji graficznej. (**4 pkt**)

6) Dla każdego leku w bazie danych podać liczbę szlaków, z którymi dany lek wchodzi w interakcje. Przedstawić wyniki w postaci histogramu z odpowiednio opisanymi osiami.
(**4 pkt**)

7) Utworzyć ramkę danych zawierającą informacje o białkach, z którymi poszczególne leki wchodzą w interakcje. Białka te to tzw. targety. Ramka danych powinna zawierać przynajmniej DrugBank ID targetu, informację o zewnętrznej bazie danych (ang. *source*, np. Swiss-Prot), identyfikator w zewnętrznej bazie danych, nazwę polipeptydu, nazwę genu kodującego polipeptyd, identyfikator genu GenAtlas ID, numer chromosomu, umiejscowienie w komórce. (**4 pkt**)

8) Utworzyć wykres kołowy prezentujący procentowe występowanie targetów w różnych częściach komórki. (**4 pkt**)

9) Utworzyć ramkę danych, pokazującą ile leków zostało zatwierdzonych, wycofanych, ile jest w fazie eksperymentalnej (ang. *experimental* lub *investigational*) i dopuszczonych w leczeniu zwierząt. Przedstawić te dane na wykresie kołowym. Podać liczbę zatwierdzonych leków, które nie zostały wycofane. (**4 pkt**)

10) Utworzyć ramkę danych opisującą w jaki sposób dany lek wchodzi w interakcje ze swoimi targetami. (**4 pkt**)

11) Opracować według własnego pomysłu graficzną prezentację zawierającą informacje o konkretnym genie lub genach, substancjach leczniczych, które z tym genem/genami wchodzą w interakcje, oraz produktach farmaceutycznych, które zawierają daną substancję leczniczą. Wybór dotyczący tego, czy prezentacja graficzna jest realizowana dla konkretnego genu, czy wszystkich genów jednocześnie pozostawiamy Państwa decyzji. Przy dokonywaniu wyboru należy kierować się czytelnością i atrakcyjnością prezentacji graficznej. (**7 pkt**)

12) Zaproponować własną analizę i prezentację danych dotyczących leków. Można w tym celu pozyskiwać dodatkowe informacje z innych biomedycznych i bioinformatycznych baz danych dostępnych online. Należy jednak upewnić się, czy dana baza danych pozwala na zautomatyzowane pobieranie danych przez program. Na przykład baza danych GeneCards wprost tego zabrania, co zostało na czerwono podkreślone na tej stronie. Przykładowe bazy danych to: UniProt (https://www.uniprot.org/), Small Molecule Pathway Database (https://smpdb.ca/), The Human Protein Atlas (https://www.proteinatlas.org/). (**7 pkt**)

13) Stworzyć symulator, który generuje testową bazę 20000 leków. Wartości generowanych 19900 leków w kolumnie “DrugBank Id” powinny mieć kolejne numery, a w pozostałych kolumnach wartości wylosowane spośród wartości istniejących 100 leków. Zapisz wyniki w pliku drugbank_partial_and_generated.xml. Przeprowadź analizę według punktów 1-12 testowej bazy. (**7 pkt**)

14) Przygotować testy jednostkowe z pomocą biblioteki pytest. (**7 pkt**)

15) Zrealizować punkt 6 tak, aby możliwe było wysłanie id leku na Twój serwer, który zwróci wynik w odpowiedzi (skorzystaj z fastapi i uvicorn; wystarczy zademonstrowanie przesłania danych metodą POST, przez Execute w dokumentacji) **(4 pkt)**