

Michał Kopyt, Eryk Mroczko, Adrian Prorok

TASS, projekt 2

Koncepcja wykonania

1. Temat projektu

Wyszukiwarka popularnych miejsc turystycznych: na podstawie danych o trasach lotów i danych demograficznych znaleźć miasta o nieproporcjonalnie dużym w stosunku do liczby mieszkańców ruchu lotniczym.

2. Wstępne wytypowanie źródeł danych oraz metodyka ich łączenia

Dane dotyczące tras lotów będą pochodziły z pliku CSV dostępnego pod adresem:

https://figshare.com/articles/dataset/flights_csv/9820139/1

Na podstawie tych danych możemy stwierdzić, że miał miejsce lot z lotniska X na lotnisko Y, ponieważ każdy rekord reprezentujący przelot zawiera informacje o kodach lotnisk (ORIGIN_AIRPORT, DESTINATION_AIRPORT). Mamy tutaj też dostępne między innymi informacje o tym, kiedy odbywały się loty.

Aby dowiedzieć się, jakich dokładnie miast dotyczyły loty, potrzebujemy danych zawierających przyporządkowania kodów lotnisk do konkretnych miast. Takie dane dostępne są w pliku CSV pod adresem: <https://www.kaggle.com/datasets/mike90/airport-codes>. Każdy rekord zawiera kolumnę IATA przechowującą ten sam rodzaj kodu, który znajduje się w kolumnach ORIGIN_AIRPORT, DESTINATION_AIRPORT, więc możemy połączyć rekordy na podstawie tych kodów.

Na tym etapie mamy już dane lotów połączone z danymi miast. Do zrealizowania projektu potrzebujemy jeszcze tylko danych demograficznych dla miast. Takie dane są dostępne w pliku CSV pod adresem <https://www.kaggle.com/datasets/viswanathanc/world-cities-datasets>. Dane połączymy na podstawie nazw miast w kolumnie CITY w dotychczas utworzonym zbiorze danych oraz kolumny CITY lub CITY_ASCII dostępnej w danych demograficznych dla miast.

3. Technologia pobrania i przygotowania danych

Udało się znaleźć wszystkie potrzebne dane w plikach, więc wystarczy zwyczajnie pobrać je z wymienionych wcześniej adresów.

Zawartość plików powinna być oczyszczona z nadmiarowych danych. Oprócz tego chcielibyśmy, żeby nasza wyszukiwarka nie musiała za każdym razem przetwarzać masy danych, dlatego zamierzamy zagregować nasze dane.

Dane w znalezionych plikach są statyczne, dlatego zostaną przetworzone z wykorzystaniem skryptu napisanego w języku Python. Wynikowy zbiór danych będzie zawierał dane pozwalające na stworzenie wyszukiwarki będącej tematem zadania. Uzyskany zbiór zostanie zapisany w jednym pliku po to, by móc go potem wygodnie wczytać.

4. Technologia przechowywania danych

Dane z przygotowanego pliku zamierzamy załadować do bazy danych PostgreSQL. Dzięki temu będziemy mogli w łatwy sposób dokonywać wyszukiwania interesujących miejsc turystycznych.

W bazie danych zamierzamy przechowywać pola w formie zagregowanej dla każdego miasta. Poniżej znajduje się lista przykładowych pól, które mogą być istotne dla naszej wyszukiwarki. Jest to tylko przedstawione w sposób poglądowy i zostanie doprecyzowane na dalszych etapach rozwoju projektu:

- Nazwa miasta
- Nazwa kraju
- Populacja miasta
- Liczba lotów do danego miasta
- Liczba lotnisk, z których można dolecieć do miasta
- Liczba krajów, z których można dolecieć do miasta
- Miesiąc z największą liczbą lotów do danego miasta

5. Aplikacja - wyszukiwarka

Utworzona aplikacja będzie serwerem napisanym w języku Java z wykorzystaniem frameworku Spring. W celu zwiększenia ergonomii rozwiązania planujemy udostępnić różnego rodzaju zapytania dotyczące popularnych miejsc turystycznych w formie REST API, do którego będzie można kierować odpowiednie żądania HTTP i tym samym uzyskiwać odpowiedzi wyszukiwarki w czytelnej, wygodnej do analizy formie. Dzięki temu z aplikacji będzie mógł korzystać zarówno człowiek (np. wysyłając żądania bezpośrednio z przeglądarki internetowej lub z aplikacji Postman), jak i inne aplikacje wysyłające żądania do udostępnionego API.

6. Scenariusz użycia aplikacji

1. Wysyłamy żądanie HTTP do aplikacji. Celem żądania jest uzyskanie listy popularnych miejsc turystycznych spełniającej podane kryteria wyszukiwania (np. filtrowanie, sortowanie).
2. W odpowiedzi otrzymujemy listę wyników spełniających kryteria w formacie JSON.

7. Raport końcowy

Po stworzeniu wyszukiwarki zamierzamy zbadać jej zachowanie, korzystając z narzędzia Postman oraz przeanalizować odpowiedzi na interesujące z naszego punktu widzenia zapytania po to, by móc wysnuć wnioski na temat wybranych przez nas zbiorów danych. Wyniki zamieścimy w sprawozdaniu końcowym.