Opóźnienia pociagów

Wiktor Piela

2022-10-28

Wstęp i cel badania

Na stronie internetowej https://infopasazer.intercity.pl/ udostępniane są informacje na temat aktywnych połączeń kolejowych w czasie rzeczywistym realizowanych zarówno przez regionalnych przewoźników jak i PKP Intercity. Pasażerowie mogą sprawdzić gdzie aktualnie znajduje się interesujący ich pociąg oraz ewentualne prognozowane opóźnienie jego przyjazdu do okreslonej stacji. Ponadto, poziom szczegółowości prezentowanych danych pozwala na sprawdzenie liczby przystanków, a także godzinę planowego przyjazdu i odjazdu na każdej z nich.

Zebranie odpowiedniej ilości takich danych pozwoliłoby pomóc w szukaniu odpowiedzi na pytanie - co wpływa na opóźnienia pociągów w Polsce? Dodatkowo - przy pomocy modelowania styatystycznego i zebranym danym można byłoby pokusić się o predykcję - czy i ile minut dana relacja będzie opóźniona - a zatem, jakie cechy połączenia kolejowego mogłby sugerować, że przyjedzie ono do stacji końcowej z okreslonym opóźnieniem.

Metodologia pozyskiwania danych i założenia badania

Niemożliwe byłoby pozyskanie danych o absolutnie wszystkich połączeniach kolejowych mających miejsce na teerenie Polski w danym okresie - zatem, dość intuicyjnym pomysłem byłoby zebranie odpowiedniej próby. Dla celów tego badania zostaną pozyskane dane o połączeniach przejeżdzających przez dworce główne 16 wybranych miast w Polsce - zarówno loklanych jak i ogólnopolskich.

Miasta zostały dobrane w taki sposób, aby były jednymi z większych miast w swoich regionach oraz były równomiernie rozłożone na terenie Polski. Dworce główne w większych miastach obsługują najwięcej pasażerów, zatem próba danych pochodząca z tym miast powinna być reprezentatywna, a dalsze wnioski wynikające z danych będa odnosiły się do największego odsetka pasażerów podróżujących polskimi kolejami.

Wspomniana wcześniej strona internetowa nie pozwala jednak na wybranie kilku stacji jednocześnie, dlatego w celu uzyskania aktywnych połączeń ze wszystkich stacji, program zbierający dane będzie musiał iteracyjnie otworzyć strony dla wszystkich 16 stacji oddzielnie - każda z nich ma swoje id, co ułatwia zadanie.

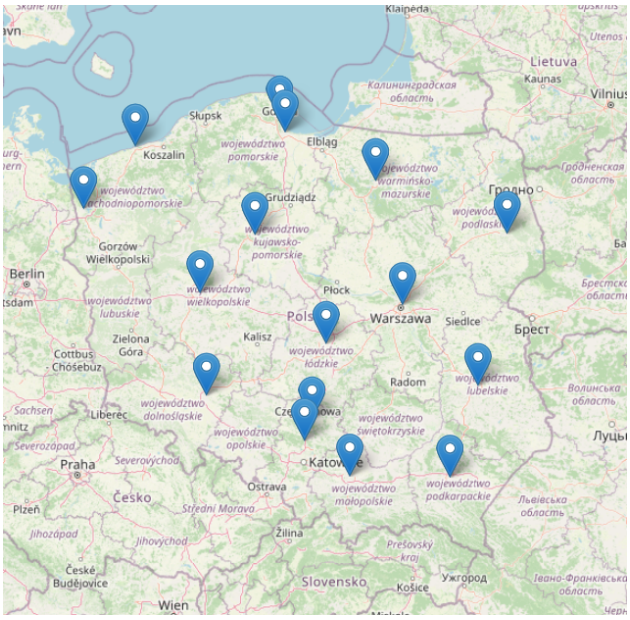

Figure 1: Lokalizacja miast dworców głównych wybranych do badania

Table 1: Miasta i odpowiadające im ID

miasto	id
Wrocław Główny	60103
Kraków Główny	80416
Warszawa Centralna	33605
Szczecin Główny	273
Rzeszów Główny	82669
Gdańsk Wrzeszcz	7534
Łódź Widzew	46409
Kołobrzeg	4101
Poznań Główny	30601
Katowice	73312
Białystok	24000
Częstochowa Stradom	62802
Lublin Główny	50500
Bydgoszcz Główna	18408
Olsztyn Główny	9209
Gdynia Główna	5900

Zasada działania algorytmu zbierania danych

Zbieranie danych takiej skali oraz ilości to nie jest zadanie manualne. Interesuje nas bowiem zebranie jak najwiekszej ilości danych w zadanym okresie oraz obszarze. Nawet ograniczając się do samej próby - codziennie w Polsce odbywają się setki kursów, w ciągu dnia,a także w nocy. Trudno zatem wyobrazić sobie, żeby ktoś całą dobę zajmował się tylko pozyskiwaniem tych danych i to przez pewien okres czasu. Ponadto, często zdarza się sytuacja, że kilka pociagów kończy bieg dokładnie o tej samej godzinie w różnych cześciach kraju, a strona przechowuje informacje tylko o aktywnych połączeniach, zatem znikają informacje o nich po godzinie przyjazdu. Ten fakt tym bardziej uniemożliwia ręczne scrapowanie danych.

W tym celu zaprojektowany został schemat pozyskiwania danych wraz z narzędziem zaimplementowanym w języku Python - https://github.com/wiktorpiela/train_delays/blob/main/01_data_scraping/scraper.py

Działa ono dwuetapowo:

1. Etap I

a. po pierwsze, funkcja `get_trains()` zasilana jest wcześniej zdefiniowanym słownikiem miasto-id. Iteracyjnie przechodzi po stronach z aktywnymi relacjami wszystkich miasts pozyskując odnośniki do stron z konkretnymi połączeniami
b. następnie, usunięte zostają duplikaty, tak, aby pociąg przejeżdzający przez więcej niż jedno miasto z listy pojawił sie tylko raz
c. uzyskanie informacji o przyjeździe do stacji końcowej każdej relacji, bo dokładnie wtedy ostatecznie scrapowane będą dane
d. konieczność korekty błędu właściciela strony - pociąg wyjeżdzający wieczorem danego dnia a przyjeżdżający do stacji końcowej po północy posiada datę wyjazdu (np. wyjazd 28.10.2022 godz. 19:50, przyjazd 28.10.2022 godz. 00:12)
e. wprowadzenie losowego szumu do dokładnej chwili przyjazdu, aby uniknąć kolizji w przypadku pozyskiwania danych dokładnie o tej samej godzinie co do sekundy
f. filtrowanie połączeń - wybór tylko przyszłych, ale nie dalszych niż 24 godziny - wtedy następi kolejna iteracja zadania
g. ostatecznie funkcja `get_trains()` zwraca listę tupli, każda tupla zawiera bezpośredni link do połączenia wraz z dokładną godziną przyjazdu

2. Etap II

a. funkcji `scrape_data()` dostarczany jest argument w postaci wygenerowanej listy tupli przez wcześniejszą funkcję - iteruje ona po tej liście, pozyskując dane o połaczeniu o dokładnej godzinie
b. takie dane zapisywane są lokalnie do pliku `.parquet` o nazwie godziny pobrania z dodatkowym suffixem, aby zminimalizować ryzyko tej pojawienia się tej samej nazwy dla więcej niż jednego pliku, a co za tym idzie, nadpisania pierwszego z nich
c. w ciele funkcji znajduje się także wątek dotyczący obsługi błedów, aby program nie przestał pracować wobec ewentualnego nie napotkania tabeli pod odnośnikiem do relacji