

POLITECHNIKA WARSZAWSKA
WYDZIAŁ MATEMATYKI I NAUK
INFORMACYJNYCH

METODY DATA SCIENCE

Koncepcja projektu

Autorzy:

Piotr IZERT
Przemysław RZĄD
Anna ZAWADZKA

24 marca 2016

1 Wstęp

Niniejszy projekt ma na celu wykonanie uczącego się systemu, który na podstawie pobieranych danych po odpowiednim ich przetworzeniu przygotowuje analizę w postaci prognozowania.

2 Dane

Źródłem danych są zbiory udostępnione przez ministerstwo transportu rządu Wielkiej Brytanii na stronach internetowych:

- <https://data.gov.uk/dataset/road-accidents-safety-data>
- <https://data.gov.uk/dataset/dft-eng-srn-routes-journey-times>

Dane obejmują informacje na temat wypadków drogowych na terenie Wielkiej Brytanii oraz średnich prędkościach przejazdu samochodów na danych odcinkach dróg w latach 2009-2014.

Dane pobierane są jako zamknięty zbiór rekordów, natomiast wykorzystane będą w taki sposób, że symulowany będzie ich stopniowy (tygodniowy lub miesięczny) napływ. Dzięki temu system będzie przystosowany do obsługi danych pozyskiwanych na bieżąco. Dane będą napływały z pięciu kanałów - informacje o wypadkach, o pojazdach, ich markach i modelach, o poszkodowanych oraz o średnich prędkościach przejazdu samochodów na danych odcinkach dróg. Zostaną one odpowiednio przefiltrowane zgodnie z potrzebami projektu, po czym będą skonsolidowane na podstawie unikalnego identyfikatora wypadku oraz numeru drogi. Do wstępnej fazy nauki, która poprzedza stopniowy napływ danych, system otrzyma do interpretacji pewną część danych (na przykład z całego 2009 roku).

3 Cel projektu

Głównym celem projektu jest zapewnienie jednostkom policji informacji na temat przewidywanych wypadków w każdym z rejonów. System na podstawie takich danych jak na przykład pogoda, dzień tygodnia, stan drogi, lokalizacja geograficzna itp. przewiduje ilość wypadków. Rezultatem działania systemu będą tygodniowe lub dzienne liczby wypadków z podziałem na pewne kategorie, takie jak na przykład region kraju, aktualna pogoda, rodzaj pojazdu itp. Dzięki temu służby będą mogły odpowiednio przygotować się do możliwych wydarzeń.

System ma również możliwość przewidywania stopnia obrażeń uczestników wypadków z wykorzystaniem danych o średnich prędkościach przejazdu samochodów na danych odcinkach dróg. Natychmiastowa prognoza jest użyteczna dla służb ratunkowych, ponieważ zakładamy, że dane o poszkodowanych napływają z kilkunastogodzinnym opóźnieniem.

Po każdej prognozie system będzie miał możliwość zweryfikowania jej poprawności na podstawie napływających nowych danych, informacja ta będzie podstawą do dalszej nauki.

4 Sposób wykonania

Do prezentacji danych wyjściowych planowane jest zastosowanie kostek OLAP na przykład w technologii SAS. Do przetwarzania danych wykorzystany zostanie między innymi Hadoop oraz Apache Spark.