Politechnika Warszawska
Wydział Matematyki i Nauk Informacyjnych

Metody Data Science

Koncepcja projektu

Autorzy:

Piotr Izert
Przemysław Rząd
Anna Zawadzka

19 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 jest zbiór udostepniony przez ministerstwo transportu rządu Wielkiej Brytanii na stronie internetowej
`https://data.gov.uk/dataset/road-accidents-safety-data`.
Dane obejmują informacje na temat wypadków drogowych na terenie Wielkiej Brytanii 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. Dzieki temu system będzie przystosowany do obsługi danych pozyskiwanych na bieżąco. Dane będą napływały z czterech kanałów - informacje o wypadkach, o pojazdach, ich markach i modelach oraz o poszkodowanych. Zostaną one odpowiednio przefiltrowane zgodnie z potrzebami projektu, po czym będą skonsolidowane na podstawie unikalnego indentyfikatora wypadku. 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 oraz ich typ (z udziałem samochodów, motorów, śmiertelne, czy nie itp.). Rezultatem działania systemu będą tygodniowe lub miesięczne liczby wypadków z podziałem na pewne kategorie, takie jak na przykład region kraju, aktualna pogoda, rodzaj pojazdu itp. 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, a także wykorzystanie Hadoop Spark.