Problem big data

iBigWorld:
Innovations for Big Data in a Real World

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Problem big data

Big data to zasadniczo zbiory danych, które są zbyt duże dla tradycyjnych systemów przetwarzania danych i dlatego wymagają nowych technologii przetwarzania.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Problem big data

Jedną z podstawowych zasad nauki o danych jest to, że dane i zdolność czerpania z nich przydatnej wiedzy należy postrzegać jako:

kluczowe aktywa strategiczne.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Problem big data

- Te platformy, które wykorzystują przetwarzanie rozproszone, wymagają złożonych projektów do budowy i utrzymania
- Z drugiej strony platformy big data wymagają również dodatkowych algorytmów, które zapewniają wsparcie dla odpowiednich zadań, takich jak wstępne przetwarzanie i analiza big data.

Co-funded by the Erasmus+ Programme of the European Union

Problem big data

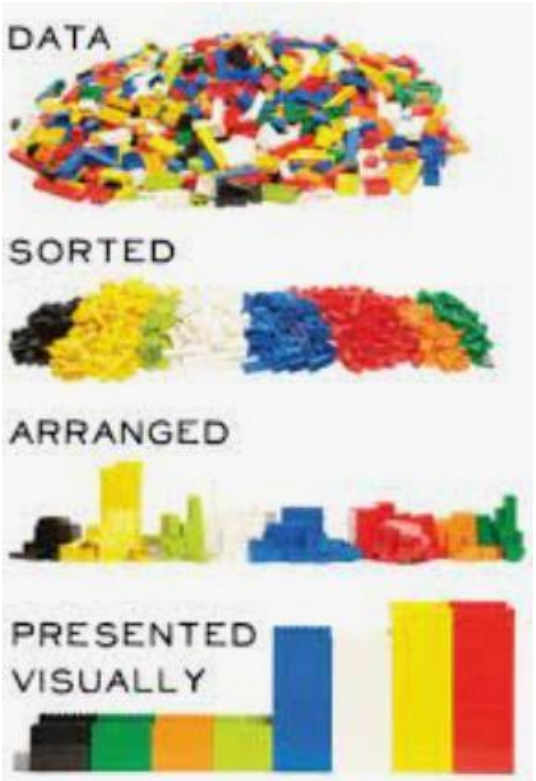
- Standardowe algorytmy dla tych zadań również muszą zostać przeprojektowane w celu badania zbiorów danych na dużą skalę. Nie jest to błaha sprawa i stanowi duże wyzwanie dla badaczy

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Problem big data: podstawowe koncepcje

- Pozyskiwanie przydatnej wiedzy z danych do rozwiązywania problemów biznesowych może być rozważane systematycznie, podążając za procesem z wystarczająco dobrze zdefiniowanymi krokami



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Problem big data: podstawowe koncepcje

- Jeśli przyjrzysz się zestawowi danych zbyt uważnie, znajdziesz coś, ale może to nie uogólnić danych, na które patrzysz.Nazywa się to przesyceniem zestawu danych.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Problem big data: podstawowe koncepcje

- Formułowanie rozwiązań do eksploracji danych i ocena wyników wymaga starannego rozważenia kontekstu, w jakim będą używane. Jeśli naszym celem jest uzyskanie potencjalnie przydatnej wiedzy, jak możemy wyrazić, co jest przydatne?

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Problem big data: podstawowe koncepcje

- Identyfikacja atrybutów informacyjnych — tych, które korelują lub dają nam informacje o nieznanej wartości
- Dopasowanie modelu funkcji numerycznej do danych poprzez wybór celu i znalezienie zestawu parametrów na podstawie tego celu

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Problem big data: podstawowe koncepcje

- Kontrola złożoności jest konieczna, aby znaleźć dobry kompromis między uogólnieniem a przesyceniem
- Obliczanie podobieństwa między obiektami opisanymi przez dane

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Jakość i przydatność danych

Niestety negatywne czynniki, takie jak szum, brakujące wartości, sprzeczne i nadmiarowe dane oraz ogromne wymiary w przykładach i funkcjach mają duży wpływ na dane wykorzystywane do uczenia się i uczenia się (patrz Wykład 9.2.)
Wiadomo, że niskiej jakości dane będą prowadzić do niskiej jakości wiedzy.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Bazy danych mogą zawierać

- Pola, które są przestarzałe lub zbędne,
- Brakujące wartości,
- Wartości odstające,
- Dane w postaci nieodpowiedniej dla modeli analizy danych,
- Wartości, które nie odpowiadają polityce ani zdrowemu rozsądkowi.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Problem big data

Głównym celem jest zminimalizowanie GIGO, zminimalizowanie śmieci, które trafiają do naszego modelu, abyśmy mogli zminimalizować ilość śmieci emitowanych przez nasze modele.
W zależności od zestawu danych, samo wstępne przetwarzanie danych może zająć 10-60% całkowitego czasu i wysiłku całego procesu eksploracji danych.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Frameworks

- MapReduce
- Apache Hadoop
- Apache Spark
- Apache Storm
- Apache Flink

Co-funded by the Erasmus+ Programme of the European Union

Dziękuję za uwagę!

Co-funded by the
Erasmus+ Programme
of the European Union