POZNAN UNIVERSITY OF TECHNOLOGY

Tematy projektów
Edycja 2019

Robert Wrembel
Poznan University of Technology
Institute of Computing Science
Robert.Wrembel@cs.put.poznan.pl
www.cs.put.poznan.pl/rwrembel

Temat 1

- **Implementacja modelu predykcji w Netezza**
- **Założenia:**
  - **implementacja w R**
  - **przykładowe modele predykcji**
    - **random forrest, SVM epsilon regression, SVM nu regression**
    - **LS-SVM, feedforward neural network, bayesian additive regression trees, bayesian regularized neural networks**
    - **dane dot. zużycia energii https://trynthink.github.io/buildingsdatasets/show.html?title_id=long-term-energy-environment-data-for-ornl-research-house-3**
- **L. osób: 3**

Temat 2 (Santander)

- **Hurtownia danych w chmurze**
- **Zadania**
  - **przygotowanie środowiska developerskiego w chmurze na potrzeby hurtowni danych**
  - **uruchomić Teradata lub Snowflake**
  - **przygotować bazę danych**
  - **przygotować i uruchomić mechanizm do przenoszenia schematu i danych do chmury**
  - **sprawdzenie wydajności rozwiązania**
  - **opracowanie mechanizmu automatyzacji i skryptów**
  - **przygotowanie prezentacji i dokumentacji**
- **Opiekun: Marcin Kurek (Santander)**
- **Uwagi: projekt realizowany w środowisku Santander**
- **L. osób: 3**

Temat 3 (Santander)

- **System Snowflake**
- **Zadania**
  - **przygotowanie środowiska lokalnego lub chmurowego Snowflake**
  - **przygotowanie struktur bd**
  - **załadowanie danych**
  - **wykonanie testów wydajnościwych**
  - **rozpoznanie możliwości bazy**
  - **analiza języka i składni SQL**
  - **analiza connectorów**
  - **analiza porównawcza z Teradata (preferowane) lub Oracle**
  - **przygotowanie prezentacji i dokumentacji**
- **Opiekun: tbd (Santander)**
- **Uwagi: projekt realizowany w środowisku Santander**
- **L. osób: 3**

Temat 4 (Santander)

- **Profilowanie danych w AbInitio**
- **Zadania**
  - **zapoznanie się z Ab Initio Data Profiler**
  - **zapoznanie się i przygotowanie algorytmów do profilowania danych**
  - **przygotowanie paczek danych testowych**
  - **profilowanie danych i przetestowanie algorytmów**
  - **przygotowanie prezentacji i dokumentacji**
- **Opiekun: Hubert Półtorak (Santander)**
- **Uwagi: projekt realizowany w środowisku Santander**
- **L. osób: 3**

Temat 5 (Roche)

- **Profilowanie danych w formacie JSON**
- **Zadania**
  - **przygotowanie narzędzia bądź (preferowane) rozwinięciu biblioteki pandas-profiling dla plików w formacie JSON**
    - **parsowanie JSON**
    - **wyznaczanie charakterystyk (ich lista do uzgodnienia)**
    - **wizualizacja – minimum to obecny kształt raportu, idealnie rozwiązanie takie jak proponuje Attacama bądź parsery oparte o xml-a (do ustalenia)**
- **Opiekun: Fabian Wiktorowski (Roche)**
- **Wymagania:**
  - **laptop z zainstalowanym środowiskiem Python (anaconda)**
  - **nie jest wymagany dostęp do środowiska Roche**
- **L. osób: 4**

Temat 6 (Roche)

- **Konwersja pliku JSON do postaci tabelarycznej z uwzględnieniem tablic Jsonowych**
- **Zadania**
  - **przygotowanie narzędzia/frameworku, które mając zadany plik JSON przygotuje wstępnie zestaw tabel płaskich (relacyjnych) które można zasilić źródłowym JSONem**
    - **szczególne istotne jest rozparsowanie tabel JSON (Array)**
- **Opiekun: Krzysztof Walkiewicz**
  - **laptop z zainstalowanym środowiskiem Python (anaconda)**
  - **nie jest wymagany dostęp do środowiska Roche**
- **L. osób: 2**

Temat 7

- **Ocena narzędzi open-source do profilowania danych**
- **Zadania**
  - **przegląd narzędzi**
  - **wybór narzędzi do testów**
  - **opracowanie testów**
  - **wykonanie testów na danych rzeczywistych**
    - **zanieczyszczenie powietrza w Lombardii (1968-2019)**
    - **https://dati.lombardia.it/stories/s/auv9-c2sj**
  - **ocena narzędzi**
- **Uwagi**
  - **należy przetestować top-4 narzędzia**
- **L. osób: 4**

Temat 8

- **Profilowanie danych w Netezza**
- **Zadania**
  - **zainstalowanie emulatora Netezza**
  - **przygotowanie bazy danych**
  - **opracowanie zapytań profilujących**
  - **wykonanie profilowania na danych rzeczywistych**
    - **zanieczyszczenie powietrza w Lombardii (1968-2019)**
    - **https://dati.lombardia.it/stories/s/auv9-c2sj**
    - **w szczególności:**
      - min-max czas odczytu danych, z podziałem na lata
      - min-max czas odczytu danych, z podziałem na mierniki
      - wykres czasu pracy poszczególnych mierników w latach
- **Uwagi: implementacja w SQL i j. proceduralnym**
- **L. osób: 2**

Temat 9 (PKO BP)

- **Ocena technik i narzędzi de-duplikacji danych**
- **Zadania**
  - **przegląd technik i narzędzi**
  - **wybór narzędzi open-source do testów**
  - **opracowanie testów**
  - **przygotowanie zbioru danych testowych**
  - **wykonanie testów**
  - **ocena narzędzi**
- **Uwagi**
  - **należy przetestować top-6 narzędzi**
- **L. osób: 6**

Temat 10 (Kogeneracja Zachód)

- **Ocena algorytmów predykcji zużycia energii**
- **Zadania**
  - **analiza istniejących rozwiązań**
  - **wybór 2 algorytmów do testowania**
  - **wykonanie testów**
    - **jakość predykcji**
    - **wydajność**
  - **opracowanie wyników**
- **Uwagi:**
  - **impementacja: Python, R, .Net**
- **L.osób: 2**