

Big Data [223090-0421]

dr Mariusz Rafało

<http://mariuszrafalo.pl>

mrafalo@sgh.waw.pl

Zasady zaliczenia przedmiotu

Zasady ogólne

Zaliczenie opiera się na następujących elementach:

- Test - można otrzymać 40 punktów, trzeba zdobyć co najmniej 24 punkty
- Indywidualny projekt zaliczeniowy - można otrzymać 40 punktów, trzeba zdobyć co najmniej 24 punkty
- Prace domowe/na zajęciach - można otrzymać 20 punktów
- Praca podczas zajęć: każda nieobecność skutkuje obniżeniem całkowitej liczby punktów o 10

Ocena z przedmiotu jest wyznaczana zgodnie z tabelą:

Punkty od	Punkty do	Ocena
0	59	niedostateczny
60	67	dostateczny
68	75	dostateczny plus
76	84	dobry
85	92	dobry plus
93	100	bardzo dobry

Test zaliczeniowy

Test wyboru, oparty na pytaniach zamkniętych; 4 możliwe warianty odpowiedzi, jedna odpowiedź prawidłowa.

Projekt zaliczeniowy

Informacje ogólne

Należy przygotować projekt w technologii Databricks. Dokument w formie raportu eksportujemy do formatu HTML. Celem raportu jest dokonanie przetwarzania i ogólnej analizy danych za pomocą technik poznanych na zajęciach (pySpark). Na wstępie określamy cel analizy a następnie dokumentujemy i wizualizujemy poszczególne kroki. W pracy koncentrujemy się na przetwarzaniu danych, mniej na analizie i wizualizacji. Struktura raportu:

- Tematyka raportu oraz cel przetwarzania danych
- Źródło danych (opis, informacje o źródle)
- Podstawowe transformacje danych (agregowanie, wyznaczenie wskaźników i kalkulacji)
- Analiza eksploracyjna danych (wykresy i tabele)
- Podsumowanie i wnioski

Raport przygotowany w ten sposób przesyłamy mailem.

Dane

- Dane pobieramy z ogólnodostępnego repozytorium zbiorów danych, przykładowo Kaggle: <https://www.kaggle.com/datasets>
- Najlepiej gdyby dane dotyczyły działalności biznesowej, życia społecznego lub podobnych
- W oparciu o dane definiujemy problem, który chcemy zbadać, np.: analiza bankowych transakcji marketingu bezpośredniego, analiza uwarunkowań zarobków pracowników w różnych krajach, itp.
- Pracujemy na platformie *Databricks* (<https://community.cloud.databricks.com>). Należy założyć sobie konto na tej platformie.

- Pracujemy w języku *Python (pySpark)* i *SQL*; pracę dokumentujemy w *Markdown*
- Korzystamy z technologii poznanych na zajęciach: *Databricks, Hive, Spark*

Transformacja danych

Zakres techniczny projektu powinien obejmować minimum elementy:

- Załadowanie danych do *Data Frame* (format danych dowolny: CSV, JSON, XML, itp.)
- Zapisanie danych w bazie danych Hive
- Wykonanie agregacji i transformacji danych w SQL
- Wykonanie agregacji i transformacji danych w Python (biblioteka pySpark)
- Przedstawienie wyników na kilku wykresach opartych na Databricks