

Program studiów podyplomowych

BIG DATA – Analytisc & Society

rok akademicki 2024/2025

Nazwa przedmiotów	Liczba godzin
Wprowadzenie do technologii Big Data	6
Podstawy programowania (Python, R)	40
Metody przetwarzania i analizy danych	20
Algorytmy i struktury danych	20
Bazy danych	30
Machine Learning	18
Analiza sieci społecznych	15
Wizualizacja danych	18
Internetowe narzędzia analityczne	15
Big Data w chmurze obliczeniowej	6
Sztuczna Inteligencja w Big Data	6
Psychologia Big Data	6
Prawne i etyczne aspekty gromadzenia i przetwarzania danych	6
Seminarium	9
Suma godzin	215

1. Wprowadzenie do Big Data i Data Science (6h)

- Czym jest big data – filozofia, technologia, analityka,
- Model 42V
- Podejścia do problemów skali Big Data, algorytm MapReduce, koncepcja BigTable, NoSQL oraz rozproszone systemy plikowe HDFS
- Projektowanie rozwiązań klastrowych i chmurowych oraz integracja systemów Big Data z istniejącymi systemami i oprogramowaniem.
- Big data: definicje i taksonomia.
- Historia procesów informacyjnych, na których bazuje Big Data oraz sposoby gromadzenia, weryfikacji i przetwarzania dużych ilości danych.
- Wartość big data dla biznesu i życia społecznego.
- Big Words – najważniejsze pojęcia Big data (Data science, Data mining, Machine learning, Map Reduce, Hadoop).
- Narzędzia do analizy Big data oraz wizualizacja danych.
- Case – architektura systemu wspomagania decyzji wykorzystująca technologie przetwarzania danych Big data.
- Nowoczesna analityka w przedsiębiorstwie cyfrowym.

-
- Zastosowania Big data w życiu codziennym (case'y: Jak wygrać wybory w USA?; Weisbeerger – zamiana drinka w Big data.; Global Forest Watch.; In Google we trust.).

2. Podstawy programowania 40h (20h Python/ 20h R)

Język programowania R

- Wprowadzenie do R (Praca w R , Przestrzeń robocza w RStudio, Pakiety)
- Struktury danych (Wektor, Macierz, Faktor/czynnik, Lista,)Ramka danych
- Przetwarzanie danych (Pakiet tidyverse, Import danych, Filtrowanie, Wybieranie kolumn, Tworzenie nowej zmiennej, Zmiana nazwy zmiennej, Podsumowanie danych, Grupowanie, Sortowanie, Łączenie zbiorów, Wąska i szeroka reprezentacja danych, Eksport danych)
- Wizualizacja danych (Korzystanie z wbudowanych funkcji, Pakiet ggplot2, Wykres punktowy, Wykres pudełkowy, Wykres liniowy, Wykres kolumnowy)
- Programowanie w R (Tworzenie funkcji, Pętle, Instrukcje warunkowe)

Język programowania Python

- Wprowadzenie do języka Python
- Składnia języka Python
- Wyrażenia regularne w języku Python
- Klasy, obiekty, moduły
- Skrypty i kontrola przepływu programu
- Zastosowanie bibliotek naukowych: NumPy, SciPy, Matplotlib
- Rozwiązywanie praktycznych problemów Big Data z wykorzystaniem języka Python

3. Metody przetwarzania i analizy danych (20h)

- Eksploracja danych – wstęp, techniki eksploracji danych (zmienne jakościowe i numeryczne, relacje wielowymiarowe, dyskretyzacja, regresja wielokrotna, algorytm najbliższych sąsiadów, funkcje decyzyjne, drzewa klasyfikacyjne i regresyjne)
- Metody przetwarzania i analizy danych (regresja logistyczna, regresja liniowa, uogólnione modele liniowe (GLM), modele mieszane (Mixed models), algorytm locality-sensitive hashing)
- Przetwarzanie Big Data za pomocą Apache Hadoop, implementacja algorytmów w paradygmacie MapReduce,
- wykorzystanie narzędzi analitycznych Pig i Hive
- Przetwarzanie Big Data z użyciem Apache Spark
- Praca z platformą analityczną SAS LASR Server

4. Algorytmy i struktury danych (20h)

- Podstawowe definicje, schematy zapisów algorytmów

-
- Podstawowe struktury danych: zmienne, wskaźniki i referencje, rekordy, tablice i macierze. Struktury listowe, grafowe, słownikowe i programistyczne
 - Algorytmy iteracyjne i rekurencyjne. Sortowanie: bąbelkowe, przez selekcję i wstawianie, quicksort. Wyszukiwanie liniowe i binarne. Grafowe: przeszukiwanie w głąb i w szerz. Cykle: Eulera i Hamiltona, przeszukiwanie drzew.

5. Bazy danych (30h)

- Relacyjny model baz danych
- Wprowadzenie do języka strukturalnego SQL: składnia SQL : DML, DDL, DCL, DQL
- Praktyczne zastosowanie SQL – zapytania analityczne
- Metody efektywnego tworzenia modeli danych i ich pozyskiwania z wykorzystaniem SQL
- Czym jest baza danych i hurtownia danych. Rys historyczny, silniki baz danych; konceptualne modelowanie danych z wykorzystaniem diagramów ER, zbiory encji, atrybuty, związki, klucze, krotność i obligatoryjność
- Modelowanie danych, relacyjny model danych, podstawowe operacje na bazie danych z wykorzystaniem języka SQL: dodawanie, usuwanie, modyfikowanie, wyświetlanie danych, określenie schematu bazy danych, rozszerzenia relacyjnych baz danych, elementy baz danych NoSQL
- Definicja hurtowni danych wg. B. Inmona, R. Kimballa, analiza wymiarów hurtowni danych, zasady budowania hurtowni danych, model koncepcyjny, model logiczny oraz implementacja fizyczna hurtowni danych, architektura hurtowni danych, kopie zapasowe hurtowni, archiwizacja, ekstrakcja i ładowanie danych, Aplikacje CRT – analiza przykładowego oprogramowania oraz jego obsługa, przygotowanie danych dla hurtowni danych, proces ETL., wymiary w hurtowni danych, analiza danych (OLAP cube, analiza w bazach NoSQL)
- Konfiguracja środowiska do analizy Big Data w technologii Open Source (Apache Cassandra, Apache HBase, Apache Spark, Apache Mahout, Apache MLib, Apache Hadoop)

6. Machine Learning (18h)

- Charakterystyka sztucznej inteligencji
- Metody i zastosowania uczenia maszynowego
- Sieć neuronowa jako klasyczny przykład SI w zakresie ML
- Reprezentacja wiedzy i jej konsekwencje
- Błędy ludzi, które powielają lub których unikają maszyny
- Dobór metody z zakresu SI i ML do klasy zadania
- Optymalizacja i walidacja rozwiązań

7. Analiza sieci społecznych (15h)

- SNA jako metoda i narzędzie badawcze
- Zastosowanie w biznesie (obserwowanie zmian społeczno-gospodarczych, analiza aktywności użytkowników internetu)

-
- Oprogramowanie badawcze i jego praktyczne zastosowanie
 - Definicja sieci społecznych, aktora sieci społecznej, połączenia i zbioru relacji w sieciach społecznych, zastosowanie w biznesie (obserwowanie zmian społeczno-gospodarczych, analiza aktywności użytkowników internetu)
 - Analiza sieci społecznej jako metoda badawcza, sieci społeczne, a wirtualne sieci społeczne – różnice i podobieństwa, znaczenie analizy sieci społecznych (SNA), poziomy analizy sieci społecznych (makro, mezo i mikro)
 - Podstawy analizy sieci społecznych (matematyczna teoria grafów, algebra macierzowa, statystyka, miary w analizie sieci społecznych: centralność, prestiż, pośrednictwo, bliskość, gęstość, kapitał społeczny, zastosowanie analizy sieci społecznych – case study: wyszukiwanie kluczowych użytkowników, marketing skierowany do grupy docelowej, systemy rekomendacyjne, tworzenie grup projektowych w sieciach społecznych, propagacja nowych idei, kolaboratywne zarządzanie wiedza.

8. Wizualizacja danych (18h)

Wizualizacja danych - teoria

- Co to jest wizualizacja danych i dlaczego jest potrzebna,
- Czym jest Business Intelligence,
- Narzędzia wizualizacji danych - QlikView, Tableau itp,
- Tableau - model biznesowy (podział na desktop, serwer, public, online), dostępność szkoleń oraz filmików szkoleniowych online,
- Ogólne zasady i dobre praktyki wizualizacji danych,
 - Metody wizualizacji danych: wykresy, infografiki, dashboardy i story,
 - Typy wykresów używanych do wizualizacji danych - kiedy jakiego używać, kiedy i jakich nie powinno się używać.

Przygotowanie danych do wizualizacji

- Rodzaje zmiennych,
- Cel porządkowania danych,
- Cechy uporządkowanych danych,
- Praktyczne przykłady porządkowania zbiorów danych,
- Porządkowanie a oczyszczanie danych,
- Wymagania Tableau dotyczące struktury danych,
- Alteryx – wstęp, dostępność szkoleń oraz filmików szkoleniowych online.

Alteryx - prezentacja aplikacji oraz jej możliwości

- Metody podłączania danych oraz rodzaje źródeł danych,
- Podstawowe narzędzia (Tools),
- Konstruowanie workflow'u,

-
- Zapisywanie tabel wynikowych do plików,
 - Ćwiczenia praktyczne w Alteryx'ie.

Tableau - prezentacja aplikacji wraz z praktycznym tworzeniem:

- wykresów - 13 typów wykresów, kiedy używać, jak budować itp,
- dashboardów,
- story.

9. Internetowe narzędzia analityczne (15h)

- Narzędzia służące do: monitoringu mediów, zarządzania social media, influencer marketingu, badania zaangażowania użytkowników, content marketingu itp.
- Analityka internetowa: wprowadzenie, konfiguracja narzędzi analitycznych, budowa kompetencji w zespole, tworzenie ciągłych procesów analizy danych
 - Przegląd wybranych narzędzi służących do: monitoringu mediów, zarządzania social media, influencer marketingu, badania zaangażowania użytkowników, content marketingu.
 - Analityka internetowa w działaniu: Google Analytics; Moz Analytics; TrueSocial Metrics; Woopra; BrightEdge; CrazyEgg; Heatmap; Hootsuite, Hotjar
 - Yandex Metrica; Adobe Analytics; Webtrends; IBM Digital Analytics; Piwik/Matomo

10. Big Data w chmurze obliczeniowej (6h)

- Wprowadzenie do technologii chmur danych
- Architektura systemów Big Data w chmurze
- Modele przetwarzania danych w chmurze
- Rozwiązania chmurowe (np. AWS, Azure, Google Cloud) i ich rola w obszarze Big Data
- Techniki przetwarzania strumieniowego
- Bezpieczeństwo i zarządzanie danymi w chmurze
- Trendy i przyszłość Big Data w chmurze

11. Sztuczna Inteligencja w analizie Big Data (6h)

- Wprowadzenie do technologii AI
- Algorytmy uczenia maszynowego w analizie Big Data
- Deep Learning w analizie Big Data
- Zastosowanie AI w analizie biznesowej
- Zastosowanie AI w przemyśle
- AI w badaniach społecznych i marketingowych
- Sztuczna inteligencja w badaniach naukowych
- Dylematy poznawcze i etyczne AI
- Perspektywy rozwoju sztucznej inteligencji

12. Psychologia Big Data (6h)

- Psychologiczny model OCEAN i jego wykorzystanie w tworzeniu psychograficznych profili konsumentów/wyborców w powiązaniu z danymi z FB. Przewidywanie cech i atrybutów na podstawie cyfrowej rejestracji ludzkich zachowań. Rozwój segmentacji psychograficznej
- Wybrane badania wykorzystujące Big Data w analizie ludzkich zachowań i ich konfrontacja z dotychczasowymi ustaleniami psychologii
- Laboratorium psychologiczne oparte na Big Data: The Psychometric Centre, University of Cambridge
- Redukcjonistyczne ujęcie (psychologii) człowieka jako skończonego zbioru danych behawioralnych i neurodanych – możliwości i konsekwencje
- Wybrane elementy cyberpsychologii
- Interakcja człowiek-media-technologia w ujęciu psychologicznym i neurokognitywnym. Multitasking i efektywność w zakresie przetwarzania i zapamiętywania informacji. Konsekwencje zacierania granic pomiędzy światem wirtualnym i realnym, psychologiczne i społeczne funkcjonowanie w rzeczywistości hybrydowej. Hipoteza cyfrowej demencji. Wpływ cyfrowych reprezentacji ludzi i środowiska na postawy i zachowania społeczne. Media społecznościowe: następstwa psychologiczne w kontekście najnowszych badań.
- Poznawcze, emocjonalne i społeczne zagrożenia wynikające z nadużywania mediów cyfrowych; wybrane zaburzenia: m.in. FOMO (“fear of missing out”), DA (Disconnectivity Anxiety), nomofobia (“no mobile phobia”)
- Are you a thinker, a feeler, a sensor or intuitor? – jak nasz dominujący styl komunikacji determinuje przekaz i odbiór informacji
- Prezentacja danych i analiz – case study i best practice – perspektywa managera.
- Procesy poznawcze – jak nasz mózg radzi sobie z procesami poznawczymi i informacjami.
- Media, technologie i duże zbiory danych a procesy poznawcze, emocjonalne

13. Prawne i etyczne aspekty gromadzenia i przetwarzania danych (6h)

- Big Data a prawo do prywatności
- Ustawa o ochronie danych osobowych versus praktyka przetwarzania danych
- tajemnice sektorowe
- Implikacje etyczne gromadzenia i przetwarzania informacji

14. Seminarium (9h)

Czas trwania: 2 semestry, 12 zjazdów, 215 godziny