Załącznik Nr 5 do Zarz. Nr 33/11/12

(pieczęć wydziału) **KARTA PRZEDMIOTU**

|  |  |
| --- | --- |
| **1. . Nazwa przedmiotu:** Data science - przetwarzanie i analiza dużych zbiorów danych | **2. Kod przedmiotu:**  |
| **3. Karta przedmiotu ważna od roku akademickiego:** 2018/2019 |
| **4. Forma kształcenia:** studia trzeciego stopnia |
| **5. Forma studiów**: studia stacjonarne / niestacjonarne |
| **6. Kierunek studiów**: Interdyscyplinarne studia doktoranckie Symulacje w Inżynierii |
| **7. Profil studiów:** akademicki |
| **8. Dyscyplina: biocybernetyka i inżynieria biomedyczna** |
| **9. Semestr:** przedmiot obieralny |
| **10. Jednostka prowadząca przedmiot: WAEI** |
| **11. Prowadzący przedmiot**: dr hab. Marek Sikora |
| **12. Przynależność do grupy przedmiotów:** moduł podstawowy |
| **13. Status przedmiotu: przedmiot wspólny** |
| **14. Język prowadzenia zajęć:** polski/angielski |
| **15. Przedmioty wprowadzające oraz wymagania wstępne:**  |
| **16. Cel przedmiotu:** Celem kształcenia jest zapoznanie studentów z realizacją podstawowych algorytmów analizy i eksploracji danych w środowisku przeznaczonym do przetwarzania dużych zbiorów danych (Big Data). Przedmiot prezentuje zasady działania ekosystemu Apache Hadoop oraz oprogramowania Apache Spark, w kontekście realizacji zadań analitycznych uruchamianych dla dużych zbiorów danych. Podstawowym językiem programowania w jakim realizowane są obliczenia jest R. Zaprezentowane zostaną przykłady realizacji zadań analitycznych za pomocą metod symbolicznych (drzew decyzji), niesymbolicznych (regresja logistyczna) oraz zespołowych (boosting). |
| **17. Efekty kształcenia:[[1]](#footnote-1)** |
| Nr | Opis efektu kształcenia | Metoda sprawdzenia efektu kształcenia | Forma prowadzenia zajęć | Odniesienie do efektów dla kierunku studiów |
| 1 | Zna problematykę analizy danych, w szczególności metodykę CRIPS DM oaz problemy związane z analiza danych zbiorów danych | Podczas wykładu,konsultacje | W | SYMIN\_W02SYMIN\_W03 |
| 2 | Zna architekturę rozwiązania Apache Hadoop. Zna zalety, wady i wymagania sprzętowe związane z wykorzystaniem poszczególnych składników eksposystemu Hadoop  | Podczas wykładu,konsultacje | W | SYMIN\_U10SYMIN\_U10 |
| 3 | Wie w jaki sposób uruchomić przetwarzanie danych (w szczególności zadania analityczne) w oparciu o środowisko Apache Hadoop | Podczas wykładu,konsultacje | W | SYMIN\_W02SYMIN\_W05 |
| 4 | Zna zasady działania paradygmatu MapReduce. Wie czym różnią się standardowe realizacje wybranych algorytmów analizy danych od ich realizacji w paradygmacie MapReduce | Podczas wykładu,konsultacje | W | SYMIN\_W03SYMIN\_U10 |
| 5 | Zna co najmniej trzy metody analizy danych zawarte w środowisku Apache Spark. Potrafi poprawnie je stosować i konfigurować (indukcja drzew. SVN liniowy, boosting). | Podczas wykładu,konsultacje | W | SYMIN\_W07SYMIN\_U12 |
| **18. Formy zajęć dydaktycznych i ich wymiar (liczba godzin)** **W. 10 Ćw. - L. - P. - Sem. -** |
| **19** **Treści kształcenia:**Wykład prowadzony będzie metodą tradycyjną. Podczas wykładu podawane są schematy blokowe algorytmów konfiguracji środowisk oraz algorytmów analizy danych. Prezentowane są zasady działania platformy Hadoop i środowiska Spark (w szczególności prezentowane są krypty konfiguracyjne – które udostępnione zostaną studentom). Przedstawione zostaną przypadki użycia zaprezentowanych technologii i metod do analizy danych**Wykład**:1. Wprowadzenie do tematyki analizy i eksploracji danych. Podstawowe pojęcia związane z teriologią Big Data. Trudności wynikające z analizy dużych zbiorów danych. Apache Hadoop i Apache Spark na tle konkurencji.
2. Analiza danych metody dla big data
3. Hadoop 1 – Architektura i narzędzia środowiska ekposystemu (HDFS, Pig, Hive, MapReduce, Yarn, HBase, Spark, Solr, Ambari, Zookeeper, Oozie, Kafka, Falcon, Knox)
4. Hadoop, 2 – Konfiguracja i użycie do realizacji zadań analitycznych wybranych składników ekosystemu.
5. Przypadki użycia (klasyfikacja, regresja – drzewa, SVN, boosting)

W materiałach przygotowanych w ramach wykładu znajdzie się również opis realizacji zadań analizy dużych zbiorów danych za pomocą konkurencyjnego środowiska RapidMier i serwera Radoop Serwer Radoop to ekosystemu obliczeń rozproszonych dla środowiska analitycznego RapidMiner zrealizowany również w oparciu o Apache Hadoop – warstwa pośrednicząca pomiędzy Hadoop, a RapidMiner..  |
| **20. Egzamin:** brak |

|  |
| --- |
| **21. Literatura podstawowa:**1. Jurney R.: Zwinna analiza danych. Apache Hadoop dla każdego. Wydawnictwo Helion 2015.
2. Cichosz P.: Data mining algorithms: Explained usign R. John Wiley &Sons 2014.
3. Morzy T. Eksploracja danych. Wydawnictwo Naukowe PWN, Warszawa 2013.
 |
| **22. Literatura uzupelniająca:**1. Witten I.H., Frank E.: Data mining: practical machine learning tools and techniques. Morgan Kaufmann, 2011.2 . M. North. Data Mining for the masses. A Global Text Project Book, 2012. |
| **23. Nakład pracy studenta potrzebny do osiągnięcia efektów kształcenia**

|  |  |  |
| --- | --- | --- |
| Lp. | Forma zajęć  | Liczba godzin kontaktowych / pracy studenta |
| 1 | Wykład | 10/10 |
| 2 | Ćwiczenia | / |
| 3 | Laboratorium | / |
| 4 | Projekt | / |
| 5 | Seminarium | / |
| 6 | Inne (przygotowanie do zajęć) | 0 /15 |
|  | Suma godzin | 10 / 25 |

 |
| **24. Suma wszystkich godzin:** 10 |
| **25. Liczba punktów ECTS: 1** |
| **26. Liczba punktów ECTS uzyskanych na zajęciach z bezpośrednim udziałem nauczyciela akademickiego: 1** |
| **27. Liczba punktów ECTS uzyskanych na zajęciach o charakterze praktycznym (laboratoria, projekty):** |
| **26. Uwagi:** |

 Zatwierdzono:

 …………………………………………………

 (*data i podpis kierownika studiów doktoranckich)*

1. należy wskazać ok. 4 – 5 efektów kształcenia [↑](#footnote-ref-1)