



3. KARTA PRZEDMIOTU

| | |
|--------------------------------------|---|
| Kod przedmiotu | Z-ID-U-609a |
| Nazwa przedmiotu | Analiza danych niestrukturalnych |
| Nazwa przedmiotu w języku angielskim | Unstructured Data Analysis |
| Obowiązuje od roku akademickiego | 2019/2020 |

USYTUOWANIE MODUŁU W SYSTEMIE STUDIÓW

| | |
|----------------------------------|---|
| Kierunek studiów | INŻYNIERIA DANYCH |
| Poziom kształcenia | I stopień |
| Profil studiów | Praktyczny |
| Forma i tryb prowadzenia studiów | Studia stacjonarne |
| Zakres | Analityka danych i modelowanie |
| Jednostka prowadząca przedmiot | Katedra Informatyki i Matematyki Stosowanej |
| Koordinator przedmiotu | Dr hab. Marzena Nowakowska Mgr inż. Karolina Bęben |
| Zatwierdził | Dr hab. inż. Artur Bartosik, prof. PŚk |

OGÓLNA CHARAKTERYSTYKA PRZEDMIOTU

| | |
|---|---|
| Przynależność do grupy/bloku przedmiotów | Przedmiot specjalnościowy |
| Status przedmiotu | Obowiązkowy |
| Język prowadzenia zajęć | Polski |
| Usytuowanie modułu w planie studiów - semestr | Semestr VI |
| Wymagania wstępne | Bazy danych, Wstępna eksploracja i przygotowanie danych do analiz, Odkrywanie związków w danych wielowymiarowych |
| Egzamin (TAK/NIE) | NIE |
| Liczba punktów ECTS | 3 |

| Forma prowadzenia zajęć | wykład | ćwiczenia | laboratorium | projekt | inne |
|---------------------------|-----------|-----------|--------------|---------|------|
| Liczba godzin w semestrze | 10 | | 30 | | |

EFEKTY UCZENIA SIĘ

| Kategoria | Symbol efektu | Efekty kształcenia | Odniesienie do efektów kierunkowych |
|-----------------------|---------------|--|-------------------------------------|
| Wiedza | W01 | Student zna podstawowe metody <i>Text Mining</i> do strukturyzacji informacji tekstowych i automatycznej analizy dokumentów tekstowych oraz zna ich zastosowanie. | ID1_W05 |
| | W02 | Student zna narzędzie SAS Text Miner oraz wybrane darmowe oprogramowanie do eksploracji danych. | ID1_W13 |
| | W03 | Student ma wiedzę na temat poszczególnych metod statystycznych przydatnych w analizie danych niestrukturnalnych oraz zna ich przykładowe zastosowania prowadzące do znalezienia zależności pomiędzy tymi danymi. | ID1_W02 ID1_W09 |
| | W04 | Student rozumie konieczność przekształcania dokumentów tekstowych oraz zna różne rodzaje reprezentacji tekstu. | ID1_W13 |
| Umiejętności | U01 | Student potrafi wykonać proste zadania konwersji dokumentów tekstowych do wektorów cech. | ID1_U05 |
| | U02 | Student potrafi dokonać klasyfikacji zbioru dokumentów tekstowych z wykorzystaniem odpowiednich algorytmów i narzędzi. | ID1_U06 ID1_U16 |
| | U03 | Student potrafi dokonać grupowania zbioru dokumentów tekstowych za pomocą odpowiednich algorytmów i narzędzi. | ID1_U06 ID1_U16 |
| | U04 | Student potrafi samodzielnie pozyskiwać odpowiednie dane tekstowe do analizy. | ID1_U01 |
| Kompetencje społeczne | K01 | Student rozumie potrzebę ciągłego poszerzania wiedzy z obszaru <i>Data Mining</i> i <i>Text Mining</i> . | ID1_K01 |

TREŚCI PROGRAMOWE

| Forma zajęć | Treści programowe |
|--------------|--|
| wykład | 1. Wprowadzenie do metod analizy danych niestrukturnalnych. Techniki <i>Data Mining</i> , <i>Text Mining</i> , <i>Web Mining</i> i ich zastosowanie. Wprowadzenie do SAS Text Miner. |
| | 2. Wstępna analiza danych tekstowych. Proces doskonalenia reprezentacji dokumentów (tokenizacja, stopwords, stemming/lematyzacja). |
| | 3. Modele reprezentacji tekstu. Model przestrzeni wektorowej (macierz TFM, funkcje ważące macierz TFM, miary odległości dla reprezentacji wektorowej). Transformacja danych tekstowych (redukcja wymiarów macierzy częstości). |
| | 4. Metody eksploracji danych – grupowanie dokumentów. Hierarchiczne i niehierarchiczne metody grupowania. Grupowanie dokumentów tekstowych za pomocą algorytmów aglomeracyjnych i podziałowych. |
| | 5. Metody eksploracji danych – klasyfikacja tekstu. Tworzenie tezausów. |
| | 6. Zapoznanie z innym oprogramowaniem do analiz dokumentów tekstowych (np. Rapid Miner, R). |
| | 7. Sprawdzian końcowy, podsumowujący. |
| laboratorium | 1. Zapoznanie ze środowiskiem SAS Text Miner. Identyfikacja źródeł oraz pozyskiwanie danych tekstowych do analiz. Wczytywanie danych tekstowych w różnych formatach i przetwarzanie zbioru dokumentów w celu utworzenia repozytorium do analiz – korpusu. |
| | 2. Wstępne przetwarzanie pozyskanych danych tekstowych (korpusu) w środowisku SAS. Eliminacja nieistotnych wyrazów (stop word) oraz redukcja do rdzenia lub formy podstawowej (stemming, lematyzacja) w celu doskonalenia reprezentacji danych tekstowych. |

| |
|--|
| 3. Wektorowa reprezentacja danych tekstowych w środowisku SAS. Metoda N – gram. Macierz term-frequency (TF). Ocena ważności słów w macierzy – przekształcenia częstości występowania słów (funkcje ważące). |
| 4. Opracowanie raportu. Sprawozdanie z wykonania prac z ćwiczeń nr 1-3. Porównanie wpływu testowanych metod przetwarzania dokumentów na jakość uzyskiwanych wyników. Wnioski i zalecenia. |
| 5. Analiza dokumentów tekstowych w środowisku SAS – grupowanie dokumentów korpusu. Ekstrakcja cech dokumentów, wybór miary odległości. Grupowanie z wykorzystaniem algorytmów hierarchicznych. |
| 6. Analiza dokumentów tekstowych w środowisku SAS – odkrywanie tematów w korpusie. Modelowanie tematyczne. |
| 7. Opracowanie raportu. Sprawozdanie z wykonania prac z ćwiczeń nr 5-6. Porównanie testowanych algorytmów grupowania. Wnioski i zalecenia. |
| 8. Analiza dokumentów tekstowych w środowisku SAS – tworzenie modelu predykcyjnego w celu przypisania tematu do każdego z dokumentów korpusu. Przygotowanie zbioru danych uczących, ocena błędu klasyfikacji. |
| 9. Analiza dokumentów tekstowych w środowisku SAS – wpływ parametrów konfiguracyjnych procesu wektoryzacji dokumentów tekstowych na jakość klasteryzacji i klasyfikacji. |
| 10. Opracowanie raportu. Sprawozdanie z wykonania prac z ćwiczeń nr 8-9. Porównanie testowanych algorytmów klasyfikacji. Wnioski i zalecenia. |
| 11. Zapoznanie z innym oprogramowaniem do analiz dokumentów tekstowych (np. Rapid Miner, R). |
| 12. Eksperymenty badawcze dla porównania analiz TM realizowanych w wybranych środowiskach programistycznych. |
| 13. Realizacja projektu indywidualnego – opracowanie modelu analizy danych tekstowych. Określenie celu analizy, zebranie i przygotowanie zbioru danych tekstowych dotyczących wybranego zagadnienia. Import plików do środowiska SAS i utworzenie zasobów SAS. |
| 14. Realizacja projektu indywidualnego – sporządzenie podstawowego modelu Text Mining, ocena wstępnych rezultatów, zastosowanie wybranej metody klasteryzacji. |
| 15. Opracowanie raportu. Sprawozdanie z wykonania prac z ćwiczeń nr 12-13. Wnioski i zalecenia. |

METODY WERYFIKACJI EFEKTÓW UCZENIA SIĘ

| Symbol efektu | Metody sprawdzania efektów kształcenia (zaznaczyć X) | | | | | |
|---------------|--|-----------------|-----------|---------|--------------|------|
| | Egzamin ustny | Egzamin pisemny | Kolokwium | Projekt | Sprawozdanie | Inne |
| W01 | | | X | | | |
| W02 | | | X | | | |
| W03 | | | X | | | |
| W04 | | | X | | | |
| U01 | | | | | X | |
| U02 | | | | | X | |
| U03 | | | | | X | |
| U04 | | | | | X | |
| K01 | | | | | | X |

FORMA I WARUNKI ZALICZENIA

| Forma zajęć | Forma zaliczenia | Warunki zaliczenia |
|--------------|--------------------|---|
| wykład | zaliczenie z oceną | Uzyskanie co najmniej 50% punktów ze sprawdzianu końcowego. |
| laboratorium | zaliczenie z oceną | Uzyskanie co najmniej 50% punktów z realizacji i raportowania zadań na ćwiczeniach, będąc członkiem zespołu dwuosobowego. |

NAKŁAD PRACY STUDENTA

| Bilans punktów ECTS | | | | | | | |
|---------------------|--|---------------------|---|----|---|---|-----------|
| Lp. | Rodzaj aktywności | Obciążenie studenta | | | | | Jednostka |
| | | W | C | L | P | S | |
| 1. | Udział w zajęciach zgodnie z planem studiów | 10 | | 30 | | | h |
| 2. | Inne (konsultacje, egzamin) | 2 | | 2 | | | h |
| 3. | Razem przy bezpośrednim udziale nauczyciela akademickiego | 44 | | | | | h |
| 4. | Liczba punktów ECTS, którą student uzyskuje przy bezpośrednim udziale nauczyciela akademickiego | 1,8 | | | | | ECTS |
| 5. | Liczba godzin samodzielnej pracy studenta | 31 | | | | | h |
| 6. | Liczba punktów ECTS, którą student uzyskuje w ramach samodzielnej pracy | 1,2 | | | | | ECTS |
| 7. | Nakład pracy związany z zajęciami o charakterze praktycznym | 56 | | | | | h |
| 8. | Liczba punktów ECTS, którą student uzyskuje w ramach zajęć o charakterze praktycznym | 2,3 | | | | | ECTS |
| 9. | Sumaryczne obciążenie pracą studenta | 75 | | | | | h |
| 10. | Punkty ECTS za moduł <i>1 punkt ECTS=25 godzin obciążenia studenta</i> | 3 | | | | | ECTS |

LITERATURA

1. Delen D., Miner G., Fast A., *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, Academic Press, 2012.
2. Dzieciatko M., Spinczyk D., *Text Mining: metody, narzędzia i zastosowania*, Wydawnictwo Naukowe PWN, 2016.
3. *Getting Started with SAS® Text Miner 12.1*, SAS Institute Inc, 2012.
4. Larose D.T., *Metody i modele eksploracji danych*, Wydawnictwo Naukowe PWN, 2012.
5. Markov Z., Larose D.T., *Eksploracja zasobów internetowych*, Wydawnictwo Naukowe PWN, Warszawa 2009.
6. Weiss S. M., Indurkha N., Zhang T., Damerau F., *Text Mining: Predictive Methods for Analyzing Unstructured Information*, Springer Science and Business Media, 2005.