Nazwa przedmiotu	Analityka danych tekstowych i źródeł internetowych
Kod przedmiotu	*do wypełnienia przez dziekanat*
Profil studiów	ogólnoakademicki
Stopień studiów	pierwszy
Kierunek studiów	Ekonomia
Specjalność (jeśli dotyczy)	Analityka Biznesowa
Forma studiów	studia stacjonarne i niestacjonarne wieczorowe
Rok studiów	II
Semestr	letni
Liczba godzin	30
Forma/typ zajęć	konwersatorium
Punkty ECTS	3
Rodzaj przedmiotu (przedmiot obowiązkowy / kierunkowy do wyboru / fakultatywny / seminarium)	obowiązkowy
Sposób realizacji przedmiotu (stacjonarny / zdalny / kurs internetowy / kurs realizowany metodą blended learning)	stacjonarny
Język wykładowy	polski
Wymagania formalne	Zaliczenie kursu “Wprowadzenie do programowania w języku Python”.
Założenia wstępne	Student posiada wiedzę z zakresu podstaw programowania w języku Python.
Skrócony opis przedmiotu	Zajęcia są dedykowane omówieniu technik zautomatyzowanego pobierania danych z sieci oraz podstawowych metod analizy danych tekstowych. Uczestnicy kursu nauczą się oceniać legalność scrapowania, jak również dobierać i implementować stosownie do potrzeb narzędzia temu służące. Studenci nauczą się stosować różnorodne techniki analizy danych tekstowych, począwszy od tokenizacji, przez stemming, lematyzację, usuwanie stopwords, stosowanie n-gramów, wykorzystanie statystyk TF, IDF, TF-IDF, aż po narzędzia pozwalające na ekstrakcję informacji, analizę sentymentu, modelowanie tematyczne, czy generowanie tekstu. Tym samym, kursanci nauczą się jak analizować tekst w celu identyfikacji potencjalnie istotnych wzorców, tendencji. Dzięki temu będą w stanie przetwarzać, kategoryzować oraz wizualizować dane tekstowe, co docelowo pozwoli im na lepsze zrozumienie odzwierciedlonych w danych tekstowych procesów ekonomicznych.
Treści kształcenia dla przedmiotu	Celem zajęć jest zapoznanie studentów z technikami zautomatyzowanego pobierania danych z sieci oraz podstawowymi metodami analizy danych tekstowych w stopniu wystarczającym do samodzielnego i efektywnego gromadzenia oraz analizowania danych nieustrukturyzowanych, jak również interpretowania i wyciągania wniosków z analiz tychże. Na kurs składają się następujące bloki tematyczne: 1. Wprowadzenie do web scrapingu: ● aspekty prawne automatycznego gromadzenia danych z sieci: standardy plików robots.txt, charakter prawny regulaminów

	korzystania ze stron internetowych, odpowiedzialność odszkodowawcza i karna, ● parsowanie struktur języka html, standardy XPath i CSS i rozszerzenia przeglądarek je zwracające, ● scrapowanie stron statycznych i dynamicznych, ● przegląd najpopularniejszych bibliotek języka programistycznego Python umożliwiających scrapowanie: BeautifulSoup, Selenium, Scrapy. 2. Podstawy pracy z danymi tekstowymi: ● wstępne przetwarzanie danych tekstowych: tokenizacja, stemming i lemmatyzacja, usuwanie stopwords, ● budowa i zastosowania n-gramów, ● najważniejsze statystyki opisowe dla filtrowania tokenów i reprezentowania tekstów: TF, IDF, TF-IDF, ● macierz dokumentów-terminów, ● wizualizacja danych tekstowych. 3. Podstawy ekstrakcji informacji: ● dopasowywanie wzorców: automaty skończone, wyrażenia regularne, ● podstawy modelowania sekwencji: rozpoznawanie nazwanych jednostek, ekstrakcja relacji i wydarzeń, ● typologia metod ekstrakcji informacji; metody oparte na wzorcach, słownikach, składni, statystykach korpusu, grafach, wprowadzenie do metod statystycznych i omówienie wybranych algorytmów. 4. Analiza sentymentu: ● proste metody słownikowe, ich specyfika i ograniczenia, zależność od domeny, omówienie wybranych metod, w tym: AFINN, Pattern, VADER, ● metody korpusowe, w szczególności oparte na skalowaniu manualnie przypisanych etykiet przez podobieństwo, oraz metody półnadzorowane, w tym algorytm propagacji etykiet, ● wprowadzenie do metod nadzorowanych, modelowanie charakterystyk na poziomie dokumentów, omówienie na przykładzie implementujących podstawowe algorytmy umożliwiające predykcję. 5. Modelowanie tematów: ● przegląd algorytmów modelowania tematów: algorytmy semantyczne, probabilistyczne i ich rozszerzenia, ● Latent Dirichlet Allocation jako najpopularniejszy algorytm modelowania tematów. 6. Wprowadzenie do dużych modeli językowych: ● specyfika i mechanizm działania dużych modeli językowych, encodery i decodery, mechanizm samouwagi, ● przykładowe aplikacje dużych modeli językowych w ekstrakcji informacji, analizie sentymentu, modelowaniu tematów, ● wiarygodność i ocena jakości wyników uzyskiwanych z implementacji rozwiązań opartych o duże modele językowe.
Przedmiotowe efekty uczenia się	Po ukończeniu przedmiotu, student: **W ZAKRESIE WIEDZY:** ● zna i rozumie regulacje prawne dotyczące scrapowania danych, ● zna i rozumie podstawowe operacje służące przygotowaniu danych tekstowych do analiz, ● zna i rozumie podstawowe techniki składające się na ekstrakcję informacji z danych tekstowych, analizę sentymentu, modelowanie tematów. **W ZAKRESIE UMIEJĘTNOŚCI:**

	● potrafi dobrać i zaimplementować odpowiednie dla danego problemu narzędzia celem scrapowania danych, ● potrafi przygotować dane tekstowe do analiz, ● potrafi przeprowadzić prostą analizę tekstu, tzn. dokonać ekstrakcji informacji, skwantyfikować sentyment tekstu, modelować występujące w dokumentach tematy. **W ZAKRESIE KOMPETENCJI:** ● wykazuje świadomość zalet, jak również ograniczeń wybranych podejść do analizy danych tekstowych, w szczególności jest świadomy potencjału, ale też potencjalnej niskiej interpretowalności i możliwej zawodności rozwiązań opartych o duże modele językowe, ● jest gotów samodzielnie prowadzić podstawowe analizy na danych tekstowych, jak również wie gdzie i w jaki sposób szukać informacji o bardziej zaawansowanych technikach przetwarzania danych tego typu.
Powiązanie efektów przedmiotowych z efektami kierunkowymi/specjalnościowymi (oznaczonymi kodami z programu nauczania)	kierunek Ekonomia: S_W02 S_W04 S_U01 S_U02 S_K01 S_K02
Nakład pracy studenta	Szacunkowy nakład pracy studenta: 3ECTS x 25h = 75h (K) - godziny kontaktowe (S) - godziny pracy samodzielnej zajęcia: 30h (K) 0h (S) konsultacje: 8h (K) 0h (S) zaliczenie: 2h (K) 0h (S) przygotowanie do ćwiczeń: 0h (K) 10h (S) przygotowanie do wykładów: 0h (K) 5h (S) przygotowanie do zaliczenia: 0h (K) 20h (S) Razem: 40h (K) + 35h (S) = 75h
Metody i kryteria oceniania	Uzyskanie zaliczenia przedmiotu wymaga: 1. uzyskania co najmniej połowy punktów z kolokwium pisemnego; 2. uzyskania co najmniej połowy punktów z projektu przygotowanego w domu; 3. uzyskania co najmniej połowy punktów z prezentacji wyżej wspomnianego projektu; 4. uzyskania co najmniej połowy z łącznej sumy punktów uzyskanych ze wszystkich wymienionych powyżej elementów. 5. Skala ocen: [0%-50%) - ndst [50%-60%) - dst [60%-70%) - dst+ [70%-80%) - db [80%-90%) - db+ [90%-100%] - bdb
Literatura	● Jurafsky, D., & Martin, J. H. (2025). Speech and Language Processing (3rd ed. draft). Stanford University. Retrieved from: web.stanford.edu/~jurafsky/slp3/. ● R. Mitchell (2018). Web Scraping with Python: Collecting Data from the Modern Web. 2nd Edition. O'Reilly Media.

	● Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media. ● Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press. ● Srivastava, A. N., & Sahami, M. (Eds.). (2009). Text Mining: Classification, Clustering, and Applications. Chapman & Hall/CRC. ● Tunstall, L., Von Werra, L., & Wolf, T. (2022). Natural language processing with transformers. Building Language Applications with Hugging Face. 1st Edition. O'Reilly Media.
Metody dydaktyczne	W ramach konwersatorium wykorzystywane są następujące metody dydaktyczne: -prezentacja wykładowcy, -prezentacja studencka, -analiza źródeł danych, -analiza przypadku, -metoda projektu, -ćwiczenia rachunkowe i analityczne, -metoda projektów zespołowych, -prezentacja multimedialna, -analiza danych w programach komputerowych.
Oprogramowanie	Visual Studio Code, wtyczki: Jupyter, Python