SYLABUS PRZEDMIOTU

***Ekstrakcja informacji***

1. **Informacje ogólne**

Nazwa przedmiotu Ekstrakcja informacji

Kod przedmiotu **EKI**

Rodzaj przedmiotu: specjalistyczny

Kierunek studiów: Informatyka

Poziom kształcenia: uzupełniające

Profil kształcenia: ogólnoakademicki

Rok studiów: II

Rodzaje zajęć i liczba godzin

Wykład **30**

Ćwiczenia **0**

Laboratoria **30**

Praktyki **0**

Liczba punktów ECTS **6**

Imię, nazwisko, tytuł/stopień naukowy, adres e-mail wykładowcy (wykładowców)/ prowadzących zajęcia

* dr Filip Graliński [filipg@amu.edu.pl](mailto:filipg@amu.edu.pl)
* mgr Jakub Pokrywka [jp40407@st.amu.edu.pl](mailto:jp40407@st.amu.edu.pl)

Język wykładowy polski

Przedmiot prowadzony zdalnie (e-learning) **częściowo**

1. **Informacje szczegółowe**
2. **Cele przedmiotu**

Przedmiot stawia następujące cele:

* poznanie problematyki wyszukiwania informacji tekstowych od strony teoretycznej i praktycznej,
* poznanie problematyki klasycznych technik ekstrakcji informacji w zakresie teoretycznym i praktycznym,
* zaznajomienie z neuronowymi technikami ekstrakcji informacji,
* zdobycie wiedzy na temat sumaryzacji tekstu z wykorzystaniem modeli neuronowych,
* nabycie wiedzy i umiejętności wykorzystania modeli językowych do praktycznych zastosowań ekstrakcji informacji.

1. **Wymagania wstępne w zakresie wiedzy, umiejętności oraz kompetencji społecznych**

Podstawowa umiejętność programowania w języku Python 3

1. **Efekty uczenia się (EU) dla zajęć i odniesienie do efektów uczenia się (EK) dla kierunku studiów**

|  |  |  |
| --- | --- | --- |
| **Symbol EU dla przedmiotu** | **Symbol EK dla kierunku studiów** | **Po zakończeniu modułu i potwierdzeniu osiągnięcia EU student/ka:** |
| EKI\_01 | KINF2\_W01 | Rozumie sposób działania tradycyjnych wyszukiwarek tekstowych. Zna sposoby przetwarzania tekstu na potrzeby wyszukiwarek tekstowych (lowercasing, tokenizacja, tworzenie wektorów TF-IDF). Potrafi wyjaśnić sposób, w jaki rankingowane są dokumenty w tradycyjnej wyszukiwarce (TF-IDF scoring, Okapi BM25). Zna pojęcie *inverted index* i jego znaczenie w optymalizacji zapytań. |
| EKI\_02 | KINF2\_U01 | Umie przetwarzać tekst w celu stworzenia wyszukiwarki tekstowej przy pomocy gotowych bibliotek. Potrafi stworzyć ranking dokumentów dla zapytania. |
| EKI\_03 | KINF2\_W03,  KINF2\_W05,  KINF2\_W06,  KINF2\_W07 | Zna popularne silniki wyszukiwarek. Zna sposoby ewaluacji wyszukiwarek internetowych (ludzka ewaluacja, *cumulative gain* i jej pochodne, testy A/B), ich zalety i słabości. Potrafi wyjaśnić pojęcie relewantnej wyszukiwarki. Zna techniki poprawiania jakości wyszukiwarki oraz cykl budowy i utrzymania wyszukiwarki. Rozumie komercyjne i społeczne znaczenie wyszukiwarek internetowych. |
| EKI\_04 | KINF2\_U02,  KINF2\_U05 | Potrafi stworzyć wyszukiwarkę przy pomocy gotowych silników wyszukiwarek. Umie poprawić jakość wyszukiwarki dla konkretnych przypadków użycia. Jest w stanie zwiększyć precyzję lub pokrycie wyników wyszukiwania. Umie wyznaczać popularne metryki rankingu przy wykorzystaniu gotowych narzędzi. |
| EKI\_05 | KINF2\_W02,  KINF2\_W05 | Zna obecny stan badań w wyszukiwarkach tekstowych. Rozumie pojęcie wyszukiwania semantycznego. Zna obecne mechanizmy interakcji wyszukiwarki z użytkownikiem (podpowiadanie frazy, *ghosting*, rekomendacje, korekta frazy). |
| EKI\_06 | KINF2\_U02 | Potrafi stworzyć wyszukiwarkę bazującą na podobieństwie semantycznym frazy do dokumentu. |
| EKI\_07 | KINF2\_W02,  KINF2\_W03 | Rozumie pojęcie ekstrakcji informacji i zadania, które należą do tej dziedziny. Zna praktyczne zastosowania ekstrakcji informacji. |
| EKI\_08 | KINF-U01 | Potrafi stworzyć system ekstrakcji fraz z określonego korpusu tekstów używając połączenia metod lingwistycznych i statystycznych. |
| EKI\_09 | KINF2\_W01,  KINF2\_W02 | Zna pojęcie zadania *sequence labeling* oraz jego zastosowania: *named entity recognition (NER)*, *part-of-speech tagging (POS)* i inne. Rozumie różne sposoby ewaluacji zadania *sequence labeling*. Potrafi wyjaśnić podstawy działania statystycznych modeli *sequence labeling*: *hidden markov models* (HMM) oraz *conditional random fields* (CRF). |
| EKI\_10 | KINF2\_U01 | Umie stworzyć i ewaluować statystyczny model *sequence labeling* i zastosować go w zadaniu NER. |
| EKI\_11 | KINF2\_W02 | Zna neuronowe podejścia do zadania *sequence labeling* bazujące na sieciach rekurencyjnych. |
| EKI\_12 | KINF2\_U01,  KINF2\_U05 | Potrafi zaimplementować architekturę rekurencyjnego neuronowego modelu NER przy użyciu biblioteki sieci neuronowych. Potrafi wykonać trening tak zdefiniowanego modelu w celu uzyskania jak najlepszej jakości modelu. |
| EKI\_13 | KINF2\_W02 | Zna neuronowe podejścia do zadania *sequence labeling* bazujące na sieciach neuronowych typu *transformer*. |
| EKI\_14 | KINF2\_U02,  KINF2\_U05,  KINF2\_U07 | Potrafi użyć pretrenowany model typu *transformer* w celu jego dostrojenia do zadania NER. Potrafi porównać różne modele NER (statystyczne, rekurencyjne sieci neuronowe, sieci neuronowe typu transformer). |
| EKI\_15 | KINF2\_W02 | Rozumie zadanie sumaryzacji tekstu i sposoby jego ewaluacji. Zna aktualny stan badań w zakresie sumaryzacji tekstu, w szczególności modele *transformer* i *transformer text-to-text*. |
| EKI\_16 | KINF2\_U02 | Potrafi wykorzystać neuronowy model typu transformer do zadania sumaryzacji tekstu. |
| EKI\_17 | KINF2\_W02 | Zna zadanie i obecny stan badań odnośnie neuronowej ekstrakcji informacji z nieustrukturyzowanych dokumentów należących do jednej dziedziny. |
| EKI\_18 | KINF2\_U02 | Potrafi wykorzystać neuronowy model oparty o sieci typu *transformer* do ekstrakcji informacji z nieustrukturyzowanych dokumentów należących do jednej dziedziny. |

1. **Treści programowe zapewniające uzyskanie efektów uczenia się (EU) z odniesieniem do odpowiednich efektów uczenia się (EU) dla przedmiotu**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Lp.** | **Symbol EU dla przedmiotu** | **Godzin Wykład** | **Godzin  ĆW/ LAB/ SEM** | **Godzin pracy własnej** | **Opis treści kształcenia modułu zajęć/przedmiotu** |
| **Suma** | | **30** | **30** | **90** |  |
| 1. | EKI\_01 | 3 |  | 5 | Wprowadzenie do wyszukiwarek tekstowych, przetwarzanie tekstu na potrzeby wyszukiwarek tekstowych, algorytmy rankingowania wyszukiwarek tekstowych (TF-IDF, Okapi BM25) optymalizacja działania wyszukiwarek *(inverted index*). |
| 2. | EKI\_02 |  | 3 | 5 | Przetwarzanie tekstu i stworzenie prostego systemu rankingującego dokumenty dla określonego zapytania przy wykorzystaniu prostych narzędzi i bibliotek (innych niż silniki wyszukiwarki). |
| 3. | EKI\_03 | 4 |  | 5 | Wyszukiwarki tekstowe – część zaawansowana. Dostępne komercyjnie silniki wyszukiwarek (*Solr, Elasticsearch*), sposoby ewaluacji wyszukiwarek, cykl rozwoju i utrzymania wyszukiwarki i dopasowanie pod potrzeby biznesowe klienta, znaczenie społeczne wyszukiwarek internetowych. |
| 4. | EKI\_04 |  | 4 | 5 | Stworzenie wyszukiwarki przy użyciu narzędzia *ElasticSearch*. Optymalizacja skuteczności wyszukiwania pod daną frazę. Liczenie metryki rankingu nDCG przy pomocy gotowych narzędzi. |
| 5. | EKI\_05 | 3 |  | 5 | Najnowsze badania i trendy w wyszukiwarkach tekstowych, wyszukiwanie semantyczne, mechanizmy interakcji między wyszukiwarką i użytkownikiem. |
| 6. | EKI\_06 |  | 3 | 5 | Implementacja prostej wyszukiwarki semantycznej. |
| 7. | EKI\_07 | 3 |  | 5 | Ekstrakcja informacji – wprowadzenie, nakreślenie problemów i zadań, zastosowania. |
| 8. | EKI\_08 |  | 3 | 5 | Implementacja systemu ekstrakcji fraz dla specjalistycznego korpusu tekstowego przy użyciu metod lingwistycznych i statystycznych. |
| 9. | EKI\_09 | 4 |  | 5 | Zadanie *sequence labelling* i jego zastosowania (NER, POS). Ewaluacja zastosowań s*equence labeling*, statystyczne modele *sequence labeling* (HMM i CRF). |
| 10. | EKI\_10 |  | 4 | 5 | Stworzenie statystycznego modelu NER wykorzystującego HMM lub CRF. |
| 11. | EKI\_11 | 4 |  | 5 | Podstawy sieci neuronowych, neuronowy rekurencyjny model *sequence labeling.* |
| 12. | EKI\_12 |  | 4 | 5 | Zapoznanie z sieciami neuronowymi w bibliotece *pytorch*, implementacja neuronowego rekurencyjnego modelu NER i trenowanie modelu. |
| 13. | EKI\_13 | 3 |  | 5 | Neuronowe modele *sequence labeling* wykorzystujące sieci typu transformer. |
| 14. | EKI\_14 |  | 3 | 5 | Dostrajanie pretrenowanego modelu typu *transformer* do zadania NER. Porównanie jakości różnych modeli NER stworzonych na zajęciach (statystycznego, neuronowego bazującego na sieciach rekurencyjnych, neuronowego bazującego na sieci typu *transformer*). |
| 15. | EKI\_15 | 3 |  | 5 | Zadanie sumaryzacji tekstu i jego ewaluacja, aktualny stan badań w zakresie sumaryzacji tekstu ze szczególnym uwzględnieniem modeli typu transformer i transformer text-to-text. |
| 16. | EKI\_16 |  | 3 | 5 | Wykorzystanie pretrenowanego modelu typu transformer do zadania sumaryzacji tekstu. |
| 17. | EKI\_17 | 3 |  | 5 | Zadanie ekstrakcji informacji z nieustrukturyzowanych dokumentów domenowych – definicja zadania, modele neuronowe typu transformer, obecny stan wiedzy naukowej. |
| 18. | EKI\_18 |  | 3 | 5 | Wykorzystanie modelu neuronowego typu *transformer* do zadania ekstrakcji informacji z nieustrukturyzowanych dokumentów należących do jednej dziedziny. |

1. **Zalecana literatura**
2. Doug Turnbull, John Berryman. 2016. Relevant Search: With applications. for Solr and Elasticsearch. Manning Publications.
3. Marie-Francine Moens. 2006. Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer.
4. Alex Graves. 2012. Supervised sequence labelling. Studies in Computational Intelligence, vol 385. Springer. Berlin, Heidelberg.
5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Association for Computational Linguistics (NAACL).
6. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research vol 21, number 140, pages 1-67.
7. Flip Graliński, Tomasz Stanisławek, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, Przemysław Biecek. 2020. Kleister: A novel task for information extraction involving long documents with complex layout. URL <https://arxiv.org/abs/2003.02356>
8. Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Filip Graliński. 2020. LAMBERT: Layout-Aware (Language) Modeling using BERT. URL <https://arxiv.org/pdf/2002.08087>

1. **Informacje dodatkowe**
2. **Metody i formy prowadzenia zajęć umożliwiające osiągnięcie założonych EU (proszę wskazać z proponowanych metod właściwe dla opisywanych zajęć lub/i zaproponować inne)**

|  |  |
| --- | --- |
| **Realizacja** | **Metody i formy prowadzenia zajęć** |
| ✔ | Wykład z prezentacją multimedialną wybranych zagadnień |
|  | Wykład konwersatoryjny |
|  | Wykład problemowy |
|  | Dyskusja |
|  | Praca z tekstem |
| ✔ | Metoda analizy przypadków |
|  | Uczenie problemowe (Problem-based learning) |
|  | Gra dydaktyczna/symulacyjna |
|  | Rozwiązywanie zadań (np.: obliczeniowych, artystycznych, praktycznych) |
|  | Metoda ćwiczeniowa |
| ✔ | Metoda laboratoryjna |
| ✔ | Metoda badawcza (dociekania naukowego) |
|  | Metoda warsztatowa |
| ✔ | Metoda projektu |
|  | Pokaz i obserwacja |
|  | Demonstracje dźwiękowe i/lub video |
|  | Metody aktywizujące (np.: „burza mózgów”, technika analizy SWOT, technika drzewka decyzyjnego, metoda „kuli śniegowej”, konstruowanie „map myśli”) |
|  | Praca w grupach |
| ✔ | Wykład zdalny w czasie rzeczywistym |
|  | Wykład zdalny asynchroniczny uzupełniony spotkaniem w czasie rzeczywistym |
|  | Wykład zdalny asynchroniczny z aktywnością studenta uzupełniony spotkaniem w czasie rzeczywistym |
| ✔ | Ćwiczenia/laboratoria/konwersatoria zdalne w czasie rzeczywistym |
|  | Ćwiczenia zdalne asynchroniczne z pracą indywidualną studenta uzupełnione spotkaniem w czasie rzeczywistym |
|  | Ćwiczenia zdalne asynchroniczne z pracą grupową studentów uzupełnione spotkaniem w czasie rzeczywistym |
|  | Laboratorium cyfrowe zdalne uzupełnione spotkaniem w czasie rzeczywistym |
|  | Konwersatorium asynchroniczne zdalne uzupełnione spotkaniem w czasie rzeczywistym |
|  | Seminarium zdalne w czasie rzeczywistym |
|  | Seminarium asynchroniczne zdalne ze spotkaniem w czasie rzeczywistym |
|  | Inne (jakie?) - |

1. **Sposoby oceniania stopnia osiągnięcia EU (proszę wskazać z proponowanych sposobów właściwe dla danego EU lub/i zaproponować inne**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sposoby oceniania** | **Symbole EU**  **dla modułu** **zajęć/przedmiotu** | | | | | | | | | | |
| **EKI: 01,03,05,07,09,11,13,15,17** | **EKI:02,04,06,08,10,12,14,16,18** |  |  |  |  |  |  |  |  |
| Egzamin pisemny | ✔ |  |  |  |  |  |  |  |  |  |
| Egzamin ustny |  |  |  |  |  |  |  |  |  |  |
| Egzamin z „otwartą książką” |  |  |  |  |  |  |  |  |  |  |
| Kolokwium pisemne |  |  |  |  |  |  |  |  |  |  |
| Kolokwium ustne |  |  |  |  |  |  |  |  |  |  |
| Test |  |  |  |  |  |  |  |  |  |  |
| Projekt |  | ✔ |  |  |  |  |  |  |  |  |
| Esej |  |  |  |  |  |  |  |  |  |  |
| Raport |  |  |  |  |  |  |  |  |  |  |
| Prezentacja multimedialna |  |  |  |  |  |  |  |  |  |  |
| Egzamin praktyczny (obserwacja wykonawstwa) |  |  |  |  |  |  |  |  |  |  |
| Portfolio |  |  |  |  |  |  |  |  |  |  |
| Zadania cząstkowe na wykładzie |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |

1. **Nakład pracy studenta i punkty ECTS**

|  |  |  |
| --- | --- | --- |
| **Forma aktywności** | | **Średnia liczba godzin na zrealizowanie aktywności** |
| Godziny zajęć (wg planu studiów) z nauczycielem | | 60 |
| Praca własna studenta\* | Przygotowanie do zajęć | 20 |
| Czytanie wskazanej literatury | 10 |
| Przygotowanie pracy pisemnej, raportu, prezentacji, itp. | 10 |
| Przygotowanie projektu | 30 |
| Przygotowanie pracy semestralnej | 0 |
| Przygotowanie do egzaminu/zaliczenia | 0 |
| Praca z materiałem do samokształcenia (np. Jupyter Notebook) | 20 |
| Praca z laboratorium cyfrowym (np. Code Runner) | 0 |
| Inne (jakie?) |  |
| **SUMA GODZIN** | | 150 |
| LICZBA PUNKTÓW ECTS DLA PRZEDMIOTU | | **6** |

\* proszę wskazać z proponowanych przykładów pracy własnej studenta właściwe dla opisywanego modułu lub/i zaproponować inne

1. **Kryteria oceniania wg skali stosowanej w UAM**

|  |  |
| --- | --- |
| **Ocena** | **Kryterium** |
| bardzo dobry (bdb; 5,0) | od 90% punktów |
| dobry plus (+db; 4,5) | od 80% punktów |
| dobry (db; 4,0) | od 70% punktów |
| dostateczny plus (+dst; 3,5) | od 60% punktów |
| dostateczny (dst; 3,0) | od 50% punktów |
| niedostateczny (ndst; 2,0) | poniżej 50% punktów |