Uczenie głębokie w przetwarzaniu tekstu

Sylabus zajęć

Informacje podstawowe

Kierunek studiów Informatyka	**Cykl dydaktyczny** 2023/24
Specjalność -	**Kod zajęć** 06INFS.42S.01029.23
Jednostka organizacyjna Wydział Matematyki i Informatyki	**Języki wykładowe** polski
Poziom studiów studia drugiego stopnia poinżynierskie	**Obligatoryjność** Fakultatywny
Forma studiów studia stacjonarne	**Blok zajęciowy** Przedmioty specjalnościowe
Profil studiów profil ogólnoakademicki	

Koordynator zajęć	Karol Kaczmarek
Prowadzący zajęcia	Karol Kaczmarek

Okres	**Forma zajęć / liczba godzin / forma zaliczenia**	**Liczba punktów ECTS**
Semestr 2	• Laboratorium: 30, Zaliczenie z oceną	3

Cele kształcenia dla zajęć

Kod	Cel
C1	Nabycie umiejętności wykorzystania neuronowych modeli języka w przetwarzaniu tekstu.
C2	Poznanie architektury Transformer.
C3	Poznanie modeli typu dekoder: GPT-2, GPT-3, Longformer, GPT-NeoX, XGLM, FLAN, OPT.
C4	Poznanie modeli typu enkoder: BERT, RoBERTa, Polish RoBERTa, DistilBERT.
C5	Poznanie modeli typu enkoder-dekoder: T5, Polish T5, LongT5, T0, Switch Transformer, FlanT5.
C6	Poznanie wielojęzycznych modeli: XLM, XLM-RoBERTa, mT5, ByT5, XGLM.
C7	Poznanie słowników opartych o algorytm Byte Pair Encoding oraz uni-gramowy model języka.
C8	Poznanie podejścia few-show, one-shot oraz zero-shot learning.

Wymagania wstępne

Podstawowa umiejętność programowania w języku Python.

Efekty uczenia się dla zajęć

Kod	Efekty uczenia się dla zajęć w zakresie	Efekty uczenia się dla kierunku	Metody weryfikacji osiągnięcia efektów uczenia się dla zajęć
Wiedzy – Student/ka:			
W1	Zna neuronowe modele języka oparte na architekturze Transformer. Potrafi wskazać różnicę pomiędzy modelami opartymi na architekturze enkoder, dekoder oraz enkoder-dekoder.	INF_K4_W02	Projekt, Zadania wykonywane podczas zajęć
W2	Zna oraz potrafi wskazać cech charakterystyczne neuronowych modeli języka: BERT, RoBERTa, GPT-2, T5, T0, OPT.	INF_K4_W03	Projekt, Zadania wykonywane podczas zajęć
W3	Zna oraz potrafi wskazać cechy charakterystyczne wielojęzycznych neuronowych modeli języka: XLM, XLM-RoBERTa, mT5, ByT5, XGLM.	INF_K4_W03	Projekt, Zadania wykonywane podczas zajęć
W4	Zna problem długości sekwencji w tekście oraz potrafi wskazać sposoby na rozwiązanie tego problemu. Zna neuronowe modele języka: Longformer, LongT5 oraz FlashAttention.	INF_K4_W03	Projekt, Zadania wykonywane podczas zajęć
W5	Zna różnicę pomiędzy modelami dziedzinowymi i ogólnymi (jednojęzyczne oraz wielojęzyczne) oraz umie wskazać wady i zalety modeli.	INF_K4_W02	Projekt, Zadania wykonywane podczas zajęć
W6	Zna różnicę pomiędzy podejście few-shot, one-shot oraz zero-shot learning.	INF_K4_W02	Projekt, Zadania wykonywane podczas zajęć
Umiejętności – Student/ka:			
U1	Zna oraz potrafi wskazać sposoby rozwiązania problemu "rzadkich słów" – Byte Pair Encoding (model GPT-2) oraz SentencePiece (model T5).	INF_K4_U02	Projekt, Zadania wykonywane podczas zajęć

Kod	Efekty uczenia się dla zajęć w zakresie	Efekty uczenia się dla kierunku	Metody weryfikacji osiągnięcia efektów uczenia się dla zajęć
U2	Potrafi wytrenować oraz dostroić neuronowy model języka oraz zna pojęcia "pretraining" oraz "fine-tuning".	INF_K4_U01, INF_K4_U02, INF_K4_U06	Projekt, Zadania wykonywane podczas zajęć
U3	Potrafi wykorzystać neuronowy model języka w problemie klasyfikacji oraz zadaniu typu QA (Pytanie-Odpowiedź).	INF_K4_U01, INF_K4_U02, INF_K4_U06	Projekt, Zadania wykonywane podczas zajęć
U4	Potrafi wykorzystać neuronowy model języka w trybie: few-shot, one-shot oraz zero-shot learning.	INF_K4_U01, INF_K4_U02, INF_K4_U06	Projekt, Zadania wykonywane podczas zajęć

Treści programowe dla zajęć

Lp.	Treści programowe dla zajęć	Efekty uczenia się dla zajęć	Formy zajęć
1.	Rozwiązanie problemu "rzadkich słów" z wykorzystaniem algorytmu Byte Pair Encoding oraz uni-gramowego modelu języka.	U1	Laboratorium
2.	Charakterystyka architektury Transformer, modeli opartych na architekturze enkoder, dekoder oraz enkoder-dekoder.	W1	Laboratorium
3.	Charakterystyka neuronowych modeli języka: BERT, RoBERTa, GPT-2, T5, T0, OPT.	W2	Laboratorium
4.	Charakterystyka neuronowych modeli języka: XLM, XLM-RoBERTa, mT5, ByT5, XGLM.	W3	Laboratorium
5.	Rozwiązywanie problemu długości sekwencji, wykorzystanie neuronowych modeli języka: Longformer, LongT5 oraz FlashAttention.	W4	Laboratorium
6.	Charakterystyka dziedzinowych, ogólnych (jednojęzycznych oraz wielojęzycznych) modeli.	W5	Laboratorium
7.	Trenowanie oraz dostrojenie neuronowych modeli języka z wykorzystaniem biblioteki transformers.	U2	Laboratorium
8.	Wykorzystanie podejścia few-shot, one-shot oraz zero-shot learning z neuronwymi modelami języka.	W6, U4	Laboratorium
9.	Rozwiązanie zadania klasyfikacji oraz typu QA (Pytania-Odpowiedź) z wykorzystaniem neuronowych modeli języka.	U3	Laboratorium

Informacje dodatkowe

Forma zajęć	Metody i formy prowadzenia zajęć
Laboratorium	Wykład z prezentacją multimedialną wybranych zagadnień, Uczenie problemowe (Problem-based learning), Metoda laboratoryjna, Metoda warsztatowa, Metoda projektu, Praca w grupach

Forma zajęć	Warunki zaliczenia zajęć
Laboratorium	Końcowa ocena składa się z następujących elementów: 1. projekt – 90%, 2. zadania wykonywane podczas zajęć – 10%. Skala ocen: 1. bardzo dobry (bdb; 5,0) – od 90% punktów, 2. dobry plus (db plus; 4,5) – od 80% punktów, 3. dobry (db; 4,0) – od 70% punktów, 4. dostateczny plus (dst plus; 3,5) – od 60% punktów, 5. dostateczny (dst; 3,0) – od 50% punktów, 6. niedostateczny (ndst; 2,0) – poniżej 50% punktów.

Literatura

Obowiązkowa

1. Philipp Koehn, "Statistical Machine Translation - Draft of Chapter 13: Neural Machine Translation"
2. A. Vaswani and et al., “Attention is all you need”, 2017
3. Sennrich, Rico and Haddow, Barry and Birch, Alexandra, "Neural Machine Translation of Rare Words with Subword Units”, 2016
4. Taku Kudo, "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates”, 2018
5. Alec Radford and Jeffrey Wu and et al, "Language Models are Unsupervised Multitask Learners”, 2019
6. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah and et al., "Language Models are Few-Shot Learners”, 2020
7. Devlin, Jacob and et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 2018
8. Yinhan Liu and et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach”, 2019
9. Colin Raffel and et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”, 2019

Dodatkowa

1. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah and et al., „Language Models are Few-Shot Learners”, 2020
2. Sławomir Dadas and et al., "Pre-training Polish Transformer-based Language Models at Scale”, 2020
3. Guillaume Lample, Alexis Conneau, "Cross-lingual Language Model Pretraining”, 2019
4. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov, „Unsupervised Cross-lingual Representation Learning at Scale”, 2019
5. Linting Xue and et al., "mT5: A massively multilingual pre-trained text-to-text transformer”, 2020
6. Iz Beltagy and Matthew E. Peters and Arman Cohan, "Longformer: The LongDocument Transformer”, 2020
7. Sinong Wang and Belinda Z. Li and Madian Khabsa and Han Fang and Hao Ma, "Linformer: Self-Attention with Linear Complexity”, 2020
8. Thomas Wolf and et al., "HuggingFace's Transformers: State-of-the-art Natural Language Processing”, 2019
9. Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, „DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”, 2019
10. Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, Ming Zhou. „MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers”, 2020

Nakład pracy studenta i punkty ECTS

Rodzaje zajęć studenta	Średnia liczba godzin* przeznaczonych na zrealizowane rodzaje zajęć
Laboratorium	30
Przygotowanie do zajęć	5

Czytanie wskazanej literatury	5
Przygotowanie projektu	30
Inne	15
Łączny nakład pracy studenta	**Liczba godzin** 85
Liczba punktów ECTS	**ECTS** 3

* godzina (lekcyjna) oznacza 45 minut

Efekty uczenia się dla kierunku

Kod	Treść
INF_K4_U01	Absolwent/ka potrafi zastosować zaawansowaną wiedzę matematyczną do formułowania, analizowania i rozwiązywania złożonych i nietypowych zadań związanych z informatyką
INF_K4_U02	Absolwent/ka potrafi adaptować istniejące oraz tworzyć nowe metody informatyczne do rozwiązywania nieszablonowych problemów praktycznych i teoretycznych
INF_K4_U06	Absolwent/ka potrafi rozwiązywać złożone problemy z wybranych obszarów informatyki oraz proponować nowe algorytmy, narzędzia i metody wykorzystując odpowiednio dobrane źródła, które poddaje krytycznej analizie, syntezie i twórczej interpretacji
INF_K4_W02	Absolwent/ka zna i rozumie współczesny stan badań i tendencje rozwojowe w wiodących obszarach informatyki
INF_K4_W03	Absolwent/ka zna i rozumie w pogłębionym stopniu współczesne metody, narzędzia i technologie informatyczne właściwe dla wybranych obszarów zastosowań niezbędne przy budowie złożonych systemów informatycznych oraz przy prowadzeniu prac badawczo-rozwojowych