Силабус освітнього компонента
Програма навчальної дисципліни

Обробка та аналіз текстової інформації

Шифр та назва спеціальності
113 – Прикладна математика

Інститут
ННІ Комп'ютерних наук та інформаційних технологій

Освітня програма
Інтелектуальний аналіз даних

Кафедра
Комп'ютерна математика і аналіз даних

Рівень освіти
Бакалавр

Тип дисципліни
Спеціальна (фахова), Вибіркова

Семестр
8

Мова викладання
Українська

Викладачі, розробники

Костюк Ольга Василівна

Olha.Kostiuk@khpi.edu.ua
Кандидат технічних наук, доцент кафедри комп'ютерної математики і аналізу даних НТУ «ХПІ»

Автор наукових та навчально-методичних праць. Провідний лектор з дисциплін: «Нечіткі моделі та методи», «Теорія прийняття рішень», «Вступ до спеціальності та інженерної діяльності»
Детальніше про викладача на сайті кафедри

Загальна інформація

Анотація

Головною метою навчальної дисципліни «Обробка та аналіз текстової інформації» є отримання базових знань в області обробки та аналізу текстової інформації, а також отримання навичок вирішення завдань, що виникають при розробці систем обробки лінгвістичних даних. Зміст курсу спрямовано на ознайомлення студентів з базовими поняттями обробки природної мови, класичним підходом до обробки тексту, методами попередньої обробки, підходів щодо морфологічного, синтаксичного та семантичного аналізів. Також у курсі будуть розглянуті сучасні моделі обробки текстової інформації, які засновані на нейронних мережах.

Мета та цілі дисципліни

Мета цього курсу - надати учасникам глибокі знання та навички в галузі обробки текстової інформації. Учасники отримають поглиблене розуміння основних методів аналізу, обробки, та витягування інформації з текстів, а також навчаться використовувати сучасні інструменти та програмні засоби для автоматизації цих процесів. Курс спрямований на підготовку учасників до ефективного використання обробки текстової інформації у різних галузях, включаючи природну мову, аналітику соціальних медіа, автоматизовану обробку документів, тощо. Після завершення курсу учасники матимуть необхідні навички для впровадження сучасних технік обробки текстової інформації в практичні завдання та дослідження.

Формат занять

Лекції, лабораторні роботи, самостійна робота, консультації. Підсумковий контроль – іспит.

Компетентності

ЗК 1. Здатність учитися й оволодівати сучасними знаннями.
ЗК 2. Здатність застосовувати знання у практичних ситуаціях.
ЗК 6. Здатність до абстрактного мислення, аналізу і синтезу.
ЗК 7. Здатність до пошуку, оброблення й аналізу інформації з різноманітних джерел.
ЗК 8. Знання і розуміння предметної області та розуміння професійної діяльності.
СК 3. Здатність обирати та застосовувати математичні методи для розв'язання прикладних задач, моделювання, аналізу, проєктування, керування, прогнозування, прийняття рішень.
СК 10. Здатність до проведення математичного і комп'ютерного моделювання, аналізу та обробки даних, обчислювального експерименту, розв'язання формалізованих задач за допомогою спеціалізованих програмних засобів.
СК 14. Здатність зрозуміти постановку завдання, сформульовану мовою певної предметної галузі, здійснювати пошук та збір необхідних вихідних даних.
СК 20. Здатність до розробки та експлуатації програмних засобів інтелектуального аналізу даних вимірювань та спостережень, текстів, сигналів і зображень.

Результати навчання

РН 3. Формалізувати задачі, сформульовані мовою певної предметної галузі; формулювати їх математичну постановку та обирати раціональний метод вирішення; розв'язувати отримані задачі аналітичними та чисельними методами, оцінювати точність та достовірність отриманих результатів.
РН 12. Розв'язувати окремі інженерні задачі та/або задачі, що виникають принаймні в одній предметній галузі: в соціології, економіці, екології та медицині.
РН 22. Знати та розуміти методи розв'язання математичних задач інтелектуального інформаційного пошуку та видобування знань.
РН 24. Вміти застосовувати існуючі та розробляти нові алгоритми і програмні засоби обробки даних вимірювань та спостережень, текстів, сигналів та зображень.

Обсяг дисципліни

Загальний обсяг дисципліни 120 год. (4 кредитів ECTS): лекції – 20 год., лабораторні роботи – 20 год., самостійна робота – 80 год.

Передумови вивчення дисципліни (пререквізити)

Для успішного проходження курсу необхідно мати знання та практичні навички з наступних дисциплін: «Алгоритмізація та програмування», «Комп'ютерна дискретна математика», «Теорія ймовірностей», «Математична статистика», «Нейромережеві технології», «Методи та засоби машинного навчання».

Особливості дисципліни, методи та технології навчання

Лекції та лабораторні проводяться онлайн з використання сучасних програмних засобів. На лабораторних роботах використовується програмне середовище Colab, акцентується увага сама на практичному застосуванні різних моделей та методів або фреймворків для обробки текстової інформації.

Програма навчальної дисципліни

Теми лекційних занять

Тема 1. Вступ до обробки природної мови (ОПМ)

Основні поняття. Сфери застосування ОПМ: існуючі завдання, додатки та реальні приклади з життя.

Тема 2. Передобробка текстової інформації

Базова термінологія та алгоритми, які допомагають підготувати дані для лінгвістичного аналізу: токенізація, видалення стоп-слів, стемінг та лематизація, морфологічній аналіз.

Тема 3. Моделювання мови, мовних процесів

Статистичні мовні моделі, n-грами та нейронні моделі мови (наприклад, Word2Vec, GloVe).

Тема 4. Інформаційний пошук та текстова класифікація

Базове розуміння моделей векторного простору та їх реалізації. Що таке термін "частота", обернена до документу частота (TF-IDF) і для чого використовується TF-IDF та як її обчислювати. Поняття схожість документів і класифікації текстів, алгоритми класифікації текстів (наприклад, Naive Bayes, Support Vector Machines) та приклади їх використання.

Тема 5. Аналіз настроїв та думок

Аналіз настроїв на рівні документа, аналіз настроїв на основі аспектів та аналіз думок на основі даних із соціальних мереж.

Тема 6. Розпізнавання іменованих об'єктів та зв'язків між сутностями

Підходи до розпізнавання іменованих сутностей, зв'язування сутностей та використання баз знань (напр., DBpedia, Wikidata) для вилучення сутностей.

Тема 7. Тематичне моделювання

Виявлення прихованих тем у колекції документів. Прихований розподіл Діріхле (Latent Dirichlet Allocation, LDA). Невід'ємна матрична факторизація (NMF).

Тема 8. Реферування тексту

Методи стискання текстів у стислі та зв'язні резюме. Екстрактивне та абстрактне реферування. Метрики для оцінки якості згенерованих рефератів, такі як ROUGE, BLEU та METEOR.

Тема 9. Машинний переклад

Моделі автоматичного перекладу тексту з однієї мови на іншу. Нейронний машинний переклад, архітектури глибокого навчання.

Тема 10. Системи "питання-відповідь", чат-боти

Архітектури систем автоматичної відповіді на запитання (Question Answering). Моделі глибокого навчання, такі як BERT та архітектури на основі трансформерів.

Теми практичних занять

Практичні роботи в рамках дисципліни не передбачені.

Теми лабораторних робіт

Тема 1. Початковий аналіз текстової інформації

Методи передобробки текстової інформації, використання регулярних виразів для обробки текстової інформації.

Тема 2. Статистичні методи обробки текстової інформації

Використання бібліотеки NLTK, Pymorphy, WordNet, n-grams. Методи лематизації, стемінга та морфологічного аналізу (part-of-speach tagging).

Тема 3. Пошук лінгвістичних властивостей у тексті

Вилученні лінгвістичних особливостей з текстів, таких як POS-теги, синтаксичний розбір залежностей, та лематизація. Використання таких пакетів як Spacy, NLTK та Stanza. Робота зі словосполучення та фразовими дієсловами.

Тема 4. Векторне представлення слів

Використання та побудова векторних представлень слів, зокрема Fastext, GloVe та Word2Vec. Розв'язання різних задач з використанням векторних представлень слів.

Тема 5. Вилучення іменованих сутностей

Задачі з розпізнавання іменованих об'єктів (NER), зокрема, розпізнавання об'єктів у резюме та пошук місць (локацій) з використанням різних підходів, таких як POS-based, пакетів Spacy, Stanza, та пошуку залежностей.

Тема 6. Класифікація текстів

Класифікація текстів за допомогою методів класичного машинного навчання. Реалізація задачі сентимент аналізу.

Тема 7. Тематичне моделювання

Реалізація класичних алгоритмів тематичного моделювання (LDA) на своїх даних.
Тема 8. Побудова системи машинного перекладу.
Реалізація системи машинного перекладу для української мови на основі класичних методів "sequence to sequence" та за допомогою нейромережевих підходів.
Тема 9. Побудова свого чат-боту
Використання вже натренованої моделі, яка базується на трансформерах, у застосунку telegram. Або донавчення такої моделі на своїх даних та використання у якості чат-боту в telegram.

Самостійна робота

Самостійна робота передбачає індивідуальне завдання за вибраною студентом темою. Також студентам рекомендуються додаткові ресурси (відео та статті) для самостійного ознайомлення.

Література та навчальні матеріали

Книжки:
"Speech and Language Processing" by Daniel Jurafsky and James H. Martin.
"Natural Language Processing in Action" by Lane, Howard, and Hapke.
"Foundations of Statistical Natural Language Processing" by Christopher D. Manning and Hinrich Schütze.
"Deep Learning for NLP and Speech Recognition" by Palash Goyal, Sumit Pandey, Karan Jain.
Online курси:
Coursera: Natural Language Processing Specialization
Udemy: Natural Language Processing with Deep Learning in Python
Наукові статті:
"Attention is All You Need" by Vaswani et al. (Transformer model).
"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" by Devlin et al.
"Word2Vec" by Mikolov et al.
"GloVe: Global Vectors for Word Representation" by Pennington et al.
Блогі та навчальні матеріали:
The Annotated Transformer
Jay Alammar's Visualizing Neural Machine Translation
Chris Olah's Understanding LSTM Networks
Наукові журнали на конференції:
Association for Computational Linguistics (ACL): https://www.aclweb.org/
Conference on Empirical Methods in Natural Language Processing (EMNLP): https://www.emnlp2022.org/
Додаткові ресурси:
Hugging Face Transformers - A library for working with state-of-the-art natural language processing.
AllenNLP - An open-source NLP research library, built on PyTorch.

Система оцінювання

Критерії оцінювання успішності студента та розподіл балів

100% підсумкової оцінки складаються з результатів оцінювання у вигляді іспиту (40%) та поточний практичних (60%).

Іспит: письмове завдання у вигляді тесту (з відкритими та закрити завданнями) та усна доповідь.
Поточне оцінювання: лабораторні заняття.

Шкала оцінювання

Сума балів	Національна оцінка	ECTS
90–100	Відмінно	A
82–89	Добре	B
75–81	Добре	C
64–74	Задовільно	D
60–63	Задовільно	E
35–59	Незадовільно (потрібне додаткове вивчення)	FX
1–34	Незадовільно (потрібне повторне вивчення)	F

Норми академічної етики і політика курсу

Студент повинен дотримуватися «Кодексу етики академічних взаємовідносин та доброчесності НТУ «ХПІ»: виявляти дисциплінованість, вихованість, доброзичливість, чесність, відповідальність. Конфліктні ситуації повинні відкрито обговорюватися в навчальних групах з викладачем, а при неможливості вирішення конфлікту – доводитися до відома співробітників дирекції інституту. Нормативно-правове забезпечення впровадження принципів академічної доброчесності НТУ «ХПІ» розміщено на сайті: http://blogs.kpi.kharkov.ua/v2/nv/akademichna-dobrochesnist/

Погодження

Силабус погоджено

Дата погодження, підпис
31.08.2023 р.

Завідувач кафедри
Олена АХІЄЗЕР

Дата погодження, підпис
31.08.2023 р.

Гарант ОП
Олена АХІЄЗЕР