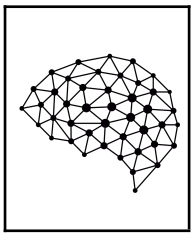
Національний технічний університет України
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»



Кафедра штучного інтелекту

ОБРОБКА ПРИРОДНИХ МОВ З ВИКОРИСТАННЯМ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	Другий (магістерський)
Галузь знань	12 Інформаційні технології
Спеціальність	122 Комп'ютерні науки
Освітня програма	Системи і методи штучного інтелекту
Статус дисципліни	За вибором студентів
Форма навчання	очна(денна)/дистанційна/змішана
Рік підготовки, семестр	1 курс, весняний семестр
Обсяг дисципліни	5 кредитів ЄКТС, 150 годин, 36 г. лекцій, 36 г. лабораторних робіт, 78 г. СРС
Семестровий контроль/ контрольні заходи	Екзамен
Розклад занять	Лекції один раз на тиждень. Практичні заняття один раз на 2 тижні.
Мова викладання	Українська
Інформація про керівника курсу / викладачів	Лектор: Баздирев Антон Андрійович, bazdyrev.anton@gmail.com Лабораторні: Баздирев Антон Андрійович, bazdyrev.anton@gmail.com
Розміщення курсу	Кампус КПІ ecampus.kpi.ua

Програма навчальної дисципліни

1. Опис навчальної дисципліни, її мета, предмет вивчення та результати навчання

Програма дисципліни "Обробка природних мов за допомогою трансформерів" передбачає вивчення статистичних методів обробки тексту, методів глибокого навчання для обробки текстів, класифікації, кластеризації та генерації тексту. В процесі вивчення

студенти використовують мову програмування Python.

Дисципліна "Обробка природних мов за допомогою трансформерів" вивчається в другому семестрі на 1 курсі другого (магістерського) рівня вищої освіти і їй передують дисципліни базової і професійної та практичної підготовки, які вивчаються в період навчання на першому рівні (бакалаврська підготовка) напряму підготовки (спеціальності) 122 «Комп'ютерні науки» і забезпечують можливість її вивчення.

Серед них можна виділити наступні дисципліни: «Математичний аналіз», «Дискретна математика», «Алгоритмізація та програмування», «Алгоритми та структури даних», «Об'єктно-орієнтоване програмування», «Проектування та аналіз обчислювальних алгоритмів».

Кредитний модуль доповнює формування у студентів таких загальних та фахових компетентностей:

ЗК 1: Здатність до абстрактного мислення, аналізу та синтезу.

ЗК 2: Здатність застосовувати знання у практичних ситуаціях.

ЗК 5: Здатність вчитися й оволодівати сучасними знаннями.

ФК 5: Здатність розробляти, описувати, аналізувати та оптимізувати архітектурні рішення інформаційних та комп'ютерних систем різного призначення.

ФК 6: Здатність застосовувати існуючі і розробляти нові алгоритми розв'язування задач у галузі машинного навчання.

та програмних результатів навчання:

ПРН 1: Мати спеціалізовані концептуальні знання, що включають сучасні наукові здобутки у сфері комп'ютерних наук і є основою для оригінального мислення та проведення досліджень, критичне осмислення проблем у сфері комп'ютерних наук та на межі галузей знань.

ПРН 2: Мати спеціалізовані уміння/навички розв'язання проблем комп'ютерних наук, необхідні для проведення досліджень та/або провадження інноваційної діяльності з метою розвитку нових знань та процедур.

ПРН 11: Створювати нові алгоритми розв'язування задач у сфері комп'ютерних наук, оцінювати їх ефективність та обмеження на їх застосування.

ПРН 26 Розробляти адекватні методи навчання та самонавчання, включаючи методи глибокого навчання (Deep Learning) та використовувати їх для налаштування нейронних мереж для вирішення конкретних задач прогнозування, керування, класифікації та інтелектуального аналізу даних.

ПРН 28 Розробляти та використовувати алгоритми розпізнавання зображень та мовних сигналів в системах розпізнавання образів та класифікації в різних предметних областях.

2. Пререквізити та постреквізити дисципліни (місце в структурно-логічній схемі навчання за відповідною освітньою програмою)

Кредитні модулі, які передують вивченню дисципліни і забезпечують її вивчення, відносяться до дисциплін циклу професійної підготовки:

- Математичний аналіз;
- Дискретна математика;
- Алгоритмізація та програмування;
- Алгоритми і структури даних;
- Об'єктно-орієнтоване програмування;
- Проектування та аналіз обчислювальних алгоритмів.

3. Зміст навчальної дисципліни

№ з/п	Назва теми лекції та перелік основних питань (перелік дидактичних засобів, завдання на СРС з посиланням на літературу)
1	Лекція 1. Огляд основних завдань та методів обробки природних мов Завдання на СРС: Повторення лекційного матеріалу.
2	Лекція 2. Техніки попередньої обробки текстів: токенізація, стемінг та лематизація. Завдання на СРС: Повторення лекційного матеріалу.
3	Лекція 3. Моделі мови: статистичні та нейронні підходи Завдання на СРС: Повторення лекційного матеріалу.
4	Лекція 4. Побудова та використання словників та корпусів текстів Завдання на СРС: Повторення лекційного матеріалу.
5	Лекція 5. Векторне представлення слів та текстів. Завдання на СРС: Повторення лекційного матеріалу.
6	Лекція 6. Методи машинного навчання для розпізнавання частин мови та іменованих сутностей Завдання на СРС: Повторення лекційного матеріалу.
7	Лекція 7. Класифікація текстів: наївний баєсовський класифікатор, SVM та нейронні мережі Завдання на СРС: Повторення лекційного матеріалу.
8	Лекція 8. Кластеризація текстів та тематичне моделювання Завдання на СРС: Повторення лекційного матеріалу.
9	Лекція 9. Пошук та підсумовування інформації в текстах Завдання на СРС: Повторення лекційного матеріалу.
10	Лекція 10. Аналіз емоцій та настроїв у текстах Завдання на СРС: Повторення лекційного матеріалу.
11	Лекція 11. Машинний переклад текстів: методи та моделі
12	Лекція 12. Глибинне навчання у задачах обробки природних мов Завдання на СРС: Повторення лекційного матеріалу.
13	Лекція 13. Автоматична генерація текстів: методи та застосування Завдання на СРС: Повторення лекційного матеріалу.
14	Лекція 14. Обробка мови для соціальних медіа та інтернету речей Завдання на СРС: Повторення лекційного матеріалу.
15	Лекція 15. Етичні та соціальні питання в обробці природних мов Завдання на СРС: Повторення лекційного матеріалу.
16	Лекція 16. Оптимізація інференсу та квантизація моделей. Завдання на СРС: Повторення лекційного матеріалу.
17	Лекція 17. Розгортання моделей, MLOps. Завдання на СРС: Повторення лекційного матеріалу.
18	Лекція 18. Контрольна робота.

4. Навчальні матеріали та ресурси

Базова література:

1. Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing (3rd ed.). Pearson. ISBN: 978-0131873216.
2. Goldberg, Y. (2017). Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers. ISBN: 978-1627052986.
3. Rao, D., & McMahan, B. (2019). Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning. O'Reilly Media. ISBN: 978-1491978238.
4. Kaggle Inc., "Natural Language Processing Competitions," Kaggle, [Online]. Available: https://www.kaggle.com/competitions?sortBy=grouped&group=featured&segmentId=278.
5. The Stanford Natural Language Processing Group, "Stanford Natural Language Processing Group," Stanford University, [Online]. Available: https://nlp.stanford.edu/.
6. The Association for Computational Linguistics, "ACL Digital Library," ACL, [Online]. Available: https://www.aclweb.org/anthology/.

Навчальний контент

5. Методика опанування навчальної дисципліни (освітнього компонента)

Лекційні заняття

Назви розділів і тем	Кількість годин				
	Всього	у тому числі			
		Лекції	Практичні	Лабораторні	СРС
Розділ 1. Вступ до дисципліни					
Тема 1. Огляд основних завдань та методів NLP	6	2			4
Тема 2. Техніки попередньої обробки текстів	6	2			4
Разом за розділом 1	12	4			8
Розділ 2. Підготовка даних					
Тема 1. Статистичні та нейронні моделі обробки мови	8	4			8
Тема 2. Побудова та використання словників	6	2			8
Разом за розділом 2	14	6			16
Розділ 3. Статистична обробка тестів					
Тема 1. Векторне представлення слів та текстів	8	2		4	8
Разом за розділом 3	8	2		4	8
Розділ 4. Класифікація текстів					
Тема 1. Задача розпізнавання сутностей	8	2		4	4

Назви розділів і тем	Кількість годин				
	Всього	у тому числі			
		Лекції	Практичні	Лабораторні	СРС
Тема 2. Класифікація текстів: наївний баєсовський класифікатор, SVM та нейронні мережі	8	4			4
Разом за розділом 4	16	6		4	8
Розділ 5. Глибинне навчання для обробки текстів					
Тема 1. Кластеризація текстів та тематичне моделювання	6	2			4
Тема 2. Пошук та підсумовування інформації в текстах	7	2			5
Разом за розділом 5	13	4			9
Розділ 6. Рекурентні моделі глибинного навчання					
Тема 1. Аналіз емоцій та настроїв у текстах	13	4		8	5
Разом за розділом 6	13	4		8	5
Розділ 7. Трансформери					
Тема 1. Машинний переклад текстів	10	2		8	4
Тема 2. Глибинне навчання (BERT, RoBERTa)	6	2			4
Разом за розділом 7	16	4		8	8
Розділ 8. Генерація тексту					
Тема 1. Автоматична генерація текстів (LLM, GAN)	7	2			5
Тема 2. Обробка мови для соціальних медіа та інтернету речей (GPT-J, GPT-4)	11	2		8	5
Тема 3. Етичні та соціальні питання в обробці природних мов	4	2			2
Разом за розділом 8	22	6		8	12
Модульна контрольна робота з розділів 2, 3, 4, 6	6	2			4
Залік	2			4	
Всього годин	150	36		36	78

Лабораторні заняття

Основні завдання циклу лабораторних занять полягають у закріпленні теоретичного матеріалу та отриманні студентами практичних навичок з обробки природних мов.

№ з/п	Назва лабораторної роботи	Кількість ауд. годин
1	Розробка моделі автоматичного розпізнавання іменованих сутностей (NER) за допомогою методів машинного навчання. У цій роботі використані різні підходи до розв'язання цієї задачі, включаючи методи на основі правил, методи на основі статистики та методи на основі глибокого навчання.	4
2	Розробка системи автоматичного машинного перекладу (MT), використовуючи моделі глибокого навчання, такі як нейронні мережі з довільною рекурсією (LSTM) або трансформери.	4
3	Розробка моделі класифікації текстів за допомогою методів машинного навчання, таких як наївний баєсовський класифікатор, метод опорних векторів (SVM) або глибокі нейронні мережі.	4
4	Використання техніки трансферного навчання для дотренування великих лінгвістичних моделей	4
5	Екзамен	2

6. Самостійна робота студента

№ з/п	Назви тем і питань, що виноситься на самостійне опрацювання та посилання на навчальну літературу	Кількість годин СРС
1	Огляд основних завдань та методів обробки природних мов Завдання на СРС: Побудувати гістограму частоти слів в корпусі текстів та знайти найбільш часто вживані слова.	4
2	Техніки попередньої обробки текстів: токенізація, стемінг та лематизація. Завдання на СРС: Застосувати токенізацію, стемінг та лематизацію до тексту та порівняти результати	4
3	Моделі мови: статистичні та нейронні підходи Завдання на СРС: Застосувати метод TF-IDF для порівняння документів за темою. Навчити наївний баєсовський класифікатор для класифікації текстів на задані категорії.	4
4	Побудова та використання словників та корпусів текстів Завдання на СРС: Повторення лекційного матеріалу.	4
5	Векторне представлення слів та текстів. Завдання на СРС: Використовуючи векторне представлення слів, побудувати словник слів та вектори речень та порівняти їх за схожістю.	4
6	Методи машинного навчання для розпізнавання частин мови та іменованих сутностей Завдання на СРС: Порівняти різні види NER.	4

№ з/п	Назви тем і питань, що виноситься на самостійне опрацювання та посилання на навчальну літературу	Кількість годин СРС
7	Класифікація текстів: наївний баєсовський класифікатор, SVM та нейронні мережі Завдання на СРС: Порівняти результати роботи наївного баєсовського класифікатора, SVM та нейронної мережі для задачі класифікації текстів.	4
8	Кластеризація текстів та тематичне моделювання Завдання на СРС: Побудувати програму для автоматичного підсумовування довгих текстів.	4
9	Пошук та підсумовування інформації в текстах Завдання на СРС: Побудувати програму для автоматичного підсумовування довгих текстів.	5
10	Аналіз емоцій та настроїв у текстах Завдання на СРС: Реалізувати програму для аналізу емоцій у тексті..	5
11	Машинний переклад текстів: методи та моделі Реалізувати просту систему для машинного перекладу текстів методом заміни слів.	4
12	Глибинне навчання у задачах обробки природних мов Завдання на СРС: Побудувати графік зміни настрою користувачів за допомогою аналізу текстів з соціальних мереж.	4
13	Автоматична генерація текстів: методи та застосування Завдання на СРС: Побудувати графік зміни настрою користувачів за допомогою аналізу текстів з соціальних мереж.	5
14	Обробка мови для соціальних медіа та інтернету речей Завдання на СРС: Повторення лекційного матеріалу.	5
15	Етичні та соціальні питання в обробці природних мов Завдання на СРС: Авторське право згенерованого тексту.	2
16	Модульна контрольна робота Завдання на СРС: підготовка до контрольної роботи	4

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

Вимоги, яких має дотримуватися студент в рамках даної дисципліни:

- відвідування лекційних та лабораторних занять є бажаним, але не обов'язковим;
- під час проведення занять мобільні телефони мають бути переведені у беззвучний режим;
- дозволяється, при необхідності, використання засобів зв'язку для пошуку потрібної інформації на платформі дистанційного навчання та/або в інтернеті;
- лабораторні роботи мають бути виконані та захищені особисто, під час захисту студент повинен відповісти на питання викладача, що стосуються як самої лабораторної роботи, так і теоретичного матеріалу, на якому вона базується;

- заохочувальні бали можуть призначатися за активність на лекціях та нестандартні рішення при виконанні лабораторних робіт;
- штрафні бали можуть призначатися за несвоєчасне виконання лабораторних робіт;
- при виконанні лабораторних робіт потрібно дотримуватися графіка, який доводиться до відома студентів викладачем на початку семестру;
- обов'язковим є дотримання академічної доброчесності.

Таблиця 7.1. Система рейтингових (вагових) балів та критеріїв оцінювання

Категорія оцінювання	Мінімальна оцінка в балах	Максимальна оцінка в балах
Лабораторна робота 1	6	10
Лабораторна робота 2	6	10
Лабораторна робота 3	6	10
Лабораторна робота 4	6	10
Модульна контрольна робота	12	20
Екзамен	24	40

8. Види контролю та рейтингова система оцінювання результатів навчання (РСО)

Поточний контроль: лабораторна робота № 1 (2 частини по 5 балів кожна), лабораторна робота № 2 (2 частини по 5 балів кожна), лабораторна робота № 3 (2 частини по 5 балів кожна), лабораторна робота № 4 (2 частини по 5 балів кожна), МКР (20 балів).

Поточний контроль проводиться методом опитування студентів під час приймання лабораторних робіт. За кожний тиждень затримки виконання лабораторної роботи нараховуються штрафні бали: – 2 бали (але не більш ніж 10 балів на обидві частини лабораторної роботи).

Для проведення модульного контролю передбачено проведення однієї модульної контрольної роботи з розділів 2, 3, 4, 6. Метою роботи є перевірка засвоєння студентами знань, що стосуються обробки текстів. Передбачено 8 варіантів індивідуальних завдань для проведення контрольної роботи, по 4 завдання у кожному. Максимальна кількість балів за кожне завдання МКР – 5. Критерії оцінки кожного завдання модульної контрольної роботи:

- відповідь абсолютно вірна, наведено всі необхідні пояснення – 5 балів;
- хід розв'язання правильний, але наявні деякі помилки – 4 бали;
- присутня правильна ідея розв'язання задачі, але не доведена до кінця – 3 бали;
- відповідь неправильна або відсутня, задача розв'язана неправильно – 0 балів.

Календарний контроль: провадиться двічі на семестр як моніторинг поточного стану виконання вимог силабусу.

Умова першого календарного контролю – зарахування першої лабораторної роботи, умова другого календарного контролю – зарахування двох лабораторних робіт.

Семестровий контроль: екзамен.

Умова допуску до семестрового контролю: зарахування усіх лабораторних робіт.

Сума рейтингових балів, отриманих студентом протягом семестру і набраних балів за екзамен, переводиться до підсумкової оцінки згідно з таблицею.

Таблиця відповідності рейтингових балів оцінкам за університетською шкалою:

Кількість балів	Оцінка
100-95	Відмінно
94-85	Дуже добре
84-75	Добре
74-65	Задовільно
64-60	Достатньо
Менше 60	Незадовільно
Не виконані умови допуску	Не допущено

9. Додаткова інформація з дисципліни (освітнього компонента)

При вивченні матеріалу курсу рекомендується особливу увагу приділити закріпленню базових положень, поступово ускладнюючи матеріал із наведенням відповідних прикладів задач. Під час МКР дозволяється використання конспекту, що зменшує психологічний тиск на студентів та стимулює більш активну роботу на лекціях. Постійний зворотній зв'язок зі студентами через сучасні засоби комунікації дозволяє швидко вирішувати труднощі в навчанні та є засобом індивідуального навчання.

За погодженням з викладачем, студент має можливість пройти дистанційні чи онлайн курси за відповідною тематикою та зарахувати отримані сертифікати як додаткові бали до рейтингу (не більше 10 балів).

Приклади питань, які виносяться на модульну контрольну роботу:

- Які техніки та алгоритми використовуються для побудови моделей обробки природних мов?
- Які основні методи та техніки для обробки тексту можуть бути використані для вирішення задач в NLP?
- Які гіперпараметри можуть впливати на якість роботи моделей обробки природних мов?
- Як можна оцінити якість роботи моделі обробки природних мов?
- Які методи можуть бути використані для зменшення впливу англійської мови на розробку моделей обробки природних мов для інших мов?
- Які основні проблеми можуть виникати при розробці моделей обробки природних мов, які працюють з іншими мовами?
- Як можна застосовувати моделі обробки природних мов для різних завдань, таких як розпізнавання іменованих сутностей, класифікація текстів або машинний переклад?

Робочу програму навчальної дисципліни (силабус):

Складено Баздиревом Антоном Андрійовичем

Ухвалено кафедрою штучного інтелекту (протокол № 14 від 11.06.2024)

Погоджено Методичною комісією ННІПСА (протокол № 10 від 24.06.2024)