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

Кафедра інженерії програмного забезпечення в енергетиці

МЕТОДИ ОБРОБКИ ПРИРОДНОЇ МОВИ

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	**Перший (бакалаврський)**
Галузь знань	12 Інформаційні технології
Спеціальність	121 Інженерія програмного забезпечення
Освітня програма	Інженерія програмного забезпечення інтелектуальних кібер-фізичних систем в енергетиці
Статус дисципліни	Вибіркова
Форма навчання	очна (денна)
Рік підготовки, семестр	3 курс, весняний семестр
Обсяг дисципліни	4,0 кредити ECTS /120 годин (36 годин лекцій, 18 годин практичних занять, 66 годин — самостійна робота студента)
Семестровий контроль/ контрольні заходи	Залік, МКР, представлення робіт комп'ютерного практикуму
Розклад занять	http://rozklad.kpi.ua
Мова викладання	Українська
Інформація про керівника курсу / викладачів	Лектор: к.т.н., доцент, Стативка Юрій Іванович, statyvka-yu@lll.kpi.ua Практичні заняття: к.т.н., доцент, Стативка Юрій Іванович, statyvka-yu@lll.kpi.ua
Розміщення курсу	https://ecampus.kpi.ua

Програма навчальної дисципліни

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

В курсі розглядаються як концептуальні основи комп'ютерної лінгвістики, необхідні розробнику систем обробки природної мови (NLP – Natural language Processing), так і методи, алгоритми та технології для практичної їх побудови.

Значна увага приділяється розгляду класичних та новітніх методів і засобів лексичного, морфологічного, синтаксичного, семантичного аналізу та моделювання природної мови, та їх застосуванню до української мови.

Метою дисципліни є опанування студентами теоретичних знань та набуття практичного досвіду проєктування і побудови NLP-застосунків.

Предмет дисципліни — методи та засоби розробки NLP-застосунків.

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

Пререквізити дисципліни. Знання, отримані при вивченні дисциплін: «Основи програмування» «Об’єктно-орієнтований аналіз та конструювання програмних систем», «Комп'ютерна дискретна математика», «Алгоритми та структури даних».

Постреквізити дисципліни. Компетенції, отримані студентами в процесі вивчення дисципліни можуть бути використані при виконанні дипломної роботи.

3. Зміст навчальної дисципліни

Розділ 1 Моделювання мові і мовлення

Тема 1.1. Вступ до обробки природної мови.
Тема 1.2. Мовлення: розпізнавання та синтез.
Тема 1.3. Методи обробки елементів тексту.
Тема 1.4. Мовні ресурси і завдання NLP.
Тема 1.5. Діалогова система на основі правил.
Тема 1.6. Штучний інтелект (ШІ) в контексті NLP.
Тема 1.7. Мовні моделі і великі мовні моделі.

Розділ 2 Часткові завдання NLP-систем

Тема 2.1. Пошук шаблонів (pattern) у тексті.
Тема 2.2. Частиномовний та синтаксичний аналіз.
Тема 2.3. Семантичний аналіз.
Тема 2.4. Представлення тексту.
Тема 2.5. Генерування тексту засобами ШІ.
Тема 2.6. Генерування тексту засобами LLM.

Розділ 3 Застосування NLP-систем

Тема 3.1. Класифікація текстів.
Тема 3.2. Анотування тексту.
Тема 3.3. Вилучення даних.
Тема 3.4. Переклад.
Тема 3.5. Методи NLP: підсумки.

4. Навчальні матеріали та ресурси

Основна література

1. D. Jurafsky, J.H. Martin: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 3nd ed. https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf
2. Методи обробки природної мови: Феномен ChatGPT [Текст]: навч. посіб. для здобувачів ступеня бакалавра за освітньою програмою «Інженерія програмного забезпечення інтелектуальних кібер-фізичних систем в енергетиці» / автор: Ю. І. Стативка; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 621 кбайт). – Київ: КПІ ім. Ігоря Сікорського, 2023. – 67 с.

3. Hobson Lane. Natural Language Processing in Action. Understanding, analyzing, and generating text with Python/ Hobson Lane, Cole Howard, Hannes Hapke. – Manning Publications Co., 2019. – 544 p.

Додаткова література

4. J. Thanaki: Python Natural Language Processing. // Packt Publishing, 2017.
5. K. Bhavsar, N. Kumar, P. Dangeti: Natural Language Processing with Python Cookbook. // Packt Publishing, 2017.
6. Lewis Tunstall. Natural Language Processing with Transformers, Revised Edition / Lewis Tunstall, Leandro von Werra, Thomas Wolf. – Beijing-Boston-Farnham-Sebastopol-Tokyo : O'Reilly Media, Inc., 2022. – 383 p.
7. Sowmya Vajjala. Practical Natural Language Processing / Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana / - O'Reilly Media, 2020 — 456 p.
8. D. Chopra, N. Joshi, I. Mathur: Mastering Natural Language Processing with Python. // Packt Publishing, 2016.
9. N. Hardeniya, J. Perkins, D. Chopra, N. Joshi, I. Mathur: Natural Language Processing: Python and NLTK. // Packt Publishing, 2016.
10. N. Hardeniya: NLTK Essentials. // Packt Publishing, 2015.
11. S. Bird, E. Klein, E. Loper: Natural Language Processing with Python. // Packt Publishing, 2015.
12. J. Perkins: Python 3 Text Processing with NLTK 3 Cookbook. // Packt Publishing, 2014.
13. A. Clark, C. Fox, S. Lappin (eds.): The Handbook of Computational Linguistics and Natural Language Processing. // Wiley-Blackwell, 2010.

Інформаційні ресурси

14. https://github.com/LinguisticAndInformationSystems/mphdict
15. https://lang.org.ua/uk/
16. https://github.com/UniversalDependencies/UD_Ukrainian-IU
17. https://mova.institute/
18. https://spacy.io/models/uk

Навчальний контент

5. Методика опанування навчальної дисципліни (освітнього компонента)

Лекційні заняття

Розділ 1 Моделювання мови і мовлення

Лекція 1. Вступ до обробки природної мови.
Обробка природної мови (Natural language Processing, NLP). Завдання обробки природної мови. Природна мова як система. Методи обробки природної мови. Сучасний стан NLP: класичні підходи і штучний інтелект, великі мовні моделі (Large Language Model, LLM). Мотиваційний приклад — діалогова система.

Лекція 2. Мовлення: розпізнавання та синтез.
Поняття про фонетико-акустичні та просодичні характеристики мовлення. Методи розпізнавання та синтезу мовлення. Використання у програмних застосунках засобів розпізнавання та синтезу усного мовлення.

Лекція 3. Методи обробки елементів тексту.
Токенізація, стемінг, лематизація, сегментація. Фонетичні алгоритми. Нечіткий пошук.

Лекція 4. Мовні ресурси і завдання NLP.
Мовні ресурси української мови: первинні - словники, корпуси текстів; вторинні — мовні моделі. словники, корпуси текстів. Словники і корпуси текстів української мови — методи доступу і варіанти використання.

Лекція 5. Діалогова система на основі правил.
Чат-боти і віртуальні асистенти. Архітектура. Функції. Засоби реалізації.

Лекція 6. Штучний інтелект (ШІ) в контексті NLP.
Архітектура. Функції. Засоби реалізації.

Лекція 7. Мовні моделі і великі мовні моделі.
Мовні моделі і великі мовні моделі для української мови — методи доступу і варіанти використання.

Розділ 2 Часткові завдання NLP-систем

Лекція 8. Пошук шаблонів (pattern) у тексті.
Пошук у послідовності слів, у послідовності властивостей. Пошук у дереві.

Лекція 9. Частиномовний та синтаксичний аналіз.
POS-тегування. Граматика з фразовою структурою. Граматика залежностей.

Лекція 10. Семантичний аналіз.
Розв’язання анафори. Розпізнавання іменованих сутностей.

Лекція 11. Представлення тексту.
Векторне представлення. Набір слів. n-грами. Нейромережеві моделі.

Лекція 12. Генерування тексту засобами ШІ.
Застосування методів ШІ до завдань генерування тексту.

Лекція 13. Генерування тексту засобами LLM.
На прикладах: GPT (Generative Pre-trained Transformer), LLaMA (Large Language Model Meta AI), BERT (Bidirectional Encoder Representations from Transformers). Переваги і недоліки. Галюцинації LLM.

Розділ 3 Застосування NLP-систем

Лекція 14. Класифікація текстів.
Методи класифікації: класичні і нейромережеві.

Лекція 15. Анотування тексту.
Застосування методів ШІ до завдань анотування текстів.

Лекція 16. Вилучення даних.
Методи вилучення даних з неструктурованого тексту.

Лекція 17. Переклад.
Методи і засоби автоматизованого перекладу.

Лекція 18. Методи NLP: підсумки.
Огляд курсу: концепти, технології, інструменти. Набутий досвід у контексті методів обробки української мови.

Практичні заняття

Практичне заняття 1. Інструменти та технології для розробки NLP-застосунків.
Текстовий і голосовий echo-бот (діалоговий застосунок). Українські SST і TTS.
БД для віртуального асистента.

*Практичне заняття 2.*Діалог з використанням LLM.
Представлення роботи комп'ютерного практикуму № 1 (**КП 1**).

*Практичне заняття 3.*Діалог, заснований на правилах.
Пошук шаблонів. Представлення **КП 2**.

*Практичне заняття 4.*Обробка словникових ресурсів української мови.
Визначення лем, грамем, синонімів. Визначення частоти слів, словоформ, словосполучень.
Відстань між рядками. Фонетичні алгоритми. Нечіткий збіг

Практичне заняття 5. Мовні моделі в діалогових застосунках.
Частиномовний аналіз. Іменовані сутності. Синтаксична структура репліки.
Представлення **КП 3**.

Практичне заняття 6. Великі мовні моделі в діалогових застосунках.
Використання LLM у проєкті віртуального асистента. Тонке налаштування LLM/

Практичне заняття 7. LLM в діалогових застосунках.
Варіанти використання LLM. Представлення **КП 4**.

Практичне заняття 8. Недіалогові завдання NLP.
Виявлення термінів. Класифікація текстів. Анотування. Автоматичний переклад.

Практичне заняття 9. Підсумкове заняття.

Перелік робіт комп'ютерного практикуму

1. КП 1. Діалогові програми: echo-бот і чат з використанням LLM.
2. КП 2. Віртуальний асистент, заснований на правилах.
3. КП 3. Віртуальний асистент з використанням мовних моделей.
4. КП 4. Віртуальний асистент з використанням LLM.

Контрольна робота

Метою контрольної роботи є закріплення та перевірка теоретичних знань із освітнього компонента, набуття студентами практичних навичок самостійного вирішення задач.

Модульна контрольна робота (МКР) виконується на шістнадцятому тижні семестру. Кожне завдання містить одне теоретичне питання і одне практичне завдання.

Самостійна робота студента

Приблизний розподіл часу СРС

№ з/п	*Вид самостійної роботи*	*Кількість*	*Кількість годин СРС*
1	Виконання робіт комп'ютерного практикуму	4	24
2	Підготовка до практичних занять	9	36
3	Підготовка до МКР	1	6
		Разом	66

Політика та контроль

6. Політика навчальної дисципліни (освітнього компонента)

Система вимог при вивченні дисципліни:

- правила відвідування занять: заборонено оцінювати присутність або відсутність здобувача на аудиторному занятті, в тому числі нараховувати заохочувальні або штрафні бали. Відповідно до РСО даної дисципліни бали нараховують за відповідні види навчальної активності на лекційних та практичних заняттях.
- правила поведінки на заняттях: студент має можливість отримувати бали за відповідні види навчальної активності на лекційних та практичних заняттях, передбачені РСО дисципліни. Використання засобів зв'язку для пошуку інформації на гугл-диску викладача, в інтернеті, в дистанційному курсі на платформі Сікорський здійснюється за умови вказівки викладача;
- політика дедлайнів та перескладань: якщо студент не проходив або не з'явився на МКР (без поважної причини), його результат оцінюється у 0 балів. Перескладання результатів МКР не передбачено;
- політика щодо академічної доброчесності: Кодекс честі Національного технічного університету України «Київський політехнічний інститут» https://kpi.ua/files/honorcode.pdf встановлює загальні моральні принципи, правила етичної поведінки осіб та передбачає політику академічної доброчесності для осіб, що працюють і навчаються в університеті, якими вони мають керуватись у своїй діяльності, в тому числі при вивченні та складанні контрольних заходів з дисципліни «Методи обробки природної мови»;
- при використанні цифрових засобів зв'язку з викладачем (мобільний зв'язок, електронна пошта, переписка на форумах та у соцмережах тощо) необхідно дотримуватись загальноприйнятих етичних норм, зокрема бути ввічливим та обмежувати спілкування робочим часом викладача.

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

Система рейтингових (вагових) балів та критерії оцінювання

***Поточний контроль**: активність — питання і відповіді на лекційних заняттях, питання і відповіді та участь у ревю на практичних заняттях; МКР; виконання завдань до практичних занять; виконання та представлення робіт комп'ютерного практикуму.*
***Календарний контроль**: провадиться двічі на семестр як моніторинг поточного стану виконання вимог силабусу ("атестація").*
Семестровий контроль: *залік.*
***Умови допуску до семестрового контролю**: семестровий рейтинг — не менше 40 балів.*
Таблиця відповідності рейтингових балів оцінкам за університетською шкалою:

Бали	**Оцінка**
95 - 100	Відмінно
85 - 94	Дуже добре
75 - 84	Добре
65 - 74	Задовільно
60 - 64	Достатньо
Менше 60	Незадовільно
R < 40	Не допущено

- Активність

Активність (частота, змістовність) участі студента у процесі обговорення відповідних тем на заняттях — питання і відповіді на лекційних заняттях, питання і відповіді та участь у ревю на практичних заняттях; виконання завдань до практичних занять; оцінюється, максимум, 20 балами, які може отримати кожен студент за семестр.

- Роботи комп'ютерного практикуму

Максимальна кількість балів за виконання кожної роботи комп'ютерного практикуму становить 15 балів.

Критерії оцінювання:
Виконання робіт комп'ютерного практикуму:

- виконана у повному обсязі – 15 балів;
- виконана частково — відповідна частка від максимальної кількості балів.

- Модульна контрольна робота

Максимальна кількість балів за модульну контрольну роботу дорівнює 20 балів.
Якість виконання роботи:

- виконана у повному обсязі з необхідними текстовими поясненнями дій та результатів – максимальна кількість балів;
- виконана частково з поясненнями — відповідна частка від максимальної кількості балів;
- виконана без текстових пояснень дій та результатів – не більше чотирьох балів.

- Творче завдання

Студент може обрати додаткове завдання творчого характеру, результати виконання якого можуть бути опубліковані у наукових виданнях, або повідомлені на студентській науковій конференції з публікацією тез. Максимальна кількість балів за виконання творчого завдання — 30 балів, за умови, що загальна кількість балів не перевищує 100 балів.

Розрахунок шкали (R) рейтингу:

Сума вагових балів контрольних заходів протягом семестру (шкала рейтингу) складає:

$$R = r_{АКТ} + r_{КП} + r_{МКР} + r_{ТвЗ} = 20 + 60 + 20 + (\text{до } 30) = 100 \text{ балів.}$$

Робочу програму навчальної дисципліни (силабус) «Методи обробки природної мови»:

Склав доцент кафедри ІПЗЕ, к.т.н., доц. Стативка Юрій Іванович

Ухвалено кафедрою ІПЗЕ (протокол № 34 від 10.05.2024 р)

Погоджено Методичною комісією НН ІАТЕ ім.Ігоря Сікорського
(протокол № 9 від 31.05.2024 р.)