НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
Центр наукових досліджень та викладання іноземних мов

«ЗАТВЕРДЖУЮ»
Директор Центру наукових досліджень
та викладання іноземних мов НАН України
к.філол.н., доцент
ЖАЛАЙ В.Я.

«21» листопада 2023 р.

СИЛАБУС
ВК 05 ДВА

Основи комп'ютерної лінгвістики

підготовки здобувачів третього (освітньо-наукового) рівня вищої освіти – доктора філософії – спеціальності 035 «Філологія»

Київ 2023

1. Загальна інформація про навчальну дисципліну	
Повна назва навчальної дисципліни	Основи комп'ютерної лінгвістики
Повна офіційна назва закладу вищої освіти	Центр наукових досліджень та викладання іноземних мов НАН України (м. Київ)
Повна назва структурного підрозділу	кафедра іноземних мов
Розробник	Крамар Н.А., доктор філософії з філології
Рівень вищої освіти	Третій (освітньо-науковий) рівень НРК України – 9 рівень QF for EHEA – третій цикл, EQF for LLL – 8 рівень
Семестр вивчення навчальної дисципліни	1-2 семестр 3 року навчання (5-6 семестри)
Обсяг навчальної дисципліни	3 кредити ЄКТС / 90 годин, з яких 10 годин – лекції 8 годин – практичні заняття 72 годин – самостійна робота
Мова(и) викладання	англійська, українська
Викладач(і)	Крамар Н.А., доктор філософії з філології
2. Місце навчальної дисципліни в освітній програмі	
Статус дисципліни	вибірковий компонент, дисципліна вибору аспіранта ВК 05 ДВА
Передумови для вивчення дисципліни	Другий рівень вищої освіти (диплом магістра). Аспірант повинен знати: англійську мову на рівні не нижче С1 та мати навички пошуку та аналізу наукової інформації.
Додаткові умови	Додаткові умови відсутні.
Обмеження	Обмеження відсутні.
3. Мета навчальної дисципліни	
Мета: курс передбачає ознайомлення студентів з основами комп'ютерної лінгвістики та надання їм практичних навичок у застосуванні інструментів і методів аналізу тексту для різноманітних завдань, таких як екстракція даних (data mining), аналіз настроїв (sentiment analysis), тематичне моделювання, класифікація текстів тощо.	
4. Зміст навчальної дисципліни	

1-й модульний цикл
Тема 1. Місце комп'ютерної лінгвістики та NLP у сучасному лінгвістичному ландшафті.
Тема 2. Основи програмування у Python для NLP-задач.
2-й модульний цикл
Тема 3. Попередня обробка текстових даних у Python: бібліотеки, методи, підходи.
Тема 4. Обробка текстових даних: від морфологічного до семантичного рівня.
Тема 5. Основи машинного навчання для NLP-задач.
3-й модульний цикл
Тема 6. Векторна репрезентація слів.
Тема 7. Тематичне моделювання та класифікація текстових даних.

5. Очікувані результати навчання навчальної дисципліни

У результаті вивчення навчальної дисципліни аспірант повинен знати:

- роль комп'ютерної лінгвістики у сучасному лінгвістичному ландшафті
- історію становлення комп'ютерної лінгвістики
- основні поняття комп'ютерної лінгвістики такі як: корпус, датасет, токен, лема, стемінг, n-грам (n-gram), парсинг, синтаксичний аналіз, морфологічний аналіз, векторизація, частотний аналіз, вбудування слів (word embedding).
- особливості аналізу тексту та екстрагування інформації у Python
- особливості лематизації та стемінгу мов з різних мовних сімей
- специфіку NLP-бібліотек Python і їхню доцільність для різних завдань комп'ютерної лінгвістики
- методи word embedding (вбудування слів)
- методи векторної репрезентації слів
- основні алгоритми тематичного моделювання (LDA, LSA, PLSA)
- відмінність між класифікацією на основі правил (rule-based classification) та класифікацією на основі машинного навчання
- етичні проблеми, пов'язані з веб-скрейпінгом та аналізом чутливих даних (sensitive data)
- принципи функціонування великих мовних моделей (large language models)
- можливості доналаштування (fine-tuning) наявних мовних моделей
- основні корпуси української, англійської та інших мов та способи їх використання для цілей NLP

вміти:

• використовувати основні NLP-бібліотеки Python (Spacy, NLTK, Textblob, Stanza, Pandas та інші)
• створювати функції у Python
• обробляти дані різних форматів та виконувати конвертацію з одного формату в інший (txt, pdf, csv, xml тощо)
• виконувати аналіз частотності та аналіз ключових слів
• виділяти основні n-грами у текстових даних
• виконувати попередню обробку (preprocessing) даних (видалення пунктуації та нерелевантних символів, видалення стоп-слів, нормалізація правопису тощо)
• здійснювати лематизацію та стемінг мовних даних
• застосовувати SpacyMatchers та Regex для екстрагування інформації потрібного формату
• виконувати тематичне моделювання та класифікацію даних
• етично використовувати чатботи на основі великих мовних моделей та створювати ефективні промти для них
• візуалізувати результати аналізу у вигляді діаграм та схем
• чітко викладати та презентувати результати застосування методів комп'ютерної лінгвістики в усній та письмовій формах

6. Роль навчальної дисципліни у досягненні програмних результатів

Завдання навчальної дисципліни ВК 05 ДВА «Основи комп'ютерної лінгвістики» полягає у формуванні та набутті таких компетентностей: загальні компетентності: ЗК: 1, 3, 4, 5 (відповідно до переліку загальних компетентностей ОНП). Фахові компетентності: ФК 2, 4, 5, 6, 7, 8, 9, 10, 13, 14 (відповідно до переліку фахових компетентностей ОНП). Програмні результати навчання: ПРН 1.4, 2.1, 2.2, 2.3, 4.2, 4.4 (відповідно до переліку програмних результатів навчання ОНП).

6.1 Види навчальних занять

Видами навчальних занять при вивченні дисципліни є лекції (Л), практичні (індивідуальні) заняття (П) та самостійна робота (С).

Навчально-тематичний план лекцій і практичних занять

5-6 семестри 3 року навчання

Тема	Назва теми	Кількість годин

		Лекції	Практичні (індивідуальні)	Самостійна робота
1.	Тема 1. Місце комп'ютерної лінгвістики та NLP у сучасному лінгвістичному ландшафті.	1	1	10
2.	Тема 2. Основи програмування у Python для NLP-задач.	2	2	20
3.	Тема 3. Попередня обробка текстових даних у Python: бібліотеки, методи, підходи.	2	1	20
4.	Тема 4. Обробка текстових даних: від морфологічного до семантичного рівня.	2	1	10
5.	Тема 5. Основи машинного навчання для NLP-задач.	1	1	5
6.	Тема 6. Векторна репрезентація слів та вбудування слів (word embedding).	1	1	2
7.	Тема 7. Тематичне моделювання та класифікація текстових даних.	1	1	5
		10	8	72
Усього		90		

Змістовий модуль 1

Тема 1

Л-1
Місце комп’ютерної лінгвістики та NLP у сучасному лінгвістичному ландшафті
П-1
Можливості застосування комп’ютерної лінгвістики для наукових цілей та бізнес-цілей

Змістовий модуль 2

Тема 2

Л-2
Основи програмування у Python для NLP-задач: змінні, цикли, функції
П-2
Робота з різними типами даних у Python: рядки, списки, числа, словники
Л-3
Оператори, цикли for і while, функції
П-3
Завантаження текстових даних та імпортування бібліотек у Python

Тема 3

Л-4
Принципи попередньої обробки текстових даних у Python
Л-5
Основи роботи з регулярними виразами

П-4
Практикум з попередньої обробки тексту з використанням бібліотеки Spacy

Тема 4

Л-6
Морфологічний аналіз текстових даних: tokenization, POS tagging, n-grams.
Л-7
Синтаксичний та семантичний аналіз текстових даних: dependency parsing, named entity recognition
П-5
Практикум з частиномовної розмітки та побудови дерева залежностей в NLTK та Spacy

Змістовий модуль 3

Тема 5

Л-8
Основи машинного навчання для NLP-задач
П-6

Тренування ML моделей

Тема 6

Л-9

Векторна репрезентація слів та вбудування слів (word embedding)

П-7

Практикум із застосування TF-IDF

Тема 7

Л-10

Тематичне моделювання та класифікація текстових даних

П-8

Порівняння провідних видів тематичного моделювання (LDA, NMF, BERTopic) на корпусі текстів

7.2 Види навчальної діяльності

Підготовка до лекцій.
Підготовка до практичних занять.
Підготовка до виконання індивідуальних завдань за модулями.
Підготовка до іспиту.

8. Методи викладання, навчання

Модульне навчання.
Лекції-візуалізації.
Інтерактивні лекції.
Flipped learning.
Обговорення-дискусії.
Проблемно-пошукові методи.
Міждисциплінарне навчання.

9. Методи та критерії оцінювання

9.1. Критерії оцінювання

Оцінка	Визначення	Національна шкала оцінювання	Рейтингова бальна шкала оцінювання	
ВІДМІННО	відмінне виконання лише з незначною кількістю помилок	5, 0 (відмінно)	$90 \leq RD \leq 100$	

ДОБРЕ	в цілому правильна робота з певною кількістю помилок	4, 0 (добре)	$74 \leq RD \leq 89$	
ЗАДОВІЛЬНО	виконання задовольняє мінімальні критерії	3,0 (задовільно)	$60 \leq RD \leq 73$	
НЕЗАДОВІЛЬНО	можливе повторне складання	2 (незадовільно)	$35 \leq RD \leq 59$	
НЕЗАДОВІЛЬНО	необхідний повторний курс з навчальної дисципліни	2 (незадовільно)	$RD < 35$	

9.2 Методи поточного формативного оцінювання

За дисципліною передбачені такі методи поточного формативного оцінювання:
експрес-контроль, настанови викладача в процесі виконання індивідуальних практичних завдань, перевірка, обговорення та оцінювання виконаної роботи.

9.3 Методи підсумкового сумативного оцінювання

Оцінювання протягом семестру проводиться у формі усних опитувань, письмового тестування, перевірки виконання індивідуальних завдань та дискусій.
Робота повинна бути виконана самостійно.
Оцінювання аспіранта здійснюється таким чином:
1. Експрес-контроль – 30 балів
2. Практичні (індивідуальні заняття) та самостійна робота – 30 балів
3. Залік – 40 балів
Заохочувальні та штрафні бали:
1. Відсутність на лекції без поважних причин - (-) 2 бали.
2. Відсутність на практичних (індивідуальних) заняттях без поважних причин (-) 2
бали.
3. Підготовка публікації до друку та/або виступу на конференції (+) 10 балів
Сума як штрафних, так і заохочувальних балів не має перевищувати 0,1R=10 балів.
Форма підсумкового контролю – іспит.

10. Ресурсне забезпечення навчальної дисципліни

10. 1 Засоби навчання	Навчальний процес потребує використання мультимедійних, друкованих та електронних освітніх ресурсів.
10.2 Інформаційне та навчально методичне забезпечення	**Основна література** 1. Bengfort, B., Bilbro, R., & Ojeda, T. (2018). *Applied Text Analysis with Python*. O'Reilly Media, Inc. 2. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media. 3. Hapke, H., Lane, H., & Howard, C. (2019). *Natural Language Processing in Action*. Manning Publications. 4. Hovy, D. (2012). Programming in Python for Linguists. A Gentle Introduction. Retrieved from http://www.dirkhovy.com/portfolio/papers/download/pfl_handout.pdf **Допоміжна література** 1. Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *6*(7), 886498. 2. Johnson, M. (2011). How relevant is linguistics to computational linguistics? Linguistic Issues in Language Technology, 6(7), 1–23. 3. Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing, 2nd edition*. Pearson Prentice Hall. 4. Lukeš, D., & Rosa, R. (2020). An Introduction to Python for Linguists. Retrieved from https://v4py.github.io/intro.html 5. Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press. 6. Mitkov, R. (2009). The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press. 7. Panggabean, H., & Tobing, A. (2015). Computational Linguistics Application Using Python Programming. IOSR Journal of Humanities and Social Science (IOSR-JHSS), 20(7), 18-30. 8. Roth, B., & Wiegand, M. (2021). Python for Linguists. *Computational Linguistics*, *47*(1), 217–220. 9. Дарчук Н. П. (2008). Комп’ютерна лінгвістика (автоматичне опрацювання тексту): підручник. К.: Видавничо-поліграфічний центр “Київський університет”.

	10. Жуковська, В. (2013). Вступ до корпусної лінгвістики: навчальний посібник. Житомир: Вид-во ЖДУ ім. І. Франка.

Силабус підготувала:

доктор філософіїї з філології,
ст.н.с. Крамар Н.А.