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.

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ ВК 05 ДВА

Основи комп'ютерної лінгвістики

для аспірантів спеціальності 035 – «Філологія»
третього освітньо-наукового рівня вищої освіти – доктора філософії

2023

Розробник

доктор філософії з філології

Крамар Н.А.

Робочу програму затверджено Вченою радою Центру науковихдосліджень та викладання іноземних мов НАНУ

протокол № 7 від 21 листопада 2023 року

1. Загальні відомості

Найменування показників	Характеристика дисципліни за денною формою навчання
Назва дисципліни	Основи комп'ютерної лінгвістики
Вид дисципліни	вибірковий компонент, дисципліна вибору аспіранта ВК 05 ДВА
Мова викладання, навчання та оцінювання	англійська, українська
Загальний обсяг кредитів / годин	3/90
Курс	3
Семестр	5-6
Кількість змістових модулів	3
Обсяг кредитів	3
Обсяг годин,	90
в тому числі:	
Лекції	10
Практичні (індивідуальні) заняття	8
Самостійна робота	72
Форма підсумкового контролю	іспит

2. Мета, завдання та очікувані результати навчальної дисципліни

Робоча програма навчальної дисципліни ВК 05 ДВА «Основи комп'ютерної лінгвістики» є нормативним документом, який розроблено на основі освітньо-наукової програми (далі ОНП) підготовки здобувачів третього рівня відповідно до навчального плану спеціальності 035 «Філологія», затвердженої Вченою радою Центру наукових досліджень та викладання іноземних мов НАНУ, протокол № 5 від 05.09.2023 року. Дисципліна «Основи комп'ютерної лінгвістики» належить до важливих навчальних дисциплін за освітньо-науковим рівнем «доктор філософії». Вона покликана допомогти аспірантові розширити свої знання про методи лінгвістичних досліджень; зорієнтуватися у сучасних корпусах української мови та іноземних мов; опанувати основи програмування у мові Python; посилити навички компілювання власних корпусів та екстракції інформації із них завдяки використанню спеціалізованого програмного забезпечення.

Мета: Метою цього курсу є 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. Курс спрямовано на те, щоб допомогти аспіранту обирати найдоцільніші методи та інструменти для роботи з текстовою інформацією великого обсягу.

Передумова вивчення.

Попередні вимоги

Аспірант повинен знати: українську та англійську мови на рівні не нижче С1, основні прийоми пошуку та аналізу інформації, основи лінгвістичної методології; основи статистики; мати навички критичного мислення.

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).

Фахові програмні результати навчання (вимоги до знань та вмінь):
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

3. Тематичний план навчальної дисципліни

Навчальний матеріал дисципліни складається із 7 тем.
Курс викладається під час п’ятого та шостого семестрів третього року навчання в аспірантурі.
Тема 1. Місце комп’ютерної лінгвістики та NLP у сучасному лінгвістичному ландшафті.
Тема 2. Основи програмування у Python для NLP-задач.
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.
Тема 6. Векторна репрезентація слів.
Тема 7. Тематичне моделювання та класифікація текстових даних.

4. Програма навчальної дисципліни

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
	Тема 7. Тематичне моделювання та класифікація текстових даних.	1	1	5
		10	8	72
Усього		90		

5. Структура навчальної дисципліни

Змістовий модуль 1.

1.1. Співвідношення понять «комп'ютерна лінгвістика», «NLP» та «language technology».
1.2. Основні цілі та завдання комп'ютерної лінгвістики.
1.3. Історія становлення комп'ютерної лінгвістики від 1950х рр. до сьогодні.
1.4. Зв'язок комп'ютерної лінгвістики з іншими лінгвістичними напрямками.
1.5. Подальші перспективи розвитку технологій обробки природної мови та їхній вплив на суспільне життя.
1.6. Основні поняття комп'ютерної лінгвістики: токен, лемма, стемінг, парсинг, частиномовна розмітка.

Змістовий модуль 2.

2.1. Змінні, типи даних та операції.
2.2. Цикли for та while.
2.3. Функції та об'єкти.
2.4. Робота з файлами: відкриття, читання та запис файлів текстового та бінарного типу.
2.5. Операції зі структурами даних: списки, кортежі, множини, словники.
2.6. Принципи роботи з зовнішніми бібліотеками Python.
2.7. Основи регулярних виразів (regex).
2.8. Компіляція та підготовка датасету для обробки у Python.
2.9. Методи попередньої обробки текстових даних: нормалізація правопису, видалення зайвої пунктуації та спеціальних символів, видалення стоп-слів, токенізація.
2.10. Стемінг та лематизація: як обрати доцільний підхід?
2.11. Синтаксичний аналіз текстових даних: dependency parsing vs. constituency parsing.
2.12. Частиномовна розмітка.
2.13. Принципи роботи зі Spacy Matchers.
2.14. Розпізнання іменованих сутностей (Named Entity Recognition).

Змістовий модуль 3.

3.1. Типи машинного навчання: контрольоване навчання, неконтрольоване навчання та навчання з підкріпленням.
3.2. Feature engineering для машинного навчання в NLP.
3.3. Методи векторної репрезентації слів: one-hot encoding, bag of words, TF-IDF.

3.4. Сучасні підходи до вбудування слів (word embedding): Word2Vec, GloVe, FastText.
3.5. Класифікація тексту на основі правил vs. класифікація тексту на основі машинного навчання.
3.7. Методи тематичного моделювання текстових даних: LDA, LSA, GSDMM та інші.
3.8. Великі мовні моделі (large language models) та принципи роботи з ними.
3.9. Архітектура нейронних мереж: RNN, CNN, Transformer.

Індивідуальні заняття
Мета практичних (індивідуальних) занять - практичне закріплення питань, пов'язаних з темами лекцій та дослідженнями аспіранта.

Самостійна робота аспіранта, її зміст та обсяг

№	Зміст самостійної роботи аспіранта	Обсяг СР (годин)
1	Опрацювання лекційного матеріалу	30
2	Підготовка до індивідуальних (практичних) занять	42
	Усього за навчальною дисципліною	72

6. Рейтингова система оцінювання

Рейтинг аспіранта першого року зі спеціальності 035 «Філологія» складається з балів, що їх отримано за:
1. Експрес-контроль – 30 балів
2. Практичні (індивідуальні) заняття та самостійна робота – 30 балів
3. Іспит – 40 балів
Заохочувальні і штрафні бали:
1. Відсутність на лекції без поважних причин - (-) 2 бали.
2. Відсутність на індивідуальних (практичних) 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.

Система рейтингових балів та критерії оцінювання
Експрес-контроль (30 балів) проводиться з метою перевірки якості роботи аспіранта в аудиторії та самостійної роботи в позааудиторний час за

допомогою усного описування або перевірочних робіт тривалістю 10–30 хвилин або індивідуальних домашніх завдань протягом семестру, завдяки чому перевіряються набуті знання та розуміння того як матеріали курсу можна застосувати у власних дослідженнях за темою дисертаційної роботи та при підготовці публікацій за темою дисертації.

Розрахункова шкала рейтингу складає:

RC = 30 + 30 +40 = 100 (балів).

Рейтинг RD аспіранта складається з рейтингу, одержаного протягом семестру з урахуванням заохочувальних і штрафних балів. Необхідною умовою допуску аспіранта до іспиту з дисципліни є позитивний рейтинг з усіх

форм семестрової атестації. Аспіранти, які набрали протягом семестру менше 30 балів, зобов'язані підвищити свій рейтинг для допуску до іспиту.

Шкала оцінок

Оцінка	**Визначення**	**Національна шкала оцінювання**	**Рейтингова бальна шкала оцінювання**
ВІДМІННО	відмінне виконання лише з незначною кількістю помилок	**5, 0 (відмінно)**	**90 ≤ RD ≤100**
ДОБРЕ	в цілому правильна робота з певною кількістю помилок	**4, 0 (добре)**	**74 ≤ RD ≤ 89**
ЗАДОВІЛЬНО	виконання задовольняє мінімальні критерії	**3,0 (задовільно)**	**60 ≤ RD ≤ 73**
НЕЗАДОВІЛЬНО	можливе повторне складання	**2 (незадовільно)**	**35 ≤ RD ≤ 59**
НЕЗАДОВІЛЬНО	необхідний повторний курс з навчальної дисципліни	**2 (незадовільно)**	**RD < 35**

7. Орієнтовний перелік питань на іспит:

1. Які основні цілі та завдання комп’ютерної лінгвістики?

2. Які етапи можна виділити у розвитку комп’ютерної лінгвістики від 1950х років до сьогодні?

3. Що таке лінгвістичний корпус і які завдання можна вирішувати з його допомогою?

4. Які переваги та обмеження пов'язані із застосуванням корпусів для досліджень?

5. Які методи анотування корпусу існують?

6. Переваги та недоліки різних бібліотек Python для вирішення NLP-задач на основі англомовних та україномовних даних.

7. Поясніть поняття «токен» і «лемма».

8. Поясніть різницю між стемінгом та лематизацією. Для яких мов та NLP-задач доцільніший один чи інший?

9. Опишіть основні етапи попередньої обробки лінгвістичних даних.

10. Опишіть алгоритм роботи з модулем Spacy Matchers.

11. Опишіть алгоритм здійснення аналізу настроїв (sentiment analysis).

12. Поясніть різницю між класифікацією на основі правил та класифікацією на основі машинного навчання.

13. Які існують основні методи векторної репрезентації та вбудування слів?

14. Які існують типи машинного навчання?

15. Поясніть специфіку таких методів тематичного моделювання як LDA, LSA та BERTopic.

16. Поясніть принцип створення та функціонування великих мовних моделей на кшталт GPT, Bloom та Llama.

8. Список рекомендованих джерел:

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. 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.
4. Hapke, H., Lane, H., & Howard, C. (2019). *Natural Language Processing in Action*. Manning Publications.
5. Hovy, D. (2012). Programming in Python for Linguists. A Gentle Introduction. Retrieved from http://www.dirkhovy.com/portfolio/papers/download/pfl_handout.pdf
6. Johnson, M. (2011). How relevant is linguistics to computational linguistics? Linguistic Issues in Language Technology, 6(7), 1–23.
7. Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing, 2nd edition*. Pearson Prentice Hall.
8. Lukeš, D., & Rosa, R. (2020). An Introduction to Python for Linguists. Retrieved from https://v4py.github.io/intro.html
9. Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
10. Mitkov, R. (2009). The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press.
11. Panggabean, H., & Tobing, A. (2015). Computational Linguistics Application Using Python Programming. IOSR Journal of Humanities and Social Science (IOSR-JHSS), 20(7), 18-30.
12. Roth, B., & Wiegand, M. (2021). Python for Linguists. *Computational Linguistics, 47*(1), 217–220.
13. Дарчук Н. П. (2008). Комп’ютерна лінгвістика (автоматичне опрацювання тексту): підручник. К.: Видавничо-поліграфічний центр “Київський університет”.
14. Жуковська, В. (2013). Вступ до корпусної лінгвістики: навчальний посібник. Житомир: Вид-во ЖДУ ім. І. Франка.