МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Львівський національний університет імені Івана Франка
Факультет електроніки та комп'ютерних технологій
Кафедра оптоелектроніки та інформаційних технологій

Затверджено
На засіданні кафедри оптоелектроніки та інформаційних технологій
факультету електроніки та комп'ютерних технологій
Львівського національного університету імені Івана Франка
(протокол №296 від 29.08. 2023 р.)

Завідувач кафедри _____ Олег КУШНІР

Силабус з навчальної дисципліни «Комп'ютерна лінгвістика та обробка природної мови», що викладається в межах ОПП «Інженерія програмного забезпечення» першого (бакалаврського) рівня вищої освіти для здобувачів зі спеціальності 121 – Інженерія програмного забезпечення

Львів 2023

Назва дисципліни	Комп'ютерна лінгвістика та обробка природної мови
Адреса викладання дисципліни	Корпуси факультету електроніки та комп'ютерних технологій, Львівський національний університет імені Івана Франка: вул. Драгоманова 50, 79005 м. Львів вул. ген. Тарнавського 107, 79011 м. Львів
Факультет та кафедра, за якою закріплена дисципліна	Факультет електроніки та комп'ютерних технологій, кафедра оптоелектроніки та інформаційних технологій
Галузь знань, шифр та назва спеціальності	12 – Інформаційні технології 121 – Інженерія програмного забезпечення
Викладачі дисципліни	Кушнір Олег Степанович, докт. фіз.-мат. наук, професор, професор
Контактна інформація викладачів	oleh.kushnir@lnu.edu.ua https://electronics.lnu.edu.ua/employee/kushnir-o-s
Консультації з питань навчання по дисципліні відбуваються	Консультації в день проведення лекційних занять (за попередньою домовленістю): кімн. 215, корпус факультету електроніки та комп'ютерних технологій, м. Львів, вул. Тарнавського, 107. Також можливі онлайн-консультації через Zoom або Telegram. Для погодження часу онлайн-консультацій слід писати на електронну пошту викладача або на Telegram.
Сторінка дисципліни	https://electronics.lnu.edu.ua/course/komp-iuterna-linhvistyka-its http://194.44.208.156/moodle/course/view.php?id=59
Інформація про дисципліну	Дисципліна «Комп'ютерна лінгвістика та обробка природної мови» є вибірковою дисципліною зі спеціальності 121 – Інженерія програмного забезпечення для освітньо-професійної програми «Інженерія програмного забезпечення», яка викладається в 8 семестрі в обсязі 5,0 кредитів (за Європейською Кредитно-Трансферною Системою – ECTS).
Коротка анотація дисципліни	Навчальну дисципліну розроблено для одержання студентами теоретичних знань з комп'ютерної лінгвістики та обробки природної мови, а також для формування в них практичних навичок алгоритмізації та розробки ефективного програмного забезпечення для вирішення стандартних прикладних задач опрацювання природної мови.
Мета та цілі дисципліни	*Метою* вивчення дисципліни «Комп'ютерна лінгвістика та обробка природної мови» є ознайомлення студентів з теоретичними основами, алгоритмами та методами комп'ютерної лінгвістики та опрацювання природної мови. *Цілями* дисципліни є формування в студентів практичних навичок, які би дали змогу використовувати засвоєні знання, а також обирати, обґрунтовувати та ефективно застосовувати алгоритми, методи і прикладні програми для опрацювання природної мови.
Література для вивчення дисципліни	Основна література: 1. Волошин В. Г. Комп'ютерна лінгвістика / В. Г. Волошин. – Суми : Університетська книга, 2004. – 382 с. 2. Delmonte R. Computational Linguistic Text Processing / New York: Nova Science Publishers, 2009 – 382 p. 3. Пасічник В. В. Математична лінгвістика. Книга 1. Квантитативна лінгвістика / В. В. Пасічник, Ю. М. Щербина, В. А. Висоцька, Т. В. Шестакевич. – Львів: Новий світ – 2000, 2012. – 359 с. 4. Bolshakov I. Computational linguistics. Models, resources, applications / I. Bolshakov, A. Gelbukh. – Mexico : Ciencia de la Computacion, 2004. – 198 p. 5. Clark A. The Handbook of Computational Linguistics and Natural Language Processing / A. Clark, C. Fox, S. Lappin. – Chichester: John Wiley and Sons, 2010. – 801 p.

	6. https://www.gutenberg.org/ Додаткова література: 7. Manning C. D. Foundations of statistical natural language processing / Manning C. D., Schutze H. – London: The MIT Press Cambridge, 1999. – 680 p. 8. Espitia D. Universal and non-universal text statistics: Clustering coefficient for language identification / D. Espitia, H. L. Ridaura // Physica A. – 2019. – Vol. 553. – 123905 (25 pp.). 9. Pilgrim C. Bias in Zipf's law estimators / C. Pilgrim, T. T. Hills // Scientific Reports. – 2019. – Vol. 11. – 17309 (12 p.). 10. Multilayer networks for text analysis with multiple data types / C. C. Hyland, Yuanming Tao, L. Azizi, M. Gerlach, T. P. Peixoto, E. G. Altmann. – EPJ Data Science. – 2019. – 16 p. https://doi.org/10.1140/epjds/s13688-021-00288-5
Обсяг курсу	Сумарно 150 год. Із них 64 год. аудиторних занять (32 год. лекцій і 32 год. лабораторних робіт) і 86 год. самостійної роботи
Очікувані результати навчання	Після завершення цього курсу студент буде: *знати* основні методи комп'ютерної лінгвістики та опрацювання природної мови, основні визначення, теорії, моделі та алгоритми опрацювання природної мови і опису лінгвістичних систем, інформаційного пошуку та інтелектуального аналізу текстових даних; *вміти* аналізувати моделі для опрацювання природної мови, створювати програмні продукти та працювати з готовими програмними продуктами, застосовувати комп'ютерну техніку для вирішення лінгвістичних задач, розробляти та реалізувати відповідні алгоритми, писати прикладні програми та користуватися ними. Після вивчення даного курсу студенти набудуть таких Загальних та фахових компетентностей і Програмних результатів навчання: ЗК1. Здатність до абстрактного мислення, аналізу та синтезу. ЗК2. Здатність застосовувати знання у практичних ситуаціях. ЗК3. Здатність спілкуватися державною мовою як усно, так і письмово. ЗК4. Здатність спілкуватися іноземною мовою як усно, так і письмово. ЗК6. Здатність до пошуку, оброблення та аналізу інформації з різних джерел. ФК13. Здатність ідентифікувати, класифікувати та формулювати вимоги до програмного забезпечення. ФК14. Здатність розробляти архітектури, модулі та компоненти програмних систем. ФК16. Здатність формулювати та забезпечувати вимоги щодо якості програмного забезпечення у відповідності з вимогами замовника, технічним завданням та стандартами. ФК27. Здатність використовувати для розробки програмного забезпечення перспективні засоби та технології, зокрема, науки про дані, штучного інтелекту, IoT, вбудованих систем тощо. ФК29. Здатність здійснювати розробку програмного забезпечення використовуючи сучасні парадигми програмування. ПРН 1. Аналізувати, цілеспрямовано шукати і вибирати необхідні для вирішення професійних завдань інформаційно-довідкові ресурси і знання з урахуванням сучасних досягнень науки і техніки. ПРН5. Знати і застосовувати відповідні математичні поняття, методи доменного, системного і об'єктно-орієнтованого аналізу та математичного моделювання для розробки програмного забезпечення. ПРН6. Уміння вибирати та використовувати відповідну задачі методо-

	логію створення програмного забезпечення.
Ключові слова	Комп'ютерна лінгвістика, статистична лінгвістика, опрацювання природної мови, машинний переклад, комп'ютерна лексикографія, аналіз і синтез мови
Формат курсу	Очний
	Проведення лекцій, лабораторних робіт та консультації для поглибленого розуміння тем
Теми	Див. СХЕМУ КУРСУ
Підсумковий контроль, форма	Залік вкінці семестру
Пререквізити	Для вивчення курсу студенти потребують базових знань з дисциплін «Вища математика», «Дискретна математика», «Основи програмування», «Теорія алгоритмів», «Методи та технології обчислень», «Прикладна статистика та ймовірнісні процеси», «Об'єктно-орієнтоване програмування», «Бази даних», «Алгоритми і структури даних».
Навчальні методи та техніки, які будуть використовуватися під час викладання курсу	Інформаційні методи (лекції, бесіди, презентації, ілюстрації, демонстрації); дедуктивні методи на основі узагальнень; евристичні методи (обговорення, проблемні лекції, дискусії); лабораторні роботи, індивідуальні практичні завдання, інтерактивні методи.
Необхідне обладнання	Мультимедіа, платформи Microsoft Teams, Moodle і Zoom, доступ до мережі Інтернет, комп'ютерне програмне забезпечення: .NET, Python 3, JDK, Qt5. Мінімальні вимоги до техніки та системи: CPU i3, RAM 4GB, Windows7. Зокрема, для проведення лекційних занять потрбіні: • монітор TFT 23"; • системний блок (процесор Intel i5-6500, 8GB оперативної пам'яті, HDD 256GB) ; • мультимедійне обладнання (проектор, проекційний екран, дошка настінна, звуковий підсилювач та аудіосистема); • комутатор мережевий для доступу до мережі Internet. Для проведення лабораторних занять: • комп'ютерна лабораторія з 12-14 робочими місцями; • монітори TFT 23"; • системні блоки (процесор Intel i5-6500, 8GB оперативної пам'яті, HDD 256GB); • мультимедійне обладнання (проектор, проекційний екран, дошка настінна, звуковий підсилювач та аудіосистема); • комутатор мережевий для доступу до мережі Internet. Базова комп'ютерна техніка, що задовольняє вимогам дисципліни, відповідає обладнанню комп'ютерного класу №7 (корпус по вул. Тарнавського, 107).
Критерії оцінювання (окремо для кожного виду навчальної діяльності)	Оцінювання проводиться упродовж семестру та під час залікової сесії за 100-бальною шкалою. Бали нараховуються за такими видами робіт із таким співвідношенням: • лабораторні (14 робіт, максимально 14x5=70 балів) або індивідуальні (1 програмістська або дослідницька робота, можлива також командна; максимально 1x70=70 балів) практичні роботи: 70% оцінки; максимальна кількість балів 70. • активність на лекціях (відвідування занять, відповіді на питання, участь у дискусіях і обговоренні, внесок у розв'язання проблемних ситуацій тощо): максимально 10 балів або 10% оцінки.

	• 2 письмові модульні контролі (на лекціях): максимально 2х10=20 балів або 20% оцінки. Загалом 100 балів. Оцінки за лабораторні заняття (5 балів) розподіляються так: повне виконання всіх лабораторних завдань – 40% (2 бали), теоретичні знання за предметом роботи та правильні відповіді на запитання викладача або тестування – 40% (2 бали), якість інтерпретації даних і оформлення звіту – 20% (1 бали). Студент отримує 2 бали за захист лабораторної роботи, якщо повністю володіє теоретичним матеріалом і в повному обсязі, аргументовано відповідає на запитання. Студент отримує 1 бал, якщо відповіді на поставленні питання поверхові, з суттєвими неточностями. Якщо студент не володіє теоретичним матеріалом лабораторної роботи та/або не вміє пояснити її програмну реалізацію, то студенту виставляється 0 балів). Загалом бали **оцінювання лабораторних робіт** нараховуються за наступним співвідношенням: 5 – студент в повному обсязі володіє матеріалом, має повне розуміння розглянутої теми, надає правильні відповіді на запитання по темі, код програми функціонує відповідно до завдання; 4 – студент достатньо розуміє матеріал та принципи написаного ним коду програми, проте присутні неточності та незначні помилки у відповідях на запитання по темі, код програми функціонує відповідно до завдання (або з несуттєвими недоліками); 3 – студент не досить добре розуміє матеріал та написаний код програми, вагається та надає неточні/не конкретні відповіді на запитання по темі, код програми функціонує неточно, або з помірними недоліками; 2 – студент погано розуміє матеріал та написаний код програми, в більшості надає помилкові відповіді на питання по темі, код програми функціонує з суттєвими недоліками; 1 – студент погано розуміє матеріал та написаний код програми, код програми не функціонує належним чином; 0 – студент зовсім не засвоїв матеріал, написаний код програми не відповідає темі/не функціонує взагалі. **Оцінювання змістових модулів** (2 змістові модулі по 10 балів за кожний) – за результатами висвітлення студентом деяких теоретичних тем тощо. Бали оцінювання змістових модулів нараховуються за наступним співвідношенням: 9–10 – розглянуту тему розкрито в повному обсязі, правильно, обґрунтовано, логічно, містить аналіз і систематизацію, аргументовані висновки. Засвідчено глибоке володіння матеріалом. Наведені приклади коду повністю робочі та відповідають темі. Можуть бути присутні несуттєві помилки та невідповідності; 7–8 – розкрито значну частину теми. Виявлено знання і розуміння основних положень навчальної дисципліни, проте присутні неточності та/або невідповідності основній темі. Наведені приклади коду частково робочі, проте загалом відповідають темі; 5–6 – відстежується загальне розуміння теми. Виявлено множинні неточності та невідповідності, пояснення наведеного коду відсутні, код функціонує із значними неточностями (або відсутні приклади запуску коду на виконання взагалі); 3–4 – студент погано розуміє тему. Виявлено суттєві неточності та невідповідності. Наведені приклади коду мають суттєві недоліки або не відповідають темі; 0–2 – студент фактично не розуміє тему. Тему не розкрито, кількість

	викладеного матеріалу не відповідає загальним нормам обраного виду роботи. Алгоритм або код не пояснено. У плані імплементації **неформальної освіти** здобувач за бажанням може додатково здобути максимально 20 балів за самостійну роботу, пред'явивши сертифікати зі споріднених курсів («Комп'ютерна лінгвістика», «Опрацювання природної мови», «NLTK», «OpenNLP», «Stanford CoreNLP», «Lingpipe» тощо) та написавши есе за відповідною тематикою. _____. **Академічна доброчесність**: Очікується, що лабораторні та контрольні роботи студентів будуть їхніми оригінальними дослідженнями або міркуваннями. Відсутність посилань на використані джерела, фабрикування джерел, списування, втручання в роботу інших студентів становлять, але не обмежують, приклади можливої академічної недоброчесності. Виявлення ознак академічної недоброчесності в роботі сутдента є підставою для її незарахуванння викладачем, незалежно від масштабів плагіату або спроб обману. **Відвідування занять** є важливою складовою навчання. Очікується, що всі студенти відвідають усі лекції та лабораторні заняття курсу. Студенти мають інформувати викладача про неможливість відвідати заняття. Студенти зобов'язані дотримуватися всіх термінів, визначених для виконання видів робіт, передбачених курсом. **Література.** Уся література, яку студенти не зможуть знайти самостійно, буде надана викладачем виключно в освітніх цілях без права її передачі третім особам. Студенти також заохочуються до використання іншої літератури та джерел, зокрема наукової літератури, яка відсутня серед обов'язкової та рекомендованої. **Політика виставлення балів.** Враховуються бали, набрані на поточному опитуванні, самостійній роботі та бали підсумкового контролю знань. Обов'язково враховуються присутність на заняттях та активність студента під час лабораторних занять; наголошується на неприпустимості пропусків або запізнень на заняття, користування мобільним телефоном, планшетом або іншими мобільними пристроями під час занять з метою, не пов'язаною з навчанням, списування та плагіату, несвоєчасного виконання поставлених завдань і т. ін. Жодні форми порушення академічної доброчесності не толеруються.
Питання до заліку	Перелік питань і завдань для проведення підсумкової оцінки знань усіх тем курсу до контрольних робіт розміщено на сторінці https://drive.google.com/drive/folders/1Kamy1aZ080cxg0ZwuDX9W8kT2rADpYDF?usp=sharing
Опитування	Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенню курсу.

СХЕМА КУРСУ

Тиж.	Тема, план, короткі тези	Форма діяльності (заняття)	Література. Ресурси в Інтернеті	Завдання (лабораторна робота), год.	Термін виконання
1, 2	**Вступ. Лінгвістика та її структура. Базові поняття лінгвістики** Зв'язки комп'ютерної лінгвістики з інформатикою та системами штучного інтелекту. Лінгвістика та її структура. Загальні поняття.	Лекція	2, 3, 7	Вступне заняття. Академічна доброчесність. Препроцесинг текстових документів Визначення семантичного навантаження тексту за параметрами кластеризації слів	1, 2 тиж. семестру

3, 4	**Методи та продукти комп'ютерної лінгвістики** Розвиток ідей, теорій, підходів і методів комп'ютерної лінгвістики. Продукти комп'ютерної лінгвістики.	Лекція	2, 4, 7	Опрацювання природних мов на різних лінгвістичних рівнях за допомогою засобів Python Фонетичний аналіз і силабіфікація текстів	3, 4 тиж. семестру
5, 6	**Теорії, моделі та алгоритми комп'ютерної лінгвістики** Лінгвістичні знаки та моделі. Поняття тексту і змісту. Способи представлення змісту. Розкладання та «атомізація» змісту.	Лекція	1, 2, 8	Програмування задач морфологічного синтезу дієслівних форм	5, 6 тиж. семестру
7, 8	**Статистична лінгвістика** Тексти як складні системи та мережі. Основні поняття лінгвістичної статистики та складних систем. Методика вивчення лінгвістичної.	Лекція	1, 4, 8, 9	Рангові залежності та розподіли ймовірності для прізвищ та імен Мережеві методи визначення ключових слів у текстах	7, 8 тиж. семестру
9, 10	**Основні закони лінгвістики.** Закони Ціпфа, Гіпса та Парето. Статистика інших складних систем. Механізми появи степеневих розподілів	Лекція	2, 4, 7	Закони Ціпфа, Парето та Гіпса для слів у текстах Практичні рецепти бінування для словника корпусу текстів	9, 10 тиж. семестру
11,12	**Інші закони лінгвістичної статистики. Стохастичні гіпотези в лінгвістиці** Статистичні закони для n-грам. Довжина слова та речення. Закон Манцерата–Альтмана. Рандомні моделі мови. Розрізнення природних і рандомних текстів.	Лекція	1, 2, 10	Статистика n-грам у текстах і корпусах текстів Вивчення рандомних моделей текстів	11, 12 тиж. семестру
13, 14	**Основи опрацювання природної мови** Пошук ключових слів у текстах. Встановлення мови, стилістики та плагіату. Мережеві властивості текстів. Інформаційний пошук.	Лекція	1, 3, 4, 7, 10	Метод TF-IDF Кластеризаційні методи встановлення ключових слів у текстах	13, 14 тиж. семестру
15, 16	**Аналіз, розпізнавання та синтез природної мови. Машинний переклад та комп'ютерна лексикографія** Автоматичне введення мови, аналіз та розпізнавання мови. Синтез мови. Машинний переклад і комп'ютерна лексикографія.	Лекція	1, 7, 8, 10	Вирішення задач розпізнавання за допомогою пакету FineReader	15, 16 тиж. семестру