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 №6 від 29 серпня 2023 р.)

Завідувач кафедри _____ Олег КУШНІР

Силабус з навчальної дисципліни «Опрацювання природної мови / Natural Language Processing», що викладається в межах ОП «Комп'ютерні науки» другого (магістерського) рівня вищої освіти для здобувачів зі спеціальності 122 – Комп'ютерні науки

Львів 2023

Назва дисципліни	Опрацювання природної мови / Natural Language Processing
Адреса викладання дисципліни	Корпус факультету електроніки та комп'ютерних технологій, Львівський національний університет імені Івана Франка м. Львів, вул. Тарнавського, 107
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 122 Комп'ютерні науки
Викладачі дисципліни	Кушнір Олег Степанович, докт. фіз.-мат. наук, проф., проф.
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/opratsiuvannia-pryrodnoi-movy-122-komp-iuterni-nauky http://194.44.208.156/moodle/course/view.php?id=59 https://drive.google.com/drive/folders/1Kamy1aZ080cxg0ZwuDX9W8kT2rADpYDF?usp=sharing
Інформація про дисципліну	Дисципліна «Опрацювання природної мови» є нормативною дисципліною зі спеціальності 122 Комп'ютерні науки для освітньої програми «Комп'ютерні науки», яка викладається в 2 семестрі в обсязі 6,0 кредитів (за Європейською Кредитно-Трансферною Системою – ECTS).
Коротка анотація дисципліни	Навчальну дисципліну розроблено для одержання студентами теоретичних знань з комп'ютерної лінгвістики та опрацювання природної мови, а також для формування в них навичок ефективного застосовування засвоєних знань і методів у розв'язанні прикладних задач опрацювання природної мови. Представлено теоретичні основи комп'ютерної, статистичної та математичної лінгвістики, класифікація та огляд особливостей відомих продуктів для опрацювання природної мови, основи машинного перекладу та комп'ютерної лексикографії, а також відповідні комп'ютерні алгоритми і засоби опрацювання даних.
Мета та цілі дисципліни	*Метою* вивчення дисципліни «Опрацювання природної мови» є ознайомлення студентів з теоретичними основами комп'ютерної лінгвістики та опрацювання природної мови, а *ціллю* – формування в них практичних навичок, які б дали змогу ефективно застосовувати засвоєні знання, алгоритми, методи та прикладні програми для опрацювання природної мови.
Література для вивчення дисципліни	**Основна:** 1. Bolshakov I. Computational linguistics. Models, resources, applications / I. Bolshakov, A. Gelbukh. – Mexico : Ciencia de la Computacion, 2004. – 198 p. 2. Bird S. Natural language processing with Python / S. Bird, E. Klein, E. Loper. – Sebastopol : O'Reilly. – 2009. – 504 p.

	3. Manning C. D. Foundations of statistical natural language processing / Manning C. D., Schutze H. – London : The MIT Press Cambridge, 1999. – 680 p. 4. Кушнір О. С. Основи комп'ютерної лінгвістики (конспект лекцій) / О. С. Кушнір. – Львів : Видавн. Львів. ун-ту, 2023. – 292 с. 5. Jurafsky D. Speech and language processing / D. Jurafsky, J. H. Martin. – New Jersey : Prentice Hall, 2023. – 628 p. 6. Clark A. The handbook of computational linguistics and natural language processing / A. Clark, C. Fox, S. Lappin. – Chichester : John Wiley & Sons, 2010. – 801 p. 7. Kracht M. Introduction to probability theory and statistics for linguistics / M. Kracht. – Oakland : UCLA, 2005. – 137 p. 8. Delmonte R. Computational linguistic text processing / New York : Nova Science Publishers, 2009. – 382 p. 9. Kornai A. Mathematical linguistics / A. Kornai. – London : Springer, 2007. – 300 p. 10. Web information retrieval / S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali, S. Quarteroni. – Berlin : Springer, 2013. – 287 p. 11. de Araújo L. C. Statistical analyses in language usage / L. C. de Araújo. – Belo Horizonte : Universidade Federal de Minas Gerais, 2013. – 199 p. 12. 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. 13. Волошин В. Г. Комп'ютерна лінгвістика / В. Г. Волошин. – Суми : Університетська книга, 2004. – 382 с. **Додаткова:** 14. Zanette D. H. Statistical patterns in written language / Zanette D. H. – Centro Atomico Bariloche, 2012. – 87 p. http://fisica.cab.cnea.gov.ar/estadistica/2te/ 15. Складні мережі // Ю. Головач, О. Олємской, К. фон Фербер, Т. Головач, О. Мриглод, І. Олємской, В. Пальчиков // Журн. фіз. дослідж. – 2006. – Т. 10, №4. – С. 247–289. 16. Newman M. E. J. Power laws, Pareto distributions and Zipf's law / Newman M. E. J. // Contemporary Phys. – 2005. – Vol. 46. – P. 323–351. 17. Ferrer i Cancho R. Zipf's law from a communicative phase transition / R. Ferrer i Cancho // Eur. Phys. J.: B. – 2005. – Vol. 47. – P. 449–457. 18. Pilgrim C. Bias in Zipf's law estimators / C. Pilgrim, T. T. Hills // Sci. Rep. – 2021. – Vol. 11. – 17309 (12 pp.). 19. Espitia D. Universal and non-universal text statistics: Clustering coefficient for language identification / D. Espitia, H. L. Ridaura // Physica A. – 2020. – Vol. 553. – 123905 (25 pp.). 20. Altmann E. G. Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words / E. G. Altmann, J. B. Pierrehumbert, A. E. Motter // PLOS ONE. – 2009. – Vol. 4. – e7678 (7 pp.).
Обсяг курсу	Аудиторні години – 64, з них лекції – 32 години, лабораторні роботи – 32 години і 116 годин самостійної роботи
Очікувані результати навчання	Після завершення цього курсу студент буде: - знати основні методи комп'ютерної лінгвістики та опрацювання природної мови, основні 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, працювати з відповідними програмними продуктами, застосовувати комп'ютерну техніку для вирішення лінгвістичних задач, розробляти та реалізувати відповідні алгоритми, писати прикладні програми та користуватися ними. Після вивчення курсу здобувачі набудуть таких компетентностей і програмних результатів: 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. СК2. Здатність формалізувати предметну область певного проєкту у вигляді відповідної інформаційної моделі. СК3. Здатність використовувати математичні методи для аналізу формалізованих моделей предметної області. СК6. Здатність застосовувати існуючі та розробляти нові алгоритми розв'язування задач у галузі комп'ютерних наук. СК7. Здатність розробляти програмне забезпечення відповідно до сформульованих вимог із урахуванням наявних ресурсів і обмежень. СК8. Здатність розробляти та реалізовувати проєкти зі створення програмного забезпечення, у т. ч. в непередбачуваних умовах, за нечітких вимог і необхідності застосовувати нові стратегічні підходи, використовувати програмні інструменти для організації командної роботи над проєктом. СК11. Здатність ініціювати, планувати та реалізовувати процеси розробки інформаційних і комп'ютерних систем та програмного забезпечення, включно з його розробкою, аналізом, тестуванням, системною інтеграцією, впровадженням і супроводом. СК13. Здатність застосовувати методи і підходи штучного інтелекту, інтелектуального аналізу та науки про дані та підходів оптимізації до розв'язання конкретних проблем комп'ютерних наук. РН1. 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. РН2. 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. РН5. Оцінювати результати діяльності команд та колективів у сфері інформаційних технологій, забезпечувати ефективність їх діяльності. РН8. Розробляти математичні моделі та методи аналізу даних (включно з великими). РН9. Розробляти алгоритмічне та програмне забезпечення для аналізу даних (включно з великими). РН14. Тестувати програмне забезпечення. РН16. Виконувати дослідження у сфері комп'ютерних наук. РН17. Виявляти та усувати проблемні ситуації в процесі експлуатації програмного забезпечення, формулювати завдання для його модифікації або реінжинірингу. РН19. Аналізувати сучасний стан і світові тенденції розвитку комп'ютерних наук та інформаційних технологій. РН20. Володіти методами та засобами штучного інтелекту, інженерії та аналізу даних, розпізнавання образів і адаптивного опрацювання інформації, аналізу та обробки природної мови, моделювання та оптимізації.
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. СХЕМА КУРСУ
Підсумковий контроль, форма	Іспит вкінці семестру

Пререквізити	Для вивчення курсу студенти потребують базових знань у галузі 12 – Інформаційні технології.
Навчальні методи та техніки, які будуть використовуватися під час викладання курсу	Лекції, презентації, лабораторні роботи, індивідуальні та командні практичні завдання програмістського та дослідницького характеру, обговорення, дискусії, самостійна робота.
Необхідне обладнання	Мультимедіа, платформи Microsoft Teams, Moodle і Zoom, доступ до мережі Інтернет комп'ютерне програмне забезпечення: .NET, Python 3, JDK.
Критерії оцінювання (окремо для кожного виду навчальної діяльності)	Оцінювання проводиться упродовж семестру та під час екзаменаційної сесії за 100-бальною шкалою. Бали нараховуються за такими видами робіт із таким співвідношенням: • лабораторні (8 робіт, максимально 8x5=40 балів) або індивідуальні (1 програмістська або дослідницька робота, можлива також командна; максимально 1x40=40 балів) практичні роботи: 40% оцінки; максимальна кількість балів 40. • 1 письмовий модульний контроль (на лекціях): 10% оцінки; максимально 1x10=10 балів. • іспит: 50% оцінки; максимальна кількість балів 50. Загалом 100 балів. 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). _____. **Контрольні заміри знань проводять у формі стандартних практичних завдань і теоретичних питань.** **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 студента є підставою для її незарахування 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	Перелік питань і завдань для проведення підсумкової оцінки знань усіх тем курсу

контрольних робіт	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, самостійна робота *), год.	Термін виконання
1, 2	**Вступ. Лінгвістика та її структура. Базові поняття лінгвістики та опрацювання природної мови** Зв'язки комп'ютерної лінгвістики з галузями інформатики та систем штучного інтелекту. Роль опрацювання природної мови. Лінгвістика та її структура. Поняття та межі галузі. Загальні поняття про мову. Мова і мислення. Фонетика. Морфологія. Синтаксис. Семантика. Семантичні мережі. Прагматика.	Лекція	1, 3, 4, 6–10, 14	Вступне заняття. Академічна доброчесність. Встановлення наявності семантики текстів на основі мережевих параметрів, кластеризації та кореляції їхніх лінгвістичних елементів	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	1, 4, 13, 14	Кластеризація та класифікація текстів Побудова запитів до баз лінгвістичних даних	3, 4 тиж. семестру
5, 6	**Теорії та моделі в основі алгоритмів комп'ютерної лінгвістики** Мова як двонаправлений перетворювач змісту та тексту. Лінгвістичні знаки. Лінгвістичні моделі. Поняття тексту і змісту. Способи представлення змісту. Розкладання і «атомізація» змісту. Неоднозначність «картування».	Лекція	1, 4, 9, 10, 13	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. Методичні особливості вивчення лінгвістичної статистики.	Лекція	4, 7–9, 12, 13, 17, 18, 19, сайт курсу	Мережеві методи визначення ключових слів у текстах	7, 8 тиж. семестру
9, 10	**Основні статичні та динамічні закони лінгвістики.** Закони Ціпфа, Гіпса та Парето. Статистика інших складних систем. Механізми появи степеневих розподілів	Лекція	4, 8, 12, 14, 17, 20–22, 25, 26, сайт курсу	Вивчення статистики лексичних n-грам для окремих текстів і корпусів текстів	9, 10 тиж. семестру
11,12	**Інші закони лінгвістичної статистики. Нульові стохастичні гіпотези в статистичній лінгвістиці** Статистичні закони для n-грам. Середня довжина слова та речення. «Шуми» Закон Менцерата–Альтмана. Типи рандомних текстів. Стохастичні моделі. Рандомізовані природні тексти. Проблеми розрізнення природних і рандомних текстів.	Лекція	4, 6, 8, 23, 24, 27, сайт курсу	Генерування марковських ланцюжків і вивчення їхніх властивостей Дослідження закону Менцерата–Альтмана для довжин складів, слів і речень Характеристика повторюваності лінгвістичних елементів у тексті	11, 12 тиж. семестру
13, 14	**Основи опрацювання природної мови** Пошук ключових слів і абстрагу-	Лекція	2–4, 6, 7, 11, 18, 23, 27, сайт	Вивчення розподілів імовірності часів очікування лінгвістичних елементів і відносні методи визначення ключових	13, 14 тиж. семестру

	вання текстів. Встановлення мови, авторства, стилістики та плагіату. Скейлінг у лінгвістиці та інших складних системах. Мережеві властивості текстів. Інформаційний пошук.		курсу	слів у текстах, засновані на кластеризації Флуктуації та кореляції в текстах. Метод FA Флуктуаційний аналіз на текстових базах Метод DFA дослідження кореляцій для часових послідовностей із трендами Дослідження подібності текстів і плагіату	
15, 16	**Аналіз, розпізнавання та синтез природної мови. Машинний переклад та комп'ютерна лексикографія** Автоматичне введення звуків мови, аналіз та розпізнавання мови. Синтез мови та мовні технології. Машинний переклад і комп'ютерна лексикографія. Письмовий модуль	Лекція	2, 3, 5, 7, 9, 11, 15, 16	Типові шляхи вирішення проблем розпізнавання письмової мови	15, 16 тиж. семестру

* Типові завдання самостійної роботи містяться на сторінці з матеріалами навчальної дисципліни: https://drive.google.com/drive/folders/1Kamy1aZ080cxg0ZwuDX9W8kT2rADpYDF?usp=sharing