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 №8 від 20 серпня 2024 р.)

Завідувач кафедри _____

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

Львів 2024

Назва дисципліни	Опрацювання природної мови / Natural Language Processing
Адреса викладання дисципліни	м. Львів, вул. Тарнавського, 107
Факультет та кафедра, за якою закріплена дисципліна	Факультет електроніки та комп'ютерних технологій, кафедра опто-електроніки 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). Також можливі онлайн-консультації через Zoom. Для погодження часу онлайн-консультацій слід писати на електронну пошту викладача.
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
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 також 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. 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.

	6. Волошин В. Г. Комп'ютерна лінгвістика / В. Г. Волошин. – Суми : Університетська книга, 2004. – 382 с. 7. Jurafsky D. Speech and language processing / D. Jurafsky, J. H. Martin. – New Jersey : Prentice Hall, 2023. – 628 p. 8. Clark A. The handbook of computational linguistics and natural language processing / A. Clark, C. Fox, S. Lappin. – Chichester : John Wiley & Sons, 2010. – 801 p. 9. Hausser R. Foundations of computational linguistics: Man–machine communication in natural language / R. Hausser. – Berlin : Springer, 1999. – 468 p. 10. Kracht M. Introduction to probability theory and statistics for linguistics / M. Kracht. – Oakland : UCLA, 2005. – 137 p. 11. Delmonte R. Computational linguistic text processing / New York : Nova Science Publishers, 2009. – 382 p. 12. Kornai A. Mathematical linguistics / A. Kornai. – London : Springer, 2007. – 300 p. 13. Web information retrieval / S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali, S. Quarteroni. – Berlin : Springer, 2013. – 287 p. 14. de Araújo L. C. Statistical analyses in language usage / L. C. de Araújo. – Belo Horizonte : Universidade Federal de Minas Gerais, 2013. – 199 p. **Додаткова:** 15. Zanette D. H. Statistical patterns in written language / Zanette D. H. – Centro Atomico Bariloche, 2012. – 87 p. URL: http://fisica.cab.cnea.gov.ar/estadistica/2te/ 16. 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. 17. Newman M. E. J. Power laws, Pareto distributions and Zipf's law / Newman M. E. J. // Contemporary Phys. – 2005. – Vol. 46. – P. 323–351. 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. Zanette D. H. Dynamics of text generation with realistic Zipf distribution / D. H. Zanette, M. A. Montemurro // J. Quant. Linguist. – 2005. – Vol. 12. – P. 29–40. 21. 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.). 22. Repetition characteristic for single texts / O. S. Kushnir, L. B. Ivanitskyi, A. I. Kashuba, M. R. Mostova, V. B. Mykhaylyk // Proc. 5th Int. Conf. on Computational Linguistics and Intelligent Systems (COLINS-2021). Vol. I: Main Conference. Ed. by N. Sharonova et al. – Lviv, Ukraine, April 22-23, 2021. – CEUR Workshop Proc., Vol. 2870. – P. 629–641. 23. Glatch S. Repetition definition: types of repetition in poetry and prose / S. Glatch – "writers.com", 2024. URL: https://writers.com/repetition-definition 24. System of automatic determination of Ukrainian text tone / I. Olenych, M. Prytula, O. Sinkevych, O. Khamar // Proc. IEEE XII-th International Conference on Electronics and Information Technologies. – Lviv, Ukraine, 19-21 May 2021. – P. 80–83.
Обсяг курсу	Сумарно 180 годин. Із них 32 години лекцій, 32 годин лабораторних

	робіт 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.
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 з дисциплін «Вища математика», «Дискретна математика», «Алгоритми та структури даних», «Чисельні методи», «Теорія ймовірності та математична статистика», «Об'єктно-орієнтоване програмування», «Бази даних та знань», «Основи комп'ютерної лінгвістики» та ін.
Навчальні методи та техніки, які будуть використовуватися під час викладання курсу	Лекції, презентації, лабораторні роботи, індивідуальні практичні завдання, обговорення, дискусії.
Необхідне обладнання	Мультимедіа, платформи Microsoft Teams, Moodle і Zoom, комп'ютерне програмне забезпечення
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 або індивідуальні практичні роботи: 40% оцінки; максимальна кількість балів 40. • модульні контрольні заміри знань на лекціях: 10% оцінки; максимальна кількість балів 2x5 = 10. • іспит: 50% оцінки; максимальна кількість балів 50. Загалом 100 балів. _____. **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 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 контрольних робіт	Перелік питань і завдань для проведення підсумкової оцінки знань усіх тем курсу до контрольних робіт розміщено на веб-сторінці: 1. Лінгвістика та її структура. 2. Загальні поняття про мову. Мова і мислення. Мова як знакова система. Мова і мовлення. Мовна структура. Рівні мови. 3. Фонологія. Поняття про фонеми. Звуки і букви. Поняття про алфавіт. 4. Морфологія. Основні поняття. Морфема. Види морфем. 5. Синтаксис. Основні поняття. Словосполучення. Речення. Синтаксичні відносини. Керування. Узгодження. Комунікативна організація речення. 6. Семантика. Семантичні відносини. Парадигматичні та синтагматичні відносини. Формальні моделі семантики. Семантичні мережі. Фрейми. 7. Лексикографія. Основи комп'ютерної лексикографії. 8. Прагматика. Основні поняття. Проблема розуміння текстів. 9. Історична (компаративна) лінгвістика. Синхронія і діахронія. 10. Основні родини мов. Приклади взаємних впливів. 11. Контрастивна лінгвістика або лінгвістична топологія. 12. Соціолінгвістика та діалектологія. Психолінгвістика. 13. Зв'язки комп'ютерної лінгвістики із системами штучного інтелекту та складними системами. 14. Основні завдання опрацювання природної мови. 15. Комп'ютерна лінгвістика. Квантитативна, статистична та математична лінгвістика. 16. Комп'ютерна лексикографія. 17. Приклади «lingware». 18. Контекстно-вільна граматика. Трансформаційна граматика. 19. Теорія «зміст–текст» у лінгвістиці. 20. Поняття «слово» в лінгвістиці. Словоформа, слововживання і лексема. Типи та токени. 21. Поняття систем. Системи із взаємодіючими елементами. Складні системи і мережі. Приклади. Лінгвістичні системи та мережі. 22. Статистичні підходи в лінгвістиці. Детермінізм і стохастичний підхід. Випадкові величини. Дискретні та неперервні величини. 23. Генеральна сукупність і вибірка. Репрезентативність вибірки. Ефекти «скінченного розміру» вибірки. Середнє значення та дисперсія вибірки. 24. Розподіли ймовірності випадкових величин. Гістограми. Масова функція розподілу ймовірності. Густина розподілу ймовірності. 25. Кумулятивна функція розподілу. Приклад однорідного розподілу. 26. Частотні таблиці слів. Алгоритмізація побудови частотних лексичних таблиць. 27. Статистичні моменти розподілів. Проблема існування моментів. Середнє значення випадкової змінної для «несправедливих» розподілів. 28. Середнє та середньоквадратичне відхилення для неперервних і дискретних випадкових величин. Зважені параметри. 29. Приклади розподілів випадкових величин. Зріст людини. Населення міст. 30. Приклади розподілів випадкових величин. Сила землетрусів. Розподіл багатства між людьми. Відвідування сайтів. 31. Частота слів і букв. Нульові гіпотези в статистичній лінгвістиці.

	32. Розподіл Пуасона. 33. Експоненційний і «розширений» експоненційний розподіли. 34. Розподіл Вейбуля. «Логарифмічні» рангові залежності. 35. Нормальний розподіл. Причини універсальності нормального розподілу. Правило «трьох сигм». 36. Лог-нормальний розподіл. «Оманливі» степеневі розподіли. 37. Розподіл Ґуда. Частоти речень і слів різних довжин. 38. Розподіли з «важким хвостом». Степеневі розподіли. Розподіл Парето. 39. «Справедливі» та «несправедливі» розподіли. Правило Парето. 40. Степеневі розподіли. Скейлінг. Візуалізація степеневих розподілів. 41. Препроцесинг текстів. Парсинг. 42. Тегування, стеммінг і лемматизація. 43. Методика досліджень степеневих розподілів. Шуми функції $p(x)$. Бінування. Типи бінування. 44. Алгоритми визначення показника степеневих розподілів. Нелінійна апроксимація. 45. Метод максимальної схожості в оцінюванні показника степеневої залежності. 46. Критерій хі-квадрат у статистичній лінгвістиці. Інші статистичні критерії та тести. 47. Рангові залежності частоти слів. Абсолютна та відносна частота. 48. Перший і другий закони Ціпфа для лексики. Поняття про лексичний спектр. 49. Кумулятивна ймовірність появи слів з різними частотами. Закон Парето. 50. Лексичний словник. Закон Гіпса–Гердана. 51. Синтетичні та аналітичні мови. Порівняння словників синтетичних і аналітичних мов. 52. Метод біжучого вікна для розрахунку залежності словника від розмірів тексту. 53. Корпуси текстів. Закон Гіпса для корпусів текстів.. 54. Альтернативи степеневих законів Ціпфа в лінгвістиці. Логарифмічні та експоненційні рангові розподіли і їхня графічна візуалізація. Ієрогліфічні мови з обмеженим словником. 55. Рангові залежності для частоти букв і типів складів. Залежність розмірів «словника» букв від довжини тексту. Рангові залежності в спортивних турнірах. 56. Взаємні зв’язки експонент степеневих законів для лінгвістики. Ілюстрації для статистики букв і східних мов. 57. Приклади статистичних розподілів поза межами лінгвістики. Розподіли ймовірності для імен і прізвищ. Прізвища в Кореї. Популярність наукових журналів, статей і авторів. Індекс Герша. 58. Показник степеневого розподілу лексичної ймовірності для текстів і проблема психічних захворювань. Особливості лексичних спектрів для дітей. 59. Основні теорії для пояснення законів Ціпфа. Нульові стохастичні моделі. «Фазовий перехід» комунікація–відсутність комунікації. 60. Закони лінгвістики для текстів програм. 61. Особливості статистики графем і фонем. Рангові та частотні залежності для букв і знаків. «Шуми» в рангових залежностях. «Словник» літер і закон Гіпса для літер. 62. n-грами в лінгвістиці. Буквені, символьні та лексичні n-грами. Узагальнення законів статистичної лінгвістики для n-грам. 63. Рангові та частотні залежності для слів різних довжин. Середня довжина слова в різних мовах. Особливості флуктуацій частоти літер. 64. Інформаційна ентропія Шеннона. Ентропії для залежностей $F(r)$ і $p(F)$. «Довжина тексту» з урахуванням ентропії. 65. Закони, що пов’язують лінгвістичні елементи різних рівнів. Закон Менцерата–Альтмана. 66. Розподіли ймовірності першої появи слів і лексичних n-грам у тексті. 67. Рандомні тексти. Типи рандомних текстів. 68. Тексти Хомського та Саймона. Модель зростання Саймона. Модифіковані моделі Саймона. 69. Тексти «мавпи Міллера». Тексти на основі ланцюжків Маркова. 70. Рандомізовані тексти. Локальна та глобальна рандомізація. Рандомізація та кореляції в текстах. 71. Властивості рандомних текстів. Проблема розпізнавання природних і рандомних текстів. Інвертування текстів. 72. Основні підходи в поясненні законів статистичної лінгвістики. Наближене відтворення законів Ціпфа, Парето та Гіпса.

	73. Принцип «багатий стає багатшим» у складних системах. 74. Модель «дискурс–оточення». Поняття тригерування інновацій. 75. Принцип переважного приєднання в складних мережах. 76. Характеристика повторюваності в текстах. Основні означення. 77. Алгоритм та режими обчислень характеристики повторень. 78. Характеристика повторень для різних типів текстів. Пошук «самоплагіату» в текстах. 79. Порівняння і класифікація текстів. Лінгвістичні маркери. 80. Категоризація та класифікація текстів. Індексування текстів. 81. Проблеми визначення мови і встановлення авторства, жанру та стилю. Стилістика та її статистичні ознаки. 82. Проблема розпізнавання текстового плагіату. «Кусковий» плагіат. 83. Статистичні ваги маркерів текстів. Методика порівняння і класифікації текстів із машинним навчанням. 84. Параметри «recall» і «precision» в інформаційному пошуку. 85. Поняття подібності текстів і «відстаней» між текстами. Дефініції «відстаней». «Відстані» між ранговими залежностями і лексичними спектрами. 86. Опис подібності текстів на основі векторного простору. Скалярний добуток текстів. 87. Поняття ключових слів і фраз (n-грам). Частотні ознаки ключових слів. 88. «Абсолютні» та «відносні» методи визначення ключових слів. 89. Метод TF-IDF. Ключові слова в інформаційному пошуку. Метод Бріна та Пейджа. 90. Закон рідкісних подій і розподіл Пуасона. 91. Час очікування рідкісної події. Приклади. Аварії та технічні поламки, землетруси, серцеві ритми. 92. Час очікування рідкісної події в лінгвістиці. Експоненційний розподіл для часів очікування рідкісних подій. 93. Відхилення від експоненційного розподілу часів очікування рідкісних подій. Функціональні, змістові та ключові слова. 94. Семантика тексту. Розрізнення природних і рандомних текстів на основі кластеризації слів. 95. Ключові слова в текстах комп'ютерних програм. 96. Флуктуації. Приклади зі статистичної механіки. Відносні флуктуації. 97. Мікроскопічне, мезоскопічне і макроскопічне наближення в статистичній лінгвістиці. 98. Кореляції в статистичній лінгвістиці. Короткосяжні та довгосяжні кореляції. 99. Поняття кореляцій. Флуктуації за наявності кореляцій. 100. Поняття скейлінгу флуктуацій. Самоусереднювані статистичні величини. 101. Флуктуації в лінгвістиці. Відображення тексту на числовий ряд. 102. Закон Тейлора. 103. Часова мова в описі флуктуацій. Метод рандомних прогулянок. Нормальна та аномальна дифузія. Підхід біжучого вікна в дослідженні флуктуацій у лінгвістиці. 104. Ресурси флуктуаційного аналізу: кореляції, ключові слова та семантика тексту. 105. Флуктуації частоти слів у текстах програм. 106. Флуктуаційний аналіз на корпусах текстів. 107. Мережі. Основні параметри мереж. 108. Властивості складних мереж: перколяція, ефект тісного світу та механізм переважного приєднання. 109. Приклади складних мереж та їхні властивості. 110. Лінгвістичні мережі. Означення зв'язків між вузлами мережі. 111. Ефект «тісного світу» в лінгвістичних мережах. 112. Мережеві методи пошуку ключових слів текстів.
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 лінгвісти-**	Лекція	1, 3, 4, 6–10, 14	Встановлення наявності семантики текстів на основі мережевих парамет-	1, 2 тиж. семестру

	ки 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.			рів, кластеризації та кореляції їхніх лінгвістичних елементів	
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	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	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	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 абстрагування текстів. Встановлення мови, авторства, стилістики та плагіату. Скейлінг у лінгвістиці та інших складних системах. Мережеві властивості текстів. Інформаційний пошук.	Лекція	2–4, 6, 7, 11, 18, 23, 27	Вивчення розподілів імовірності часів очікування лінгвістичних елементів і відносні методи визначення ключових слів у текстах, засновані на кластеризації Флуктуації та кореляції в текстах. Метод FA Флуктуаційний аналіз на текстових базах	13, 14 тиж. семестру

				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.	Лекція	2, 3, 5, 7, 9, 11, 15, 16	Типові шляхи вирішення проблем розпізнавання письмової мови	15, 16 тиж. семестру