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 29.08 2023 р.)

Завідувач кафедри:

_____ Олег КУШНІР

Силабус з навчальної дисципліни "Аналіз текстової інформації", що викладається в межах ОПП "Комп'ютерні науки" першого (бакалаврського) рівня вищої освіти для здобувачів з спеціальності 122 – Комп'ютерні науки

Львів 2023 р.

Назва дисципліни	Аналіз текстової інформації
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): 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	http://194.44.208.156/moodle/course/view.php?id=59 https://drive.google.com/drive/folders/1Kamy1aZ080cxg0ZwuDX9W8kT2rADpYDF?usp=sharing
Інформація про дисципліну	Дисципліна «Аналіз текстової інформації» є вибірковою дисципліною студентів зі спеціальності 122 Комп'ютерні науки для освітньої програми «Комп'ютерні науки», 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 опрацювання природної мови та аналізу текстової інформації, а також для формування в них навичок ефективного застосовування засвоєних знань і методів у розв'язанні прикладних задач такого аналізу. Представлено теоретичні основи комп'ютерної лінгвістики та опрацювання природної мови, класифікація та огляд особливостей відомих продуктів у цій галузі, а також відповідні алгоритми і засоби опрацювання даних.
Мета та цілі дисципліни	*Метою* вивчення дисципліни «Аналіз текстової інформації» є ознайомлення студентів з теоретичними основами опрацювання природної мови, а *цілями* – формування в студентів практичних навичок, які б дали змогу ефективно застосовувати засвоєні знання, алгоритми, методи та прикладні програми.
Література для вивчення дисципліни	*Основна література*: 1. Кушнір О. С. Основи комп'ютерної лінгвістики (конспект лекцій) / О. С. Кушнір. – Львів: Видавн. Львів. ун-ту, **2023**. – 292 с. 2. Волошин В. Г. Комп'ютерна лінгвістика / В. Г. Волошин. – Суми : Університетська книга, 2004. – 382 с. 3. Bird S. Natural language processing with Python / S. Bird, E. Klein, E. Loper. – Sebastopol : O'Reilly. – 2009. – 504 p. 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 & Sons, 2010. – 801 p. 6. Kracht M. Introduction to probability theory and statistics for linguistics / M. Kracht. – Oakland : UCLA, 2005. – 137 p.

	7. 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. 8. https://www.gutenberg.org/ *Додаткова література*: 9. Pilgrim C. Bias in Zipf's law estimators / C. Pilgrim, T. T. Hills // Sci. Rep. – 2021. – Vol. 11. – 17309 (12 pp.). 10. 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/ 11. 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.).
Обсяг курсу	Аудиторні години – 80, з них лекції – 32 години, лабораторні роботи – 48 годин і 70 годин самостійної роботи.
Очікувані результати навчання	Після завершення цього курсу студент буде: - знати основні методи аналізу текстової інформації та опрацювання природної мови, основні теорії, моделі та алгоритми галузі і опису лінгвістичних систем, інформаційного пошуку та інтелектуального аналізу текстових даних; - вміти аналізувати моделі 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, опрацювання природної мови, текстова інформація, текстові дані, статистична лінгвістика, квантитативна лінгвістика, штучний інтелект
Формат курсу	Очний
	Проведення лекцій, лабораторних та індивідуальних практичних робіт, а також консультації для поглибленого розуміння тем
Теми	Див. СХЕМА КУРСУ
Підсумковий контроль, форма	Залік 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», «Системи штучного інтелекту», «Аналіз даних», «Машинне навчання».
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, 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 (10 робіт, максимально 10х6=60 балів) або індивідуальні (1 програмістська або дослідницька робота, можлива також командна; максимально 1х60=60 балів) практичні роботи: 60% оцінки; максимальна кількість балів 60. • активність на лекціях (відвідування занять, відповіді на питання, участь у дискусіях і обговоренні, внесок у розв'язання проблемних ситуацій тощо): максимально 20 балів або 20% оцінки. • 1 письмовий модульний контроль (на лекції): максимально 1х30=30 балів або 20% оцінки. Загалом 100 балів. Оцінки за лабораторні заняття (6 балів) розподіляються так: повне виконання всіх лабораторних завдань – 33% (2 бали), теоретичні знання за предметом роботи та правильні відповіді на запитання викладача або тестування – 33% (2 бали), якість інтерпретації даних і оформлення звіту – 33% (2 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: 6 – студент в повному обсязі володіє матеріалом, має повне розуміння розглянутої теми, надає правильні відповіді на запитання по темі, код програми функціонує відповідно до завдання; 5 – 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); 4 – 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; 3 – студент погано розуміє матеріал та написаний код програми, в більшості надає помилкові відповіді на питання по темі, код програми функціонує з суттєвими недоліками; 1–2 – 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** – за результатами висвітлення студентом деяких теоретичних тем тощо. Бали оцінювання змістових модулів нараховуються за наступним співвідношенням: 18–20 – 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; 13–17 – 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; 9–12 – 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); 4–8 – студент погано розуміє тему. Виявлено суттєві неточності та невідповідності. Наведені приклади коду мають суттєві недоліки або не відповідають темі; 0–3 – 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. _____. **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 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 модульного контролю	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. Комп'ютерна лінгвістика** Місце комп'ютерної лінгвістики в галузі комп'ютерних наук. Лінгвістика та її структура. Базові поняття лінгвістики.	Лекція	2, 3, 7	Вступне заняття. Академічна доброчесність. Препроцесинг текстових документів Фонетичний аналіз і силабіфікація на основі методу сонорності	1, 2 тиж. семестру
3, 4	**Основні ідеї та методи комп'ютерної лінгвістики** Мова як двонаправлений перетворювач зміст<->текст. Лінгвістичні знаки та лінгвістичні моделі. Структурний підхід Н. Хомського. Граматики.	Лекція	1, 2, 4, 7, 8	Програмування задач морфологічного синтезу дієслівних форм	3, 4 тиж. семестру
5, 6	**Лінгвістичні знаки та лінгвістичні моделі** Лінгвістичні знаки. Лінгвістичні моделі. Поняття тексту і змісту. Способи представлення змісту. Розкладання і атомізація змісту. Неоднозначність «картування» змісту.	Лекція	1, 4, 8, 9	Параметр повторюваності лінгвістичних елементів у тексті	5, 6 тиж. семестру
7, 8	**Системний аналіз текстів. Поняття статистичної лінгвістики** Системи та мережі. Тексти як складні системи та мережі. Основні поняття статистичної лінгвістики.	Лекція	2, 4, 7, 10	Вивчення розподілів імовірності часів очікування різних лінгвістичних елементів	7, 8 тиж. семестру
9, 10	**Методика та закони статистичної лінгвістики** Методичні особливості вивчення лінгвістичної статистики. Основні закони лінгвістики для слів. Статичні та динамічні закони.	Лекція	1, 2, 10	Закони Ціпфа та Парето для слів і лексичних n-грам у текстах Закон Гіпса для слів в окремих текстах. Зростання словника для корпусу текстів	9, 10 тиж. семестру
11, 12	**Степеневі розподіли та інші закони лінгвістики** Механізми степеневих розподілів у лінгвістиці. Закони лінгвіс-	Лекція	1, 7, 8, 11, 12, 13	Закони статистичної лінгвістики на лінгвістичних рівнях букв (символів) і символьних n-грам для окремих текстів	11, 12 тиж. семестру

	тичної статистики для інших лінгвістичних рівнів.			Вивчення розподілів імовірності часів очікування різних лінгвістичних елементів	
13, 14	**Нульові гіпотези. Пошук ключових слів** Нульові стохастичні гіпотези в статистичній лінгвістиці. Пошук ключових слів у текстах. Встановлення мови, авторства, стилістики та плагіату.	Лекція	7, 8, 11, 12	Генерування та вивчення властивостей рандомних текстів різних типів Методи визначення ключових слів у текстах	13, 14 тиж. семестру
15, 16	**Кореляції лінгвістичних елементів у текстах** Автокореляція. Скейлінг флуктуацій у лінгвістиці та інших складних системах. Мережеві властивості природних і рандомних текстів**.**	Лекція	1, 5, 6, 8, 13	Флуктуації та кореляції в текстах	15, 16 тиж. семестру