

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 18 серпня 2021 р.)

Завідувач кафедри _____

Силабус з навчальної дисципліни
«Обробка природної мови»,
що викладається в межах ОПП «Інформаційні системи та
технології» першого (бакалаврського) рівня вищої освіти
для здобувачів зі спеціальності
126 Інформаційні системи та технології

Назва дисципліни	Обробка природної мови
Адреса викладання дисципліни	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, 126 Інформаційні системи та технології
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/obrobka-pryrodnoi-movy-ist http://194.44.208.156/moodle/course/view.php?id=59
Інформація про дисципліну	Дисципліна «Обробка природної мови» є нормативною дисципліною спеціалізації зі спеціальності 126 Інформаційні системи та технології для освітньої програми «Інформаційні системи та технології», яка викладається в 5 семестрі в обсязі 3,5 кредитів (за Європейською Кредитно-Трансферною Системою – ECTS).
Коротка анотація дисципліни	Навчальну дисципліну розроблено для одержання студентами теоретичних знань з основ обробки природної мови та формування в них практичних навиків застосування засвоєних знань і методів у розв'язанні прикладних задач. Представлено теоретичні основи комп'ютерної, статистичної та математичної лінгвістики, класифікація та огляд особливостей відомих продуктів комп'ютерної лінгвістики та обробки природної мови.
Мета та цілі дисципліни	<i>Метою</i> вивчення дисципліни «Обробка природної мови» є ознайомлення студентів з основами комп'ютерної лінгвістики та обробки природної мови, а також 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. Волошин В. Г. Комп'ютерна лінгвістика / В. Г. Волошин. – Суми : Університетська книга, 2004. – 382 с. 2. Bolshakov I. Computational linguistics. Models, resources, applications / I. Bolshakov, A. Gelbukh. – Mexico : Ciencia de la Computacion, 2004. – 198 p. 3. Кушнір О. С. Основи комп'ютерної лінгвістики (конспект лекцій) / О. С. Кушнір. – Львів: Видавн. Львів. ун-ту, 2023. – 292 с. 4. 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. 5. https://www.gutenberg.org/ Додаткова література: 6. Jurafsky D. Speech and language processing / D. Jurafsky, J. H. Martin. – New Jersey : Prentice Hall, 2023. – 628 p. 7. Clauset A. Power-law distributions in empirical data / Clauset A., Shalizi C. R., Newman M. E. J. // SIAM Rev. – 2009. – Vol. 51. – P. 661–703. 8. 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	Сумарно 105 годин. Із них 32 години лекцій, 32 години лабораторних робіт, а також 41 година самостійної роботи
Очікувані результати навчання	Після завершення цього курсу студент буде: <ul style="list-style-type: none"> - знати основні методи комп'ютерної лінгвістики, основні теорії, моделі та алгоритми комп'ютерної лінгвістики і опису лінгвістичних систем, інформаційного пошуку та інтелектуального аналізу текстових даних; - вміти аналізувати моделі комп'ютерної лінгвістики, 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, обробка природної мови, комп'ютерна лексикографія, аналіз і синтез мовлення
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».
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)	<p>Оцінювання проводиться упродовж семестру та під час екзаменаційної сесії за 100-бальною шкалою. Бали нараховуються за такими видами робіт із таким співвідношенням:</p> <ul style="list-style-type: none"> • лабораторні роботи: 40% оцінки; максимальна кількість балів 40. • контрольні заміри знань на лекціях: 10% оцінки; максимальна кількість балів $2 \times 5 = 10$. • іспит: 50% оцінки; максимальна кількість балів 50. <p>Загалом 100 балів.</p> <hr/> <p>Контрольні заміри знань проводять у формі стандартних практичних завдань і теоретичних питань.</p> <p>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.</p> <p>Відвідування занять є важливою складовою навчання. Очікується, що всі студенти відвідають усі лекції та лабораторні заняття курсу. Студенти мають інформувати викладача про неможливість відвідати занят-</p>

	<p>тя. Студенти зобов'язані дотримуватися всіх термінів, визначених для виконання видів робіт, передбачених курсом.</p> <p>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.</p> <p>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 запізнь 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.</p> <p>Жодні форми порушення академічної доброчесності не толеруються.</p>
Питання до контрольних робіт	Перелік питань і завдань для проведення підсумкової оцінки знань усіх тем курсу до контрольних робіт розміщено на веб-сторінці.
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	Комп'ютерна лінгвістика. Структура лінгвістики. Основні поняття Зв'язки комп'ютерної лінгвістики з галузями інформатики та систем штучного інтелекту. Лінгвістика та її структура. Загальні поняття про мову. Семантика. Прагматика.	Лекція	1, 2, 4	Попереднє опрацювання текстів. Дослідження швидкодії програм Фонетичний і силабічний аналіз текстів східнослов'янськими мовами. Програмування задачі визначення співвідношення кількості голосних і приголосних	1, 2 тиж. семестру
3, 4	Ідеї, теорії та методи комп'ютерної лінгвістики Мова як двонаправлений перетворювач зміст->текст. Лінгвістичні знаки та лінгвістичні моделі. Структурний підхід Н. Хомського. Граматики.	Лекція	1, 2, 3, 4	Характеристика повторюваності символів у текстах Генерування фрагментів тексту: формальні контекстно-вільні граматики Побудова дерев синтаксичних залежностей для стверджувальних речень	3, 4 тиж. семестру
5, 6	Лінгвістика та системний підхід Системи та мережі. Складні системи та мережі. Приклади. Поняття статистичної лінгвістики. Методика вивчення лінгвістичної статистики.	Лекція	3, 5, 7, 8	Побудова гістограм словника текстової бази	5, 6 тиж. семестру
7, 8	Основні закони лінгвістики Статичні та динамічні закони лінгвістики. Статистика інших складних систем. Механізми степеневих розподілів.	Лекція	3, 4, 7	Закони статистичної лінгвістики для окремих текстів. Статистика лексичних н-грам	7, 8 тиж. семестру
9, 10	Інші закони лінгвістики Закони лінгвістичної статистики для інших лінгвістичних рівнів.	Лекція	3, 4, 5, 7	Визначення середньої довжини слів і речень	9, 10 тиж. семестру

	Нульові стохастичні моделі лінгвістики			Дослідження базових лінгвістичних властивостей рандомних символічних послідовностей	
11, 12	Основи опрацювання природної мови Пошук ключових слів у текстах. Встановлення мови, авторства, стилістики та плагіату.	Лекція	2, 3, 4, 6, 7, 8	Абсолютні методи визначення ключових слів у текстах. Метод TF-IDF Визначення семантичного навантаження текстів на основі параметрів кластеризації	11, 12 тиж. семестру
13, 14	Скейлінг і мережеві властивості в лінгвістиці Скейлінг у лінгвістиці та інших складних системах. Мережеві властивості природних і рандомних текстів.	Лекція	3, 5, 6, 8	Робота з базами лінгвістичних даних. Побудова запитів з урахуванням лінгвістичних міркувань	13, 14 тиж. семестру
15, 16	Огляд сучасних продуктів комп'ютерної лінгвістики Класифікація прикладних лінгвістичних систем. Мовна перевірка. Переклад. Інформаційні пошукові системи. Генерація тексту. Аналіз і синтез мовлення.	Лекція	2, 3, 4, 6, 8	Обробка природних мов за допомогою засобів Python Програмні пакети машинного перекладу Stylus і MemoQ	15, 16 тиж. семестру