МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Львівський національний університет імені Івана Франка
Факультет прикладної математики та інформатики
Кафедра прикладної математики

Затверджено
На засіданні
кафедри прикладної математики
факультету прикладної математики та інформатики
Львівського національного університету імені Івана Франка
(протокол № ____ від _____ 2023р.)

Завідувач кафедри

_____Юрій ЯЩУК

Силабус з навчальної дисципліни "Статистичні моделі в комп'ютерній лінгвістиці", що викладається в межах ОПП Прикладна математика другого (магістерського) рівня вищої освіти для здобувачів з спеціальності 113 – прикладна математика

Львів 2023 р.

Назва дисципліни	Статистичні моделі в комп'ютерній лінгвістиці
Адреса викладання дисципліни	Головний корпус ЛНУ ім. І. Франка м. Львів, вул. Університетська 1
Факультет та кафедра, за якою закріплена дисципліна	Факультет прикладної математики та інформатики Кафедра прикладної математики
Галузь знань, шифр та назва спеціальності	11 – математика та статистика 113– прикладна математика
Викладачі дисципліни	Чирун Любомир Вікторович, доцент кафедри прикладної математики
Контактна інформація викладачів	lyubomyr.chyrun@lnu.edu.ua; chyrunlv@gmail.com; Головний корпус ЛНУ ім. І. Франка, каб. 278. м. Львів, вул. Університетська, 1
Консультації з питань навчання по дисципліні відбуваються	Консультації в день проведення лекцій/практичних занять (за попередньою домовленістю).
Сторінка курсу	https://ami.lnu.edu.ua/course/statystyni-modeli-v-komp-iuterniy-linhvistytsi-prykladna-matematyka
Інформація про дисципліну	Дисципліна "Статистичні моделі в комп'ютерній лінгвістиці" є вибірковою дисципліною з спеціальності 113 – прикладна математика для освітньої програми Прикладна математика, яка викладається в 2-му семестрі в обсязі 3-ох кредитів (за Європейською Кредитно-Трансферною Системою ECTS).
Коротка анотація дисципліни	Курс розроблено таким чином, щоб надати учасникам знання про статистичні моделі комп'ютерної лінгвістики, принципів їх побудови, а також про основні задачі опрацювання природної мови.
Мета та цілі дисципліни	Метою вивчення вибіркової дисципліни "Статистичні моделі в комп'ютерній лінгвістиці" є освоєння студентами теоретичних і практичних основ в галузі комп'ютерної лінгвістики та основ опрацювання природної мови.
Література для вивчення дисципліни	1. https://www.python.org 2. C Manning, P Raghavan, H Schutze - Introduction to information retrieval, 2008 3. Christopher D Manning, Hinrich Schütze - Statistical natural Language processing, Cambridge, MA: MIT Press ,1999
Обсяг курсу	Загальний обсяг: 90 години. Аудиторних занять: 32 год., з них 16 год. лекцій, 16 год. практичних. Самостійної роботи: 58 год.
Очікувані результати навчання	Після завершення цього курсу студент буде : Знати: - Основні задачі опрацювання природної мови; - Ймовірнісні лінгвістичні моделі; - Головні підходи до побудови відповідних моделей; - Поняття лематизації та токенізації корпусу тексту; - Поняття уніграм, біграм та методи роботи з ними. Вміти:

	- Будувати алфавітно частотні словники згідно досліджуваного корпусу тексту та зберігати їх для подальшого опрацювання; - Нормалізувати алфавітно-частотні словники; - Реалізовувати ймовірнісні лінгвістичні моделі.
Ключові слова	Комп'ютерна лінгвістика, лематизація, токенізація, NLP
Формат курсу	Очний Проведення лекцій і консультацій, лабораторних робіт.
Теми	1. Основна задача опрацювання природної мови (NLP). 3 типи NLP застосувань. Аналіз. Аналіз відчуттів, настрою користувача. Визначення сарказму. Типи персонажів серед користувачів. Машинний переклад. Машинний переклад. 2. Типи NLP застосувань. Системи перетворень мова-текст, текст-мова. Відповіді на запитання. Анотування документів. Чат-боти. Застосунки для вивчення мови. Задача завершення оповіді. 3. Які завдання вирішує комп'ютерний лінгвіст. Задача парсування текстів. Поняття токена та токенізації. Основні помилки для вирішення при корекції помилок. Завдання для застосунків виправлення помилок. Підходи до задачі виправлення помилок. 4. Робота з патернами при виправленні помилок. Статистика при виправленні помилок. Поняття N-грам. 5. Машинне навчання при виправленні помилок. Round-Trip. 6. Регулярні вирази. Пошук. Пошук по діапазонах. Пошук по запереченнях. Пошук з використанням операторів ? * + . Пошук з використанням операторів ^ $. Типи помилок. 7. Задача нормалізації тексту. Задача токенізації тексту. Поняття леми та словоформи. 8. Токени та типи. Поняття алфавітно-частотного словника. Навіщо його будувати. 9. Токенізація. Особливості деяких мов. Токенізація китайської мови. Підходи 10. Алгоритм максимуму співпадінь. Нормалізація тексту. Навіщо. Власні назви та абревіатури при токенізації. Що робити. Лематизація. Що це таке. Навіщо. Морфологія. Стеммінг. 11. Алгоритм Портера. 12. Сегментація речень. 13. Дерева прийняття рішень в задачі сегментації речень. 14. Реалізація дерев прийняття рішень. 15. Інші підходи окрім дерев прийняття рішень для класифікаторів. 16. Ймовірнісні лінгвістичні моделі. Задачі для вирішення. 17. Ймовірнісні лінгвістичні моделі. Мета та завдання. 18. Правило Чена. 19. Правило Чена для послідовності слів у реченні. 20. Припущення Маркова. 21. Припущення Маркова. Уніграми. 22. Припущення Маркова. Біграми. 23. Оцінка ймовірностей біграм. 24. Перехід від ймовірностей до логарифмів. Чому. 25. Оцінка моделі мови. 26. Поняття навчальної та тестової множин. 27. Поняття perplexity. 28. Метод Шенона. 29. Встановлення авторства на триграмах. Що робити з нульовими ймовірностями. 30. Оцінка add-1.

	31. Оцінка максимальної подібності. 32. Лінійна інтерполяція. 33. Проблема визначення коефіцієнтів при інтерполяції. Що робити. 34. Згладжування N-грам. 35. Згладжування Кнесера-Нея. 36. Задача класифікації текстів. Постановка задачі. 37. Задача класифікації текстів. Правила вручну. Переваги та недоліки. 38. Припущення Наїва-Баєса. Класифікатори Наїва-Баєса.
Підсумковий контроль, форма	Залік.
Пререквізити	Для вивчення курсу студенти потребують базових знань з - Математики; - Логіки; - Математичної статистики; - Теорії ймовірності.
Навчальні методи та техніки, які будуть використовуватися під час викладання курсу	Презентації, лекції Індивідуальні завдання Групові проекти, менторство
Необхідне обладнання	Комп'ютер із програмним забезпеченням Visual Studio 2017/2019, Internet доступ.
Критерії оцінювання (окремо для кожного виду навчальної діяльності)	Оцінювання проводиться за 100-бальною шкалою. Бали нараховуються за наступним співвідношенням: • індивідуальні завдання : 50% семестрової оцінки; максимальна кількість балів 50 • підсумковий тест: 50% семестрової оцінки; максимальна кількість балів 50 Підсумкова максимальна кількість балів 100. **Письмові роботи:** Очікується, що студенти виконають одну письмову роботу (тест з теоретичних завдань) і звіт про виконання індивідуальних та командних завдань. **Академічна доброчесність**: Очікується, що роботи студентів будуть їх оригінальними дослідженнями чи 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 письмовій роботі студента є підставою для її незарахуванння викладачем, незалежно від масштабів плагіату чи обману. **Відвідання занять** є важливою складовою навчання. Очікується, що всі студенти відвідають усі лекції та лабораторні зайняття курсу (дистанційно). Студенти повинні інформувати викладача про неможливість відвідати заняття. У будь-якому випадку студенти зобов'язані дотримуватися термінів визначених для виконання всіх видів письмових робіт та індивідуальних завдань, 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 заохочуються до використання також й іншої літератури та джерел, яких немає серед рекомендованих. **Політика виставлення балів.** Враховуються бали набрані за індивідуальні завдання та підсумковий тест. При цьому обов'язково

	враховуються присутність на заняттях та активність студента під час практичного заняття; недопустимість пропусків та запізнень на заняття; користування мобільним телефоном, планшетом чи іншими мобільними пристроями під час заняття в цілях не пов'язаних з навчанням; списування та плагіат; несвоєчасне виконання поставленого завдання і т. ін. Жодні форми порушення академічної доброчесності не толеруються.
Питання до заліку чи екзамену.	Машинне навчання при виправленні помилок. Round-Trip. Ймовірнісні лінгвістичні моделі. Задачі для вирішення. Типи NLP застосувань. Припущення Маркова. Біграми. Оцінка ймовірностей біграм. Перехід від ймовірностей до логарифмів. Чому. Оцінка моделі мови. Поняття навчальної та тестової множин. Встановлення авторства на триграмах. Що робити з нульовими ймовірностями.
Опитування	Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенню курсу.