

**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 ЯЦУК**

**Силабус з навчальної дисципліни**  
**“Статистичні моделі в комп’ютерній лінгвістиці”,**  
**що викладається в межах ОПІ Прикладна математика**  
**другого (магістерського) рівня вищої освіти для здобувачів з**  
**спеціальності 113 – прикладна математика**

**Львів 2023 р.**

<b>Назва дисципліни</b>	Статистичні моделі в комп'ютерній лінгвістиці
<b>Адреса викладання дисципліни</b>	Головний корпус ЛНУ ім. І. Франка м. Львів, вул. Університетська 1
<b>Факультет та кафедра, за якою закріплена дисципліна</b>	Факультет прикладної математики та інформатики Кафедра прикладної математики
<b>Галузь знань, шифр та назва спеціальності</b>	11 – математика та статистика 113– прикладна математика
<b>Викладачі дисципліни</b>	Чирун Любомир Вікторович, доцент кафедри прикладної математики
<b>Контактна інформація викладачів</b>	<a href="mailto:lyubomyr.chyrun@lnu.edu.ua">lyubomyr.chyrun@lnu.edu.ua</a> ; <a href="mailto:chyrunlv@gmail.com">chyrunlv@gmail.com</a> ; Головний корпус ЛНУ ім. І. Франка, каб. 278. м. Львів, вул. Університетська, 1
<b>Консультації з питань навчання по дисципліні відбуваються</b>	Консультації в день проведення лекцій/практичних занять (за попередньою домовленістю).
<b>Сторінка курсу</b>	<a href="https://ami.lnu.edu.ua/course/statystyni-modeli-v-komp-iuterniy-linhvistytsi-prykladna-matematyka">https://ami.lnu.edu.ua/course/statystyni-modeli-v-komp-iuterniy-linhvistytsi-prykladna-matematyka</a>
<b>Інформація про дисципліну</b>	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).
<b>Коротка анотація дисципліни</b>	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.
<b>Мета та цілі дисципліни</b>	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.
<b>Література для вивчення дисципліни</b>	<ol style="list-style-type: none"> <li><a href="https://www.python.org">https://www.python.org</a></li> <li>C Manning, P Raghavan, H Schütze - Introduction to information retrieval, 2008</li> <li>Christopher D Manning, Hinrich Schütze - Statistical natural Language processing, Cambridge, MA: MIT Press ,1999</li> </ol>
<b>Обсяг курсу</b>	Загальний обсяг: 90 години. Аудиторних занять: 32 год., з них 16 год. лекцій, 16 год. практичних. Самостійної роботи: 58 год.
<b>Очікувані результати навчання</b>	Після завершення цього курсу студент буде : Знати: <ul style="list-style-type: none"> <li>- Основні задачі опрацювання природної мови;</li> <li>- Ймовірнісні лінгвістичні моделі;</li> <li>- Головні підходи до побудови відповідних моделей;</li> <li>- Поняття лематизації та токенізації корпусу тексту;</li> <li>- Поняття уніграм, біграм та методи роботи з ними.</li> </ul> Вміти:

	<ul style="list-style-type: none"> <li>- Будувати алфавітно частотні словники згідно досліджуваного корпусу тексту та зберігати їх для подальшого опрацювання;</li> <li>- Нормалізувати алфавітно-частотні словники;</li> <li>- Реалізовувати ймовірнісні лінгвістичні моделі.</li> </ul>
<b>Ключові слова</b>	Комп'ютерна лінгвістика, лематизація, токенизація, NLP
<b>Формат курсу</b>	Очний Проведення лекцій і консультацій, лабораторних робіт.
<b>Теми</b>	<ol style="list-style-type: none"> <li>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.</li> <li>2. Типи NLP застосувань. Системи перетворень мова-текст, текст-мова. Відповіді на запитання. Анотування документів. Чат-боти. Застосунки для вивчення мови. Задача завершення оповіді.</li> <li>3. Які завдання вирішує комп'ютерний лінгвіст. Задача парсування текстів. Поняття токена та токенизації. Основні помилки для вирішення при корекції помилок. Завдання для застосунків виправлення помилок. Підходи до задачі виправлення помилок.</li> <li>4. Робота з патернами при виправленні помилок. Статистика при виправленні помилок. Поняття N-грам.</li> <li>5. Машинне навчання при виправленні помилок. Round-Trip.</li> <li>6. Регулярні вирази. Пошук. Пошук по діапазонах. Пошук по запереченнях. Пошук з використанням операторів ? * + . Пошук з використанням операторів ^ \$. Типи помилок.</li> <li>7. Задача нормалізації тексту. Задача токенизації тексту. Поняття леми та словоформи.</li> <li>8. Токени та типи. Поняття алфавітно-частотного словника. Навіщо його будувати.</li> <li>9. Токенизація. Особливості деяких мов. Токенизація китайської мови. Підходи</li> <li>10. Алгоритм максимуму співпадінь. Нормалізація тексту. Навіщо. Власні назви та аббревіатури при токенизації. Що робити. Лематизація. Що це таке. Навіщо. Морфологія. Стеммінг.</li> <li>11. Алгоритм Портера.</li> <li>12. Сегментація речень.</li> <li>13. Дерева прийняття рішень в задачі сегментації речень.</li> <li>14. Реалізація дерев прийняття рішень.</li> <li>15. Інші підходи окрім дерев прийняття рішень для класифікаторів.</li> <li>16. Ймовірнісні лінгвістичні моделі. Задачі для вирішення.</li> <li>17. Ймовірнісні лінгвістичні моделі. Мета та завдання.</li> <li>18. Правило Чена.</li> <li>19. Правило Чена для послідовності слів у реченні.</li> <li>20. Припущення Маркова.</li> <li>21. Припущення Маркова. Уніграми.</li> <li>22. Припущення Маркова. Біграми.</li> <li>23. Оцінка ймовірностей біграм.</li> <li>24. Перехід від ймовірностей до логарифмів. Чому.</li> <li>25. Оцінка моделі мови.</li> <li>26. Поняття навчальної та тестової множин.</li> <li>27. Поняття perplexity.</li> <li>28. Метод Шенона.</li> <li>29. Встановлення авторства на триграмах. Що робити з нульовими ймовірностями.</li> <li>30. Оцінка add-1.</li> </ol>

	<p>31. Оцінка максимальної подібності.</p> <p>32. Лінійна інтерполяція.</p> <p>33. Проблема визначення коефіцієнтів при інтерполяції. Що робити.</p> <p>34. Згладжування N-грам.</p> <p>35. Згладжування Кнесера-Нея.</p> <p>36. Задача класифікації текстів. Постановка задачі.</p> <p>37. Задача класифікації текстів. Правила вручну. Переваги та недоліки.</p> <p>38. Припущення Наїва-Баєса. Класифікатори Наїва-Баєса.</p>
<b>Підсумковий контроль, форма</b>	Залік.
<b>Пререквізити</b>	<p>Для вивчення курсу студенти потребують базових знань з</p> <ul style="list-style-type: none"> <li>- Математики;</li> <li>- Логіки;</li> <li>- Математичної статистики;</li> <li>- Теорії ймовірності.</li> </ul>
<b>Навчальні методи та техніки, які будуть використовуватися під час викладання курсу</b>	<p>Презентації, лекції</p> <p>Індивідуальні завдання</p> <p>Групові проекти, менторство</p>
<b>Необхідне обладнання</b>	Комп'ютер із програмним забезпеченням Visual Studio 2017/2019, Internet доступ.
<b>Критерії оцінювання (окремо для кожного виду навчальної діяльності)</b>	<p>Оцінювання проводиться за 100-бальною шкалою. Бали нараховуються за наступним співвідношенням:</p> <ul style="list-style-type: none"> <li>• індивідуальні завдання : 50% семестрової оцінки; максимальна кількість балів 50</li> <li>• підсумковий тест: 50% семестрової оцінки; максимальна кількість балів 50</li> </ul> <p>Підсумкова максимальна кількість балів 100.</p> <p><b>Письмові роботи:</b> Очікується, що студенти виконають одну письмову роботу (тест з теоретичних завдань) і звіт про виконання індивідуальних та командних завдань.</p> <p><b>Академічна доброчесність:</b> 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 незарахування викладачем, незалежно від масштабів плагіату чи обману.</p> <p><b>Відвідання занять</b> 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.</p> <p><b>Література.</b> 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.</p> <p><b>Політика виставлення балів.</b> Враховуються бали набрані за індивідуальні завдання та підсумковий тест. При цьому обов'язково</p>

	<p>враховуються присутність на заняттях та активність студента під час практичного заняття; недопустимість пропусків та запізнь 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.</p> <p>Жодні форми порушення академічної доброчесності не толеруються.</p>
<b>Питання до заліку чи екзамену.</b>	<p>Машинне навчання при виправленні помилок. Round-Trip.</p> <p>Ймовірнісні лінгвістичні моделі. Задачі для вирішення.</p> <p>Типи NLP застосувань.</p> <p>Припущення Маркова. Біграми.</p> <p>Оцінка ймовірностей біграм.</p> <p>Перехід від ймовірностей до логарифмів. Чому.</p> <p>Оцінка моделі мови.</p> <p>Поняття навчальної та тестової множин.</p> <p>Встановлення авторства на триграмах. Що робити з нульовими ймовірностями.</p>
<b>Опитування</b>	<p>Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенню курсу.</p>