МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Львівський національний університет імені Івана Франка
Механіко-математичний факультет
Кафедра математичної статистики і диференціальних рівнянь

Затверджено
На засіданні
кафедри математичної статистики і диференціальних рівнянь
механіко-математичного факультету
Львівського національного університету імені Івана Франка
(протокол № 1 від 22.06.2023 р.)

Завідувач кафедри: Олег БУГРІЙ

Силабус з навчальної дисципліни
“ Обробка природної мови ”,
що викладається в межах ОПП “ Статистичний аналіз даних ”
першого (бакалаврського) рівня вищої освіти для здобувачів з спеціальності 112 - Статистика

Львів 2023 р.

Назва дисципліни	Обробка природної мови
Адреса викладання дисципліни	Головний корпус ЛНУ ім. І. Франка м. Львів, вул. Університетська 1, 79000
Факультет та кафедра, за якою закріплена дисципліна	Механіко-математичний факультет Кафедра математичної статистики і диференціальних рівнянь
Галузь знань, шифр та назва спеціальності	11 - Математика та статистика 112 - Статистика
Викладачі дисципліни	Бугрій О.М., доктор фізико-математичних наук, професор, завідувач кафедри математичної статистики і диференціальних рівнянь, Власов В.А., кандидат фізико-математичних наук, асистент кафедри математичної статистики і диференціальних рівнянь
Контактна інформація викладачів	oleh.buhrii@lnu.edu.ua , http://new.mmf.lnu.edu.ua/employee/buhrii_o_m ; vitaly.vlasov@lnu.edu.ua , https://new.mmf.lnu.edu.ua/employee/vlasov_v_a
Консультації з питань навчання по дисципліні відбуваються	Консультації в день проведення лекцій/практичних занять (за попередньою домовленістю). Головний корпус ЛНУ ім. І. Франка, каб. 267. м. Львів, вул. Університетська, 1
Сторінка курсу	https://new.mmf.lnu.edu.ua/course/opmovy-112-bak23
Інформація про дисципліну	Дисципліна “Обробка природної мови” є нормативною дисципліною зі спеціальності 112 – Статистика для освітньої програми “Статистичний аналіз даних”, яка викладається в 6-му семестрі в обсязі 3-х кредитів (за Європейською Кредитно-Трансферною Системою ECTS).
Коротка анотація дисципліни	Курс розроблено для ознайомлення студентів з підходами до обробки природної мови за допомогою машинного навчання
Мета та цілі дисципліни	*Мета:* ознайомити з основними поняттями та методами обробки природної мови за допомогою машинного навчання *Цілі:* викласти основні положення обробки природної мови, показати підходи до дизайну застосувань обробки природної мови для аналізу емоційного забарвлення текстів, побудови чатботів, перекладу текстів з однієї мови на іншу.
Література для вивчення дисципліни	1) Dan Jurafsky and James H. Martin. *Speech and Language Processing. 3rd ed. draft.* Published online, 2023. https://web.stanford.edu/~jurafsky/slp3/ 2) Jacob Eisenstein. *Introduction to Natural Language Processing*. MIT Press, 2019. 3) Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009. 4) Chris Manning and Hinrich Schütze. *Foundations of Statistical Natural*

	Language Processing. MIT Press, 1999. https://nlp.stanford.edu/fsnlp/ 5) Hobson Lane, Cole Howard, Hannes Hapke. *Natural Language Processing in Action*. Manning, 2019. 6) Chris Manning, Anna Goldie. *Natural Language with Deep Learning: Stanford Artificial Intelligence Professional Program.* Stanford, 2021. https://web.stanford.edu/class/cs224n/ 7) Younes Bensouda Mourri, Lukasz Kaiser, Eddy Shyu. *Natural Language Processing Specialization*. deeplearning.ai, 2022. https://www.deeplearning.ai/program/natural-language-processing-specialization
Обсяг курсу	Загальний обсяг: 90 годин. Аудиторних занять: 48 год., з них 16 год. лекційних та 32 години лабораторних занять. Самостійної роботи: 42 год.
Очікувані результати навчання	У результаті вивчення даного курсу студент буде: **знати**: формулювання основних задач обробки природної мови, методи зведення тексту до числового вектора, підходи до автокорекції та автозаповнення незакінчених речень, підходи до перекладу текстів з однієї мови на іншу. **вміти**: використовувати алгоритми машинного навчання до аналізу емоційного забарвлення текстів, застосувати динамічне програмування, ланцюги Маркова для автокорекції слів, визначення типу слів. Застосовувати рекурентні нейронні мережі, сіамські мережі для генерації тексту, визначення дублікатів серед текстів. Використовувати енкодер-декодер архітектури для побудови чатботів, підсумовування тексту. В результаті засвоєння матеріалу даного курсу студент набуде таких **загальних (ЗК) і спеціальних (фахових) (СК) компетентностей**: ЗК-2. Здатність застосовувати знання у практичних ситуаціях. ЗК-3. Знання й розуміння предметної області та професійної діяльності. ЗК-5. Здатність спілкуватися іноземною мовою. ЗК-6. Навички використання інформаційних і комунікаційних технологій. ЗК-7. Здатність вчитися і оволодівати сучасними знаннями. ЗК-8. Здатність до пошуку, обробки та аналізу інформації з різних джерел. ЗК-10. Здатність працювати в команді. ЗК-11. Здатність до професійного спілкування з представниками інших професійних груп різного рівня (з експертами в інших галузях знань. СК-2. Здатність застосовувати у професійній діяльності знання та навички в галузях теорії ймовірностей, математичної статистики, теорії випадкових процесів. СК-7. Здатність робити якісні висновки з кількісних даних. СК-8. Уміння працювати з інформаційними базами даних. СК-9. Здатність розробляти експериментальні та спостережувальні дослідження та аналізувати дані цих досліджень. СК-10. Здатність проводити дослідження ймовірнісно-статистичних моделей та інтерпретувати одержані результати. СК-11. Здатність використання обчислювальної техніки, спеціалізованих мов програмування та програмних засобів для розв'язання задач і здобуття додаткової інформації. СК-12. Здатність застосовувати ймовірнісно-статистичні методи в міждисциплінарному контексті. СК-13. Здатність подавати статистичні процедури та результати їхнього

	застосування у формі, придатній для цільової аудиторії, до якої звертаються, як усно, так і письмово. СК-15. Здатність аналізувати основи і властивості базових економічних та фінансових структур, інтерпретувати показники фінансової діяльності, користуватися методами оптимального керування економічних та природних процесів. СК-16. Здатність застосовувати у професійній діяльності знання та навички з машинного навчання, обробки зображень і природної мови. СК-17. Здатність моделювати та пояснювати дані просторових і часових вибірок за допомогою знань і навичок з регресійного аналізу. і здобуде такі **результати навчання (РН)**: РН-2. Вміти працювати зі спеціальною літературою іноземною мовою. РН-12. Вміти збирати та обробляти дані, застосовувати статистичні процедури для аналізу даних за допомогою обчислювальної техніки та програмних засобів. РН-14. Володіти сучасними інформаційними технологіями для створення презентацій, роботи з базами даних, пошуку інформації та обміну нею. РН-16. Вміти використовувати в практичній діяльності спеціалізоване статистичне програмне забезпечення. РН-21. Вміти застосовувати у професійній діяльності знання та навички з машинного навчання, обробки зображень і природної мови, інших галузей науки про дані.
Ключові слова	Обробка природної мови, аналіз емоційного забарвлення текстів, автокорекція слів, автозаповення, генерація тексту, чатбот, розпізнавання дублікатів, підсумовування тексту, системи питань-відповідей.
Формат курсу	Очний
Теми	Див. **Схема курсу**
Підсумковий контроль, форма	Залік
Пререквізити	Для вивчення даного курсу студенту потрібні базові знання з: - Програмування; - Машинного навчання.
Навчальні методи та техніки, які будуть використовуватися під час викладання курсу	Інформаційні методи (лекція, бесіда, ілюстрація, демонстрація); дедуктивні методи на основі узагальнень; евристичні методи (проблемна лекція); інтерактивні методи (дискусія)
Необхідне обладнання	Для проведення лекційних занять: комп'ютер (мінімальні характеристики: процесор Intel Core i3, 4ГБ оперативної пам'яті), доступ до мережі Internet, засоби мультимедіа (в т.ч. проектор). Для проведення практичних/лабораторних занять: комп'ютер (мінімальні характеристики: процесор Intel Core i3, 4ГБ оперативної пам'яті), доступ до мережі Internet. Необхідне програмне забезпечення включає в себе ОС Windows 10, програмні додатки (MS Teams, MS Excel, Jupyter Notebook з вбудованим компілятором мови програмування Python).
Критерії оці-	Оцінювання проводиться за 100-бальною шкалою. Бали нараховуються за

нювання (окремо для кожного виду навчальної діяльності)	наступним співвідношенням: • Змістовий модуль 1: 8% семестрової оцінки за активну роботу на заняттях, 32% семестрової оцінки за виконання практичних аудиторних і домашніх завдань, максимальна кількість балів 40. • Змістовий модуль 2: 8% семестрової оцінки за активну роботу на заняттях, 32% семестрової оцінки за виконання практичних аудиторних і домашніх завдань, максимальна кількість балів 40. • контрольне тестування: 20% семестрової оцінки, максимальна кількість балів 20. Підсумкова максимальна кількість балів 100. **Академічна доброчесність**: Очікується, що роботи студентів будуть оригінальними дослідженнями чи міркуваннями. Списування та втручання в роботу інших студентів становлять, але не обмежують, приклади можливої академічної недоброчесності. Виявлення ознак академічної недоброчесності в написанні завдань є підставою для її незарахуванння викладачем, незалежно від масштабів плагіату чи обману. Жодні форми порушення академічної доброчесності не толеруються. **Відвідання занять** є важливою складовою навчання. Очікується, що всі студенти відвідають усі лекції та практичні/лабораторні зайняття курсу. Студенти повинні інформувати викладача про неможливість відвідати заняття. У будь-якому випадку студенти зобов'язані дотримуватися термінів визначених для виконання всіх 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 заохочуються до використання також й іншої літератури та джерел, яких немає серед рекомендованих. **Політика виставлення балів.** Враховуються бали, набрані при поточному контролі та бали підсумкового тестування. 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. **Оцінювання практичних робіт** (2 змістових модулі містять по 8 практичних робіт кожен, загалом 16 практичних робіт, максимальна кількість балів: 64) відбувається шляхом оцінки роботи студента під час проведення практичної роботи в аудиторії (0-2 балів за одну роботу) та захисту написаної студентом вдома практичної роботи (0-2 балів за одну роботу). До 1 бала студенти можуть отримати за активну роботу на заняттях. Бали оцінювання аудиторного виконання практичних робіт нараховуються за наступним співвідношенням: 2 – студент в повному обсязі володіє навчальним матеріалом, має повне розуміння розглянутої теми, надає правильні відповіді на запитання по темі, код програми функціонує відповідно до завдання; 1.5 – студент достатньо розуміє розглянутий матеріал та принципи написаного ним коду програми, присутні неточності та незначні помилки у відповідях на запитання по темі, код програми функціонує відповідно до завдання; 1 – студент не досить добре розуміє розглянутий матеріал та написаний

	ним код програми, вагається та надає неточні/не конкретні відповіді на запитання по темі, код програми функціонує з помірними недоліками; 0.5 – студент погано розуміє розглянутий матеріал та написаний ним код програми, студент в більшості надає помилкові відповіді на питання по темі, код програми не функціонує належним чином; 0 - студент зовсім не засвоїв розглянутий матеріал, написаний ним код програми не відповідає темі/не функціонує взагалі. Бали оцінювання домашнього завершення виконання практичних робіт та наданого звіту нараховуються за наступним співвідношенням: 2 – звіт цілком і повністю відображає індивідуальне завдання студента, містить правильні висновки, ілюстрований (за потреби) відповідними графіками і таблицями які правильно відображають суть виконаного завдання, студент має повне розуміння розглянутої теми, надає правильні відповіді на запитання по темі, код програми функціонує відповідно до завдання; 1.5 – звіт в достатній мірі відображає індивідуальне завдання студента, містить допустимі висновки, ілюстрований (за потреби) відповідними графіками і таблицями які частково відображають суть виконаного завдання, студент достатньо розуміє принципи написаного ним коду програми, присутні неточності та незначні помилки у відповідях на запитання по темі, код програми функціонує відповідно до завдання; 1 – звіт містить загальні формулювання завдання, висновки нечіткі, необхідні ілюстрації чи таблиці відсутні, студент не досить добре розуміє розглянутий матеріал та представлений код програми, надає неточні/не конкретні відповіді на запитання по темі, код програми функціонує з помірними недоліками; 0 – звіт не містить формулювання завдання, висновки необґрунтовані чи неповні, необхідні ілюстрації чи таблиці відсутні, студент погано розуміє розглянутий матеріал та представлений код програми, студент в більшості надає помилкові відповіді на питання по темі, код програми не функціонує належним чином; 0 – звіт відсутній/не відповідає темі, студент зовсім не засвоїв розглянутий матеріал, написаний ним код програми не відповідає темі/не функціонує взагалі. **Оцінювання контрольного тестування** (завдання з тематики кожного змістового модуля) відбувається шляхом оцінки письмових відповідей студента на поставлені запитання. Відсотки нарахування балів оцінювання відповіді на кожне запитання нараховуються за наступним співвідношенням: 75-100% – тема відтворюється в повному обсязі, правильно, обґрунтовано, логічно; 50-75% – відтворюється значна частина розглянутої теми, проте присутні неточності та/або невідповідності; 25-50% – виявлено множинні неточності та невідповідності, пояснення відсутні чи частково помилкові; 0-25% – тему майже не розкрито, кількість викладеного матеріалу не відповідає загальним нормам обраного виду роботи. **Критерії оцінювання результатів неформальної освіти:** Нарахування балів відбувається за публікацію студентом тез доповідей на конференціях, наукових статей, за участь студента у діяльності наукових

	гуртків, семінарів, круглих столів, конкурсів, участь у заходах неформальної освіти, за отримання сертифікатів про проходження навчання на різних освітніх платформах (Coursera, Prometheus тощо), курсах провідних IT компаній за тематикою навчальної дисципліни. Кількість балів визначається відсотком покриття результатів відповідної активності до вимог результатів навчання з навчальної дисципліни
Питання для контрольного тестування	методи зведення тексту до числового вектора, підходи до автокорекції та автозаповнення незакінчених речень, підходи до перекладу текстів з однієї мови на іншу.
Опитування	Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенню курсу.

Схема курсу "Обробка природної мови"
для студентів спеціальності 112 - Статистика

Тижні	Лекційний курс		Лабораторні заняття		К-сть год СР	Літе-ратура
	Назва теми	К-сть год	Назва теми	К-сть год		
1	2	3	4	5	6	7
1	Аналіз емоційного забарвлення тексту.	2	Зведення тексту до числового вектора. Створення бінарного класифікатора для твітів за допомогою логістичної регресії.	2	2	[1], [2], [3], [7], Сайт курсу
2			Створення класифікатора твітів за допомогою Naive Bayes.	2	3	[1], [2], [3], [7], Сайт курсу
3	Векторні представлення. Пошук в документах.	2	Створення векторного представлення тексту. Візуалізація залежностей між словами у 2-вимірному просторі, використовуючи PCA.	2	2	[1], [2], [4], [5], [7], Сайт курсу
4			Locality-sensitive hashing для організації ефективного пошуку.	2	3	[1], [2], [4], [5], [7], Сайт курсу
5	Автокорекція. Розпізнавання частин мови.	2	Автокорекція. Побудова алгоритму із використанням динамічного програмування.	2	2	[1]-[5], [7], Сайт курсу
6			Створення тегів з частинами мови для корпусу даних.	2	3	[1]-[5], [7], Сайт курсу
7	Автозаповнення. Векторне представлення. Рекурентні нейронні мережі.	2	Побудова моделі автозаповнення на корпусі твітів. Тренування нейронної мережі з представленнями слів GloVe для емоційного аналізу твітів.	2	2	[1], [2], [5], [6], Сайт курсу
8			Побудова генератора мови, використовуючи рекурентні нейронні мережі.	2	3	[1], [2], [5], [6], Сайт курсу
9	Розпізнавання названих сутностей. Сіамські мережі.	2	Побудова системи розпізнавання названих сутностей, використовуючи LSTM.	2	2	[1], [2], [3], [5]-[7], Сайт курсу
10			Тренування сіамської мережі для ідентифікації однакових	2	3	[1], [2], [3], [5]-[7],

			питань у корпусі.			Сайт курсу
11	Машинний переклад. Трансформери.	2	Тренування нейронної мережі для моделі з увагою для перекладу з англійської на німецьку мову.	2	2	[1], [2], [5]-[7], Сайт курсу
12			Порівняння рекурентних нейронних мереж з архітектурою трансформерів для генерації підсумку текстів.	2	3	[1], [2], [5]-[7], Сайт курсу
13	Мовні моделі для відповідей на питання.	2	Тренування T5, BERT моделей, що здатні знаходити відповіді на запитання.	2	4	[1], [2], [6], [7], Сайт курсу
14			Моделі трансформерів	2	2	[1], [2], [6], [7], Сайт курсу
15	Чатботи. Підсумкове заняття	2	Побудова чатботів, використовуючи Reformer модель.		4	[1]-[7], Сайт курсу
16			Перевірка рівня знань і умінь студентів.	2	2	[1]-[7], Сайт курсу
	Разом	**16**		**32**	**42**	
	Викладач: **Бугрій О.М.**		Викладач: **Власов В.А.**			