МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЗАХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ФАКУЛЬТЕТ КОМП'ЮТЕРНИХ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

ЗАТВЕРДЖУЮ:
В.о. декана факультету
комп'ютерних інформаційних
технологій
Ігор ЯКИМЕНКО
_____
"____"_____2023 р.

ЗАТВЕРДЖУЮ:
В.о. проректора з
науково-педагогічної роботи
Віктор ОСТРОВЕРХОВ
_____
"____"_____2023 р.

ЗАТВЕРДЖУЮ:
Директор навчально-наукового
інституту новітніх освітніх
технологій
Святослав ГІТЕЛЬ
"____"_____2023 р.

РОБОЧА ПРОГРАМА

з дисципліни «**Інтелектуальна обробка тексту та природньої мови**»
ступінь вищої освіти – **магістр**
галузь знань – **12 "Інформаційні технології"**
спеціальність – **122 „Комп'ютерні науки"**
освітньо-професійна програма – **„Комп'ютерні науки"**

Кафедра інформаційно-обчислювальних систем і управління

Форма навчання	Курс	Семестр	Лекції (год.)	Практичні заняття (год.)	ІРС (год.)	Тренінг, КПІЗ (год.)	Самост. робота студ. (год.)	Разом (год.)	Залік. (сем.)
Денна	1	2	30	15	5	4	96	150	2
Заочна	1	2	8	4	–	–	138	150	2

29.09.2023

Тернопіль – ЗУНУ
2023

Робочу програму склала доцент кафедри ІОСУ, к.т.н. Христина ЛІП'ЯНІНА-ГОНЧАРЕНКО

Робоча програма затверджена на засіданні кафедри інформаційно-обчислювальних систем і управління, протокол № 2 від 29 вересня 2023 р.

Завідувач кафедри Мирослав КОМАР

Розглянуто та схвалено групою забезпечення спеціальності „Комп'ютерні науки", протокол № 2 від 29 вересня 2023 р.

Голова групи
забезпечення спеціальності,
д-р техн. наук, професор Мирослав КОМАР

Гарант освітньо-професійної
програми "Комп'ютерні науки",
к.т.н, доцент Діана ЗАГОРОДНЯ

СТРУКТУРА РОБОЧОЇ ПРОГРАМИ НАВЧАЛЬНОЇ ДИСЦИПЛІНИ " ІНТЕЛЕКТУАЛЬНА ОБРОБКА ТЕКСТУ ТА ПРИРОДНЬОЇ МОВИ "

1. Опис дисципліни "Інтелектуальна обробка тексту та природньої мови"

Дисципліна «Інтелектуальна обробка тексту та природньої мови	**Галузь знань, спеціальність, СВО**	**Характеристика навчальної дисципліни**
Кількість кредитів – 5	Галузь знань – 12 “Інформаційні технології”	**Статус дисципліни:** вибіркова **Мова навчання:** Українська
Кількість залікових модулів – 3	Спеціальність – 122 «Комп’ютерні науки»	Рік підготовки: 1 Семестр: *Денна* – 2 *Заочна* – 2
Кількість змістових модулів – 2	Освітньо-професійна програма „Комп’ютерні науки”	Лекції: *Денна* –30 год. *Заочна* – 8 год. Практичні заняття: *Денна* – 15 год. *Заочна* – 4 год.
Загальна кількість годин – 150	Ступінь вищої освіти – магістр	Самостійна робота: *ДФН – 100 год., у т.ч. тренінг – 4 год.; ЗФН – 138 год.* Індивідуальна робота: *5 год*. (*ДФН*)
Тижневих годин – 10, з них аудиторних – 3 год.		Вид підсумкового контролю – залік

2. Мета і завдання дисципліни "Інтелектуальна обробка тексту та природньої мови"

2.1. Мета вивчення дисципліни

Мета вивчення дисципліни "Інтелектуальна обробка тексту та природної мови" - це надання знань і навичок для проектування, розробки та застосування алгоритмів і методів підсумкового аналізу природної мови і призначених для автоматизації процесу обробки тексту. Ці підходи використовуються для створення штучних інтелектуальних систем для аналізу та відображення значень з природньої мови тексту.

2.2. Завдання вивчення дисципліни

Дисципліна "Інтелектуальна обробка тексту та природної мови" є ділянкою програмної інженерії, яка заснована на комп'ютерній лінгвістиці. Це дисципліна створена для того, щоб розробити програми для обробки природної мови та текстів, що використовуються в пошуку та розумінні цих мов. Це може включати процеси, такі як лексикографічний аналіз, тезаурусний аналіз та правопис. Дисципліна також включає в себе процеси аналізу тексту, такі як процеси класифікації тексту, аналіз відносин та розпізнавання образів.

2.3. Результати навчання

В результаті вивчення навчальної дисципліни студент повинен:

знати:

- основні поняття та визначення інтелектуальної обробки тексту та природної мови;
- моделі та методи побудови моделей та аналізу залежностей у текстових даних;
- сучасні програмні засоби для проектування і розробки систем інтелектуальної обробки тексту та природної мови;
- критерії порівняння моделей і методів інтелектуальної обробки тексту та природної мови.

вміти:

- обґрунтовувати й аналізувати вибір конкретного типу моделі та методу інтелектуальної обробки тексту та природної мови при вирішенні практичних задач;
- використовувати сучасні програмні засоби для проектування та дослідження систем інтелектуальної обробки тексту та природної мови;
- аналізувати результати побудови та використання систем інтелектуальної обробки тексту та природної мови при вирішенні прикладних задач.

3. Програма навчальної дисципліни «Інтелектуальна обробка тексту та природньої мови»

Змістовий модуль 1 – Теоретичні основи

Тема 1. Основи обробки природної мови

Визначення природної мови. Поняття лінгвістики. Синтаксис і структура мови. Семантика мови. Обробка природної мови. Аналітика тексту. Машинне навчання. Глибоке навчання.

Тема 2. Python для обробки природної мови.

Реалізації та версії Python. Налаштування надійного середовища Python. Синтаксис і структура Python. Робота з текстовими даними. Основи обробки та аналізу тексту. Фреймворки обробки природної мови.

Тема 3. Опрацювання та розуміння тексту.

Токенізація тексту. Видалення символів із наголосами. Розширення скорочень. Видалення спеціальних символів. Перетворення регістру. Виправлення тексту. Лематизація. Видалення стоп-слів. Об'єднання всього разом — створення нормалізатора тексту. Розуміння синтаксису та структури тексту. Встановлення необхідних залежностей. Визначення частин мови. Неглибокий аналіз або фрагментація. Аналіз залежностей.

Тема 4. Розробка функцій для подання тексту

Розуміння текстових даних. Побудова текстового корпусу. Попередня обробка текстового корпусу. Інженерні моделі традиційних функцій обробки тексту. Розширені інженерні моделі обробки тексту.

Змістовий модуль 2 – Методи обробки природної мови та аналітики тексту

Тема 5. Класифікація тексту

Що таке класифікація тексту? Автоматизована класифікація тексту. Проект класифікації тексту. Отримання даних. Попередня обробка та нормалізація даних. Створення навчальних і тестових наборів даних. Характеристика техніки інженерії. Моделі класифікації. Оцінка моделей класифікації.

Тема 6. Конспектування тексту та тематичні моделі.

Резюмування тексту та вилучення інформації. Вилучення ключової фрази. Моделювання теми. Моделювання теми на наукових роботах. Тематичні моделі з Gensim. Тематичні моделі за допомогою Scikit-Learn. Автоматизоване підсумовування документів.

Тема 7. Подібність тексту та кластеризація.

Основні поняття. Подібність тексту. Аналіз подібності термінів. Аналіз схожості документів. Створення рекомендатора фільмів. Кластеризація документів. Кластеризація фільмів.

Тема 8. Семантичний аналіз

Семантичний аналіз. Вивчення WordNet. Розпізнавання сенсу слова. Розпізнавання іменованих сутностей. Створення тигера NER з нуля. Створення наскрізного тигера NER. Аналіз семантичних репрезентацій.

Тема 9. Аналіз настроїв

Визначення залежностей в тексті. Попередня обробка та нормалізація тексту. Неконтрольовані моделі на основі лексиконів. Класифікація настрою з

контрольованим навчанням. Традиційні керовані моделі машинного навчання. Новіші моделі керованого глибокого навчання. Розширені керовані моделі глибокого навчання. Аналіз причинно-наслідкового настрою.

Тема 10. Перспективи глибокого навчання

Вставні речення. Тенденції в моделях вставних слів. Тенденції в універсальних моделях вставних речень. Універсальні вставні речення та їх виявлення методами глибокого навчання. Трансферне навчання з різними універсальними вставними реченнями. Майбутні сфери застосування.

4. Структура залікового кредиту з дисципліни «Інтелектуальна обробка тексту та природньої мови»

Денна форма навчання

Тема	***Кількість годин***					
	Лекції	Практичні заняття	Індивідуальна робота	Тренінг	Самостійна робота	Контрольні заходи
Змістовий модуль 1 – Теоретичні основи						
Тема 1. Основи обробки природної мови	2	1	-	2	6	Опитування під час заняття
Тема 2. Python для обробки природної мови	2	2	1		10	Опитування під час заняття
Тема 3. Опрацювання та розуміння тексту	2	2	-		10	Опитування під час заняття
Тема 4. Розробка функцій для подання тексту	2	2	1		10	Опитування під час заняття
Змістовий модуль 2 – Методи обробки природної мови та аналітики тексту						
Тема 5. Класифікація тексту	2	1	1	2	10	Опитування під час заняття
Тема 6. Конспектування тексту та тематичні моделі	4	1	-		10	Опитування під час заняття
Тема 7. Подібність тексту та кластеризація	4	2	1		10	Опитування під час заняття
Тема 8. Семантичний аналіз	4	2	1		10	Опитування під час заняття
Тема 9. Аналіз настроїв	4	1			10	Опитування під час заняття
Тема 10. Перспективи глибокого навчання	4	1			10	Опитування під час заняття
Разом	30	15	5	4	96	

Заочна форма навчання

	Кількість годин		
	Лекції	**Практичні заняття**	**Самостійна робота**
Змістовий модуль 1 – Теоретичні основи			
Тема 1. Основи обробки природної мови	4	2	12
Тема 2. Python для обробки природної мови			14
Тема 3. Опрацювання та розуміння тексту			14
Тема 4. Розробка функцій для подання тексту			14
Змістовий модуль 2 – Методи обробки природної мови та аналітики тексту			
Тема 5. Класифікація тексту	4	2	14
Тема 6. Конспектування тексту та тематичні моделі			14
Тема 7. Подібність тексту та кластеризація			14
Тема 8. Семантичний аналіз			14
Тема 9. Аналіз настроїв			14
Тема 10. Перспективи глибокого навчання			14
Разом	**8**	**4**	**138**

5. Тематика практичних занять

Практичне заняття №1
Тема: Охайний текстовий формат
Питання для обговорення:
1.1 Зіставлення акуратного тексту з іншими структурами даних
1.2 Функція unnest_tokens
1.3 Прибирання творів Джейн Остін
1.4 Пакет gutenbergr
1.5 Частоти слів

Практичне заняття №2
Тема: Аналіз настроїв
Питання для обговорення:
2.1 Набори даних настроїв
2.2 Аналіз настроїв із внутрішнім з'єднанням
2.3 Порівняння трьох словників почуттів
2.4 Найпоширеніші позитивні та негативні слова
2.5 Хмари слів
2.6 Розгляд одиниць не тільки слів.

Практичне заняття №3
Тема: Аналіз частоти слів і документів: tf-idf

Питання для обговорення:
3.1 Частота термінів у романах Джейн Остін
3.2 Закон Ціпфа
3.3 Функція bind_tf_idf().
3.4 Корпус текстів з фізики

Практичне заняття №4
Тема: Зв'язки між словами: n-грами та кореляції
Питання для обговорення:
4.1 Токенізація за допомогою n-грам
4.2 Підрахунок і співвіднесення пар слів із пакетом widyr

Практичне заняття №5
Тема: Моделювання теми.
Питання для обговорення:
5.1 Латентне виділення Діріхле
5.2 Слово-тематичні ймовірності
5.3 Імовірності теми документа

6. Комплексне практичне індивідуальне завдання

Індивідуальні завдання з дисципліни «Інтелектуальна обробка тексту та природньої мови» виконується самостійно кожним студентом. КПІЗ є науковим дослідження за варіантами, прикладну область обирає студент самостійно. Метою виконання КПІЗ є оволодіння навичками застосування методів машинного навчання при розв’язуванні прикладних проблем. КПІЗ оформлюється згідно з встановленими вимогами.

Варіанти КПІЗ з дисципліни «Інтелектуальна обробка тексту та природньої мови»

№ варіанту	Тема дослідження
1.	Коронавірусні твіти НЛП - Текстова класифікація
2.	Емоції для НЛП
3.	Класифікація віршів (НЛП)
4.	NLP про супергероїв
5.	Аналіз настроїв у Twitter
6.	Класифікація електронної пошти НЛП
7.	НЛП: Звіти та класифікація новин
8.	НЛП про дослідницькі статті
9.	Стенограми доповідей TED для НЛП
10.	Класифікація спаму для базового НЛП
11.	Набір даних про скарги споживачів для НЛП
12.	Твіти про катастрофи
13.	«Питання-відповідь».
14.	Класифікації жанрів IMDb
15.	Моделювання тем для наукових статей

7. Самостійна робота

№ з/п	Тематика	К-сть годин	
		ДФН	ЗФН
1.	Основи обробки природної мови	6	12
2.	Python для обробки природної мови	10	14
3.	Опрацювання та розуміння тексту	10	14
4.	Розробка функцій для подання тексту	10	14
5.	Класифікація тексту	10	14
6.	Конспектування тексту та тематичні моделі	10	14
7.	Подібність тексту та кластеризація	10	14
8.	Семантичний аналіз	10	14
9.	Аналіз настроїв	10	14
10.	Перспективи глибокого навчання	10	14
	Разом	**96**	**138**

8. Тренінг з дисципліни

Тематика: Розробка проекту на kaggle.

№ з/п	Вид роботи	Порядок проведення тренінгу
1	Вступна частина	ознайомлення студентів з темою тренінгового заняття і видача завдання
2	Практична частина	виконання завдань студентами згідно з індивідуальним завданням; оформлення короткого звіту
3	Підведення підсумків	обговорення результатів виконаних завдань

9. Засоби оцінювання та методи демонстрування результатів навчання

У навчальному процесі використовуються: лекції; практичні заняття; індивідуальні заняття; тренінг; виконання КПІЗ, а також методи опитування, тестування, ділові ігри тощо.

У процесі вивчення дисципліни «Інтелектуальна обробка тексту та природньої мови» використовуються наступні методи оцінювання навчальної роботи студента:

- поточне тестування та опитування;
- залікове модульне тестування та опитування;
- ректорська контрольна робота;
- оцінювання виконання завдань тренінгу;
- оцінювання виконання КПІЗ.

10. Критерії, форми поточного та підсумкового контролю

Підсумковий бал (за 100-бальною шкалою) з дисципліни «Інтелектуальна обробка тексту та природньої мови» визначається як середньозважена величина, залежно від питомої ваги кожної складової залікового кредиту:

Заліковий модуль 1	Заліковий модуль 2	Заліковий модуль 3
30 %	40 %	30 %
1. Виконання та захист практичних робіт (3 роботи по 15 балів) – 45 балів 2. Модульна контрольна робота – 55 балів	1. Виконання та захист практичних робіт (2 роботи по 20 балів) – 40 балів 2. Ректорська контрольна робота – 60 балів	1. Виконання завдань під час тренінгу – 20 балів 2. Написання та захист КПІЗ – 80 балів

Шкала оцінювання:

За шкалою ЗУНУ	**За національною шкалою**	**За шкалою ECTS**
90-100	Відмінно	A (відмінно)
85-89	Добре	B (дуже добре)
75-84		C (добре)
65-74	Задовільно	D (задовільно)
60-64		E (достатньо)
35-59	Незадовільно	FX (незадовільно, з можливістю повторного складання)
1-34		F (незадовільно, з обов'язковим повторним курсом)

11. Інструменти, обладнання та програмне забезпечення, використання яких передбачає навчальна дисципліна

№	**Найменування**	**Номер теми**
1.	Мультимедійне обладнання	1-10
2.	Python	1-10

РЕКОМЕНДОВАНІ ДЖЕРЕЛА ІНФОРМАЦІЇ

1. Patel A. A. & Arasanipalai A. U. (2021). Applied natural language processing in the enterprise teaching machines to read write and understand (First). O'Reilly.
2. BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP & Association for Computational Linguistics. (2019). Second blackboxnlp workshop on analyzing and interpreting neural networks for nlp at acl 2019 : florence italy 1 august 2019. Curran.
3. St Mt A. Wiksuana I. Ramantha I. W. & Wiagustini N. L. P. (2019). Working role of technical analysis of slowstochastic indicator and commodity channel index in affecting investment decision in indonesia stock exchange. SSRN. https://doi.org/10.2139/ssrn.3336466
4. Agerri R. Aliprandi C. Alkhalifa R. Alzetta C. Angel J. Anselmi G. Appiah Balaji N. N. Aroyehun S. T. Artigas Herold M. F. & Attanasio G. (2021). Evalita evaluation of nlp and speech tools for italian - december 17th 2020. Retrieved February 11 2023 from http://books.openedition.org/aaccademia/6732.
5. Heale A. (2021). Generation panic : simple & empowering techniques to combat anxiety. O Books.
6. Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing.

Fundamentals of artificial intelligence, 603-649.

7. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 63(10), 1872-1897.
8. Galassi, A., Lippi, M., & Torroni, P. (2020). Attention in natural language processing. IEEE transactions on neural networks and learning systems, 32(10), 4291-4308.
9. Eisenstein, J. (2019). Introduction to natural language processing. MIT press.
10. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), 1-23.
11. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.