Національний технічний університет України
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

ММСА Кафедра математичних методів системного аналізу

Системний аналіз предметної галузі із використанням текстової аналітики

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	***Перший (бакалаврський)***
Галузь знань	*12 Інформаційні технології*
Спеціальність	*124 Системний аналіз*
Освітня програма	*Системний аналіз і управління*
Статус дисципліни	*Вибіркова*
Форма навчання	*очна(денна)*
Рік підготовки, семестр	*4 курс, 8 семестр (весняний)*
Обсяг дисципліни	*4 кредити ЕКТС*
Семестровий контроль/ контрольні заходи	*Залік*
Розклад занять	*2 год лекційних та 1 год практичних занять на тиждень, https://schedule.kpi.ua/*
Мова викладання	*Українська*
Інформація про керівника курсу / викладачів	Лектор: *к.т.н., ст. викладач, Савастьянов Володимир, savastiyanov.volodymyr@lll.kpi.ua* Лабораторні: *к.т.н., ст. викладач, Савастьянов Володимир,* savastiyanov.volodymyr@lll.kpi.ua
Розміщення курсу	Google classroom: l2rmjcx

Програма навчальної дисципліни

1. Опис навчальної дисципліни, її мета, предмет вивчання та результати навчання

Мета дисципліни: *Метою навчальної дисципліни є формування у студентів навичок системного підходу до аналізу та структуризації предметних галузей із використанням текстової аналітики. Студенти навчаться ідентифікувати системні зв'язки у межах предметної галузі та створювати класифікатори на основі текстових даних. Це дозволить застосовувати сучасні методи текстової аналітики для автоматизації процесів обробки інформації, створення моделей прийняття рішень та пошуку інсайтів у текстових масивах.*

Предмет вивчення: *Предметом вивчення є методи та інструменти системного аналізу для структуризації предметної галузі, а також застосування текстової аналітики для аналізу текстових даних. Дисципліна охоплює принципи системного мислення, класифікацію інформації, побудову інформаційних моделей предметної галузі, та інтеграцію результатів текстової аналітики у ці моделі для отримання практичних висновків.*

Результати навчання: *По завершенню курсу студенти будуть здатні:*

- *Розуміти та застосовувати принципи системного підходу до аналізу предметних галузей.*
- *Структурувати знання та системні зв'язки у межах обраної предметної галузі.*
- *Створювати текстові класифікатори на основі системного аналізу.*
- *Використовувати сучасні методи текстової аналітики для обробки та аналізу текстових даних у різних контекстах.*
- *Інтегрувати результати текстової аналітики у процеси прийняття рішень та автоматизації.*
- *Оцінювати ефективність текстової аналітики у межах конкретної предметної галузі та виявляти нові можливості для покращення системного підходу.*

2. Пререквізити та постреквізити дисципліни (місце в структурно-логічній схемі навчання за відповідною освітньою програмою)

При вивченні дисципліни використовуються знання з основ системного аналізу, алгоритмів та структур даних, систем баз даних, статистичного аналізу, глибокого навчання та нейронних мереж, базового рівня володіння англійською мовою не нижче А2.

На результатах навчання з даної дисципліни базуються вимоги до виконання курсових робіт за дисциплінами: системи і методи підтримки прийняття рішень, обчислювальний інтелект, у практичній самостійній роботі випускника в галузі інтелектуального аналізу даних під час аналізу великих і надвеликих баз даних та масивів тексту, при побудові прогнозів на основі статистичних даних та оцінок експертів, при розробці корпоративних інформаційно-аналітичних систем в державних і приватних управлінських структурах..

3. Зміст навчальної дисципліни

Розділ 1. *Вступ до курсу Системний аналіз предметної галузі із використанням текстової аналітики.*

Тема 1.1. Короткий огляд сучасного стану технологій текстової аналітики.

Тема 1.2. Короткий огляд сучасних технологій Scrapping.

Розділ 2. *Основи текстової аналітики.*

Тема 2.1. Основні поняття у задачах текстової аналітики.

Тема 2.2. Приклади та види задач текстової аналітики.

Розділ 3.*Системний підхід до аналізу предметної галузі.*

Тема 3.1. Класифікація та структуризація предметної області.

Тема 3.2. Класифікація та існуючі класифікатори.

Тема 3.3. Складність аналізу предметної області (на прикладі Сентимент аналізу).

Тема 3.4. Підходи щодо класифікації та огляд класифікації за допомогою правил.

Тема 3.5. Підходи щодо класифікації та огляд класифікації за допомогою правил на прикладі правил SAS.

Тема 3.6. Уточнення класифікатору за допомогою ChatGPT.

***Розділ 4.** Популярні технології текстової аналітики.*

Тема 4.1. Обробка текстів бібліотеками Python, основні проблеми..

Тема 4.2. Моделі сентимент аналізу на основі правил (SAS Sentiment Model) .

Тема 4.3. Використання булевих правил на практиці.

Тема 4.3. Використання булевих правил на практиці (просунуте/advanced).

***Розділ 5.** Представлення результатів обробки даних предметної галузі .*

Тема 5.1. Рівні обробки текстової інформації (Інформація про 4ту лабораторну).

***Розділ 6.** Новітні технології обробки знань.*

Тема 6.1.Використання LLM для обробки знань предметної області.

4. Навчальні матеріали та ресурси

Базова

1. *В.В. Савастьянов. Слайди лекцій з кредитного модуля «Системний аналіз предметної галузі із використанням текстової аналітики», Освітня програма «Системи і методи штучного інтелекту», 2024, https://drive.google.com/drive/folders/1QNzFN_FBP436p3XJSTcDEFOxUwmWhBYJ.*
2. *Chakraborty, G., M. Pagolu, S. Garla. Text Mining and Analysis; Practical Methods, Examples, and Case Studies Using SAS®. SAS Institute Inc. (2013).*
3. *Liu B., Sentiment Analysis and Opinion Mining, ISBN-10: 1608458849, ISBN-13: 978-1608458844, Morgan & Claypool Publishers, (2012).*
4. *Ланде Д.В. Основи інформаційного та соціально-правового моделювання: навч. посіб. / Д.В. Ланде, В.М. Фурашев , К.В. Юдкова. - К.: НТУУ "КПІ", 2014. - 220 с. ISBN 978-966-622-672-6*
5. *Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. arXiv preprint arXiv:2304.01852 (2023).*

Додаткова література

6. *GPT3 & Beyond: Key concepts and open questions in a golden age for natural language understanding. Режим доступу: https://www.youtube.com/watch?v=-lnHHWRCDGk Date accessed: 2024-08-01.*
7. *NLTK documentation Режим доступу: https://www.nltk.org/howto.html Date accessed: 2024-08-01.*
8. *Amazon «What are Large Language Models (LLM)?». Режим доступу: https://aws.amazon.com/what-is/large-language-model/ Date accessed: 2024-08-01.*
9. *Project Gensim Documentation. Режим доступу: https://radimrehurek.com/gensim/auto_examples/index.html#documentation Date accessed: 2024-08-01.*
10. *Project SpaCy Documentation. Режим доступу: https://course.spacy.io/en/ Date accessed: 2024-08-01.*

11. *Курс «Інтелектуальний аналіз тексту та аналітика». Режим доступу: https://www.coursera.org/learn/text-mining Date accessed: 2024-08-01.*

12. *Andrew Ng, Machine Learning Yearning, Deepllearrning.AI. Режим доступу:* https://wordpress.deeplearning.ai/wp-content/uploads/2022/03/andrew-ng-machine-learning-yearning.pdf

Навчальний контент

5. Методика опанування навчальної дисципліни (освітнього компонента)

Структура кредитного модуля

Розділ 1. *Вступ до курсу Системний аналіз предметної галузі із використанням текстової аналітики.*

Тема 1.1. Короткий огляд сучасного стану технологій текстової аналітики.

Тема 1.2. Короткий огляд сучасних технологій Scrapping.

Розділ 2. *Основи текстової аналітики.*

Тема 2.1. Основні поняття у задачах текстової аналітики.

Тема 2.2. Приклади та види задач текстової аналітики.

Розділ 3.*Системний підхід до аналізу предметної галузі.*

Тема 3.1. Класифікація та структуризація предметної області.

Тема 3.2. Класифікація та існуючі класифікатори.

Тема 3.3. Складність аналізу предметної області (на прикладі Сентимент аналізу).

Тема 3.4. Підходи щодо класифікації та огляд класифікації за допомогою правил.

Тема 3.5. Підходи щодо класифікації та огляд класифікації за допомогою правил на прикладі правил SAS.

Тема 3.6. Уточнення класифікатору за допомогою ChatGPT.

Розділ 4. *Популярні технології текстової аналітики.*

Тема 4.1. Обробка текстів бібліотеками Python, основні проблеми.

Тема 4.2. Моделі сентимент аналізу на основі правил (SAS Sentiment Model).

Тема 4.3. Використання булевих правил на практиці.

Тема 4.4. Використання булевих правил на практиці (просунуте/advanced).

Розділ 5. *Представлення результатів обробки даних предметної галузі .*

Тема 5.1. Рівні обробки текстової інформації (Інформація про 4ту лабораторну).

Розділ 6. *Новітні технології обробки знань.*

Тема 6.1.Використання LLM для обробки знань предметної області.

Лекційні заняття

No з/п	*Назва теми лекції та перелік основних питань (перелік дидактичних засобів, посилання на літературу та завдання на СРС)*
1	*Введення у предмет.*
2	*Короткий огляд сучасного стану технологій текстової аналітики.*
3	*Короткий огляд сучасних технологій Scrapping.*
4	*Основні поняття у задачах текстової аналітики.*
5	*Приклади та види задач текстової аналітики.*
6	*Класифікація та структуризація предметної області.*
7	*Класифікація та існуючі класифікатори.*
8	*Складність аналізу предметної області (на прикладі Сентимент аналізу).*
9	*Підходи щодо класифікації та огляд класифікації за допомогою правил.*
10	*Підходи щодо класифікації та огляд класифікації за допомогою правил на прикладі правил SAS.*
11	*Уточнення класифікатору за допомогою ChatGPT.*
12	*Обробка текстів бібліотеками Python, основні проблеми.*
13	*Моделі сентимент аналізу на основі правил (SAS Sentiment Model).*
14	*Використання булевих правил на практиці.*
15	*Використання булевих правил на практиці (просунуте/advanced).*
16	*Рівні обробки текстової інформації (Інформація про 4ту лабораторну).*
17	*Використання LLM для обробки знань предметної області.*
18	*Підсумок за лекційним матеріалом дисципліни.*

Практичні заняття

No з/п	*Назва теми лекції та перелік основних питань (перелік дидактичних засобів, посилання на літературу та завдання на СРС)*
1	*Особливості текстової аналітики. Визначення обсягів знань групи.*
2	*Практика з технологій scrapping, формування власного корпусу предметної галузі*
3	*Практика з технологій scrapping, формування власного корпусу предметної галузі*
4	*Класифікація та структуризація предметної області.*
5	*Класифікація та структуризація предметної області.*
6	*Використання булевих правил на практиці у ПЗ SAS.*
7	*Використання булевих правил на практиці у ПЗ SAS.*

No з/п	Назва теми лекції та перелік основних питань (перелік дидактичних засобів, посилання на літературу та завдання на СРС)
8	*Представлення результатів обробки предметної галузі.*
9	*Представлення результатів обробки предметної галузі.*

6. Самостійна робота здобувача вищої освіти

Самостійна робота студента включає підготовку до практичних / лабораторних робіт, в тому числі опрацювання окремих частин наступних тем:

1. *Ключові концепції розуміння технологій обробки природної мови*
2. *Мережеві технології. Проходження окремих розділів відкритого курсу основ безпеки у мережі від Google. [2,8,21]*
3. *Інструменти сучасних технологій програмування на прикладі Jupyter Notebook [9,10].*
4. *Теорія черг. Брокери черг. [3,12,13]*
5. *Перспективні та новітні технології інформаційних систем.[11,14,15,16,17,19]*

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

- *Пропущені контрольні заходи оцінювання. Кожен студент має право відпрацювати пропущені з поважної причини (лікарняний, мобільність тощо) заняття за рахунок самостійної роботи. Детальніше за посиланням: https://kpi.ua/files/n3277.pdf.*
- *Процедура оскарження результатів контрольних заходів оцінювання. Студент може підняти будь-яке питання, яке стосується процедури контрольних заходів та очікувати, що воно буде розглянуто згідно із наперед визначеними процедурами. Студенти мають право аргументовано оскаржити результати контрольних заходів, пояснивши з яким критерієм не погоджуються відповідно до оціночного.*
- *Календарний контроль проводиться з метою підвищення якості навчання студентів та моніторингу виконання студентом вимог силабусу.*
- *Академічна доброчесність. Політика та принципи академічної доброчесності визначені у розділі 3 Кодексу честі Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського». Детальніше: https://kpi.ua/code.*
- *Норми етичної поведінки. Норми етичної поведінки студентів і працівників визначені у розділі 2 Кодексу честі Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського». Детальніше: https://kpi.ua/code.*
- *Інклюзивне навчання. Засвоєння знань та умінь в ході вивчення дисципліни «Сталий інноваційний розвиток» може бути доступним для більшості осіб з особливими освітніми потребами, окрім здобувачів з серйозними вадами зору, які не дозволяють виконувати завдання за допомогою персональних комп'ютерів, ноутбуків та/або інших технічних засобів.*
- *Навчання іноземною мовою. У ході виконання завдань студентам може бути рекомендовано звернутися до англомовних джерел.*

8. Види контролю та рейтингова система оцінювання результатів навчання (РСО)

Рейтингова система оцінювання результатів навчання

Рейтинг студента з кредитного модуля складається з балів, які він отримує за:

1) виконання та захист 4 практичних/лабораторних робіт;

2) презентація результатів системного аналізу обраної предметної галузі.

3) контрольні опитування щодо засвоєння матеріалу (3 опитувальні форми);

1. Практичні/ лабораторні роботи. Упродовж семестру студент має виконати 5 лабораторних робіт (ЛР).

Рейтингова оцінка кожної ЛР складається з 2 частин, які оцінюються окремо. а. Якість підготовки до роботи, її виконання та оформлення звіту.

- за умови правильно оформленого звіту з точним виконанням завдання ЛР – 7 балів;

- за наявності несуттєвих неточностей в оформленні або процедурі виконання ЛР – 5-6 балів;

- за наявності порушень в оформленні, неповного або неточного виконання – 3-4 бали.

б. Якість захисту матеріалу. В цій частині оцінюється ступінь володіння теоретичним і практичним матеріалом за темою роботи.

- відмінне володіння матеріалом – 7 балів;

- добре володіння матеріалом – 5-6 балів;

- задовільне володіння матеріалом – 4 бали.

	№ роботи			
	1	*2*	*3*	*4*
Строк здачі та захисту роботи	*28.02*	*10.03*	*25.03*	*05.04*

За несвоєчасну подачу звіту з ЛР та несвоєчасний захист ЛР нараховуються штрафні бали: мінус 0.3 бали за кожний тиждень запізнення.

*Максимальна кількість балів за всі ЛР дорівнює: 5*18 =72 балів.*

2. Критерії оцінювання презентації "Результати системного аналізу обраної предметної галузі":
Змістовна повнота та відповідність системного аналізу:

- *«Відмінно» (5 балів): Аналіз повністю відповідає обраній предметній галузі, охоплює не менше 95% необхідної інформації, з чітко визначеними системними зв'язками та правильно створеними класифікаторами. Презентація структурована та логічно побудована, без помилок.*
- *«Добре» (4 бали): Аналіз достатньо повний (не менше 75% потрібної інформації) або повний, але з незначними неточностями. Системні зв'язки та класифікатори переважно коректні, але можуть бути дрібні недоліки у структурі презентації.*
- *«Задовільно» (3 бали): Аналіз неповний (не менше 60% необхідної інформації), наявні незначні помилки в визначенні системних зв'язків або класифікаторів, структура презентації слабка або не до кінця логічна.*
- *«Незадовільно» (0 балів): Презентація не відповідає вимогам для «задовільно», містить суттєві недоліки в аналізі або відсутні ключові елементи.*

Якість використання методів текстової аналітики:

- *«Відмінно» (5 балів): Методи текстової аналітики застосовані правильно та адекватно до завдань аналізу. Продемонстровані інсайти з текстових даних відповідають предметній галузі, та їх інтеграція у загальну структуру аналізу є цілісною та логічною.*
- *«Добре» (4 бали): Методи текстової аналітики застосовані з незначними помилками або неповністю. Загалом аналіз адекватний, але можливі деякі неточності в інтеграції інсайтів у загальну картину.*
- *«Задовільно» (3 бали): Методи текстової аналітики застосовані частково або з помилками. Інсайти з текстових даних не повністю інтегровані у системний аналіз.*
- *«Незадовільно» (0 балів): Методи текстової аналітики застосовані неправильно або взагалі не використані.*

Якість презентації та візуалізація результатів:

- *«Відмінно» (5 балів): Презентація має чітку структуру, всі результати візуалізовані зрозуміло та наочно. Використані відповідні графіки, схеми та таблиці, що доповнюють аналіз та полегшують сприйняття матеріалу.*
- *«Добре» (4 бали): Презентація загалом добре структурована, але можуть бути дрібні недоліки у візуалізації або логіці викладу. Візуалізація присутня, але можливо недостатньо деталізована.*
- *«Задовільно» (3 бали): Презентація не до кінця логічно побудована, візуалізація неповна або недостатньо чітка. Можливі труднощі у сприйнятті матеріалу через слабку організацію або неякісні графіки.*
- *«Незадовільно» (0 балів): Презентація неструктурована, візуалізація відсутня або значно ускладнює розуміння матеріалу.*

Ораторська майстерність та відповідь на запитання:

- *«Відмінно» (5 балів): Презентація впевнена, доповідач відповідає на всі питання чітко, демонструючи глибоке розуміння матеріалу.*
- *«Добре» (4 бали): Презентація загалом добре проведена, проте можуть бути незначні недоліки у впевненості або повноті відповідей на питання.*
- *«Задовільно» (3 бали): Презентація недостатньо впевнена, можливі труднощі з відповідями на питання, що вказують на поверхневе розуміння окремих аспектів матеріалу.*
- *«Незадовільно» (0 балів): Доповідач не впорався з презентацією або не зміг відповісти на більшість запитань.*

Загальний підсумок:

Оцінка за презентацію визначається як сума балів за всі критерії (максимум – 20 балів).

3. Критерії оцінювання контрольних опитувань щодо засвоєння матеріалу:

Повнота та правильність відповідей (до 5 балів):

- *«Відмінно» (5 балів): Всі питання контрольного опитування повністю та правильно вирішені, демонструючи глибоке розуміння матеріалу.*
- *«Добре» (4 бали): Більшість відповідей правильні (не менше 75% матеріалу засвоєно), проте можуть бути дрібні неточності або неповнота у деяких відповідях.*
- *«Задовільно» (3 бали): Неповна відповідь (не менше 60% матеріалу засвоєно), можуть бути кілька значних помилок або відсутніх частин.*
- *«Незадовільно» (0 балів): Відповіді на більшість питань неправильні або не подані, опитування не засвідчує достатнього рівня знань.*

Пунктуальність (штраф 1-2 бали):

- *1 бал штрафу: Робота здана із запізненням до 3 днів після кінцевого терміну публікації.*
- *2 бали штрафу: Робота здана із запізненням від 4 до 7 днів після публікації.*

У разі, якщо студент вже набрав 100 балів, окремі контрольні опитування щодо засвоєння матеріалу можна пропустити.

Максимальна сума балів за роботу в семестрі складає 100. Необхідною умовою допуску до заліку є отримання рейтингу 40 балів і вище.

Таблиця відповідності рейтингових балів оцінкам за університетською шкалою:

Кількість балів	*Оцінка*
100-95	Відмінно
94-85	Дуже добре
84-75	Добре
74-65	Задовільно
64-60	Достатньо
Менше 60	Незадовільно
Не виконані умови допуску	Не допущено

9. Додаткова інформація з дисципліни (освітнього компонента)

- *Сертифікати проходження дистанційних чи онлайн курсів за тематикою дисципліни можуть бути зараховані з додатковими 5 – 10 балами до загального рейтингу студента (за темами LLM, технології ML у галузі текстової аналітики).*

Робочу програму навчальної дисципліни (силабус):

Складено ст. викладач, к.т.н, Савастьянов Володимир Володимирович

Ухвалено кафедрою ММСА (протокол №13 від 05.06.2024)

Погоджено Методичною комісією НН ІПСА (протокол №10 від 24.06.2024)