Інститут телекомунікацій і глобального інформаційного простору НАН України

Відділ інформаційної безпеки

«Застосування технологій обробки великих даних та текстової аналітики для видобування знань з неструктурованих даних»

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни	
Рівень вищої освіти	***Третій (освітньо-науковий)***
Галузь знань	*Галузь 12 Інформаційні технології*
Спеціальність	*122 – Комп'ютерні науки*
Освітньо-наукова програма	*Інформаційні технології*
Статус дисципліни	*Вибіркова*
Форма навчання	*очна (денна)*
Рік підготовки, семестр	*2 рік, весняний (4-й) семестр*
Обсяг дисципліни	*90 годин / 3 кредити ЄКТС* *(лекції – 20 год., практичні заняття – 10 год., СРС – 60 год.)*
Семестровий контроль/ контрольні заходи	*залік, модульна контрольна робота*
Розклад занять	*2 год лекційних та 1 год практичних занять відповідно до розкладу*
Мова викладання	*Українська*
Інформація про керівника курсу / викладачів	Лекції та практичні заняття проводить: *к.т.н., старший дослідник Клименков Олег Анатолійович*
Розміщення курсу	Освітній простір ІТГІП: https://edu.itgip.org/ (доступ надає викладач)

Програма навчальної дисципліни

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

Дисципліна «Застосування технологій обробки великих даних та текстової аналітики для видобування знань з неструктурованих даних» *формує здатність проєктувати масштабовані й відтворювані процеси одержання знань із великих масивів текстових та інших неструктурованих даних. Курс охоплює розподілене зберігання й оброблення, створення та документування корпусів, класичні й нейромережеві методи текстової аналітики, інформаційне вилучення, семантичний пошук, графи знань і генерацію з доповненням результатами пошуку. Окрему увагу приділено українським і багатомовним даним, оцінюванню якості, простежуваності джерел, конфіденційності та відтворюваності.*

Мета: *формування здатності самостійно розробляти, реалізовувати та експериментально оцінювати масштабовані технологічні конвеєри оброблення великих масивів неструктурованих*

даних, які забезпечують вилучення, нормалізацію, зв'язування та представлення фактів, сутностей, зв'язків, подій, тем і перевірюваних знань.

Предмет вивчення – *архітектури, методи та програмні засоби розподіленого оброблення великих даних, комп'ютерного опрацювання природної мови, інформаційного пошуку, вилучення інформації та представлення знань у семантичних індексах, онтологіях і графах знань.*

Програмні результати навчання:

Загальні компетентності

ЗК01. Здатність до абстрактного мислення, аналізу та синтезу.
ЗК02. Здатність до пошуку, оброблення та аналізу інформації з різних джерел.
ЗК04. Здатність розв'язувати комплексні проблеми комп'ютерних наук на основі системного наукового світогляду та загального культурного кругозору із дотриманням принципів професійної етики та академічної доброчесності.
ЗК05. Здатність використовувати академічну українську та іноземну мови у професійній діяльності та дослідженнях, застосовувати методи та технології наукових комунікацій для оприлюднення результатів досліджень, ефективно взаємодіяти у плюрилінгвальному (багатомовному) та полікультурному середовищі, працювати у міжнародному контексті.
ЗК06. Здатність до набуття концептуальних та методологічних знань в галузі чи на межі галузей знань або професійної діяльності, використання спеціалізованих умінь/навичок і методів, необхідних для розв'язання значущих проблем у сфері професійної діяльності, науки та/або інновацій, розширення та переоцінки вже існуючих знань і професійної практики, безперервного саморозвитку та самовдосконалення.

Фахові компетентності

СК01. Здатність виконувати оригінальні дослідження, досягати наукових результатів, які створюють нові знання у комп'ютерних науках та дотичних до них міждисциплінарних напрямах і можуть бути опубліковані у провідних наукових виданнях з комп'ютерних наук та суміжних галузей.
СК02. Здатність застосовувати сучасні методології, методи та інструменти експериментальних і теоретичних досліджень у сфері комп'ютерних наук, сучасні цифрові технології, бази даних та інші електронні ресурси у науковій та освітній діяльності.
СК03. Здатність виявляти, ставити та вирішувати дослідницькі науково-прикладні задачі та/або проблеми в сфері комп'ютерних наук, оцінювати та забезпечувати якість виконуваних досліджень.
СК04. Здатність ініціювати, розробляти і реалізовувати комплексні інноваційні проєкти у галузі комп'ютерних наук та дотичних до неї міждисциплінарних проєктах, демонструвати лідерство під час їх реалізації.
СК06. Здатність аналізувати та оцінювати сучасний стан і тенденції розвитку комп'ютерних наук та інформаційних технологій.
СК07. Здатність спілкуватись українською та англійською мовами, усно і письмово презентувати результати власних наукових досліджень, розуміти англомовні наукові тексти у сфері комп'ютерних наук та інформаційних технологій, вільно спілкуватись з питань, що стосуються сфери наукових та експертних знань, з колегами, широкою науковою спільнотою, суспільством у цілому.
СК08. Здатність дотримуватись академічної та професійної доброчесності, морально-етичних правил поведінки, етики досліджень в академічному середовищі.

Програмні результати навчання

ПРН01. Мати передові концептуальні та методологічні знання з комп'ютерних наук і на межі предметних галузей, а також дослідницькі навички, достатні для проведення наукових і прикладних досліджень на рівні останніх світових досягнень з відповідного напряму, отримання нових знань та/або здійснення інновацій.
ПРН03. Формулювати і перевіряти гіпотези; використовувати для обґрунтування висновків належні докази, зокрема, результати теоретичного аналізу, експериментальних досліджень і математичного та/або комп'ютерного моделювання, наявні літературні дані.
ПРН04. Розробляти та досліджувати концептуальні, математичні і комп'ютерні моделі процесів і систем, ефективно використовувати їх для отримання нових знань та/або створення інноваційних продуктів у комп'ютерних науках та дотичних міждисциплінарних напрямах.
ПРН05. Планувати і виконувати експериментальні та/або теоретичні дослідження з комп'ютерних наук та дотичних міждисциплінарних напрямів з використанням сучасних інструментів, критично аналізувати результати власних досліджень і результати інших дослідників у контексті усього комплексу сучасних знань щодо досліджуваної проблеми.
ПРН06. Застосовувати сучасні інструменти і технології пошуку, оброблення та аналізу інформації, зокрема, статистичні методи аналізу даних великого обсягу та/або складної структури, спеціалізовані бази даних та інформаційні системи.
ПРН07. Розробляти та реалізовувати наукові та/або інноваційні інженерні проєкти, які дають можливість переосмислити наявне та створити нове цілісне знання та/або професійну практику і розв'язувати значущі наукові та технологічні проблеми комп'ютерної науки з дотриманням норм академічної етики і врахуванням соціальних, економічних, екологічних та правових аспектів.
ПРН08. Визначати актуальні наукові та практичні проблеми у сфері комп'ютерних наук, глибоко розуміти загальні принципи та методи комп'ютерних наук, а також методологію наукових досліджень, застосувати їх у власних дослідженнях у сфері комп'ютерних наук та у викладацькій практиці.
ПРН10. Відшуковувати, оцінювати та критично аналізувати інформацію щодо поточного стану та трендів розвитку, інструментів та методів досліджень, наукових та інноваційних проєктів з комп'ютерних наук.

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

Дисципліна «Застосування технологій обробки великих даних та текстової аналітики для видобування знань з неструктурованих даних» вивчається у четвертому (весняному) семестрі. Для її успішного опанування необхідні результати навчання з обов'язкових дисциплін Н.06 «Стан та перспективи розроблення інформаційних технологій та комп'ютерних систем» і Н.07 «Сучасні методи, моделі та технології оброблення даних в інформаційно-комунікаційних системах», а також базові навички програмування мовою Python, роботи з базами даних, статистики й машинного навчання. Вивчення В.02 «Використання сучасних технологій у сфері аналізу даних, машинного навчання та інтелектуальних інформаційних систем» є рекомендованим, але не обов'язковим, оскільки обидві дисципліни вибіркові. Результати курсу використовуються у дисертаційному дослідженні, наукових публікаціях і під час створення програмних прототипів видобування знань.

3. Зміст навчальної дисципліни

Тема 1. Великі та неструктуровані дані: властивості й архітектури
Тема 2. Корпуси та подання текстових даних

Тема 3. Текстова аналітика та вилучення інформації
Тема 4. Від неструктурованого тексту до структурованих знань
Тема 5. Масштабований наскрізний конвеєр та оцінювання

4. Навчальні матеріали та ресурси

Базова література та відкриті електронні ресурси (безоплатний доступ; дата звернення: 03.08.2026)

1. Олещенко Л. М. Технології оброблення великих даних: конспект лекцій. Київ: КПІ ім. Ігоря Сікорського, 2021. 227 с. https://ela.kpi.ua/items/4ad484d3-aa21-4f7e-be09-a6d8e1d1e8bc Додатковий комп'ютерний практикум: https://ela.kpi.ua/items/204a7590-44f1-4168-bfa6-d3e1fb60d47c
2. Онищенко В. В., Гавриленко О. В., Мягкий М. Ю. Методи аналізу даних та машинного навчання в інформаційно-управляючих системах. Курс лекцій. Ч. 2. Київ: КПІ ім. Ігоря Сікорського, 2025. 318 с. https://ela.kpi.ua/handle/123456789/74212
3. Leskovec J., Rajaraman A., Ullman J. D. Mining of Massive Datasets. 3rd ed. Official Stanford open edition. https://infolab.stanford.edu/~ullman/mmds/book.pdf
4. Jurafsky D., Martin J. H. Speech and Language Processing. 3rd ed. Online manuscript released January 6, 2026. https://web.stanford.edu/~jurafsky/slp3/
5. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008. Legal online edition. https://nlp.stanford.edu/IR-book/information-retrieval-book.html
6. Bird S., Klein E., Loper E. Natural Language Processing with Python. Official NLTK online edition. https://www.nltk.org/book/
7. Apache Spark Project. Official Documentation: SQL, DataFrames, Structured Streaming, MLlib and PySpark. https://spark.apache.org/docs/latest/
8. Hugging Face. LLM Course: transformers, datasets, tokenizers and language models. https://huggingface.co/learn/llm-course/en/chapter1/1
9. W3C. RDF 1.1 Primer; OWL 2 Web Ontology Language Primer. https://www.w3.org/TR/rdf11-primer/ ; https://www.w3.org/TR/owl2-primer/

Навчальний контент

5. Методика опанування навчальної дисципліни (освітнього компонента)

№ п/п	Змістові модулі / теми	Кількість годин, відведених на:			Термін виконання
		лекції	практ. заняття	сам. робота.	
1	2	3	4	5	6
Т1	Великі та неструктуровані дані: властивості й архітектури *Короткий зміст теми: Властивості великих даних та їх вплив на архітектурні рішення. Джерела неструктурованих даних, пакетна й потокова обробка, розподілені обчислення, локальність даних, відмовостійкість і узгодженість. Розподілені файлові й об'єктні сховища, бази NoSQL, озера даних та озерні сховища. Метадані, якість, походження, ліцензії, персональні й конфіденційні дані.*	4	2	12	1-й -3-й тижні

Т2	Корпуси та подання текстових даних *Короткий зміст теми: Збирання й документування корпусу, кодування Unicode, визначення мови, сегментація, токенізація, нормалізація, лематизація та морфологічна обробка українських і багатомовних текстів. Дедуплікація, анонімізація, розмітка й узгодженість анотаторів. Розріджені подання TF–IDF, статичні та контекстні векторні подання слів, речень і документів.*	4	2	12	4-й -6-й тижні
Т3	Текстова аналітика та вилучення інформації *Короткий зміст теми: Класифікація, кластеризація, тематичне моделювання, аналіз тональності й позиції. Інформаційний пошук, BM25 і семантична подібність. Розпізнавання іменованих сутностей, вилучення зв'язків і подій, зв'язування сутностей. Метрики точності, повноти та F1, метрики пошуку й ранжування, побудова базового рішення, аналіз помилок, дисбаланс і доменний зсув.*	4	2	12	7-й -9-й тижні
Т4	Від неструктурованого тексту до структурованих знань *Короткий зміст теми: Нормалізація та зв'язування сутностей, онтології, RDF/OWL і графи знань. Лексичний, щільний і гібридний пошук, векторні індекси та повторне ранжування. Генерація з доповненням пошуком (RAG), поділ документів на фрагменти, походження й цитування відповіді, перевірка опори на джерела, узгодженість фактів і людська верифікація вилучених знань.*	4	2	12	10-й -12-й тижні
Т5	*Масштабований наскрізний конвеєр та оцінювання* *Короткий зміст теми: Apache Spark/PySpark або еквівалентні відкриті засоби; поділ даних на розділи, колонкові формати, стиснення, відкладене виконання, перемішування даних і оптимізація. Оркестрація, версійність корпусу та моделей, індексування, оновлення, моніторинг і відтворюваність. Оцінювання продуктивності та якості, малоресурсність української мови, упередження, авторське право, безпека й захист даних.*	4	2	12	13-й -15-й тижні
МКР	*Модульна контрольна робота*				15й тиждень
	Всього за дисципліну	**20**	**10**	**60**	

6. Практичні заняття

Практична робота №1. Створення, документування та аудит текстового корпусу.
На наданому або погодженому навчальному корпусі, який не використовується в індивідуальному мініпроєкті, нормалізувати метадані, виконати визначення мови, очищення, дедуплікацію та

анонімізацію. Перевірити ліцензії, персональні дані, походження, репрезентативність і можливі упередження.
Очікуваний результат: Навчальний корпус, паспорт даних, відтворюваний код підготовки та звіт про якість і обмеження (5–7 сторінок).
Практична робота №2. Масштабоване опрацювання корпусу.
На навчальному корпусі з практичної роботи №1 реалізувати завантаження, поділ, перетворення й агрегування текстів у PySpark або еквівалентному відкритому середовищі. Порівняти формати зберігання, плани виконання, час і споживання ресурсів; дослідити вплив розділення даних і перемішування.
Очікуваний результат: Відтворюваний конвеєр, журнал вимірювань і технічний звіт з обґрунтуванням архітектурних рішень (5–7 сторінок).
Практична робота №3. Порівняльна текстова аналітика та вилучення інформації.
На тому самому навчальному корпусі побудувати класичне базове рішення на основі TF–IDF або BM25 і контекстну модель для класифікації, кластеризації, тематичного моделювання, розпізнавання сутностей чи вилучення зв'язків. Виконати коректне розділення даних, оцінювання метрик і аналіз помилок.
Очікуваний результат: Код, таблиця експериментів, метрики, аналіз помилок і дослідницький звіт (6–8 сторінок).
Практична робота №4. Наскрізний прототип видобування знань.
Робота виконується протягом двох практичних занять на матеріалах навчального корпусу практичних робіт №1–3, окремо від індивідуального міні-проєкту. Реалізувати семантичний або гібридний пошук, граф знань чи систему генерації з доповненням пошуком. Забезпечити локалізацію джерела, цитування, зв'язування сутностей, перевірку фактів і людську верифікацію; оцінити якість пошуку або вилучення, масштабованість, безпеку та етичні ризики.
Очікуваний результат: Навчальний прототип, репозиторій, архітектурна схема, звіт (8–12 сторінок) і коротка демонстрація.

7. Самостійна робота здобувача ступеня доктора філософії

№	Вид самостійної роботи	Кількість годин (орієнтовно)
1.	Опрацювання лекційних матеріалів і рекомендованих відкритих, зокрема англомовних, джерел	8
2.	Формування, очищення та документування текстового корпусу	8
3.	Підготовка до практичних робіт та оформлення відтворюваних звітів	16
4.	Індивідуальний наскрізний міні-проєкт видобування знань	17
5.	Аналіз помилок, упереджень, правових обмежень і відтворюваності	5
6.	Підготовка до модульної контрольної роботи	3
7.	Підготовка двомовної анотації, презентації-мінілекції та захисту міні-проєкту	3
	Всього	**60**

Політика та контроль

8. Політика навчальної дисципліни (освітнього компонента)

Вимоги, яких має дотримуватися аспірант в рамках даної дисципліни:

- *правила відвідування занять: відвідування лекцій і практичних занять є необхідною складовою опанування дисципліни; за обґрунтованої відсутності аспірант погоджує з викладачем спосіб опрацювання матеріалу та виконання контрольних заходів; сама присутність не оцінюється;*
- *правила роботи на заняттях: аспірант бере активну участь у дискусіях, аналізі наукових публікацій і взаємному рецензуванні, аргументує вибір методів та представляє результати власних експериментів;*
- *захист практичних робіт: код, дані, параметри, версії програм і звіти подаються у визначені строки; робота зараховується після перевірки відтворюваності, індивідуального захисту та усунення істотних недоліків;*
- *використання засобів штучного інтелекту: допускається лише в межах конкретного завдання з обов'язковою декларацією назви засобу, моделі або версії, дати, мети, суттєвих запитів і способу перевірки результатів; аспірант повністю відповідає за дані, код, посилання, обчислення й висновки; заборонено передавати зовнішнім сервісам персональні, конфіденційні, службові або неопубліковані дослідницькі дані без належної правової підстави;*
- *академічна доброчесність: діють Закон України «Про академічну доброчесність» № 4742-IX, Закони України «Про освіту», «Про захист персональних даних», «Про авторське право і суміжні права», Етичний кодекс ученого та внутрішні акти ІТГІП НАН України; заборонені плагіат, фабрикація, фальсифікація, приховане співавторство, вигадані джерела, несамостійне виконання і подання результатів невиконаних експериментів; запозичені дані, код і моделі супроводжуються посиланням та відомостями про ліцензію.*

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

Семестровий контроль - залік. Рейтингова система оцінювання результатів навчання передбачає оцінювання заходів поточного контролю з дисципліни впродовж семестру. Рейтингова оцінка здобувача складається з балів, отриманих здобувачем за результатами заходів поточного контролю. Рейтингова оцінка доводиться до здобувачів на передостанньому занятті з дисципліни в семестрі. Здобувачі, які виконали всі умови допуску до заліку та мають рейтингову оцінку 60 і більше балів, отримують відповідну до набраного рейтингу оцінку без додаткових випробувань

Зі здобувачами, які виконали всі умови допуску до заліку та мають рейтингову оцінку менше 60 балів, а також з тими здобувачами, хто бажає підвищити свою рейтингову оцінку, на останньому за розкладом занятті з дисципліни в семестрі викладач проводить семестровий контроль у вигляді залікової контрольної роботи.

Для посилення зацікавленості здобувачів у якісному виконанні індивідуальних семестрових завдань, передбачених індивідуальним навчальним планом здобувача, рейтингову оцінку, у разі виконання залікової контрольної роботи, можна визначати як суму балів за залікову контрольну роботу та балів за індивідуальне семестрове завдання. У цьому випадку розмір шкали оцінювання залікової контрольної роботи зменшується на максимальне значення балів, передбачених за виконання відповідного індивідуального семестрового завдання.

Після виконання залікової контрольної роботи, якщо оцінка за залікову контрольну роботу більша ніж за рейтингом, здобувач отримує оцінку за результатами залікової контрольної роботи. Якщо оцінка за залікову контрольну роботу менша ніж за рейтингом, то здобувач отримує більшу з оцінок, що отримані за результатами залікової контрольної роботи або за рейтингом.

До відомості семестрового контролю викладач заносить рейтингові бали, отримані здобувачем у семестрі або за результатами виконання залікової контрольної роботи, та оцінку відповідно до цих балів

Критерії нарахування балів:

1. Практичні заняття оцінюються виходячи з максимальної кількості балів - 20 бали кожне:

- «відмінно» – 95 відсотків максимального балу;
- «добре» –75-95 відсотків максимального балу;
- «задовільно» – 60-75 відсотків максимального балу;
- «достатньо» – 50 відсотків – робота виконана, але не захищена.

Умови допуску до підсумкового контролю є зарахування усіх практичних робіт.

Рейтинг студента з дисципліни складається з балів, що він отримує за:

– виконання практичних (лабораторних) робіт;

– виконання самостійної роботи.

За період вивчення дисципліни студент може набрати 100 балів. Їх розподіл між видами робіт наведено в таблиці 1

Таблиця 1

Бали за виконання	Номер практичної роботи або теми					Разом
	1	2	3	4	5	
Практичної роботи	10	10	10	10	10	100
Самостійної роботи	10	10	10	10	10	

2. Залікова контрольна робота оцінюється за такими критеріями:

- «відмінно» – повна відповідь (не менше 90% потрібної інформації), надані відповідні обґрунтування та особистий погляд;
- «добре» – достатньо повна відповідь (не менше 75% потрібної інформації), що виконана згідно з вимогами до рівня «умінь», або незначні неточності);
- «задовільно» – неповна відповідь (не менше 60% потрібної інформації. що виконана згідно з вимогами до «стереотипного» рівня та деякі помилки);
- «незадовільно» – незадовільна відповідь – 0 балів.

Залікова контрольна робота передбачається у вигляді тесту, критерії оцінювання тесту:

Кількість правильних відповідей	*Відсоток правильних відповідей*	*Оцінка за національною шкалою*	*Оцінка за шкалою ECTS*
48-50	*95-100*	*Відмінно*	*A*
41-47	*82-94*	*Дуже добре*	*B*
37-40	*75-81*	*Добре*	*C*
34-36	*69-74*	*Задовільно*	*D*
30-33	*60-68*	*Достатньо*	*E*
5-29	*10-13*	*Не задовільно*	*FX*

Відповідність рейтингових балів оцінкам за шкалою Інституту та *шкалою ECTS*

Рейтингова оцінка	*Оцінка за національною шкалою*	*Оцінка за шкалою ECTS*
90-100	*Відмінно*	*A*
82-89	*Дуже добре*	*B*
75-81	*Добре*	*C*
69-74	*Задовільно*	*D*
60-68	*Достатньо*	*E*
45-59	*Не задовільно*	*FX*

Невиконання умов допуску до семестрового контролю	*Не допущено*	

10. Додаткова інформація з дисципліни (освітнього компонента)

- *питання, які виносяться на семестровий контроль:*

  1. Які властивості даних зумовлюють необхідність розподіленого зберігання й оброблення?

  2. Чим пакетна обробка відрізняється від потокової та як обрати архітектуру?

  3. Як локальність, поділ на розділи, колонкові формати й перемішування впливають на продуктивність?

  4. Як документувати походження, ліцензії, якість та обмеження текстового корпусу?

  5. Які особливості має нормалізація й морфологічне опрацювання українських і багатомовних текстів?

  6. Чим TF–IDF і BM25 відрізняються від контекстних векторних подань?

  7. Як оцінювати класифікацію, тематичне моделювання, вилучення сутностей і зв'язків?

  8. Чим лексичний пошук відрізняється від щільного й гібридного семантичного пошуку?

  9. Як виконуються зв'язування сутностей, побудова онтології та формування графа знань?

  10. Які компоненти має система генерації з доповненням пошуком і як перевіряти її опору на джерела?

  11. Як запобігати витоку, доменному зсуву й мовним упередженням під час оцінювання?

  12. Як забезпечити відтворюваність, конфіденційність, авторське право та людську верифікацію в наскрізному конвеєрі?

Можливе визнання результатів неформальної освіти за тематикою дисципліни після подання підтвердження, перевірки досягнутих результатів і виконання індивідуального контрольного завдання; рішення приймається відповідно до внутрішніх актів Інституту, а підсумковий рейтинг не може перевищувати 100 балів.

Робочу програму навчальної дисципліни (силабус):

Склав: к.т.н., старший дослідник Клименков Олег Анатолійович

Затверджено:

Вченою радою Інституту телекомунікацій
і глобального інформаційного простору
НАН України Протокол № 10 від 13 серпня 2026 р.

Введено в дію:

Наказом директора
Наказ від 14 серпня 2026 р. № 12