Національний технічний університет
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ І
імені ІГОРЯ СІКОРСЬКОГО»

Емблема кафедри (за наявності)

Кафедра біомедичної кібернетики

Обробка природних мов

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	***Другий (магістерський)***
Галузь знань	*12 Інформаційні технології*
Спеціальність	*122 Комп'ютерні науки*
Освітня програма	*Комп'ютерні технології в біології та медицині*
Статус дисципліни	*Вибіркова*
Форма навчання	*очна(денна)/дистанційна/змішана*
Рік підготовки, семестр	*1 курс, весінній семестр*
Обсяг дисципліни	*120 годин, 4 кредити*
Семестровий контроль/ контрольні заходи	*Залік, МКР*
Розклад занять	
Мова викладання	*Українська*
Інформація про керівника курсу / викладачів	проф. Настенко Євген Арнольдович, e-mail: nastenko.e@gmail.comac. Матвійчук Олександр Вадимович, e-mail: matviichuk.oleksandr@lll.kpi.ua
Розміщення курсу	Google Drive: https://drive.google.com/drive/folders/1JP4g_sOE7z6ULr__JPSnrEy_5NdwWAvi

Програма навчальної дисципліни

1. ОПИС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ, ЇЇ МЕТА, ПРЕДМЕТ ВИВЧАННЯ ТА РЕЗУЛЬТАТИ НАВЧАННЯ

знання:

ЗН 21	Знання принципів, інструментальних засобів, мов веб-програмування, технологій створення баз даних, сховищ і вітрин даних та бази знань для розробки розподілених застосувань з інтеграцією баз і сховищ даних в архітектуру клієнт-сервер.

ФК 9	Здатність реалізувати багаторівневу обчислювальну модель на основі архітектури клієнт-сервер, включаючи бази даних, знань і сховища даних, виконувати розподілену
	обробку великих наборів даних на кластерах стандартних серверів для забезпечення обчислювальних потреб користувачів, у тому числі на хмарних сервісах.

уміння:

УМ 21	Використовувати методи, технології та інструментальні засоби для проектування та розробки клієнт-серверних застосувань, проектувати концептуальні, логічні та фізичні моделі баз даних, розробляти й оптимізовувати запити до них, створювати розподілені бази даних, сховища та вітрини даних, бази знань, у тому числі на хмарних сервісах.

2. ПРЕРЕКВІЗИТИ ТА ПОСТРЕКВІЗИТИ ДИСЦИПЛІНИ (МІСЦЕ В СТРУКТУРНОЛОГІЧНІЙ СХЕМІ НАВЧАННЯ ЗА ВІДПОВІДНОЮ ОСВІТНЬОЮ ПРОГРАМОЮ)

У структурно-логічній схемі програми підготовки фахівця:

- дисципліну **забезпечують** наступні дисципліни:

№ п/п	Назва кредитного модуля (дисципліни)
1	Операційні системи
2	Об’єктно-орієнтоване програмування

- дисципліна **забезпечує** наступні навчальні дисципліни:

№ п/п	Назва кредитного модуля (дисципліни)
1	Комп'ютерні мережі

Навчальна дисципліна є основою для підготовки дипломних робіт за спеціальністю та в подальшій практичній роботі за фахом.

3. ЗМІСТ НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Розділ 1: Машинне представлення природної мови

У даному розділі розбираються методи, які забезпечують перетворення природної мови (усної та письмової) у вигляд, необхідний для подальшої її обробки за допомогою ЕОМ.

Тема 1. Статистичні методи перетворення

Тема 2. Контекстно-залежні методи перетворення

Розділ 2: Основні задачі обробки природної мови

Тема 1. Класифікація текстів

Тема 2. Задача мовного моделювання

Тема 3. Пошук схожих текстів

Тема 4. Машинний переклад

Тема 5. Named-entity recognition

4. НАВЧАЛЬНІ МАТЕРІАЛИ ТА РЕСУРСИ

Базові джерела:

1. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition - Тревор Хасти, Роберт Тибширани, Джером Фридман
2. Шолле Ф. - Глубокое обучение на Python
3. Attention Is All You Need - https://arxiv.org/pdf/1706.03762.pdf
4. Efficient Estimation of Word Representations in Vector Space -https://arxiv.org/pdf/1301.3781.pdf

Навчальний контент

5. МЕТОДИКА ОПАНУВАННЯ НАВЧАЛЬНОЇ ДИСЦИПЛІНИ (ОСВІТНЬОГО КОМПОНЕНТА)

Лекційні заняття

№ з/п	Назва теми лекції та перелік основних питань
1.	**Вступ до дисципліни.** Основні поняття дисципліни. Класифікація задач обробки природної мови.
2	**Статистичні методи перетворення тексту.** Поняття формальної граматики. Методи, засновані на частотних метриках (binary term frequencies, TF-IDF, BoW frequencies).

3	**Вступ до теорії нейронних мереж. Частина 1.** Поняття нейронної мережі, модель перцептрона Розенблатта, алгоритм зворотного поширення похибки та його варіації, поняття дропауту та регуляризації.
4	**Вступ до теорії нейронних мереж. Частина 2.** Рекурентні мережі. Механізм attention. Нейронні мережі типу transformer.
5	**Контекстно-залежні методи перетворення. Частина 1.** Моделі Word2Vec, fastText, Doc2Vec, GloVe embeddings
6	**Контекстно-залежні методи перетворення. Частина 2.** Моделі сімейства BERT та Google Big Bird
7	**Класифікація текстів** Логістична регресія, SVM, Naïve Bayes classifier та їх застосування у задачах класифікації текстів
8	**Задача мовного моделювання** Визначення частини мови та вирішення задачі автодоповнення. N-gram model
9	**Пошук схожих текстів. Частина 1.** Косинусна відстань, дивергенція Йенсена-Шеннена, Word Movers Distance (WMD), Relaxed WMD, Latent Dirichlet allocation
10	**Пошук схожих текстів. Частина 2.** Рекурентні нейронні мережі. Ранкінг за допомогою нейронної мережі автокодувальника, сіамської LSTM мережі.
11	**Машинний переклад** Підходи, засновані на використанні бази правил та на основі машинного та глибокого навчання. Приховані Марківські моделі.
12	**Named-entity recognition** Застосування word embeddings та нейронних мереж у задачі визначення типу термів

Комп’ютерні практикуми:

№ з/п	Назва теми заняття та перелік основних питань (перелік дидактичного забезпечення, посилання на літературу та завдання на СРС)	Кількість ауд. годин
1	Підготовка тексту до обробки (15.09, дедлайн 22.09)	2
2	Застосування статистичних метрик (22.09, дедлайн 6.10)	2
3	Застосування Word2Vec та BERT (6.10, дедлайн 20.10)	2
4	Класифікація текстів (20.10, дедлайн 3.11)	2
5	Пошук N схожих текстів (3.11, дедлайн 17.11)	2
6	Автоматизований переклад тексту з англійської на українську (17.11, дедлайн 1.12)	2
7	Виявлення e-mail, телефонів та адрес у тексті (1.12 – 15.12)	2

6. САМОСТІЙНА РОБОТА СТУДЕНТА

Студентам рекомендовано самостійно опрацьовувати теми, які зазначені як завдання на СРС до кожної лекції у пункті 5 Силабусу (5 годин). Заплановано виконання студентами розрахунково-графічної роботи (26 годин). На підготовку до модульної контрольної роботи відводиться 5 годин.

Політика та контроль

7. **ПОЛІТИКА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ (ОСВІТНЬОГО КОМПОНЕНТА)**

- **Відвідування занять** є обов’язковим. В умовах дистанційного навчання проведення занять відбувається за допомогою Google Meet. Постійне посилання на конференцію знаходиться у заголовку курса у Google Classroom.
- **Здача** комп’ютерних практикумів та відбувається у електронному вигляді за допомогою Google Drive (протоколи та вихідний код).
- **Заохочувальні бали** можна отримати за активність на практичних заняттях та за виконання додаткових завдань у комп’ютерних практикумах та індивідуальному завданні. Обсяг таких балів є необмеженим та визначається індивідуально у кожному випадку.
- **Штрафні бали** можуть призначатись за запізнення у сдачі комп’ютерних практикумів у розмірі 50% балів за запізнення до 2-х тижнів та 100% балів за більше ніж 2 тижні запізнення.
- **Політика щодо дотримання правил академічної доброчесності**. У випадку виявлення плагіату або списування вчасно подана на перевірку робота анулюється без права перездачі, невчасно подана робота також анулюється і на

неї додатково накладається штраф за запізнення відповідно до попереднього пункту.

8. ВИДИ КОНТРОЛЮ ТА РЕЙТИНГОВА СИСТЕМА ОЦІНЮВАННЯ РЕЗУЛЬТАТІВ НАВЧАННЯ (РСО)

Поточний контроль: комп'ютерні практикуми (7), МКР (2)

Календарний контроль: провадиться двічі на семестр як моніторинг поточного стану виконання вимог силабусу. Умови першого календарного контролю вважаються виконаними, якщо студентом здано 3 комп'ютерних практикуми. Умови другого календарного контролю вважаються виконаними, якщо студентом здано 5 комп'ютерних практикумів.

Семестровий контроль: залік

Умови допуску до семестрового контролю: зарахування усіх комп'ютерних практикумів та домашньої контрольної роботи, написання МКР на позитивну оцінку (>50% правильних відповідей), семестровий рейтинг більше 40 балів. Залік проводиться за «жорсткою» системою. У разі, якщо студент набрав більше 60% від максимального рейтингу, можливе виставлення оцінки автоматом, без проведення семестрового контролю.

Таблиця розподілу балів за різні види робіт:

Кількість балів	*Оцінка*
Комп'ютерні практикуми	70 балів
Модульні контрольні роботи	30 балів

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)

Є можливість зарахування сертифікатів про проходження курсів на платформах Coursera, Udemy за тематикою Natural Language Processing.

Робочу програму навчальної дисципліни (силабус):

Складено проф. Настенко Євген Арнольдович, ас. Матвійчук О.В.

Ухвалено кафедрою БМК ФБМІ (протокол №18 від 24.06.2024 р.)

Погоджено Методичною комісією факультету[1] ФБМІ (протокол №9 від 26.06.2024 р.)