КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

ФАКУЛЬТЕТ КОМП'ЮТЕРНИХ НАУК ТА КІБЕРНЕТИКИ

Кафедра математичної інформатика

«ЗАТВЕРДЖУЮ»
Заступник декана
з навчальної роботи
_____ Олена КАШПУР
«07» травня 2021 р.

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ
ІНТЕЛЕКТУАЛЬНА ОБРОБКА ТЕКСТІВ

для студентів

галузь знань	**12 «Інформаційні технології»**
спеціальність	**122 «Комп'ютерні науки»**
освітній рівень	**магістр**
освітня програма	**«Інформатика»**
вид дисципліни	**обов'язкова**

Форма навчання	**денна**
Навчальний рік	**2021/2022**
Семестр	**2**
Кількість кредитів ECTS	**4**
Мова викладання, навчання та оцінювання	**українська**
Форма заключного контролю	**іспит**

Викладачі: **д.ф.-м.н., проф. Олександр МАРЧЕНКО**

Пролонговано: на 20__/20__ н.р. _____(_____) «__»___ 20__р.

на 20__/20__ н.р. _____(_____) «__»___ 20__р.

КИЇВ – 2021

Розробник: Олександр МАРЧЕНКО, д.ф.-м.н., професор кафедри математичної інформатики

ЗАТВЕРДЖЕНО
Завідувач кафедри математичної інформатики
_____ Василь ТЕРЕЩЕНКО
Протокол № 10 від «27» 04 2021 р.

Схвалено гарантом освітньо-наукової програми «Інформатика»
_____ Степан ШКІЛЬНЯК
«6» травня 2021 року

Схвалено науково-методичною комісією факультету комп'ютерних наук та кібернетики
Протокол від «6» травня 2021 року № 10
Голова науково-методичної комісії _____ Людмила ОМЕЛЬЧУК

1. Мета дисципліни “Інтелектуальна обробка текстів” – отримання необхідних знань з сучасних інформаційних інтелектуальних лінгвістичних технологій, методів і алгоритмів комп'ютерної лінгвістики та їх подальше застосування для досліджень та програмування процесів розв’язання складних задач обробки природномовної інформації.

2. Попередні вимоги до опанування або вибору навчальної дисципліни:

1. Знати базовий матеріал дисциплін “Програмування”, “Дискретна математика”, “Теорія ймовірностей”, “Побудова та аналіз алгоритмів”, “Штучний інтелект”.

2. Вміти застосовувати сучасні інформаційні технології та мови програмування для розв′язання прикладних задач та проведення наукових досліджень.

3. Мати елементарні навички із побудови та аналізу алгоритмів, вміти програмувати на мовах Python, Java, C++; вміти використовувати програмні бібліотеки для побудови моделей машинного навчання, обробки природної мови тощо *(Scikit-learn, spaCy, NumPy, NLTK*).

3. Анотація навчальної дисципліни:

Навчальна дисципліна “ Інтелектуальна обробка текстів” є складовою освітньо-наукової програми підготовки фахівців за другим (*магістерським*) рівнем вищої освіти *галузі знань* 12 „Інформаційні технології” зі *спеціальності 122* „Комп’ютерні науки”, *освітньо-наукової програми* „Інформатика”.

Дана дисципліна є обов’язковою навчальною дисципліною за ***програмою “Інформатика”***. Викладається у **2** семестрі (**1** курс) в **обсязі – 120 год.** (**4 кредити ECTS**); зокрема: *лекції – 38 год., консультації – 2 год., самостійна робота – 80 год.*

У курсі передбачено **2** змістові частини та 2 контрольні роботи.

Завершується дисципліна **іспитом в 2 семестрі.**

В результаті вивчення навчальної дисципліни студент повинен:

знати комп'ютерні лінгвістичні технології, методи та алгоритми смислової обробки текстів. А саме: основні методи та алгоритми комп’ютерної лінгвістики, технології побудови баз знань та методи інтелектуальної обробки природномовної інформації, розробки онтологічних моделей представлення знань, алгоритми семантичного аналізу мови.

вміти застосувати методи та технології комп’ютерної лінгвістики для розв′язання прикладних задач та проведення наукових досліджень за фахом.

Для допуску до дисципліни „Інтелектуальна обробка текстів” освітньо-професійної програми «Інформатика» студент повинен опанувати компетентності та результати навчання, які надають дисципліни „Програмування”, ”Дискретна математика”, “Теорія ймовірностей”, “Побудова та аналіз алгоритмів”, „Штучний інтелект” програми «Інформатика». Дисципліна „Інтелектуальна обробка текстів” є базовою для засвоєння дисциплін спеціалізації та дисциплін вільного вибору студента програмістського спрямування програми «Інформатика».

4. Завдання (навчальні цілі):

набуття знань, умінь та навичок (компетентностей) на рівні новітніх досягнень у комп'ютерній лінгвістиці, відповідно до кваліфікації магістр з комп'ютерних наук.

Зокрема, розвивати:

- ЗК2. Здатність застосовувати знання у практичних ситуаціях;
- ЗК13. Здатність оцінювати та забезпечувати якість виконуваних робіт;
- СК8. Здатність вирішувати складні задачі інтелектуальної обробки даних з використанням еволюційного моделювання, нейромережних технологій, застосування обчислювального інтелекту для розв'язання практичних задач в різних галузях професійної діяльності.

Це надасть змогу проектувати та розробляти програмні системи інтелектуальної обробки текстів із застосуванням різних підходів: технологій на основі баз знань, на основі машинного навчання, нейронних мереж, нечітких моделей тощо.

5. Результати навчання за дисципліною:

Результат навчання (1. знати; 2. вміти; 3. комунікація; 4. автономність та відповідальність)		**Форми (та/або методи і технології) викладання і навчання**	**Методи оцінювання та пороговий критерій оцінювання (за необхідності)**	**Відсоток у підсумковій оцінці з дисципліни**
Код	**Результат навчання**			
РН1.1	*Знати основні поняття та підходи комп'ютерної лінгвістики*	*Лекція*	*Тест, 60% правильних відповідей, іспит*	20%
РН1.2	*Знати основні моделі та методи комп'ютерної лінгвістики та штучного інтелекту*	*Лекція*	*Тест, 60% правильних відповідей, іспит*	16%
РН1.3	*Знати основні технології комп'ютерної лінгвістики, засоби здобування та представлення даних/ знань, фреймворкі*	*Лекція*	*Тест, 60% правильних відповідей, іспит*	20%
РН2.1	*Вміти застосовувати на практиці інструментальні програмні засоби побудови систем інтелектуальної обробки текстів (ІОТ).*	*Самостійна робота*	*Іспит*	26%
РН3.1	*Обґрунтовувати власний погляд на задачу, спілкуватися з колегами з питань проектування та розробки систем ІОТ, складати письмові звіти*	*Самостійна робота*	*Поточне оцінювання*	6%
РН4.1	*Організовувати свою самостійну роботу для досягнення результату*	*Самостійна робота*	*Поточне оцінювання, іспит*	6%
РН4.2	*Відповідально ставитися до виконуваних робіт, нести відповідальність за їх якість*	*Самостійна робота*	*Поточне оцінювання, іспит*	6%

6. Співвідношення результатів навчання дисципліни із програмними результатами навчання

Результати навчання дисципліни Програмні результати навчання	РН 1.1	РН 1.2	РН 1.3	РН 2.1	РН 3.1	РН 4.1	РН 4.2
(з опису освітньої програми)							
ПРН 3. Опанувати нові інструменти роботи з даними, здійснюючи обробку веб-логів, text mining і машинне навчання, для прогнозування бізнес-процесів та ситуаційного управління, сентимент-аналізу відгуків, розробки рекомендаційних систем для сфери електронної комерції, медіа, соціальних мереж, банкингу, реклами тощо.	+	+		+	+	+	+
ПРН 4. Аналізувати великі дані та моделювати високорівневі абстракції у великих наборах даних різної природи, проектувати сховища великих даних, для видобутку даних і знань, візуалізовувати великі дані, будувати і оцінювати регресивні моделі, що генеруються на основі великих даних			+		+	+	

7. Схема формування оцінки

7.1 Форми оцінювання студентів:

- семестрове оцінювання: *Самостійна*

1. Контрольна робота (тест) 1: РН1.1, РН 2.1, РН3.1 – 24 бали.

2. Контрольна робота (тест) 2: РН1.2, РН1.3, РН 2.1, РН3.1 – 30 балів.

3. Поточне оцінювання (активна робота на заняттях): РН 3.1 – 6 балів

- підсумкове оцінювання (у формі іспиту) *вказується:*

- максимальна кількість балів які можуть бути отримані студентом: 40 балів;

- результати навчання які будуть оцінюватись: РН1.1, РН1.2, РН1.3, РН2.1, РН 4.1, РН4.2;

- форма проведення і види завдань: письмова.

Види завдань: 8 тестових та 6 письмових завдань.

Критерії оцінювання на іспиті

Завдання	Тема завдання	Максимальний відсоток від 40 балів	Всього відсотків
Завдання 1	Письмове запитання з онтологій	По 7%	28%
Завдання 2, 3, 4	Письмові запитання з сентимент аналізу		
Завдання 5, 6, 9, 10, 12	Тестове завдання з кластеризації текстів	По 5%	35%
Завдання 7, 8	Тестові завдання з POS-розмітки		
Завдання 11	Ко-референтний аналіз	10%	10%
Завдання 13	Розв’язання займенникової анафори. Розпізнавання іменованих сутностей	15%	15%
Завдання 14	Перевірка граматичної коректності текстів	12%	12%
			100%

Запитання для підготовки до іспиту

1. Вступ до комп'ютерної лінгвістики. Предмет та базові поняття. Основні напрями.
2. Токенізація. Нормалізація слів та стемінг. Сегментація речень.
3. Задача орфографічної корекції слів. Редакційна відстань.
4. Лінгвістичні моделі. N-грамні моделі
5. Сентимент аналіз
6. Розпізнавання іменованих сутностей в текстах
7. Розпізнавання відношень в текстах. Ко-референтний аналіз
8. Методи синтаксичного аналізу.
9. Лексикалізований синтаксичний аналіз.
10. Задачі пошуку у великих текстових масивах
11. Розв'язання неоднозначності слів
12. Семантичний аналіз текстів природною мовою
13. Алгоритми інтелектуальної обробки текстів природною мовою
14. Системи підтримки діалогу природною мовою
15. Корпусна лінгвістика.
16. Методи латентного аналізу.
17. Тензорний підхід.
18. Методи машинного перекладу текстів на іншу мову
19. Алгоритми реферування текстів

Студент не допускається до іспиту, якщо під час семестру набрав менше ніж 30 балів.

Для отримання загальної позитивної оцінки з дисципліни оцінка за іспит не може бути меншою 24 балів.

7.2 Організація оцінювання:

Терміни проведення форм оцінювання:

1. Контрольна робота (тест): до 8 тижня семестру.

2. Контрольна робота (тест): до 18 тижня семестру.

Студент має право на одне перескладання кожної контрольної роботи із можливістю отримання максимально 80% початково визначених за цю контрольну роботу балів. Термін перескладання визначається викладачем.

У випадку відсутності студента з поважних причин відпрацювання та перескладання контрольних робіт здійснюються у відповідності до «Положення про організацію освітнього процесу» від 07.05.2018 року.

7.3 Шкала відповідності оцінок

Відмінно / Excellent	90-100
Добре / Good	75-89
Задовільно / Satisfactory	60-74
Незадовільно / Fail	0-59
Зараховано / Passed	60-100
Не зараховано / Fail	0-59

8. Структура навчальної дисципліни

Тематичний план лекцій і лабораторних занять

№ лекції	Назва лекції	Кількість годин		
		Лекції	Консультації	Сам. робота
Частина 1. **Основи комп'ютерної лінгвістики**				
1	**Тема 1.** Вступ до комп'ютерної лінгвістики. Предмет та базові поняття. Основні напрями. ***Самостійна робота.*** Опрацювання лекційного матеріалу.	2		4
2	**Тема 2.** Токенізація. Нормалізація слів та стемінг. Сегментація речень. ***Самостійна робота.*** Опрацювання лекційного матеріалу.	2		4
3	**Тема 3.** Задача орфографічної корекції слів. Редакційна відстань. ***Самостійна робота.*** Опрацювання лекційного матеріалу.	2		4
4	**Тема 4.** Лінгвістичні моделі. N-грамні моделі. ***Самостійна робота.*** Опрацювання лекційного матеріалу.	2		4
5	**Тема 5.** Сентимент аналіз ***Самостійна робота.*** Опрацювання лекційного матеріалу.	2		6
6–7	**Тема 6.** Розпізнавання іменованих сутностей в текстах. Розпізнавання відношень в текстах. Ко-референтний аналіз ***Самостійна робота.*** Опрацювання лекційного матеріалу.	3		6
	Контрольна робота 1	1		
Всього по частині 1		14		28
Частина 2. **Алгоритми лінгвістичного аналізу**				
8	**Тема 7.** Методи синтаксичного аналізу ***Самостійна робота.*** Опрацювання лекційного матеріалу.	2		4
9	**Тема 8.** Лексикалізований синтаксичний аналіз ***Самостійна робота.*** Опрацювання лекційного матеріалу.	2		4
10	**Тема 9.** Задачі пошуку у великих текстових масивах ***Самостійна робота.*** Опрацювання лекційного матеріалу.	2		4
11	**Тема 10.** Розв'язання неоднозначності слів ***Самостійна робота.*** Опрацювання лекційного матеріалу.	2		4
12	**Тема 11.** Семантичний аналіз текстів природною мовою. ***Самостійна робота.*** Опрацювання лекційного матеріалу. Автоматизація побудови онтологічних баз знань.	2		6
13.	**Тема 12.** Алгоритми інтелектуальної обробки текстів природною мовою ***Самостійна робота.*** Опрацювання лекційного матеріалу. Автоматична побудова таксономій та тезаурусів.	2		6
14	**Тема 13.** Системи підтримки діалогу природною мовою ***Самостійна робота.*** Опрацювання лекційного матеріалу.	2		4

15–17	**Тема 14.** Корпусна лінгвістика. Методи латентного аналізу. Тензорний підхід. ***Самостійна робота.*** Опрацювання лекційного матеріалу. Факторизація багатовимірних тензорів.	5		10
18	**Тема 15.** Методи машинного перекладу текстів на іншу мову ***Самостійна робота.*** Опрацювання лекційного матеріалу.	2		4
19	**Тема 16.** Алгоритми реферування текстів ***Самостійна робота.*** Опрацювання лекційного матеріалу.	2		6
	Контрольна робота 2	1		
Всього по частині 2		24		52
Консультація			2	
	ВСЬОГО	38	2	80

Загальний обсяг *120* *год.,* в тому числі:
Лекцій **– *38*** *год.*
Консультації – ***2*** *год.*
Самостійна робота **– 80** *год.*

9. Рекомендовані джерела

Основні

1. С. Рассел П. Норвиг Искусственный интеллект. Современный подход. – М.,2006.
2. Лорьер Ж.-Л. Системы искусственного интеллекта. – М.,1991.
3. Dan Jurafsky & Chris Manning: Natural Language Processing (lectures of Stanford University course) https://www.youtube.com/playlist?list=PL6397E4B26D00A269
4. Nirenburg S., Raskin V. Ontological Semantics, 2001, http://crl.nmsu.edu/stuff.pages/Techial/book/index-book.html
5. Скороходько Ф.Ф. Семантические сети и автоматическая обработка текста. - Киев: Наукова думка, 1983.
6. Анисимов А.В. Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык. -Киев: Наукова думка, 1991
7. Анисимов А.В. Информатика. Творчество. Рекурсия. - Киев: Наукова думка,1988.-234 с.
8. Искусственный интеллект: В 3-х т. – М., 1990.

Додаткові:

9. Уинстон П. Искусственный интеллект. – М., 1980.
10. Эндрю А. Искусственный интеллект. – М., 1985.
11. Нильсон Н. Принципы искусственного интеллекта. – М.,1985
12. Miller, G., Wordnet: An online lexical database, International Journal of Lexicography, 3 (4), 1990.
13. Pusteyovsky James. The Generative Lexicon. p. 69-72. MIT, London.
14. Younger D.H. Recognition and parsing of context-free languages in time n3 // Information and Control 10:2, 1967. pp. 189-208.

10. Додаткові ресурси:

https://dl.knu.ua/course/view.php?id=8045
https://drive.google.com/drive/u/0/folders/0B2M_xS1GHaxFdjV5RTBJd3FFRTg