**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**
**ІМЕНІ ТАРАСА ШЕВЧЕНКА**

**ФАКУЛЬТЕТ КОМП'ЮТЕРНИХ НАУК ТА КІБЕРНЕТИКИ**

**Кафедра математичної інформатики**

**«ЗАТВЕРДЖУЮ»**
Заступник декана
з навчальної роботи
_____ Кашпур О.Ф.
«_30_» _08_____ 2019 року

# РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ
# ОБРОБКА ПРИРОДНОЇ МОВИ
# NATURAL LANGUAGE PROCESSING

**для студентів / for students**

| | |
|---|---|
| галузь знань | **12 – Інформаційні технології / Information Technologies** |
| спеціальність | **122 – Комп'ютерні науки / Computer Science** |
| освітній рівень | **магістр / Master** |
| освітня програма | **Штучний інтелект / Artificial Intelligence** |
| вид дисципліни | **вибіркова/elective** |
| Форма навчання | **денна** |
| Навчальний рік | **2019/2020** |
| Семестр 1 | **2** |
| Кількість кредитів ECTS | **4** |
| Мова викладання, навчання та оцінювання | **англійська, українська/ Ukrainian, English** |
| Форма заключного контролю | **екзамен/exam** |

Викладачі: к.ф.-м.н., асистент Тарануха В.Ю. (лекції)

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

на 20__/20__ н.р. _____(_____) «__»___ 20__р.
(підпис, ПІБ, дата)

**КИЇВ – 2019**

Розробник: **Тарануха Володимир Юрійович**, к.ф.-м.н., асистент кафедри «Математичної Інформатики»
**Терещенко Василь Миколайович**, д.ф.-м.н., професор, завідувач кафедри математичної інформатики.

ЗАТВЕРДЖЕНО
Зав. кафедри математичної інформатики
_____ (Терещенко В. М.)
(підпис) (прізвище та ініціали)
Протокол № 10 від « 23 » 05 2019 р.

Схвалено науково-методичною комісією факультету комп'ютерних наук та кібернетики
Протокол від « 30 » серпня 2019 року № 1
Голова науково-методичної комісії _____ (Омельчук Л.Л.)
(підпис) (прізвище та ініціали)
« 30 » серпня 2019 року

## ВСТУП

**1. Мета дисципліни** – поглиблення знань з комп'ютерної лінгвістики, опанування сучасних задач комп'ютерної лінгвістики, сучасних методів та підходів.
/

**The purpose of the course** – to improve knowledge of computational linguistics, mastering modern problems of computational linguistics, modern methods and approaches.

**2. Попередні вимоги до опанування або вибору навчальної дисципліни:**

1. *Знати* дисципліни «Штучний інтелект: принципи та методи», «Розпізнавання образів», «Машинне навчання».
2. *Вміти* проводити наукові дослідження в умовах недостатньої, неповної і/або нечіткої інформації, володіти навичками створення методів на основі ідей та принципів, алгоритмів на основі методів.
3. *Володіти* технічною англійською мовою.

**Preliminary requirements to master or choosing of the course**

*1. Know* the disciplines "Artificial Intelligence: principles and methods", "Pattern Recognition", "Machine Learning".

*2. Be able* to conduct research in conditions of insufficient, incomplete and / or fuzzy information, have the skills to create methods based on ideas and principles, algorithms based on methods.

*3. Knowledge* of technical English is required.

**3. Анотація навчальної дисципліни**:

Навчальна дисципліна «Обробка природної мови/Natural Language Processing» є складовою освітньо-професійної програми підготовки фахівців за другим (*магістерським*) рівнем вищої освіти *галузі знань* 12 «Інформаційні технології» *спеціальності 122* «Комп'ютерні науки», *освітньо-професійної* програми «Штучний інтелект».

Дана дисципліна є вибіркова навчальною дисципліною за *програмою "Штучний інтелект"*. Викладається у 2 семестрі 1 курсу магістратури в обсязі – 4 кредити ETCS год.

У курсі передбачено 2 *змістових модулі*, 1 *модульна контрольна робота, 2 лабораторні роботи*. Завершується дисципліна – екзаменом в 2 семестрі 1 курсу магістратури.
/

**Synopsis of the course:**

The discipline "Natural Language Processing" is a component of the educational-professional training program for the second (*master's*) level of higher education in the field of knowledge 12 "Information Technology" specialty 122 "Computer Science", educational-professional program "Artificial Intelligence" ».

This discipline is an elective course in the program "Artificial Intelligence". It is taught in the 2nd semester of the 1st year of master's degree in the amount of *4 ETCS credits*.

The course provides *2 modules, 1 module test, 2 laboratory work*. The discipline ends with an exam in the 2nd semester of the 1st year of master's degree study.

**4. Завдання (навчальні цілі):**

Набуття знань, умінь та навичок (компетентностей) на рівні новітніх досягнень у комп'ютерній лінгвістиці, відповідно до кваліфікації фахівець з інформаційних технологій. Зокрема, розвивати:

- здатність аналізувати та використовувати інтелектуальні інформаційні технології.;
- здатність до проектування та реалізації систем штучного інтелекту на сучасних обчислювальних системах.

/

**Learning objectives:**

Acquisition of knowledge, skills and abilities (competencies) at the level of the latest advances in computational linguistics, according to the qualification of an information technology specialist. In particular it aims to develop:

• ability to analyze and use intelligent information technologies .;

• ability to design and implement artificial intelligence systems on modern computer systems.

**5. Результати навчання за дисципліною/ Results of learning:**

| **Результат навчання (1. знати; 2. вміти; 3. комунікація; 4. автономність та відповідальність) Results of learning** | | **Форми (та/або методи і технології) викладання і навчання / Forms (and/or methods and technologies) of teaching** | **Методи оцінювання та пороговий критерій оцінювання (за необхідності) / Methods of evaluation and evaluation threshold (if necessary)** | **Відсоток у підсумковій оцінці з дисципліни/ Percentage in final mark** |
|---|---|---|---|---|
| **Код** | **Результат навчання/ Learning results** | | | |
| *РН1.1* | *Знати методи аналізу природномовних текстів та відповідні алгоритми / Know the methods of analysis of natural language texts and appropriate algorithms* | *Лекція, лабораторне заняття / Lecture, Laboratory work* | *Активна робота на лекції, усні відповіді. Тест (60% правильних відповідей) / Activity during lectures, oral answers, Test (60% correct answers)* | *35%* |
| *РН 2.1* | *Вміти аналізувати мовні явища для заданої природної мови. / Be able to analyze linguistic phenomena for a given natural language.* | *Лабораторне заняття, самостійна робота / Laboratory work, Individual work* | *Захист лабораторної роботи / Laboratory works 1,2* | *15%* |
| *РН 2.2* | *Вміти формалізувати нестрогі методи аналізу до чітких алгоритмів та підбирати оптимальний інструментарій для виконання задачі / Be able to formalize loose methods of analysis to create algorithms and select the optimal tools for the task* | | | *20%* |
| *РН 2.3* | *Вміти застосувати наявний* | | | *20%* |

| | | | | |
|---|---|---|---|---|
| | *інструментарій, перш за все програмні бібліотеки з комп'ютерної лінгвістики та машинного навчання.*<br>*/*<br>*Be able to use existing tools, especially software libraries in computer linguistics and machine learning areas* | | | |
| *РН 3.1* | *Обґрунтовувати власний погляд на задачу, спілкуватися з колегами з питань аналізу задач та проектування алгоритму*<br>*/*<br>*Be able to justify own view of the problem, communicate with colleagues in the design and development of programs, prepare written reports* | *Лабораторне заняття/ Laboratory work* | | *5%* |
| РН4.1 | *Відповідально ставитися до виконуваних робіт, нести відповідальність за їх якість*<br>*/*<br>*Responsibly treat the works performed, be responsible for their quality* | *Самостійна робота*<br>*/*<br>*Individual work* | *Поточне оцінювання самостійної роботи*<br>*/*<br>*Accomplishment of tasks assigned to Individual work* | 5% |

**6. Співвідношення результатів навчання дисципліни із програмними результатами навчання**

| **Результати навчання дисципліни**<br>**Програмні результати навчання/**<br>**Teaching results**<br>**Program results of teaching** | **1.1** | **2.1** | **2.2** | **2.3** | **3.1** | **4.1** |
|---|---|---|---|---|---|---|
| **ВПРН18.1.** Знати і застосовувати методи інтелектуального аналізу даних та штучного інтелекту, що включають методи комп'ютерної лінгвістики та комп'ютерного зору.<br>/<br>Know and apply methods of data mining and artificial intelligence, including methods of computational linguistics and computer vision. | + | + | + | + | + | + |

**7. Схема формування оцінки/Mark forming scheme.**

**7.1 Форми оцінювання студентів:**

**- семестрове оцінювання/evaluation in semester:**

*1. Активна робота на лекції, усні відповіді/ Active work on lectures, oral answers*: *РН1.1, РН2.1* – 5 балів/3 бали;

*2. Контрольна робота/Test: РН 1.1, РН 1.2,.— 15 балів/9 балів*

*3. Виконання завдань, винесених на самостійну роботу/ Tasks assigned to independent work*: *РН2.1, РН2.2, РН2.3, РН4.1* – 5 балів/3 бали

*4. Лабораторна робота 1/Laboratory work 1: РН2.1, РН 2.2, РН2.3, РН3.1 – 15 балів/9 балів.*

*5. Лабораторна робота 2/Laboratory work 2: РН2.1, РН 2.2, РН2.3,РН3.1 – 20 балів/12 балів.*

**- підсумкове оцінювання (у формі екзамену)/ final evaluation (exam)***:*

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

*- результати навчання які будуть оцінюватись: РН1.1, РН2.1, РН2.2, РН2.3, РН3.1.*

*- форма проведення і види завдань: письмова.*

*Види завдань:* 4 теоретичних та 2 практичних завдання.

/

- the maximum number of points that can be obtained by a student on the exam: - 40 points;
- learning outcomes that will be evaluated: PH1.1, PH2.1, PH2.2, PH2.3, PH3.1.
- form and types of tasks: written.

Types of tasks: 4 theoretical and 2 practical tasks.

**Загальні вимоги:**

*- Оцінки нижче від мінімального порогового рівня не додаються, що у семестрі, що на іспиті.*

*- Мінімальний пороговий рівень для сумарної оцінки за всіма компонентами становить 60% від максимально можливої кількості балів.*

**General guidelines**

- Scores below the minimum threshold are not added both in semester and during exam.
- The minimum threshold for the total assessment of all components is 60% of the maximum possible number of points.

**Критерії оцінювання на екзамені/exam scoring**

| **Завдання** | **Тема завдання** | **Максимальний відсоток від 40 балів** | **Всього відсотків** |
|---|---|---|---|
| Завдання 1, 2 / Task 1,2 | Письмові запитання зі списку наведеного нижче / Written question from the list **Exam questions** | 25% | 50% |
| Завдання 3 / Task 3 | Для заданного набору речень англійською мовою виконати його розбір за відповіними грамматиками складових та залежностей / For a given set of sentences in English to perform its analysis of the corresponding grammars of | 25% | 25% |

| | | | |
|---|---|---|---|
| | components and dependencies | | |
| Завдання 4 / Task 4 | Для заданного набору речень (української мови) скласти словник невідомих слів, виконати його розбір евристичним аналізом на морфексичні характеристики, та виконати побудову граматичної структури за грамматикою залежностей / For a given set of sentences in Ukrainian language to compile a dictionary of unknown words, perform its analysis by heuristic analysis of morphex characteristics, and perform the construction of a grammatical structure on the grammar of dependencies | 25% | 25% |
| | Всього/Total | | **100%** |

**Умови лабораторних робіт/Laboratory works:**

**Лабораторна робота 1/Learning topic structures from corpora:** Використовуючи відомий студенту метод машинного навчання побудувати модель тематичної структури текстового корпусу. / Using the method of machine learning known to the student to build a model of the thematic structure of the text corpus.

**Лабораторна робота 2/Improving topic structures OR Sequence transformations:**

А**)** На основі тематичної структури створеної в першій лабораторній роботі та використовуючи перетворення послідовностей створити на вибір:

– систему підтримки перекладу,

– чат-бота,

– систему автоматичного реферування.

Б) На основі першої лабораторної роботи побудувати точнішу систему визначення тематичної структури текстового корпусу за рахунок використання раніше не вживаних ресурсів/програмних бібліотек. Помилка $l_1$ не повинна перевищувати 0,5.

/

A) Based on the topic structure created in the first laboratory work and sequence transformation to make one of:

- translation support system,
- chatbot,
- automatic abstracting system.

B) On the basis of the first laboratory work to build a more accurate system for determining the topic structure of the text corpus through the use of previously unused resources / software libraries. Error $l_1$ should not exceed 0.5.

**Запитання для підготовки до екзамену/Exam questions:**

1. Синтаксичні і граматичні класи слів./ Syntactic and grammatical classes of words.
2. Імовірнісні контекстно-вільні граматики та граматики залежностей./ Probabilistic context-free grammars and dependencies grammars.
3. Онтології:Word Net, CYC./ Ontologies: Word Net, CYC.
4. Дискурс. Теорія риторичної структури (RST). Теорія центрування./ Discourse Theory of Rhetorical Structure (RST). The centering theory.

Виділення термінів та сутностей у тексті./ Term and entites detection.

5. Анафора, її типи. Розв’язання анафори. / Anaphora, its types. Anaphora resolution.

6. Визначення значень багатозначних слів на основі WordNet, CYC./ Word sense disambiguation based on WordNet, CYC.
7. Визначення значень багатозначних слів на основі машинного навчання./ Word sense disambiguation based on Machine Learning.
8. Побудова тематичного класифікатора великого розміру./ Construction of a thematic classifier of a large size.
9. Визначення емоційної направленості тексту для коротких текстів (Twitter). /Sentiment analysis for short texts (Twitter).
10. Автоматичне перетворення послідовностей. / Automatic sequence transformation.
11. Автоматичний переклад. /Translation.
12. Системи підтримки діалогу./ Dialogue systems.
13. Побудова онтологій./Ontology building.

**7.2 Організація оцінювання/Evaluation process:**

Обов'язковим є виконання завдань, винесених на самостійну роботу, лабораторних робіт та модульних контрольних робіт за графіком робочої програми./ It is mandatory to perform tasks assigned to independent work, laboratory work and modular tests according to the schedule of the work program.

**Терміни проведення форм оцінювання/Deadlines:**

1. *Контрольна робота (тест): до 3 тижня семестру./ Test: up to the end of 3 weeks of the semester.*
2. *Лабораторна робота 1 (проект): до 5 тижня семестру./ Laboratory work 1 (project): up to the end of 5 weeks of the semester.*
3. *Лабораторна робота 2 (проект): до 7 тижня семестру./ Laboratory work 2 (project): up to the end of 7 weeks of the semester.*

Студент має право на одне перескладання кожної контрольної роботи із можливістю отримання максимально 85% початково визначених за цю контрольну роботу балів. Термін перескладання визначається викладачем.

У випадку відсутності студента з поважних причин відпрацювання та перездачі контрольних робіт здійснюються у відповідності до „Положення про порядок оцінювання знань студентів при кредитно-модульній системі організації навчального процесу" від 1 жовтня 2010 року.

Студент має право здавати лабораторні роботи після закінчення визначеного для них терміну, але з втратою двох балів за кожен тиждень, який пройшов з моменту закінчення терміну її здачі.

/

The student can have second attempt to pass each test with (max. mark is 85% of initial course points for the test). The time of testing is set by lecturer.

Should the student be absent for valid reason re-testing and remedial training are performed according to "Regulations on evaluation of student's knowledge under credit-module learnining process" by 1 October 2010.

The student can submit lab works after deadline but with penalty to earned course points for each week after deadline.

**7.3 Шкала відповідності оцінок**

| | |
|---|---|
| **Відмінно** / Excellent | 90-100 |
| **Добре** / Good | 75-89 |
| **Задовільно** / Satisfactory | 60-74 |
| **Незадовільно** / Failed | 0-59 |

**8. Структура навчальної дисципліни. Тематичний план лекційних занять**

| № семінару | Назва лекції/ Lecture name | Кількість годин | | |
|---|---|---|---|---|
| | | Лекції / Lectures | Лабораторні/ Labs | Самост. робота./ Individual work |
| **Змістовий модуль 1. Внутрішні задачі / Module 1 Internal problems** | | | | |
| 1. | **Тема 1.** Синтаксичні і граматичні класи слів. Імовірнісні контекстно-вільні граматики та граматики залежностей. Онтології:Word Net, CYC. *Самостійна робота:* розібрати текст граматикою складових та граматикою залежностей. **Topic 1.** Syntactic and grammatical classes of words. Probabilistic context-free grammars and dependencies grammars. Ontologies: Word Net, CYC. *Individual work:* parse the text with the grammar of components and the grammar of dependencies. | 2 | 2 | 12 |
| 2. | **Тема 2.** Дискурс. Теорія риторичної структури (RST). Теорія центрування. Визначення значень багатозначних слів на основі WordNet, CYC. *Самостійна робота:* визначити значення багатозначних слів у тексті за допомогою WordNet **Topic 2.** Discourse Theory of Rhetorical Structure (RST). The centering theory. Word sense disambiguation based on WordNet, CYC. *Individual work:* determine the meaning of polysemous words in the text using WordNet | 2 | | 12 |
| 3. | **Тема 3.** Визначення значень багатозначних слів на основі машинного навчання. Виділення термінів та сутностей у тексті. Анафора, її типи. Розв'язання анафори. *Самостійна робота:*для заданного тексту встановити анафори **Topic 3.** Machine learning based word sense disambiguation. Term and entites detection. Anaphora, its types. Anaphora resolution. *Individual work:* set anaphors for a given text | 2 | 1 | 14 |
| 3. | *Підсумкова модульна контрольна робота /Test* | | 1 | |
| **Змістовий модуль 2. Системи для споживачів/ Module 2 End user systems** | | | | |
| 4. | **Тема 4.** Побудова тематичного класифікатора великого розміру. Побудова онтології *Самостійна робота:*спроектувати тематичний класифікатор для заданої області. **Topic 1.** Construction of a thematic classifier of a large size. Ontology building. *Individual work:* design a thematic classification system for a given area. | 2 | 2 | 14 |
| 5. | **Тема 5.** Визначення емоційної направленості тексту для коротких текстів (Twitter). *Самостійна робота:* спроектувати систему визначення емоційної направленості тексту | 2 | 2 | 14 |

| | | | | |
|---|---|---|---|---|
| | **Topic 1.** Sentiment analysis for short texts (Twitter).<br>*Individual work:* design a system for sentiment analysis | | | |
| 6. | **Тема 6.** Автоматичне перетворення послідовностей. Переклад.<br>*Самостійна робота:* спроектувати систему машинного перекладу<br>**Topic 1.** Automatic sequence transformation. Translation.<br>*Individual work:* design a machine translation system | 2 | 2 | 14 |
| 7. | **Тема 7.** Автоматичне перетворення послідовностей: системи підтримки діалогу.<br>*Самостійна робота:* спроектувати систему підтримки діалогу<br>**Topic 1.** Automatic sequence transformation: dialogue systems.<br>*Individual work:* design a dialog system | 2 | | 12 |
| | **ВСЬОГО/TOTAL** | **14** | **10** | **92** |

**Загальний обсяг** 120 *год.,* в тому числі/ **Total duration** *120 hours*, namely:
Лекції/ Lectures – ***14*** год./*h*.
Лабораторні роботи / Laboratory works – ***10*** *год./h.*
Консультації/ Consultations - ***4*** *год./h.*
Самостійна робота/ Individual work **- 92** *год. /h.*

**9. Рекомендовані джерела/Literature:**

***Основна:***


1. Волошин В.Г. Комп'ютерна лінгвістика: Навчальний посібник. – Суми: Університетська книга, 2004. –382 с.
2. Анисимов А.В. Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык Киев: Наук. думка, 1988.- 223 с.
3. Партико З.В. Прикладна і комп'ютерна лінгвістика, Львів, «Афіша», 2008, - 221 с.

*В тому числі й інтернет ресурси*

4. The Oxford handbook of computational linguistics R. Mitkov (Ed) Oxford University Press Ел. ресурс. Режим доступу: http://books.google.com.ua/books/about/The_Oxford_handbook_of_computational_lin.html?id=OaClhre-vW4C&redir_esc=y
5. Daniel Jurafsky and James H. Martin Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition, 2009 Ел. ресурс. Режим доступу: http://rapidlibrary.com/files/speech-and-language-processing-an-introduction-to-natural-language-processing-djv_23726564.html


***Додаткова:***


6. Український правопис / Ін-т мовознавства ім. О.О. Потебні НАН України, Ін-т укр. мови НАН України. — К. : Наук. думка, 2007. — 288 с.
7. The Oxford Handbook of Applied Linguistics / Ed. by R.Kaplan. NY: Oxford university press, 2002, 672 P.
8. J. Allen, "Natural language understanding" Menlo Park, Calif. Benjamin/Cummings 1995, 654 P.

*В тому числі й інтернет ресурси*

9. Сайт проекту WordNet Ел. ресурс. Режим доступу http://wordnet.princeton.edu/
10. Марчук Ю.Н. Компьютерная лингвистика М.: Изд-во Восток-Запад , 2007 г. , 317 с Ел. ресурс. Режим доступу: http://www.twirpx.com/file/398578/

11. Белоногов Г.Г. Компьютерная лингвистика и перспективные информационные технологии М.: Русский мир. 2004г Ел. ресурс. Режим доступу: http://www.twirpx.com/file/134393/
12. Тузов В.А. Компьютерная семантика русского языка Санкт-Петербург: Издательство Санкт-Петербургского университета, 2003 Ел. ресурс. Режим доступу: http://depositfiles.com/files/421m35w47
13. Сайт «Автоматическая обработка текстов» Ел. ресурс. Режим доступу http://aot.ru
14. Український журнал комп'ютерної лінгвістики.Ел. ресурс. Режим доступу http://franko.lviv.ua/ujcl/
15. Спільний сайт Інституту російсько мови ім. В.В. Виноградова та компанії «СЛОВАРИ.РУ» Ел. ресурс. Режим доступу http://www.slovari.ru/