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

Кафедра прикладної математики

# Статистичні методи автоматичного опрацювання текстів

## Робоча програма навчальної дисципліни (Силабус)

| Реквізити навчальної дисципліни | |
|---|---|
| **Рівень вищої освіти** | ***Другий (магістерський)*** |
| **Галузь знань** | *11 Математика та статистика* |
| **Спеціал** | *113 Прикладна математика* |
| **Освітня програма** | *Наука про дані та математичне моделювання* |
| **Статус дисципліни** | *Нормативна* |
| **Форма навчання** | *очна(денна)* |
| **Рік підготовки, семестр** | *2 курс, осінній семестр* |
| **Обсяг дисципліни** | *3 кредити* |
| **Семестровий контроль/ контрольні заходи** | *Екзамен* |
| **Розклад занять** | *Лекція – 1 раз на тиждень, практичні заняття 1 раз на 2 тижні* |
| **Мова викладання** | *Українська* |
| **Інформація про керівника курсу / викладачів** | Лектор: *канд. техн. наук, доцент, Сирота Сергій Вікторович, syrota.sergiy@lll.kpi.ua*<br>Практичні / Семінарські: канд. техн. наук, доцент, Сирота Сергій Вікторович, syrota.sergiy@lll.kpi.ua |
| **Розміщення курсу** | Дистанційний ресурс Moodle, http://moodle.eec.kpi.ua/course/view.php?id=5 |

## Програма навчальної дисципліни

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

*Вивчення навчальної дисципліни «Статистичні методи автоматичного опрацювання текстів» дає більш глибокі знання в царині сучасних технологій автоматичного аналізу текстів природної і для вирішення інших задач професійної діяльності визначених Стандартом вищої освіти НТУУ «КПІ» для другого (магістерського) рівня вищої освіти за спеціальністю 113 Прикладна математика.* ***Метою*** *курсу є вивчання статистичних методів, що застосовуються доля автоматичного аналізу текстів.* ***Предмет*** *навчальної дисципліни — застосування статистичних методів до автоматичного аналізу текстів.*

*В* ***результаті навчання*** *студенти отримають наступні компетенції*

- *Здатність учитися і оволодівати сучасними знаннями.*
- *Здатність застосовувати знання у практичних ситуаціях.*
- *Здатність генерувати нові ідеї (креативність).*
- *Здатність до абстрактного мислення, аналізу та синтезу.*
- *Здатність до пошуку, оброблення та аналізу інформації з різних джерел.*
- *Навички у використанні інформаційних і комунікаційних технологій.*
- *Здатність сформулювати математичну постановку задачі, спираючись на постановку мовою предметної галузі, та обирати метод її розв'язання, що забезпечує потрібні точність і надійність результату.*

*знання:*

- *основних положень теорії прикладної лінгвістики;*
- *основних методів видобутку та аналізу текстових даних для виявлення закономірностей;*
- *статистичних підходів, які можна загалом застосувати до довільних текстових даних будь-якої природної мови;*

*уміння:*

- *Використовувати статистичні методи аналізу текстів;*
- *Видобутку та аналізу текстових даних для виявлення закономірностей;*

*досвід:*

- *Видобування інформації з текстових корпусів;*
- *Використання спеціального програмного забезпечення для аналізу текстів*

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

*Вивченню дисципліни «Психологічні моделі прийняття управлінських рішень» повинне передувати вивчення дисциплін:*

- *«Математичний аналіз»,*
- *«Алгебра та геометрія»,*
- *«Дискретна математика»,*
- *«Теорія ймовірності»,*
- *«Математична статистика»,*
- *«Методи штучного інтелекту».*

**3. Зміст навчальної дисципліни**

Тема1 Огляд методів обробки природних мов
- Маркування частин мови (part-of-speech tagging)
- Синтаксичний аналіз
- Семантичний аналіз
- Двозначність
- Представлення текстів, представлення у вигляді набору слів
- Контекст слова; контекстна схожість.
- Синтагматичні відношення
- Парадигматичні відношення

Тема 2. Дослідження асоціацій слів
- Ентропія
- Умовна ентропія
- Спільна кількість інформації
- Тема і охоплення тема
- Термін, як тема

Тема 3. Тематичний аналіз текстів
- Змішана модель
- Компонентна модель
- Обмеження на ймовірності
- Імовірнісний Латентний Семантичний Аналіз (PLSA)
- Алгоритм максимізації очікування (EM)
- E-крок і M-крок
- Приховані змінні
- Алгоритм сходження на вершину
- Локальний максимум

- Латентний розподіл Дирихле (LDA)

Для зарахування лабораторного практикуму з дисципліни "Автоматизоване опрацювання текстів" слухачам необхідно:

- Обрати один із запропонованих інструментальних засобів автоматизованого опрацювання текстів.
- Ознайомитися з функціональним призначенням засобу і його можливостями.
- Встановити програмне забезпечення (програму/бібліотеку/toolkit/framework) на комп'ютері.
- За допомогою встановленого інструментального засобу провести аналіз довільно обраного тексту (текстів).
- Зробити висновки, написати відгук, викласти свої враження від використання данного засобу.
- Підготувати звіт обсягом 2-3 тис знаків за результатами виконаної роботи

## 4. Навчальні матеріали та ресурси

*Базова література*


1. *Cheng X. Z. Text Mining and Analytics [Електронний ресурс] / Cheng Xiang Zhai // The University of Illinois at Urbana-Champaign,. – 2021. – Режим доступу до ресурсу: https://www.coursera.org/learn/text-mining/supplement/ctYg7/welcome-to-text-mining-and-analytics.*
2. *Aggarwal C. C., Zhai C. Mining Text Data. — Springer, 2012. — 527 c. — ISBN 9781461432234.*
3. *Сирота С. В. Автоматизоване опрацювання текстів [Електронний ресурс] / С. В. Сирота, // ЦЕО КПІ. – 2021. – Режим доступу до ресурсу: http://moodle.eec.kpi.ua/course/view.php?id=10*


*Допоміжна література*


1. *Ananiadou S. Text mining for biology and biomedicine / S. Ananiadou, J. McNaught. – Boston, London: Hartec House, 2006.*
2. *Konchady M. Text mining application programming / Manu Konchady. – Boston, Mass.: Charles River Media, 2008.*


*Інформаційні ресурси*

*1. Центр електронної освіти КПІ http://moodle.eec.kpi.ua/course/view.php?id=10*

*2. Електронний кампус КПІ ім. Ігоря Сікорського. Матеріали з дисципліни «Алгоритми і структури даних». – Режим доступу : http://login.kpi.ua*

**Навчальний контент**

## 5. Методика опанування навчальної дисципліни (освітнього компонента)

*Далі в таблиці інформація за темами про лекції у формі календарного плану.*

| | *Лекційні заняття* |
|---|---|
| *Тиждень 1-2* | Тема1 Огляд методів обробки природних мов<br>• Маркування частин мови (part-of-speech tagging)<br>• СинтаксичнийаналізСторінка<br>• СемантичнийаналізСторінка<br>• Двозначність<br>• |
| *Тиждень 3-4* | • Представлення текстів, представлення у вигляді набору слів<br>• Контекст слова; контекстна схожість.<br>• Синтагматичні відношення<br>• Парадигматичні відношення |
| *Тиждень 5-6* | Тема 2. Дослідження асоціацій слів |

| | |
|---|---|
| | • Ентропія<br>• Умовна ентропія |
| *Тиждень 7-8* | • Спільна кількість інформації<br>• Тема і охоплення тема<br>• Термін, як як тема |
| *Тиждень 9-10* | Тема 3. Тематичний аналіз текстів<br>• Змішана модель<br>• Компонентна модель<br>• Обмеження на ймовірності |
| *Тиждень 11* | • Імовірнісний Латентний Семантичний Аналіз (PLSA)<br>• Алгоритм максимізації очікування (EM)<br>• Е-крок і М-крок<br>• Приховані змінні |
| *Тиждень 12* | • Алгоритм сходження на вершину<br>• Локальний максимум<br>• Латентний розподіл Дирихле (LDA) |
| *Тиждень 13-18* | Здача результатів лабораторного практикуму підготовка та здача екзамену |

### *6.* Самостійна робота студента

***Для засвоєння матеріалів курсу передбачені наступні види самостійної роботи***

- *Робота з рекомендованими джерелами і*
- *Виконання завдання лабораторного практикуму*

## Політика та контроль

### 7. Політика навчальної дисципліни (освітнього компонента)

*Відвідування*

*Відвідування лекцій не обов'язкове студентам надається можливість самостійно опрацьовувати теми занять використовуючи дистанційний курс та рекомендовані джерела.*

*Поточний контроль проходить в режимі тестування і має часові обмеження.*

*Вимоги до слухачів курсу базуються на принципах академічної доброчесності і рівності всіх студентів. У випадку виявлення випадків запозичення без відповідних посилань об'єктів авторського права, як то: програмний код, зображення, креслення, та інший мультимедійний контент або виявлення плагіату — бали за відповідні роботи будуть анульовані і нараховані штрафні бали. Повторні порушення принципів академічної доброчесності можуть призвести до недопуску до складання заліку.*

*Викладачі можуть перевіряти роботи, виконані у рамках курсу, за допомогою систем виявлення плагіату.*

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

*На семестр пропонуються наступні види діяльності за які нараховуються бали*

- *Доповідь зі звітом про виконаний лабораторний практикум 40 балів*
- *Екзамен 60 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

- *В разі проходження слухачами курсу:*
- *https://www.coursera.org/learn/text-mining/supplement/ctYq7/welcome-to-text-mining-and-analytics і наданні сертифікату студенту зараховується .60 балів*
- *Дистанційні онлайн заняття проводяться за допомогою Google Meet за запрошеннями, які публікуються в кафедральній платформі Slack та дублюються на надані студентами адреси електронної пошти;*
- *Офіційні звернення до викладача розглядаються через кафедральну платформу Slack або електронну адресу* syrota.sergiy@lll.kpi.ua.

**Робочу програму навчальної дисципліни (силабус):**

**Склав** доцент кафедри прикладної математики, канд. техн. наук, доц. Сирота Сергій Вікторович

**Ухвалено** кафедрою прикладної математики (протокол № 13 від 16.06.2022)

**Погоджено** Методичною комісією факультету[1] (протокол № 11 від 27.06.2022)

[1] Методичною радою університету – для загальноуніверситетських дисциплін.