УДК 004.8:[37.016:81-028.31
DOI: 10.31866/2617-796X.7.1.2024.307009

Костянтин Ткаченко,
кандидат економічних наук, доцент,
доцент кафедри програмного забезпечення комп'ютерних систем,
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»,
Київ, Україна
tkachenko.kostyantyn@gmail.com
https://orcid.org/0000-0003-0549-3396

ВИКОРИСТАННЯ МЕТОДІВ NLP В ІНТЕЛЕКТУАЛЬНИХ НАВЧАЛЬНИХ СИСТЕМАХ

Для ефективної організації освітніх процесів, які підтримують відповідні інтелектуальні навчальні системи, важливо обрати правильні технології, що забезпечували б індивідуалізацію навчання, адекватне сприйняття навчального контенту, так зване «розуміння» системами текстів українською мовою, які надають студенти (опис рішення завдання, відповіді, що надається власними словами, а не обирається з варіантів відповіді тесту, питань до системи тощо), створення прототипів, постійну ітерацію під час розпізнавання та обробки текстів природною мовою, максимальну надійність та ефективність процесів навчання.

Метою статті є дослідження, аналіз різних методів оброблення текстів природною мовою, концепції NLP, розгляд загальних проблем і перспектив розроблення на її основі програмного продукту оброблення українськомовного тексту в онлайн-курсах, які підтримують інтелектуальні навчальні системи.

Методами дослідження є основні методологічні підходи та технологічні засоби для аналізу текстів природною мовою в інтелектуальних навчальних системах, розроблення системи підтримки технології NLP (Natural Language Processing, оброблення природної мови) під час лінгвістичного аналізу текстів українською мовою. Такими методами, зокрема, є: системний та порівняльний аналізи – для виявлення особливостей інтелектуальних та інформаційних (з елементами інтелектуалізації) систем; метод експертних оцінок, що передбачає аналіз літературних джерел й інформаційних ресурсів, проведення інтерв'ю та опитування експертів, а також процеси розробки та тестування інтелектуальних й інформаційних систем.

Новизною проведеного дослідження є аналіз сучасних технологій розробки систем підтримки освітнього онлайн-процесу через організацію процесів сприйняття інформації, наданої студентами природною мовою, результати якого можуть застосовуватися під час розробки власного програмного продукту підтримки освітнього процесу українською мовою, забезпечення підвищення ефективності навчання на основі використання технології NLP у процесі вивчення відповідного навчального контенту.

Висновки. У роботі проаналізовано сучасні методи NLP. Проведений аналіз обумовив вибір методів токенізації, нормалізації, стемінгу та лематизації для використання в інтелектуальних навчальних системах під час лінгвістичного аналізу так званого «вільного» спілкування природною (українською) мовою студентів у процесі вивчення навчального контенту онлайн-курсів.

Під час токенізації українськомовних текстів вирішували такі проблеми, як усунення так званих «злитих» токенів, виправлення орфографічних помилок, визначення спільних префіксів у складних словах та їх впливу на семантику відповідних лексем, визначення спільних префіксів в абревіатурах, приведення слів до їхньої нормальної форми.

Лематизація особливо важлива для української мови (з її великою кількістю відмінків іменників, прикметників, словоформ тощо), потребує використання спеціально сформованих словників предметної галузі, що розглядається. У цих словниках словоформи представлені у вигляді лем (тобто іменники подано в називному відмінку).

Ключові слова: NLP (Natural Language Processing); інтелектуальна навчальна система; онлайн-курс; токенізація; стемінг; нормалізація; стоп-слова; сегментація тексту.

Вступ. Сучасні інформаційні технології та технології з використанням штучного інтелекту інтегруються в різні сфери життєдіяльності як окремих людей, так і суспільства в цілому. Однією з таких сфер є сучасна освіта (Tkachenko et al., 2024).

Саме тому інтеграція сучасних інформаційних технологій і технологій роботи з текстовою інформацією (як у вигляді навчального контенту онлайн-курсів, так і у вигляді відповідей студентів, що надаються природною мовою) обумовлює перехід до іншого (більш інтелектуального) підходу щодо організації процесів навчання, підвищення рівня індивідуалізації навчального контенту, який надається студентам, розуміння системою відповідей (запитів) студентів (Tkachenko et al., 2024).

На сьогодні технології NLP (*Natural Language Processing*, оброблення природної мови NLP) (Eisenstein, 2018; Natural Language Processing, 2023) завдяки використанню штучного інтелекту (ChatGPT, n.d.; Responsible AI, n.d.; DALL·E 2, n.d.; Rytr, n.d.), нейронних мереж (Tkachenko et al., 2024; Sutskever, Vinyals and Le, 2014), машинного навчання (Pitis, 2023; Bagui et al., 2021) усе більше проникають у процеси навчання, розширюючи його можливості через індивідуалізацію навчання, покращення взаємодії між користувачами (студентами, викладачами, авторами навчального контенту, методистами, представниками керівництва закладу вищої освіти (ЗВО), адміністрацією факультету тощо) відповідних онлайн-курсів та відповідною інтелектуальною навчальною системою. Така система може бути платформою для певних онлайн-курсів і здійснювати процеси організації, управління, контролю та моніторингу освітніх процесів у ЗВО.

Слід зауважити, що в інтелектуальних навчальних системах, які здійснюють NLP, виникають проблеми, пов'язані з неповними та/або помилковими даними, що може призводити, зокрема, до:

– побудови моделей навчального контенту, використання яких у разі неадекватної оцінки рівня початкових компетенцій студентів призводить до формування неефективних, а інколи взагалі непотрібних студентові траєкторій навчання;

– побудови моделей спілкування студентів із системою, використання яких у разі неадекватного розуміння відповідей / запитів з боку студентів призводить до неправильного вибору варіантів надання індивідуалізованого навчального контенту та/або правильного оцінювання відповідей / запитів студентів.

Штучний інтелект (ШІ) сприяє розв'язанню багатьох класів задач, пов'язаних з обробкою природної мови (NLP, *Natural Language Processing*) (Eisenstein, 2018;

Natural Language Processing, 2023). NLP передбачає використання різних методів, моделей і систем.

NLP використовують у багатьох технологічних рішеннях розпізнавання та оброблення природних мов в інтелектуальних навчальних системах. Саме тому звичайний онлайн-курс, у якому опитування студентів здійснюється за допомогою жорстко заданих опцій відповідей, поступово замінюється навчанням, яке підтримує так зване «вільне» висловлювання думки студентів, що описують свій варіант розв'язання задачі чи відповідають на запитання.

В інтелектуальних навчальних системах користувачі (студенти, викладачі, методисти, автори навчального контенту тощо) починають усе частіше спілкуватися без участі так званого «живого» посередника. Тому актуальність проблем розпізнавання текстів, наданих природною мовою, зокрема українською, не викликає сумнівів.

NLTK (*Natural Language Toolkit,* 2023) – платформа для створення NLP-програм мовою Python, яка надає зручні інтерфейси для багатьох мовних корпусів, бібліотеки для обробки текстів (класифікації, токенізації, стемінгу, розмітки, фільтрації та семантичних розмірковувань (висловлювань)).

Методи технології NLP уже широко використовуються в чатботах, але слід зауважити, що ці методи в основному стосуються текстів англійською мовою (Wu et al., 2016; Chen et al., 2018). Проблема розроблення аналогічних методів та алгоритмів NLP щодо текстів українською мовою є актуальною та потребує свого вирішення.

Аналіз і постановка проблеми. Різні предметні області формують додаткові вимоги до алгоритмів NLP, що обумовлено, зокрема, використанням у цих областях власних специфічних термінів, словоформ тощо.

Крім того, багато природних мов не мають повноцінної підтримки в NLP. Наприклад, такою мовою є українська, бо, на відміну від англійської мови, для якої розроблено багато методів NLP, які реалізовані в різноманітних програмних продуктах, оброблення текстів українською мовою ще не має такого розвинутого арсеналу програмної та алгоритмічної підтримки.

Підхід до оброблення текстів природною мовою та представлення навчального контенту онлайн-курсу, який підтримується відповідною інтелектуальною навчальною системою, обумовлює створення можливостей «вільного» спілкування користувачів системи під час навчання, які, зокрема, сприяють:

– підвищенню ефективності, індивідуалізації та якості процесів навчання;

– підвищенню рівня мотивації студентів до навчання;

– збільшенню обсягів засвоєного навчального контенту відповідного онлайн-курсу (окремої навчальної теми);

– здійсненню самоперевірки здобутих знань тощо.

Обробка тестових даних українською мовою пов'язана з вирішенням багатьох проблем, наявність яких обумовлена, зокрема, тим, що тексти природною українською мовою:

– містять термінологію (пов'язану, зокрема, з інформаційними технологіями, штучним інтелектом, програмним забезпеченням, моделюванням проєктуванням та розробкою програмного забезпечення);

– через специфіку предметної області мають проблеми щодо правильного розуміння тестових висловлювань студентів (через так звану «неохайність» у висловлюваннях, часте використання сленгових виразів тощо);

– присутність специфічних термінів, запозичених з інших природних мов, наприклад, так званих англіцизмів).

Класична попередня NLP текстів має охоплювати:

– визначення мови тексту;

– розпізнавання конкретних символів (латиниці, кирилиці, математики, інших спеціальних символів інших мов та предметних областей);

– розкладання складних слів на прості складники слова – слова-елементи;

– обробку абревіатур.

Мета і завдання дослідження. У роботі досліджено методи оброблення текстів українською мовою, які застосовуються в інтелектуальних навчальних системах у процесі навчання студентів спеціальності напряму 12 «Інформаційні технології».

Основною метою статті є дослідження, аналіз різних методів оброблення текстів природною мовою та концепції NLP, розгляд загальних проблем і перспектив розроблення на її основі програмного продукту обробки українськомовного тексту в онлайн-курсах, які підтримують інтелектуальні навчальні системи.

Завданнями дослідження, зокрема, є:

– аналіз сучасних методів попереднього оброблення текстової інформації;

– розгляд використання сучасних методів NLP в онлайн-курсах спеціальностей напряму 12 «Інформаційні технології».

Для ефективного використання природної української мови в інтелектуальних навчальних системах потрібно здійснювати, зокрема:

– комплексне очищення даних (від помилок під час написання слів, зайвих пробілів, повторів тощо);

– попереднє оброблення тексту (видалення стоп-слів, обробку абревіатур, пошук синонімів тощо);

– токенізацію (за реченнями, за словами тощо);

– лематизацію та стемінг (Saumyab271, 2022).

Результати дослідження. NLP для тексту (навчального контенту) і тексту, яким обмінюється інтелектуальна навчальна система під час організації та проведення навчання, передбачає, зокрема, використання (Tkachenko et al., 2024; Eisenstein, 2018; Awan, 2023; Mashtalir and Nikolenko, 2023):

– токенізації за реченнями;

– токенізації за словами;

– лематизації та стемінгу тексту (Saumyab271, 2022);

– стоп-слів (Мосейчук, 2013);

– регулярних виразів (Ashraf, 2023);

– «мішку слів»;

– TF-IDF.

Токенізація за реченнями (сегментація тексту за реченнями) – процес поділу тексту на фрагменти, якими є окремі речення-компоненти (Awan, 2023). У багатьох мовах поділ на речення (виділення, виокремлення речень) відбувається

кожного разу, коли згідно з відповідним алгоритмом система, що підтримує NLP, знаходить певний знак пунктуації (це може бути один з таких знаків, як крапка, знак оклику, знак питання).

Ця проблема не є простою, бо крапка може свідчити не тільки про закінчення речення, а використовуватися в скороченнях (наприклад, «т. п.»).

Щоб здійснити токенізацію за реченнями доцільно використовувати таблиці сталих скорочень, які використовуються в тій чи тій природній мові. Такі таблиці під час оброблення тексту сприяють запобіганню неправильної розстановки меж між реченнями.

На рис. 1 продемонстровано результат токенізації за реченнями, отриманої під час NLP тексту відповіді, що надав студент у процесі вивчення навчального контенту онлайн-курсу «Основи штучного інтелекту», який підтримує відповідна інтелектуальна навчальна система.

Текстовий фрагмент
Штучний інтелект надає унікальні можливості, які ще кілька років тому здавалися фантастикою. Алгоритми можуть розпізнавати голос, ідентифікувати людину, виявляти помилки, давати рекомендації, розшифровувати емоції тощо. Штучний інтелект успішно виконує різні монотонні операції та обробляє великі обсяги даних.
Токенізація за реченнями
Штучний інтелект надає унікальні можливості, які ще кілька років тому здавалися фантастикою. Алгоритми можуть розпізнавати голос, ідентифікувати людину, виявляти помилки, давати рекомендації, розшифровувати емоції тощо. Штучний інтелект успішно виконує різні монотонні операції та обробляє великі обсяги даних.

Рис. 1. Токенізації за реченнями відповіді студента в онлайн-курсі «Основи штучного інтелекту»

Токенізація за словами (сегментація тексту за словами) – процес поділу речень на слова-компоненти.

У багатьох мовах, у тому числі й українській, пробіл відіграє роль роздільника слів. Але якщо використовувати лише пробіл у ролі роздільника, між словами можуть виникнути проблеми, бо, наприклад, таку роль можуть відігравати й двокрапка, кома та крапка з комою.

На рис. 2 продемонстровано результат токенізації за словами під час вивчення навчального контенту онлайн-курсу «Основи штучного інтелекту».

Лематизація та стемінг тексту. Тексти природною мовою можуть містити різні граматичні форми одного й того ж слова і навіть можуть траплятися однокореневі слова (Saumyab271, 2022).

Текстовий фрагмент
Штучний інтелект надає унікальні можливості, які ще кілька років тому здавалися фантастикою. Алгоритми можуть розпізнавати голос, ідентифікувати людину, виявляти помилки, давати рекомендації, розшифровувати емоції тощо. Штучний інтелект успішно виконує різні монотонні операції та обробляє великі обсяги даних.
Токенізація за словами
[‘Штучний’, ‘інтелект’, ‘надає’, ‘унікальні’, ‘можливості’, ‘,’, ‘які’, ‘ще’, ‘кілька‘, ‘років‘, ‘тому’, ‘здавалися’ ,‘фантастикою’, ‘. ‘] [‘Алгоритми’, ‘можуть’, ‘розпізнавати’, ‘голос’, ‘,’, ‘ідентифікувати’, ‘людину’, ‘,’, ‘виявляти’, ‘помилки’, ‘,’, ‘давати’, ‘рекомендації’, ‘,’, ‘розшифровувати’, ‘емоції’, ‘тощо’, ‘.’] [‘Штучний’, ‘інтелект’, ‘успішно’, ‘виконує’, ‘різні’, ‘монотонні’, ‘операції’, ‘та’, ‘обробляє’, ‘великі’, ‘обсяги’, ‘даних’, ‘.’]

Рис. 2. Токенізації за словами відповіді студента в онлайн-курсі «Основи штучного інтелекту»

Лематизація і стемінг полягають у приведенні всіх словоформ, що трапляються в тексті, до однієї нормальної словникової форми (тобто до словоформи, яка представлена в словнику відповідної природної мови).

Наприклад, приведення різних словоформ до однієї може бути таким:

інтелект, інтелекту, інтелекти, інтелектом => інтелект

У такий спосіб відбувається стемінг відповідно до цілого речення.

Лематизація та стемінг є різними підходами до нормалізації тексту. Стемінг являє собою процес, під час якого відбувається видалення «зайвого» від кореня того чи того слова.

Стемінг інколи призводить до втрати суфіксів, за допомогою яких можуть утворитися нові слова (з більш уточненим значенням щодо так званого «основного» слова чи взагалі семантично іншим значенням). Лематизація ж передбачає використання словників відповідної природної мови та морфологічного аналізу, щоб у результаті можна було б відтворити слово в його канонічній формі – лемі.

Лематизація та стемінг відрізняються тим, що стемер (реалізація алгоритму стемінгу) діє без знання / використання контексту відповідного слова і тому не розуміє різницю між словами, які мають різний зміст залежно від частини мови. Але очевидною перевагою стемерів є простота їхнього використання та швидкодія.

Слово «добре» – це лема для слова «краще». Стемер не бачить цей зв’язок між словами; щоб його встановити, слід сформувати відповідний словник і під час NLP здійснювати для всіх слів перевірку на відповідність, звертаючись до такого словника (це може потребувати часу для обробки тексту природною мовою).

На відміну від стемінгу, лематизація обирає / формує правильну лему, використовуючи контекст (водночас можуть використовуватися такі формалізми, як контекстні граматики (ліво- та правосторонні)).

Слово «гра» – це базова форма слова «грати», а слово «програма» – це базова форма слова «програмувати».

Під час аналізу тексту, що використовує це слово, доцільним є використання і стемінгу, і лематизації.

Стоп-слова (шумові слова) – це слова, які не мають смислового навантаження, тому їх користь і роль для пошуку не суттєва.

Стоп-словами української мови, наприклад, є (Мосейчук, 2013):

[‘або’, ‘адже’, ‘але’, ‘багато’, ‘без’, ‘більш’, ‘більше’, ‘буде’, ‘би’, ‘був’, ‘була’, ‘були’, ‘було’, ‘бути’, ‘в’, ‘вам’, ‘вас’, ‘від’, ‘він’, ‘вона’, ‘вони’, ‘воно’, ‘втім’, ‘ви’, ‘давай’, ‘давати’, ‘де’, ‘для’, ‘до’, ‘досить’, ‘дуже’, ‘же’, ‘з’, ‘за’, ‘за винятком’, ‘завжди’, ‘замість’, ‘знову’, ‘зовсім’, ‘інший’, ‘його’, ‘йому’, ‘іноді’, ‘її’, ‘їй’, ‘є’, ‘їм’, ‘їх’, ‘коли’, ‘крім’, ‘куди’, ‘ледве’, ‘майже’, ‘мати’, ‘мене’, ‘мені’, ‘між’, ‘може’, ‘моє’, ‘мої’, ‘мій’, ‘ми’, ‘на’, ‘навколо’, ‘навіть’, ‘назавжди’, ‘над’, ‘нарешті’, ‘нас’, ‘начебто’, ‘наш’, ‘не’, ‘ні’, ‘небудь’, ‘ніколи’, ‘нічого’, ‘ну’, ‘однак’, ‘от’, ‘отут’, ‘перед’, ‘по’, ‘про’, ‘поза’, ‘під’, ‘після’, ‘потім’, ‘при’, ‘про’, ‘раз’, ‘раптом’, ‘свою’, ‘себе’, ‘сказати’, ‘так’, ‘також’, ‘такі’, ‘такий’, ‘там’, ‘тебе’, ‘теж’, ‘тепер’, ‘те’, ‘ті’, ‘тієї’, ‘тільки’, ‘ти’, ‘того’, ‘тоді’, ‘той’, ‘тому’, ‘тому що’, ‘треба’, ‘тут’, ‘увесь’, ‘уздовж’, ‘уже’, ‘униз’, ‘унизу’, ‘усередині’, ‘усі’, ‘усього’, ‘усіх’, ‘усю’, ‘хіба’, ‘хоч’, ‘хоча’, ‘хто’, ‘через’, ‘чи’, ‘чий’, ‘чиє’, ‘чия’, ‘чого’, ‘чогось’, ‘чому’, ‘ще’, ‘що’, ‘щоб’, ‘щось’, ‘ця, ‘ці’, ‘це’, ‘цю’, ‘цього’, ‘цьому’, ‘цей’, ‘якось’, ‘якщо’]

Слід розуміти, що немає універсального списку стоп-слів. Такий список залежить від мови, задачі, конкретної ситуації (текстового пояснення) тощо.

На рис. 3 наведено результат обробки тексту з метою видалення стоп-слів під час вивчення навчального контенту онлайн-курсу «Основи штучного інтелекту».

Регулярний вираз – послідовність символів, що визначає шаблон пошуку (Ashraf, 2023). Наприклад:

– • – будь-який символ (за винятком символу переходу до наступного рядка):
– \w – один символ;
– \d – одна цифра;
– \s – один пробіл;
– \W – один НЕ символ;
– \D – одна НЕ цифра;
– \S – один НЕ пробіл;
– [abc] – знаходить будь-який з указаних символів (a, b, c);
– [^abc] – знаходить будь-який символ, крім указаних (a, b, c);
– [a-i] – знаходить символ у проміжку від a до i.

Оцінка (скоринг) слів полягає, зокрема:

– в оцінці наявності слів у тексті (1 – слово є, 0 – слова немає);
– обчисленні кількості слів у текстовому документі;
– обчисленні частоти наявності слова в тексті (щодо загальної кількості слів тексту).

Текстовий фрагмент
Штучний інтелект надає унікальні можливості, які ще кілька років тому здавалися фантастикою. Алгоритми можуть розпізнавати голос, ідентифікувати людину, виявляти помилки, давати рекомендації, розшифровувати емоції тощо. Штучний інтелект успішно виконує різні монотонні операції та обробляє великі обсяги даних.
Видалення стоп-слів
[‘Штучний’, ‘інтелект’, ‘надає’, ‘ унікальні’, ‘можливості’, ‘,’, ‘ які’, ‘здавалися’, ‘фантастикою’, ‘.’] [‘Алгоритми’, ‘можуть’, ‘розпізнавати’, ‘голос’, ‘ ,’, ‘ідентифікувати’, ‘людину’, ‘ ,’, ‘виявляти’, ‘помилки’, ‘ ,’, ‘давати’, ‘рекомендації’, ‘ ,’, ‘розшифровувати’, ‘емоції’, ‘.’] [‘Штучний’, ‘інтелект’, ‘виконує’, ‘монотонні’, ‘операції’, ‘обробляє’, ‘великі’, ‘обсяги’, ‘даних’, ‘.’]

Рис. 3. Видалення стоп-слів у фрагменті навчального контенту в онлайн-курсі «Основи штучного інтелекту»

Машинне навчання починає працювати з даними, що відображають текст, наданий природною мовою, здійснивши перед цим вилучення ознак – конвертацію тексту у відповідні набори цифр (вектори). При такому підході описується входження кожного слова в текст, що аналізується. Щоб здійснити NLP треба:

– визначити словник відомих слів (токенів);

– визначити рівень (ступінь) наявності (присутності) відомих слів (використовуючи, наприклад, один з найпростіших методів скорингу (Awan, 2023), який полягає в тому, що відмічається наявність слів: 1 ставиться, якщо слово в тексті є, а 0 – у разі його відсутності);

– ігнорувати відомості щодо порядку та/або структури слів;

– ігнорувати регістр (що використовується під час написання слів тексту), пунктуацію та токени, які мають лише один символ;

– ігнорувати стоп-слова;

– здійснити ембединг – представлення слова N-мірним вектором дійсних чисел (Ghannay et al., 2016) (ембединг у такому разі можна вважати інтерфейсом між текстом і відповідною нейромережею);

– створити вектори текстового документа;

– здійснити лематизацію та стемінг, приводячи слова тексту до їхніх нормальних форм;

– виправити неправильно написані слова (наприклад, здійснивши синтаксичний аналіз тексту).

При зростанні розміру словника зростає розмір вектора, що відображає текст природною мовою (хоча в такому векторі значна частина елементів становить 0).

Розвиток NLP почався тоді, коли почалося використання глибоких нейронних мереж для обробки словоформ і речень тексту, що задається природною мовою. Серед методів NLP найбільш важливими є нормалізація, фрагментація та токенізація тексту. За допомогою штучного інтелекту під час оброблення тексту вирішуються проблеми, що пов'язані, зокрема:

– зі складністю тексту (зокрема, наповненням тексту термінами (наприклад, з невідомих користувачеві предметних областей), тексту, наданого іноземною мовою, тощо);

– з наповненням так званою «водою» (наявністю неважливої та/чи зайвої інформації);

– з граматичними й лексичними помилками.

Під час застосування NLP виникають проблеми:

– спрощення (усунення надмірності та складності тексту);

– нормалізації тексту (наприклад, через утворення абревіатур, заміну слів на більш загальні форми тощо).

Найчастіше для вирішення проблем NLP використовують рекурентні нейронні мережі (RNN) (Tarwani and Edem, 2017). RNN у процесі аналізу тексту (зокрема, відповідей, що надають студенти під час вивчення навчального контенту онлайн-курсів, які підтримують інтелектуальні навчальні системи) обробляють вхідні дані послідовно та зберігають контекст, який використовується.

Різні модифікації RNN, такі як LSTM (*long short-temp memory*) (Cheng, Dong and Lapata, 2016) та GRU (*gated recurrent units*) (Chung et al., 2014), покращували блок «пам'яті» (Martinez, 2023).

Серед програмних продуктів, які вирішують проблеми NLP та використовують при цьому штучний інтелект, слід насамперед виділити такі, як:

– Google-перекладач, який у своїй бібліотеці має більш ніж 100 мов (Перекладач, б.д.; Vivien, 2022);

– ChatGPT – чатбот, який симулює відповіді на запитання, може щось порадити, написати програмний код (зокрема, для задач, які вже є в нього у відповідній бібліотеці), симулювати деякі інформаційні об'єднання (наприклад, здійснити симулювання бази даних) (ChatGPT, n.d.; Ramponi, 2022; Shpater, 2024);

– DALL-E на основі відповідного контексту генерує зображення (DALL·E 2, n.d.; O'Connor, 2023);

– Rytr генерує тексти за темою, жанром тощо (Rytr, n.d.).

Студенти часто використовують ChatGPT як власного асистента, бо він добре справляється з основними навчальними задачами програмування, генеруючи код багатьма мовами програмування та виправляючи помилки в коді.

Попереднє оброблення тексту є основною частиною будь-якої системи, функціонал якої забезпечує вирішення проблем NLP. Є багато методів оптимізації тексту, його фрагментації, нормалізації та токенізації, а також знаходження зв'язків між токенами (Awan, 2023).

Токенізація є важливим етапом в NLP, що являє собою процес сегментування тексту в морфеми, слова та фрази за визначеними правилами відповідно до порушеної проблеми. Нормалізація тексту також стосується цього кроку, що передбачає уніфікацію лексем. Наприклад, «р.», «рік», «року» та «Рік» зводяться

до однієї форми. Водночас ураховується, що слова української мови мають такі частини, як префікси, корені, суфікси й закінчення.

Одним з методів представлення токенів у векторному просторі є використання так званого one-hot вектора, розмірність якого дорівнює розмірності відповідного словника (природної мови, професійного для окремої предметної області, мови програмування тощо) та складається з нулів, а токен кодується з урахуванням відповідної позиції в словнику, що дорівнюватиме єдиній одиниці в цьому векторі (Bagui et al., 2021).

Недоліком one-hot вектора є складність та неефективність реалізації для словників, що налічують мільйони слів. Тому таке кодування через його неефективність замінюється більш продуктивними методами ембедингу (Ghannay et al., 2016) (наприклад, такими як word2vec, GloVe, fastText (Rong, 2014; Pennington, Socher and Manning, 2014; Church, 2017)).

В інтелектуальних навчальних системах для спрощення та нормалізації тексту, що відображається природною мовою (цей текст може бути описом власного рішення проблеми, розширеної вільної відповіді на питання до навчального контенту, формулювання запиту до навчального контенту онлайн-курсу, що підтримується інтелектуальною навчальною системою, тощо), передбачаються такі дії:

– попередня обробка тексту природною мовою, яка передбачає стандартизацію речення, розбиття його на токени та додавання спеціальних токенів;

– ембединг (векторне представлення слова), за допомогою якого здійснюється перетворення токена (токенів) у відповідний векторний простір;

– створення текстової послідовності залежно від отриманого (чи наперед заданого) контексту.

NLP передбачає створення текстової бази даних, у якій відображено інформацію у вигляді пар:

<вхідний текст, бажаний результат>.

Вхідний текст – окремі речення з різних джерел, які відповідають тематиці предметної області певного онлайн-курсу. *Бажаний результат* – модифікований варіант вхідного тексту, який формується з використанням таких правил:

– комбіновані слова перетворюються на абревіатуру, наприклад: «штучний інтелект» – «ШІ», «база даних» – «БД», «експертна система» – «ЕС», «персональний комп'ютер» – «ПК» тощо;

– видалення з тексту фраз, які суттєво не впливають на його зміст (наприклад, фрази «Штучний інтелект надає унікальні можливості, які ще кілька років тому здавалися фантастикою. Алгоритми можуть розпізнавати голос, ідентифікувати людину, виявляти помилки, давати рекомендації, розшифровувати емоції тощо. Штучний інтелект успішно виконує різні монотонні операції та обробляє великі обсяги даних» перетворюються на «Алгоритми ШІ можуть розпізнавати голос, ідентифікувати людину, виявляти помилки, давати рекомендації, розшифровувати емоції, обробляти великі обсяги даних.»);

– виключення надлишкових уточнень у тексті (наприклад: «Нейромережеві технології пов'язані з нейронаукою, включаючи когнітивну нейронауку, системну нейронауку, обчислювальну нейронауку») перетворюється в «нейромережеві технології пов'язані з нейронаукою».

Розмір текстової бази даних прямо пропорційно впливає на кількість шаблонів, які спроможна розпізнавати відповідна нейромережа, та знаходження різноманітних зв'язків, зокрема між відповідними токенами.

Попередня обробка тексту природною мовою передбачає підготовку даних для навчання, зокрема виконання таких дій, як:

– переведення всіх алфавітних символів у нижній регістр;
– зведення скорочених слів до єдиної загальної форми;
– приведення різних варіантів лапок до одного виду;
– додавання додаткового пробілу перед та після широкого спектра символів, що не є числом, літерою чи дефісом;
– заміна декількох пробілів, що стоять поспіль, лише одним.

Після стандартизації тексту речення розбивається на токени (з використанням пробілів, розставлених і нормалізованих). Після цього здійснюється додаткова фільтрація отриманих текстових токенів та вилучення з їх множини порожніх елементів.

Навчальні речення (токени, з яких вони створенні) утворюють словник, у якому має своє відображення загальна кількість кожного токена в навчальних прикладах (наборах). Потім токени сортуються згідно з їхньою кількістю (від більшої кількості до меншої), отримують свій власний індекс відповідно до певної позиції та видаляються, якщо їх кількість менша за відповідну кількість, визначену експертами раніше.

Основними проблемами векторного представлення слів тексту (ембедингу), зокрема, є:

– наближення векторів слів контексту одне до одного;
– знаходження асоціативного зв'язку між окремими парами (слів, токенів, форм, речень тощо).

NLP передбачає нормалізацію та спрощення тексту, зокрема виконання таких дій і вимог:

– вихідний текст має бути меншого або рівного розміру (кількості токенів у тексті) за вхідний;
– заміна термінів, жаргонів і сленгів;
– утворення абревіатур;
– видалення уточнювальних конструкцій у круглих дужках;
– видалення посилань на джерела у квадратних дужках тощо.

Спрощення тексту складається, зокрема, з таких дій:

– підготовка навчальних даних;
– формування (створення) нейромережі для навчання ембедингу;
– формування (створення) нейромережі для генерації тексту.

Керування навчанням нейромережі охоплює такі функції:

– тренування із заданою кількістю навчальних епох;
– тестування на тренувальному наборі;
– генерація тексту з введеного контексту;
– збереження.

Після навчання можна буде впізнавати:

– знайомі контексти (тобто контексти, на яких нейромережа навчалася);
– певні шаблони та зв'язки, які допомагають генерувати нові речення.

Висновки. У роботі проаналізовано сучасні методи NLP. Проведений аналіз обумовив вибір методів токенізації, нормалізації, стемінгу та лематизації для їх використання в інтелектуальних навчальних системах під час лінгвістичного аналізу так званого «вільного» спілкування природною (українською) мовою студентів у процесі вивчення навчального контенту онлайн-курсів спеціальностей напряму 12 «Інформаційні технології».

Під час токенізації українськомовних текстів вирішували, зокрема, такі проблеми, як:

– усунення так званих «злитих» токенів;
– виправлення орфографічних помилок;
– визначення спільних префіксів у складних словах та їх впливу на семантику відповідних лексем;
– визначення спільних префіксів в абревіатурах;
– приведення слів до їхньої нормальної форми.

Лематизація особливо важлива для української мови (з її великою кількістю відмінків іменників, прикметників, словоформ тощо), потребує використання спеціально сформованих словників предметної області, що розглядається. У цих словниках словоформи представлені у вигляді лем (тобто іменники надано в називному відмінку).

СПИСОК ПОСИЛАНЬ

Мосейчук, В., 2013. Перелік стоп-слів скачати для української мови. *Книга маразмів України*, [online] 16 січня. Доступно: <https://www.marazm.org.ua/windows/50_141.html> [Дата звернення 21 березня 2024].

Перекладач, б.д. *Google*. [online] Доступно: <https://translate.google.com/> [Дата звернення 22 березня 2024].

Ashraf, A., 2023. Text Pre-Processing for NLP. *Medium*, [online] 31 August. Available at: <https://medium.com/@abdallahashraf90x/text-pre-processing-for-nlp-95cef3ad6bab> [Accessed 12 March 2024].

Awan, A.A., 2023. What is Tokenization? *Datacamp.* [blog] Available at: <https://www.datacamp.com/blog/what-is-tokenization> [Accessed 18 March 2024].

Bagui, S., Nandi, D., Bagui, S. and White, R., 2021. Machine Learning and Deep Learning for Phishing Email Classification using One-Hot Encoding. *Journal of Computer Science,* [e-journal] 7 (17), pp.610-623. https://doi.org/10.3844/jcssp.2021.610.623

ChatGPT, n.d. [online] Available at: <https://chat.openai.com/> [Accessed 12 March 2024].

Chen, M.X., Firat, O., Ankur, B., Melvin, J., Wolfgang, M., George, F., Llion, J., Mike, S., Noam, S., Niki, P., Vaswani, A., Jakob, U., Lukasz, K., Zhifeng, Ch., Yonghui, W. and Macduff, H., 2018. The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia. [e-book] Melbourne: Association for Computational Linguistics, pp.76-86. https://doi.org/10.48550/arXiv.1804.09849

Cheng, J., Dong, L. and Lapata, M., 2016. Long Short-Term Memory-Networks for Machine Reading. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language*

Processing, Austin, Texas, 01-05 November 2016. [e-book] Stroudsburg: Association for Computational Linguistics, pp.551-561. https://doi.org/10.48550/arXiv.1601.06733
Chung, J., Gulcehre, C., Cho, K. and Bengio, Y., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv:1412.3555*, [e-journal] pp.1-9. https://doi.org/10.48550/arXiv.1412.3555
Church, K.W., 2017. Word2Vec. *Natural Language Engineering*, [e-journal] 23, pp.155-162. https://doi.org/10.1017/S1351324916000334
DALL·E 2 is an AI system that can create realistic images and art from a description in natural language, n.d. *DALL·E 2*. [online] Available at: <https://openai.com/dall-e-2> [Accessed 17 March 2024].
Eisenstein, J., 2018. *Natural Language Processing*. [online]. MIT Press. Available at: <https://cseweb.ucsd.edu/~nnakashole/teaching/eisenstein-nov18.pdf> [Accessed 11 March 2024].
Ghannay, S., Favre, B., Estève, Y. and Camelin, N., 2016. Word Embedding Evaluation and Combination. In: *10th edition of the Language Resources and Evaluation Conference*. Portorož, Slovenia, 23-28 May 2016. [online] Portorož: European Language Resources Association, pp.300-305. Available at: <https://pageperso.lis-lab.fr/benoit.favre/papers/favre_lrec2016b.pdf> [Accessed 12 March 2024].
Martinez, J., 2023. Supervised Fine-tuning: customizing LLMs. *Medium*, [online] 09 August. Available at: <https://medium.com/mantisnlp/supervised-fine-tuning-customizing-llms-a2c1edbf22c3> [Accessed 13 March 2024].
Mashtalir, S.V. and Nikolenko, O.V., 2023. Data preprocessing and tokenization techniques for technical Ukrainian texts. *Applied Aspects of Information Technology*, [e-journal] 6 (3), pp.318-326. https://doi.org/10.15276/aait.06.2023.22
Natural Language Processing, 2023. *Deeplearning.ai,* [online] 11 January. Available at: <https://www.deeplearning.ai/resources/natural-language-processing/> [Accessed 02 March 2024].
Natural Language Toolkit, 2023. *NLTK Project*, [online] 02 January. Available at: <https://www.nltk.org> [Accessed 12 March 2024].
O'Connor, R., 2023. How DALL-E 2 Actually Works. *AssemblyAI*, [online] 29 September. Available at: <https://www.assemblyai.com/blog/how-dall-e-2-actually-works/> [Accessed 19 March 2024].
Pennington, J., Socher, R. and Manning, C., 2014. GloVe: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Proceedings of the Conference. Doha, Qatar, 25-29 October 2014. Stroudsburg: Association for Computational Linguistics, pp.1532-1543. Available at: <https://aclanthology.org/D14-1162.pdf> [Accessed 19 March 2024].
Pitis, S., 2023. Failure Modes of Learning Reward Models for LLMs and other Sequence Models. In: *The Many Facets of Preference-based Learning.* Workshop at the International Conference on Machine Learning (ICML) 2023. [online] Available at: <https://openreview.net/attachment?id=NjOoxFRZA4&name=pdf> [Accessed 13 March 2024].
Ramponi, M., 2022. How ChatGPT actually works. *AssemblyAI*, [online] 23 December. Available at: <https://www.assemblyai.com/blog/how-chatgpt-actually-works/> [Accessed 12 March 2024].
Responsible AI that ensures your writing and reputation shine, n.d. *Grammarly.* [online]. Available at: <https://www.grammarly.com/> [Accessed 18 March 2024].
Rong, X., 2014. word2vec Parameter Learning Explained. *arxiv: 1411.2738*, [online] pp.1-21. Available at: <https://arxiv.org/abs/1411.2738> [Accessed 12 March 2024].
Rytr, n.d. [online]. Available at: <https://rytr.me> [Accessed 21 March 2024].
Saumyab271, 2022. Stemming vs Lemmatization in NLP: Must-Know Differences. *Analytics Vidhya*. [blog] Available at: <https://www.analyticsvidhya.com/blog/2022/06/stemming-vs-lemmatization-in-nlp-must-know-differences/> [Accessed 18 March 2024].

Shpater, 2024. ChatGPT Architecture: Will ChatGPT Replace Search Engine? *OPChatGPT*. [blog] Available at: <https://opchatgpt.com/chatgpt-architecture-will-chatgpt-replace-search-engine/> [Accessed 13 March 2024].
Sutskever, I., Vinyals, O. and Le, Q., 2014. Sequence to Sequence Learning with Neural Networks. In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*. Montreal, Quebec, Canada, 8-13 December 2014. [online] Montreal, pp.3104-3112. Available at: <http://arxiv.org/abs/1409.3215v3> [Accessed 21 March 2024].
Tarwani, M.K. and Edem, S., 2017. Survey on Recurrent Neural Network in Natural Language Processing. *International Journal of Engineering Trends and Technology*, [e-journal] 48 (6), pp.301-304. https://doi.org/10.14445/22315381/IJETT-V48P253
Tkachenko, O., Tkachenko, K., Tkachenko, O., Kyrychok, R. and Yaskevych, V., 2024. Neural Networks in the Processing of Natural Language Texts in Information Learning Systems. In: *Cybersecurity Providing in Information and Telecommunication Systems 2024*. Proceedings of the Workshop Cybersecurity Providing in Information and Telecommunication Systems (CPITS 2024). Kyiv, Ukraine, 28 February 2024. [online] Kyiv, pp.73-87. Available at: <https://ceur-ws.org/Vol-3654/> [Accessed 24 March 2024].
Vivien, L., 2022. Google Translate Architecture illustrated. *La Vivien Post* [online]. Available at: <https://www.lavivienpost.com/google-translate-and-transformer-model/> [Accessed 11 March 2024].
Wu, Y., Schuster, M., Chen, Z., Le, Q., Norouzi, M. and Dean, J., 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*, [e-journal] 1. https://doi.org/10.48550/arXiv.1609.08144

REFERENCES

Ashraf, A., 2023. Text Pre-Processing for NLP. *Medium*, [online] 31 August. Available at: <https://medium.com/@abdallahashraf90x/text-pre-processing-for-nlp-95cef3ad6bab> [Accessed 12 March 2024].
Awan, A.A., 2023. What is Tokenization? *Datacamp*. [blog] Available at: <https://www.datacamp.com/blog/what-is-tokenization> [Accessed 18 March 2024].
Bagui, S., Nandi, D., Bagui, S. and White, R., 2021. Machine Learning and Deep Learning for Phishing Email Classification using One-Hot Encoding. *Journal of Computer Science*, [e-journal] 7 (17), pp.610-623. https://doi.org/10.3844/jcssp.2021.610.623
ChatGPT, n.d. [online] Available at: <https://chat.openai.com/> [Accessed 12 March 2024].
Chen, M.X., Firat, O., Ankur, B., Melvin, J., Wolfgang, M., George, F., Llion, J., Mike, S., Noam, S., Niki, P., Vaswani, A., Jakob, U., Lukasz, K., Zhifeng, Ch., Yonghui, W. and Macduff, H., 2018. The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne, Australia. [e-book] Melbourne: Association for Computational Linguistics, pp.76-86. https://doi.org/10.48550/arXiv.1804.09849
Cheng, J., Dong, L. and Lapata, M., 2016. Long Short-Term Memory-Networks for Machine Reading. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, 01-05 November 2016. [e-book] Stroudsburg: Association for Computational Linguistics, pp.551-561. https://doi.org/10.48550/arXiv.1601.06733
Chung, J., Gulcehre, C., Cho, K. and Bengio, Y., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv:1412.3555*, [e-journal] pp.1-9. https://doi.org/10.48550/arXiv.1412.3555

Church, K.W., 2017. Word2Vec. *Natural Language Engineering*, [e-journal] 23, pp.155-162. https://doi.org/10.1017/S1351324916000334
DALL·E 2 is an AI system that can create realistic images and art from a description in natural language, n.d. *DALL·E 2*. [online] Available at: <https://openai.com/dall-e-2> [Accessed 17 March 2024].
Eisenstein, J., 2018. *Natural Language Processing*. [online]. MIT Press. Available at: <https://cseweb.ucsd.edu/~nnakashole/teaching/eisenstein-nov18.pdf> [Accessed 11 March 2024].
Ghannay, S., Favre, B., Estève, Y. and Camelin, N., 2016. Word Embedding Evaluation and Combination. In: *10th edition of the Language Resources and Evaluation Conference*. Portorož, Slovenia, 23-28 May 2016. [online] Portorož: European Language Resources Association, pp.300-305. Available at: <https://pageperso.lis-lab.fr/benoit.favre/papers/favre_lrec2016b.pdf> [Accessed 12 March 2024].
Martinez, J., 2023. Supervised Fine-tuning: customizing LLMs. *Medium*, [online] 09 August. Available at: <https://medium.com/mantisnlp/supervised-fine-tuning-customizing-llms-a2c1edbf22c3> [Accessed 13 March 2024].
Mashtalir, S.V. and Nikolenko, O.V., 2023. Data preprocessing and tokenization techniques for technical Ukrainian texts. *Applied Aspects of Information Technology*, [e-journal] 6 (3), pp.318-326. https://doi.org/10.15276/aait.06.2023.22
Moseichuk, V., 2013. Perelik stop-sliv skachaty dlia ukrainskoi movy [List of stop words for Ukrainian language download]. *Knyha marazmiv Ukrainy*, [online] 16 January. Available at: <https://www.marazm.org.ua/windows/50_141.html> [Accessed 21 March 2024].
Natural Language Processing, 2023. *Deeplearning.ai*, [online] 11 January. Available at: <https://www.deeplearning.ai/resources/natural-language-processing/> [Accessed 02 March 2024].
Natural Language Toolkit, 2023. *NLTK Project*, [online] 02 January. Available at: <https://www.nltk.org> [Accessed 12 March 2024].
O'Connor, R., 2023. How DALL-E 2 Actually Works. *AssemblyAI*, [online] 29 September. Available at: <https://www.assemblyai.com/blog/how-dall-e-2-actually-works/> [Accessed 19 March 2024].
Pennington, J., Socher, R. and Manning, C., 2014. GloVe: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Proceedings of the Conference, Doha, Qatar, 25-29 October 2014. Stroudsburg: Association for Computational Linguistics, pp.1532-1543. Available at: <https://aclanthology.org/D14-1162.pdf> [Accessed 19 March 2024].
Perekladach [Translator], n.d. *Google*. [online] Available at: <https://translate.google.com/> [Accessed 22 March 2024].
Pitis, S., 2023. Failure Modes of Learning Reward Models for LLMs and other Sequence Models. In: *The Many Facets of Preference-based Learning*. Workshop at the International Conference on Machine Learning (ICML) 2023. [online] Available at: <https://openreview.net/attachment?id=NjOoxFRZA4&name=pdf> [Accessed 13 March 2024].
Ramponi, M., 2022. How ChatGPT actually works. *AssemblyAI*, [online] 23 December. Available at: <https://www.assemblyai.com/blog/how-chatgpt-actually-works/> [Accessed 12 March 2024].
Responsible AI that ensures your writing and reputation shine, n.d. *Grammarly*. [online]. Available at: <https://www.grammarly.com/> [Accessed 18 March 2024].
Rong, X., 2014. word2vec Parameter Learning Explained. *arxiv: 1411.2738*, [online] pp.1-21. Available at: <https://arxiv.org/abs/1411.2738> [Accessed 12 March 2024].
Rytr, n.d. [online]. Available at: <https://rytr.me> [Accessed 21 March 2024].
Saumyab271, 2022. Stemming vs Lemmatization in NLP: Must-Know Differences. *Analytics Vidhya*. [blog] Available at: <https://www.analyticsvidhya.com/blog/2022/06/stemming-vs-lemmatization-in-nlp-must-know-differences/> [Accessed 18 March 2024].

Shpater, 2024. ChatGPT Architecture: Will ChatGPT Replace Search Engine? *OPChatGPT*. [blog] Available at: <https://opchatgpt.com/chatgpt-architecture-will-chatgpt-replace-search-engine/> [Accessed 13 March 2024].
Sutskever, I., Vinyals, O. and Le, Q., 2014. Sequence to Sequence Learning with Neural Networks. In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*. Montreal, Quebec, Canada, 8-13 December 2014. [online] Montreal, pp.3104-3112. Available at: <http://arxiv.org/abs/1409.3215v3> [Accessed 21 March 2024].
Tarwani, M.K. and Edem, S., 2017. Survey on Recurrent Neural Network in Natural Language Processing. *International Journal of Engineering Trends and Technology*, [e-journal] 48 (6), pp.301-304. https://doi.org/10.14445/22315381/IJETT-V48P253
Tkachenko, O., Tkachenko, K., Tkachenko, O., Kyrychok, R. and Yaskevych, V., 2024. Neural Networks in the Processing of Natural Language Texts in Information Learning Systems. In: *Cybersecurity Providing in Information and Telecommunication Systems 2024*. Proceedings of the Workshop Cybersecurity Providing in Information and Telecommunication Systems (CPITS 2024). Kyiv, Ukraine, 28 February 2024. [online] Kyiv, pp.73-87. Available at: <https://ceur-ws.org/Vol-3654/> [Accessed 24 March 2024].
Vivien, L., 2022. Google Translate Architecture illustrated. *La Vivien Post* [online]. Available at: <https://www.lavivienpost.com/google-translate-and-transformer-model/> [Accessed 11 March 2024].
Wu, Y., Schuster, M., Chen, Z., Le, Q., Norouzi, M. and Dean, J., 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*, [e-journal] 1. https://doi.org/10.48550/arXiv.1609.08144

UDC 004.8:[37.016:81-028.31

Kostiantyn Tkachenko,
PhD in Economics, Associate Professor,
Associate Professor at the Department
of Computer Systems Software,
National Technical University of Ukraine
"Ihor Sikorsky Kyiv Polytechnic Institute",
Kyiv, Ukraine
tkachenko.kostyantyn@gmail.com
https://orcid.org/0000-0003-0549-3396

USING OF NLP METHODS IN INTELLIGENT EDUCATIONAL SYSTEMS

For the effective organisation of educational processes supported by relevant intelligent learning systems, it is important to choose the right technologies that would ensure individualisation of learning, adequate perception of learning content, and the so-called "understanding" of texts in Ukrainian provided by students (description of the solution to a task, answers provided in their own words, not selected from the test answer options, questions to the system, etc.), prototyping, constant iteration during natural language text recognition and processing, and maximum reliability and efficiency of learning processes.

The purpose of the article is to study and analyse various methods of natural language processing, and the concept of NLP, and to consider common problems and prospects for de-

veloping a software product for processing Ukrainian-language text in online courses that support intelligent learning systems based on it.

The research methods are the main methodological approaches and technological tools for analysing natural language texts in intelligent educational systems and developing a system for supporting NLP (Natural Language Processing) technology in the linguistic analysis of texts in Ukrainian. Such methods include, in particular: systemic and comparative analyses to identify the features of intelligence and information (with elements of intellectualisation) systems; the method of expert evaluation, which involves the study of literary sources and information resources, interviews and surveys of experts, as well as the processes of developing and testing intelligent and information systems.

The novelty of the study is the analysis of modern technologies for the development of online educational process support systems through the organisation of processes of perception of information provided by students in natural language, the results of which can be used in the development of their software product to support the educational process in Ukrainian, ensuring the improvement of learning efficiency through the use of NLP technology in the process of studying the relevant academic content.

Conclusions. The paper analyses modern NLP methods. The analysis has led to the selection of tokenisation, normalisation, stemming and lemmatisation methods for use in intelligent learning systems in the linguistic analysis of the so-called “free” communication in the natural (Ukrainian) language of students in the process of studying the educational content of online courses.

During the tokenisation of Ukrainian-language texts, we solved such problems as eliminating so-called “merged” tokens, correcting spelling mistakes, identifying common prefixes in compound words and their impact on the semantics of the corresponding lexemes, identifying common prefixes in abbreviations, and bringing words to their normal form.

Lemmatisation is especially important for the Ukrainian language (with its large number of cases of nouns, adjectives, word forms, etc.) and it requires the use of specially compiled dictionaries of the subject area under consideration. In these dictionaries, word forms are presented in the forms of lemmas (i.e., nouns are presented in the nominative case).

Keywords: NLP (Natural Language Processing); intelligent learning system; online course; tokenisation; stemming; normalisation; stop words; text segmentation.

02.04.2024