УДК 004.94

О.В. Рудський, А.М. Копп
oleksandr.rudskyi@cs.khpi.edu.ua, andrii.kopp@khpi.edu.ua
Національний технічний університет «Харківський політехнічний інститут», м. Харків

ВИКОРИСТАННЯ ТЕХНОЛОГІЙ ОБРОБКИ ПРИРОДНОЇ МОВИ ДЛЯ АНАЛІЗУ ВІДПОВІДНОСТІ МОДЕЛЕЙ БІЗНЕС-ПРОЦЕСІВ ЇХ ТЕКСТОВИМ ОПИСАМ

Моделі бізнес-процесів зарекомендували себе як ефективний засіб візуалізації та вдосконалення складних організаційних операцій. Однак створення моделей бізнес-процесів є трудомісткою справою, яка потребує значних ресурсів, тому можуть виникати ситуації в яких модель бізнес-процесу не відповідає її текстовому опису. Це може призвести до втрат у часі та значних грошових втрат. Таким чином, актуальною є задача аналізу відповідності моделей бізнес-процесів їх текстовим описам.

Для того, щоб обробляти тексти з метою аналізу відповідності моделей бізнес-процесів їх текстовим описам, пропонується застосовувати наступні технології обробки природної мови або NLP-технології (Natural Language Processing):

– токенізація;
– пошук стоп-слів;
– стемінг.

Токенізація – це перший крок у будь-якому процесі NLP. Токенізатор розбиває неструктуровані дані та текст природною мовою на фрагменти інформації, які можна розглядати як дискретні елементи. Це відразу перетворює неструктурований рядок (текстовий документ) на числову структуру даних, придатну для машинного навчання [1].

Після розділення тексту на токени, часто стає ясно, що не всі слова несуть однакову кількість інформації, якщо взагалі будь-яку інформацію для завдання прогнозного моделювання. Загальні слова, які несуть мало значної інформації, називаються стоп-словами. Стоп-слова – це слова будь-якою мовою, які не додають великого сенсу реченню. Їх можна сміливо ігнорувати, не жертвуючи змістом речення. Для деяких пошукових систем це деякі з найпоширеніших, коротких функціональних слів, таких як "the", "is", "at", "which" та "on" [2].

Стемінг є однією з найпоширеніших операцій попередньої обробки даних, яка виконується майже у всіх проектах обробки природної мови (NLP). Стемінг – це процес скорочення слова до основи шляхом відкидання допоміжних частин, таких як закінчення чи суфікс. Результати стемінгу іноді дуже схожі на визначення кореня слова, але його алгоритми базуються на інших принципах. Тому слово після обробки алгоритмом стемінгу може відрізнятися від морфологічного кореня слова [3].

Таким чином, для розв'язання задачі аналізу відповідності моделей бізнес-процесів їх текстовим описам з отриманими текстами T_1 та T_2 необхідно виконати наступні дії (рис. 1) [4]:

1) розбити отримані на вхід тексти T_1 та T_2 на окремі слова (токенізувати), отримавши відповідні мультимножини слів:

$$(W_1, m_1) = \{(t_i^1, m_1(t_i^1), t_i^1 \in W_1 \wedge i = \overline{1,n}\},$$
$$(W_2, m_2) = \{(t_j^2, m_2(t_j^2), t_j^2 \in W_2 \wedge j = \overline{1,q}\},$$

де W_1 – множина слів, отриманих в результаті токенізації тексту T_1; W_2 – множина слів, отриманих в результаті токенізації тексту T_2; $t_i^1 \in W_1$, $i = \overline{1,n}$ – слово, отримане в результаті токенізації тексту T_1; $t_j^2 \in W_2$, $j = \overline{1,q}$ – слово, отримане в результаті токенізації тексту T_2; $m_1(t_i^1)$ – відображення $m_1 : W_1 \to \{1,2,3,...\}$, яке для кожного слова $t_i^1 \in W_1$, $i = \overline{1,n}$ встановлює кількість його повторювань в тексті T_1; $m_2(t_j^2)$ – відображення $m_2 : W_2 \to \{1,2,3,...\}$, яке для кожного слова $t_j^2 \in W_2$, $j = \overline{1,q}$ встановлює кількість його повторювань в тексті T_2; n – кількість слів, отриманих в результаті токенізації тексту T_1; q – кількість слів, отриманих в результаті токенізації тексту T_2;

2) видалити стоп-слова з множин W_1 та W_2, отримавши множини лише змістовних термінів, що стосуються предметної області бізнес-процесу:

$$stop : \{W_k, k = \overline{1,r}\} \to \{W_k', k = \overline{1,r}\},$$

де W_k, $k = \overline{1,r}$ – множина слів, отримана в результаті токенізації вихідного тексту, що також містить стоп-слова; W_k', $k = \overline{1,r}$ – множина слів, отримана в результаті токенізації вихідного тексту, з якої були

видалені стоп-слова; *stop* – відображення, яке для кожної множини W_k, $k=\overline{1,r}$, яка містить стоп-слова, ставить у відповідність множину W_k', $k=\overline{1,r}$, яка не містить стоп-слова; r – кількість множин слів, що обробляються, $r=2$;

3) виконати стемінг слів у множинах W_1' та W_2', що залишились після видалення стоп-слів:

$$stemm:\{W_k', k=\overline{1,r}\}\to\{W_k'', k=\overline{1,r}\},$$

де W_k'', $k=\overline{1,r}$ – множина слів, отримана в результаті стемінгу слів, що залишились після видалення стоп-слів; $stemm$ – відображення, яке для кожної множини W_k', $k=\overline{1,r}$, з якої були видалені стоп-слова, ставить у відповідність множину W_k'', $k=\overline{1,r}$, слова в якій, що залишились після видалення стоп-слів, були скорочені до основи.

Таким чином, в результаті виконання попередній дій, будуть отримані множини слів W_1'' та W_2'':

$$W_1''\cup W_2''\subseteq\{W_k'', k=\overline{1,r}\}.$$

Обчислити подібність цих двох множин слів W_1' та W_2' можна за допомогою коефіцієнта Жаккара:

$$K_J=\frac{|W_1''\cap W_2''|}{|W_1''|+|W_2''|-|W_1''\cap W_2''|}=\frac{|W_1''\cap W_2''|}{||W_1''\cup W_2''|}.$$

Відповідно, отримане значення коефіцієнту Жаккара можна інтерпретувати як ступінь відповідності моделі бізнес-процесу її текстовому опису.

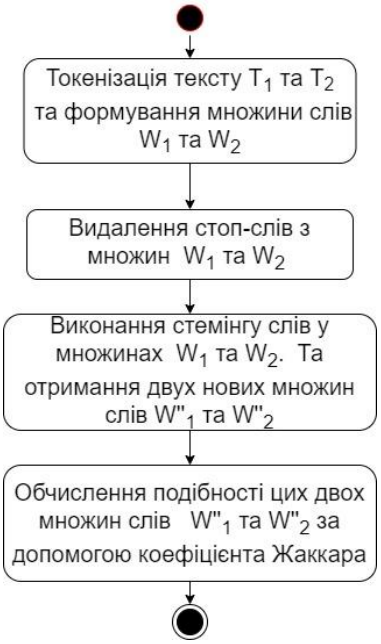

Рис. 1. Алгоритм розв'язання задачі аналізу відповідності моделей бізнес-процесів їх текстовим описам

Список літератури

1. Tokenization in NLP: Types, Challenges, Examples, Tools. URL: https://neptune.ai/blog/tokenization-in-nlp (дата звернення: 10.03.2023).

2. Rajaraman A., Ullman J. D. Mining of Massive Datasets. *Cambridge University Press*. Cambridge, 2011. P. 18–19.

3. Jongejan B., Dalianis H. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. Singapore, 2009. P. 145–153.

4. Kopp A. M., Orlovskyi D. L. The approach and the software tool to calculate semantic quality measures of business process models. *Bulletin of the National Technical University "KhPI" : System analysis, control and information technology*. 2022. No. 1 (7). P. 66–69.