УДК 004.4

Житняківський В.А., Мазурець О.В.

Хмельницький національний університет

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ КЛЮЧОВИХ СЛІВ У ТЕКСТОВИХ ПОВІДОМЛЕННЯХ ДЛЯ СОЦІАЛЬНИХ МЕРЕЖ

Досліджено проблему підвищення ефективності орієнтування користувачів соціальних мереж у великій кількості постів та текстових повідомлень, а також пошуку потрібних відомостей, для чого запропоновано використання співставлення кожному повідомленню множини ключових слів. Представлено відповідну програмну систему на платформі IOS, що реалізує запропоновану інформаційну технологію й дозволяє підтвердити її високу практичну та наукову цінність.

The problem of increasing the efficiency of targeting users of social networks in a large number of posts, text messages and search for the necessary information is investigated. For this purpose, the use of matching each messages with a set of keywords is proposed. The corresponding software system on the IOS platform is presented, which implements the proposed information technology and confirms its high practical and scientific value.

Одним із найпопулярніших сервісів, здатних утримувати увагу значної частини інтернет-аудиторії, є соціальні мережі. Виникнення віртуальної взаємодії в межах соціальної мережі фактично є історичною відповіддю на появу комунікаційного надлишку [1]. Для орієнтування в великій кількості постів та текстових повідомлень, а також пошуку потрібних відомостей, існує ефективний підхід до співставлення кожному повідомленню множини ключових слів – тегів. Зазвичай формування таких множин ключових слів виконується вручну. Проте це є трудомістким додатковим етапом при формуванні повідомлень, тому застосування технології автоматизованого визначення ключових слів у текстових повідомленнях для соціальних мереж є актуальним напрямком в розробці спеціалізованих систем спілкування та соціальних мереж [2].

Метою даної роботи є розробка інформаційної технології автоматизованого визначення ключових слів у текстових повідомленнях для соціальних мереж.

Загальна схема інформаційної технології зображена на рисунку 1. Вхідними даними для і текстове повідомлення, а вихідними даними множина наборів ключових слів.

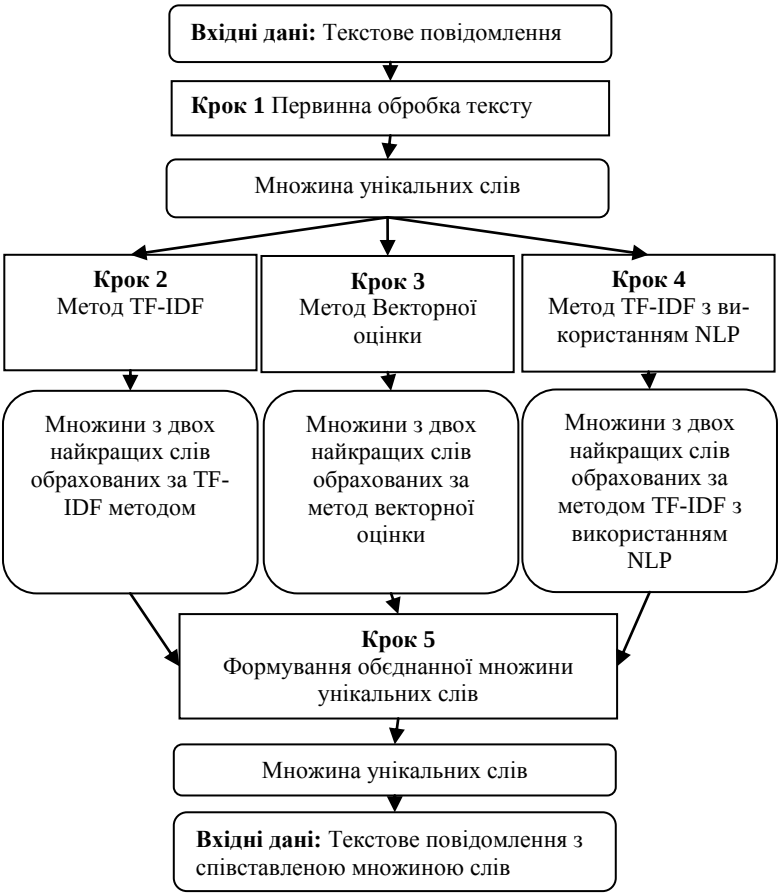

Рисунок 1 – Загальна схема інформаційної технології

На початку виконання інформаційної технології вхідні дані отримуються у вигляді текстового повідомлення, після чого виконується первинна обробка тексту (*Крок 1*). На даному етапі текст позбавляється розділових знаків та інших символів, формується загальна множина слів, опісля формується множина унікальних слів, обраховується кількість унікальних слів, та кількість появ кожного слова в текстовому документі і також обраховується загальна кількість текстових документів в яких дане слово зустрічається. На наступному етапі обраховується оцінка методом TF- IDF (*Крок 2*).

Оцінка TFIDF є добутком частоти згадувань слова у тексті TF та зворотної документарної частоти слова IDF[3]:

$$TfIdf(t_i, d) = Tf(t_i, d) * Idf(t_i, b),$$

$$Tf\ (t_i, d) = \frac{n(t_i, d)}{\sum k(t_k, d)},$$

$$Idf\ (t_i, b) = \log \frac{\sum j(d_j, b)}{\sum z(d(t_i)_z, b)},$$

де $n(t_i, d)$ – число входжень слова t_i в документ d; $\sum_k (t_k, d)$ – загальна кількість всіх слів документа d; $\sum_j (d_j, b)$ – загальна кількість документів d в вибірці b; $\sum_z (d(t_i)_z, b)$ – загальна кількість документів d, в яких зустрічається слово t_i.

За допомогою звичайного сортування отримуємо два слова з найбільшою оцінкою.

На наступному етапі обраховується векторна оцінка(*Крок 3*). Векторна оцінка за змістом близька до оцінки TF-IDF, та є оцінкою дискримінантної сили слів. Вона дозволяє відділити із загального переліку широковживаних у тексті слів слова, що розташовані рівномірно. Якщо деяке слово *T* позначається як T_k^n, де індекс k – номер появи даного слова у тексті, а n – позиція данного слова у тексті. Інтервалом між послідовними появами слова при таких позначеннях буде величина $\Delta T_k^m = T_{k+1}^m - T_k^n = m - n$, де на *m-й* та *n-й* позиції в тексті знаходиться слово *A*, яке зустрілось *k*+1-*й* і *k-й* рази.

Дана оцінка розраховується як: $DE = \sqrt{(\Delta T^2) - (\Delta T)^2} \Big/ (\Delta T)$, де (ΔT) – середнє значення послідовності $\Delta T_1, \Delta T_2, \Delta T_k$, (ΔT^2) – послідовність T_1^2, T_2^2, T_k^2, де *k* – кількість появи слова *A* у тексті [6]. Після цього за допомогою звичайного сортування отримуємо два слова з найбільшою оцінкою.

На наступному етапі оцінки важливості слова є метод TD –IDF з використанням NLP (Крок4). Даний метод відрізняється від класичного TF-IDF тим, що за допомогою фреймворка Apple NLP ми позбуваємось стоп слів.

За допомогою даного фреймворка відкидаються другорядні частини мови, всі слова приводяться до називного відмінку, а також

відкидаються відомі назви, іменна, та об'єкти [5]. Наступні дії відбуваються по аналогії з Кроком2.

Для експериментального тестування інформаційної технології було розроблено відповідну інформаційну систему (рисунок 2). Дана система створена на платформі IOS із можливістю автоматизованого пошуку ключових слів в новинах, й виконує наступні функції: авторизація і реєстрація користувачів; додавання повідомлень до власної стрічки новин користувачем; автоматизоване визначення ключових слів у доданій користувачем новині (рисунок 3); перегляд власної публічної стрічки новин; перегляд стрічок новин інших користувачів; підписка на новини від інших користувачів; перегляд об'єднаної стрічки новин від відстежуваних користувачів; перегляд об'єднаної стрічки новин від всіх користувачів системи; обмін користувачів приватними текстовими повідомленнями, геоданими, фото- та відеоматеріалами; формування стрічок приватного спілкування (чатів); можливість пересилання існуючих постів користувачів.

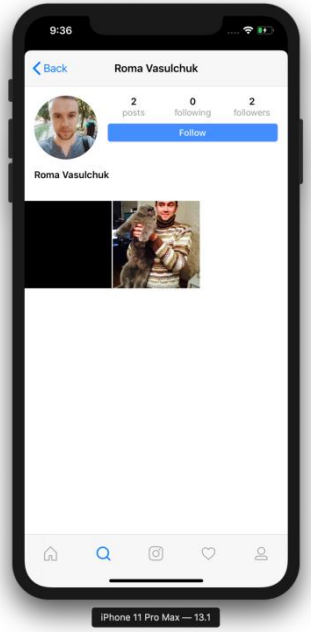

Рисунок 2 – Головна сторінка користувача інформаційної системи

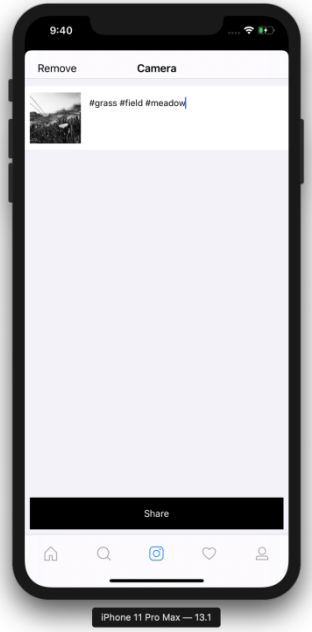

Рисунок 3 – Сторінка автоматизованого визначення ключових слів

Проведені дослідження, що включали аналіз 250 повідомлень десятьма респондентами, встановили, що результуючі множини, сформовані автоматично, містили в середньому 73% ключових слів, актуальних для користувача. При цьому серед актуальних слів частка слів, знайдених методом TF-IDF склала 41%, методом Векторної оцінки – 42%, а методом TD-IDF з використанням NLP – 57%.

Отже, було досліджено проблему підвищення ефективності орієнтування користувачів соціальних мереж у великій кількості постів та текстових повідомлень, а також пошуку потрібних відомостей, для чого запропоновано використання співставлення кожному повідомленню множини ключових слів. Розглянуто інформаційну технологію автоматизованого визначення ключових слів у текстових повідомленнях для соціальних мереж й відповідну програмну систему на платформі IOS із можливістю автоматизованого пошуку ключових слів в новинах. Проведені дослідження встановили високу ефективність запропонованої технології для допомоги користувачеві в формуванні множини ключових слів до повідомлень та новин у соціальних мережах.

Перелік посилань

1. Internet World Stats [Електронний ресурс] — Режим доступу: https://www.internetworldstats.com/stats.htm (дата звернення 30.03.2018). — Назва з екрана.
2. Ying Ch., Yilu Zh., Sencun Zh., Heng X. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety / Ch. Ying, Zh. Yilu, Zh. Sencun, X. Heng // — SOCIALCOM-PASSAT '12 Proceedings of the 2012 ASE/IEEE International Conference on Social Computing, 2012 — 2012 — Pp. 71-81. Ландэ Д.В., Снарский А.А. Компактифицированный горизонтальный граф видимости для сети слов // Труды Международной научной конференции «Интеллектуальный анализ информации ИАИ-2013. Знания и рассуждения» – КПИ, Киев: 2013. — с. 158-164.
3. Бармак О. В. Інформаційна технологія автоматизованого визначення термінів у навчальних матеріалах / О. В. Бармак, О. В. Мазурець // Вимірювальна та обчислювальна техніка в технологічних процесах. – Хмельницький, 2015. – № 2. – С. 94–102
4. Ландэ Д.В., Снарский А.А. Компактифицированный горизонтальный граф видимости для сети слов // Труды Международной научной конференции «Интеллектуальный анализ информации ИАИ-2013. Знания и рассуждения» – КПИ, Киев: 2013. — с. 158-164.
5. Stanford NLP [Електронний ресурс] — Режим доступу: https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html