# ДОСЛІДЖЕННЯ МЕТОДІВ ТА ПРОГРАМНИХ ЗАСОБІВ АНАЛІЗУ НОВИННИХ СТРІЧОК В СОЦІАЛЬНИХ МЕРЕЖАХ

Сердюк О.О., студент магістратури ПІ,
e-mail: oksana_serdiuk@nure.ua.
Науковий керівник: д.т.н., проф. Єрохін А.Л.
Харківський національний університет радіоелектроніки

The article is devoted to the methods of text mining, analyzing text and weighting words in a document in order to retrieve words those are more semantically focused. That helps to retrieve information about it main topic, tone and content.

Кожна людина зараз проводить переважну частину часу в соціальних мережах. За статистикою люди проводять у соціальних мережах мінімум годину часу. За цей час в мережу потрапляє надзвичайна кількість необробленої інформації, яка здатна впливати на свідомість людей, на їхні думки, точки зору. Аналіз усього потоку текстової інформації дозволив би вирішити декілька проблем в різних сферах.

На основі результатів аналізу тексту можна зробити висновок про його тематику, основну думку, основний мотив, тональність написаного. На рівні держави і бізнесу важливим результатом є можливість визначити найбільш обговорювані теми, загальний настрій людей у певний період з перелічених показників, або навіть прогнозування можливих реакцій на певні події. Також, у сфері соціальних мереж для кінцевих користувачів було б корисно мати фільтр з великою кількістю налаштувань, що виконує пошук згідно з даними показниками.

Для вирішення перелічених проблем можна скористатися одним методом: інтелектуальним аналізом тексту. Схема інтелектуального аналізу тексту представлена на рисунку 1. Інтелектуальний аналіз тексту призначений для обробки неструктурованого набору інформації з метою вилучення з нього корисних даних. Наприклад, він може бути використаний для виділення основної теми тексту, порівняння двох текстів на

схожість. У нашому випадку він стане в нагоді для визначення основної думки тексту і його тональності [1].

Інтелектуальний аналіз тексту обробляє текст перетворюючи його в значимі індекси, які можуть використовуватися в подальшому, наприклад, в машинному навчанні [2]. Кожне слово в тексті повинно бути проіндексовано і пораховано для складання таблиці матриці частот, яка відображає частоту вживання того чи іншого слова.

Далі побудована таблиця може бути оброблена для видалення незначущих слів, артиклів, з'єднання однакових слів, які були вжиті в різних формах (число, час, відмінок). Один з варіантів нормалізації називається стемінг. Необхідно звернути увагу, що процес стемінгу специфічний для кожної мови, так як кожна мова має свої слова, особливості їх вживання та граматичні конструкції.

Стемінг позначає процес знаходження основи слова для заданого вихідного слова так, щоб різні граматичні форми одного слова були зараховані як одне [4]. Наприклад, «читає» і «читав» має бути проіндексовано як одне слово. Тут важливо знайти синоніми і словосполучення, які мають значення відмінне від того, якби ці ж слова, присутні у фразі, були вжиті окремо. Також має сенс виключити з статистики рідкісні слова, які швидше за все не відбивають суть тексту.


Рисунок 1 – Схема інтелектуального аналізу тексту.

Зазвичай стеммером користуються для пошуку тексту з імітацією врахування морфології. Під імітацією розуміють непереборно велику кількість помилок і нерелевантних результатів, які виникають, якщо застосовувати тільки стеммер [5]. У

слов’янських мовах джерелом помилок при стемінгу є всілякі зміни кореня слова. Наочно проблеми, пов'язані з використанням стеммер, можна продемонструвати для іменника «кішка». Родовий відмінок множини має форму «кішок». Таким чином, найдовший спільний префікс всіх форм іменника «кішка» - це «кіш». Якщо виконати пошук тексту по цьому префіксу, то в результатах з може опинитися слово «Кішма», що є назвою річки. Або можливий ще один варіант помилки для цього слова, коли стеммер обирає префікс «кішк», що унеможливлює появу слова «кішок» у результатах.

Таким чином, можна виділити наступні помилки при стемінгу. Стем дає занадто велике узагальнення і тому зіставляється з граматичними формами більш ніж однієї словникової статті. Це найчисленніша група помилок стемінгу. Наприклад, якщо в стемінге братиме участь слово «вам», то в подальшому знайти певний текст дасть збіг зі словом «вампір». Компенсація помилок такого роду успішно виконується або введенням списку стоп-слів, або більш якісно – лематизатором або флексером.

Наступні помилки, пов'язані з тим, що усічення форми дає занадто довгий стем, що не зіставляється з деякими граматичними формами цього ж слова [6]. До таких помилок призводить прагнення розробника стеммеру знайти компроміс з помилками першого роду в разі, коли при словотворенні змінюється основа слова. Такі слова є навіть у англійській мові, що багата частими словозмінами (неправильні дієслова). У українській мові випадки зміни основи слова навіть не є підставою для віднесення слова до групи неправильних, настільки часте це явище.

Помилкою стемінгу третього роду є відсутність можливості побудувати стем через зміни в корені слова, яке залишає єдину букву в стемі. Або модель словозміни має на увазі використання префіксів. Приклад для першого випадку слова «шити», «вити», «лити», «пити», які мають форми «шию», «в'ють», «ллє», «п'ю». Дані слова мають корінь, що складається з однієї літери. Другий випадок виникає в рамках граматичного словника для вищого ступеня прикметників і прислівників в українській мові – наприклад «побільше» як більш висока форма порівняння для прислівника «більше».

Для ефективнішого текстового аналізу доповненням до стеммінгу може виступати лематизація [7]. Лематизацією

називається перетворення слова в словниковий вид або лемму. Даний метод використовується в алгоритмах пошукових систем при індексації сторінок. Процес дає можливість зберігання даних сторінки набором слів в індексі для зручної схематизації файлів. Це дозволяє прискорити індексацію і сформувати більш чітку відповідь на пошуковий запит, так як скорочену форму слова пошукова система аналізує швидше.

Лематизації передбачає порівняння основ слів, однак вона проводиться з урахуванням частин мови, до якої відносяться словоформи. Наприклад, стеммер для «читати», «читав», «читач», «читачі» з коренем «чит» і зробить висновок, що це одне і те саме слово. У той час як лематизатор виділить окремо слова «читати» і «читав» як форми слова «читати», і слова «читач» і «читачі» як форми слова «читач». Під лемою мається на увазі лексема, завдання лематизації ототожнити словоформи, співвідносні з однією лексемою.

Після побудови таблиці унікальних слів її можна використовувати для кластеризації, класифікації, виділення важливих слів, що несуть певне смислове навантаження, або термінів [8].

Постає питання як зрівняти смислову вагу слів посеред інших, адже просто визначити частоту вживання недостатньо для більш ефективного рішення. Тому для визначення смислової ваги слова використовують наступні терміни.

TF (term frequency) – частота вживання слова в документі. Оскільки кожен документ відрізняється за довжиною, можливо, що слово у довгих документах з'являється набагато більше разів, ніж у коротших. Таким чином, term frequency часто ділиться на довжину документу (тобто загальне число слів у документі) як спосіб нормалізації:

$$tf_{i,j} = \frac{n_i}{\sum_j n_j},$$

де $n_i$ – кількість разів, коли слово *i* з'являється в документі *j*, $n_j$ – загальна кількість слів у документі *j*.

DF (document frequency) – число документів у яких вживається і-е слово. Під час обчислення TF всі слова вважаються однаково важливими. Проте відомо, що певні слова, такі як "є", "з", "що", можуть з'являтися багато разів, але мають мало значення. Таким чином, нам потрібно зменшити важливість таких слів, збільшуючи

вагу слів, що зустрічаються не так часто, шляхом обчислення наступного:

$$df_{i,j} = \frac{d_i}{\sum_j d_j},$$

де $n_i$ – кількість документів, у яких з'являється слово $i$, $d_j$ – загальна кількість документів.

Інформація, що визначається TF (частотою слова), дає зрозуміти на скільки важливим є слово для даного тексту. Чим більше частота слова, тим більш ймовірно, що це слово добре описує зміст тесту.

Але далеко не завжди, якщо в одному тексті слово зустрічається один раз, а в іншому 3 рази, значить, що в другому воно краще описує вміст. Тому виконуються додаткові обчислення для того, щоб послабити вплив цієї частоти на результат. Можна виконати одне з наступних обчислень:

$f(tf) = \sqrt{tf}$

або

$f(tf) = 1 + log(tf)$, для $tf > 0$, де $tf$ – частота вживання слова.

Наприклад, якщо слово «машина» зустрічається у тексті 3 рази, то $tf = \sqrt{3}$ або $tf$ =1 + log 3 краще відображають важливість слова з частотою слова 3, ніж просто кількість разів. Документ з частотою слова $tf = \sqrt{3}$ або $tf$ =1 + log 3 буде дещо важливіший ніж документ з однією появою слова ($tf = \sqrt{1}$ або $tf$ =1 + $log$ 1), але не втричі важливіший.

DF (частота документу) може бути інтерпретована як індикатор інформативності. За допомогою частоти документу, можна нівелювати ще одну проблему. Припустимо в тексті слово «заняття» зустрічається 10 разів, а слово «парламент» зустрічається тільки 3 рази. Хоча, як раз слово «парламент» краще описує і відображає зміст тексту. Справа в тому, що слово «заняття» може вживатися кожен раз в новому контексті, яке ніяк не стосується теми тексту. Наприклад, «заняття у школі», «заняття як хобі». А слово «парламент» вузькоспеціалізоване і вживається у текстах переважно політичної тематики.

Це пояснюється тим, що семантично сфокусоване слово зазвичай вживається у тексті кілька разів, якщо вживається взагалі. А семантично несфокусоване слово розподілене по всьому тексту. Також, у разі якщо у тексті було вжито семантично сфокусоване

слово, велика ймовірність, що це слово буде вжите ще кілька разів [9].

Поєднати частоту слова та частоту документу в одну вагову одиницю і-го слова і j-го документа можна за допомогою наступної формули, що називається Inverse Document Frequency (IDF):

$$w(i,j) = \begin{cases} 0\text{ , якщо } tf_{ij} = 0 \\ \left(1 + log\left(tf_{ij}\right)\right) log\frac{N}{df_i}\text{, якщо } tf_{ij} \geq 1\text{'} \end{cases}$$

де $w(i,j)$ – вага і-го слова в j-ому документі, $N$ – загальна кількість документів, $df_i$ – кількість документів, які використовують і-е слово, $tf_{ij}$ – частота вживання і-го слова в j-ому документі.

Отже, можна побачити, що ця формула включає в себе як звуження простих частот слів за допомогою функції логарифму, а також включає вагових фактор, який прирівнює до 0, якщо слово зустрічається у всіх документах $(log(N/N = 1) = 0)$, і до максимального значення, коли слово зустрічається лише в одному документі $(log(N/1) = log(N))$. Цілком легко зрозуміти, як ця трансформація створює індекси, які відображають відносні частоти слів, а також їх семантичні особливості в документах, що були включені до аналізу.

Література.

1. T. Mikolov, K. Chen, G. Corrado, J. Dean, «Efficient Estimation of Word Representations in Vector Space» In Proc. of Workshop at ICLR, 2013.

2. Hill T., Lewicki P. Statistics: Methods and Applications. Comprehensive Reference for Science, Industry, and Data Mining / T. Hill, P. Lewicki –Tulsa: StatSoft, Inc, 2006. – 832 с.

3. Text Mining (Big Data, Unstructured Data). [Електронний ресурс] / TIBCO Statistica – Режим доступу: http://www.statsoft.com/Textbook/Text-Mining – 10.03.2018 р. – Загол. з екрану.

4. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных : учеб. пособие / Е.И. Большакова, К.В. Воронцов, Н.Э. Ефремова, Э.С.

Клышинский, Н.В. Лукашевич, А.С. Сапин — М.: Изд-во НИУ ВШЭ, 2017. — 269 с.

5. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие / Н. Н. Леонтьева — М.: Академия, 2006.

6. Плунгян В.А. Общая морфология. Введение в проблематику. / В.А. Плунгян — М.: Едиториал УРСС. — 2003. 384 с.

7. Лемматизация. [Електронний ресурс] / Компьютерная грамматика русского языка: лексика, морфология, синтаксис – Режим доступу: http://www.solarix.ru/for_developers/api/lemmatization.shtml – 10.03.2018 р. – Загол. з екрану.

8. Клышинский Э.С. Начальные этапы анализа текста // Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Э.С. Клышинский — М.: МИЭМ, 2011.

9. Manning C. D., Schiitze H. Foundations of Statistical Natural Language Processing/ C. D. Manning, H. Schiitze– London: The MIT Press, 1999. – 704 c.