

ДОВГОСЯЖНІ КОРЕЛЯЦІЇ ЯК КРИТЕРІЙ РЕЛЕВАНТНОСТІ СЛІВ У ТЕКСТОВИХ ДОКУМЕНТАХ

Олег Кушнір, Богдан Горон, Іван Довгань, Марта Дуфанець

LONG-RANGE CORRELATIONS AS A CRITERION FOR RELEVANCE OF WORDS IN
TEXTUAL DOCUMENTS

Oleh Kushnir, Bohdan Horon, Ivan Dovhan, Marta Dufanets

Львівський університет імені Івана Франка, вул. Тарнавського 107, м. Львів, 79017,
Україна

Abstract. *Highly efficient detection of keywords is a basis for successful information retrieval. Here we present a new criterion of relevance of words in textual documents, which is associated with the long-range autocorrelations of word-token time series. The above approach is compared with a canonical keyword detection method based on word-token clustering in a text.*

В останні роки значно зріс інтерес до алгоритмів і засобів опрацювання природної мови. Одним із важливих напрямів тут є прогрес інформаційного пошуку в полі текстових документів. Для підвищення швидкодії пошуку документів, релевантних для даного запиту, типowo використовують прийом радикального пониження розмірності задачі, аналізуючи не повні документи, а лише набори їхніх ключових слів або фраз. Успіхи тут прямо залежать від ефективності методів встановлення ключових слів текстів.

З літератури відомі багато методів знаходження ключових слів. Їх можна поділити на дві великі групи: відносні методи, що базуються на порівнянні досліджуваного тексту з великою референтною базою текстів (наприклад, метод TF-IDF), і абсолютні методи, що засновані на аналізі взаємного розміщення слововживань у єдиному досліджуваному тексті. Хоча останні методи зазвичай програють методів TF-IDF за ефективністю, проте відзначаються вкрай високою швидкістю та мають певну практичну нішу. Серед властивостей, що лежать в основі абсолютних методів, є кластеризація моментів появ кожного слова в тексті, особливості інформаційної ентропії, розрахованої для різних статистичних розподілів, мережеві властивості та ін. (див., наприклад, праці [1, 2]).

Для встановлення релевантності того чи іншого слова в цій праці ми пропонуємо використовувати такий критерій: слово релевантне до тематики та семантичного поля текстового документа, якщо його словопояви виявляють виразні довгосяжні автокореляції, тобто якщо процес повторення появ цього слова в тексті є стохастичним процесом із пам'яттю. На принципову можливість, хоча й не ефективність, такого критерію вказують результати [3], згідно з якими вияви явища «спалахів» (burstiness) у багатьох природних і залежних від людини складних системах включають не лише кластеризацію, а й довгосяжні кореляції (ДСК або пам'ять), а також дані [4] для складних лінгвістичних систем, які засвідчують присутність ДСК в розміщенні словопояв.

Для одночасного емпіричного вивчення параметра кластеризації R і степеневого параметра γ , який визначає характер флуктуаційної функції та описує ДСК, для всіх лінгвістичних елементів у тексті, а також для їхніх послідовностей (n -грам), було розроблено програму із залученням бібліотек Python. Кількісні характеристики автокореляцій у числових рядах, побудованих на послідовностях лінгвістичних елементів, вивчали в рамках стандартного підходу флуктуаційного аналізу (Fluctuation Analysis). Було досліджено низку природних англomовних текстів і тексти, побудовані на основі різних рандомних моделей природної мови. Показано, що букви, яким не притаманне жодне семантичне навантаження, практично не виявляють ефекту кластеризації ($R \approx 1$), хоча їм притаманні деякі слабкі позитивні ДСК ($\gamma > 0.5$, проте $\gamma \sim 0.5$). Для т. зв. стоп-слів із мінімальною семантикою характерна схожа

ситуація, а т. зв. змістові слова, семантичне навантаження яких вище, описуються дещо більшими параметрами ($R > 1$ і $\gamma > 0.5$). Нарешті, ключові слова, найважливіші для семантики тексту, одночасно виявляють виразні кластеризацію і ДСК, тобто найвищі параметри R і γ ($R \gg 1$ і $0.5 < \gamma < 1$, причому $\gamma \sim 1$).

Розглянемо як приклад текст «The Book of Jungle» (R. Kipling). Першими за частотою F тут є стоп-слова the, and, of та in. У першу ж десятку слів із найвищими параметрами R ($R \approx 6.4 \div 8.3$) входять Mowgli, Rikki-Tikki, Kotick, Nag, Baloo, Kala, Каа, Toomai, Bagheera, Seals. Загалом список слів тексту, ранжований за спаданням R , можна вважати якісним списком ключових слів. Водночас, у списку з найвищими ДСК ($\gamma \approx 0.89 \div 0.95$) маємо слова Kotick, Toomai, Rikki-Tikki, Mowgli, Elephants, Nag, Herd, Kala, Bandarlog, Bagheera. Можна зауважити значне перекривання списків слів із найвищими кластеризацією та ДСК, причому решта слів із першої десятки за параметром R перебувають лише незначно нижче у ранжованому списку слів за спаданням параметра γ – і навпаки. Отже, як і добре відомий з літератури параметр кластеризації R , ДСК також можуть успішно відображати релевантність слів у текстах.

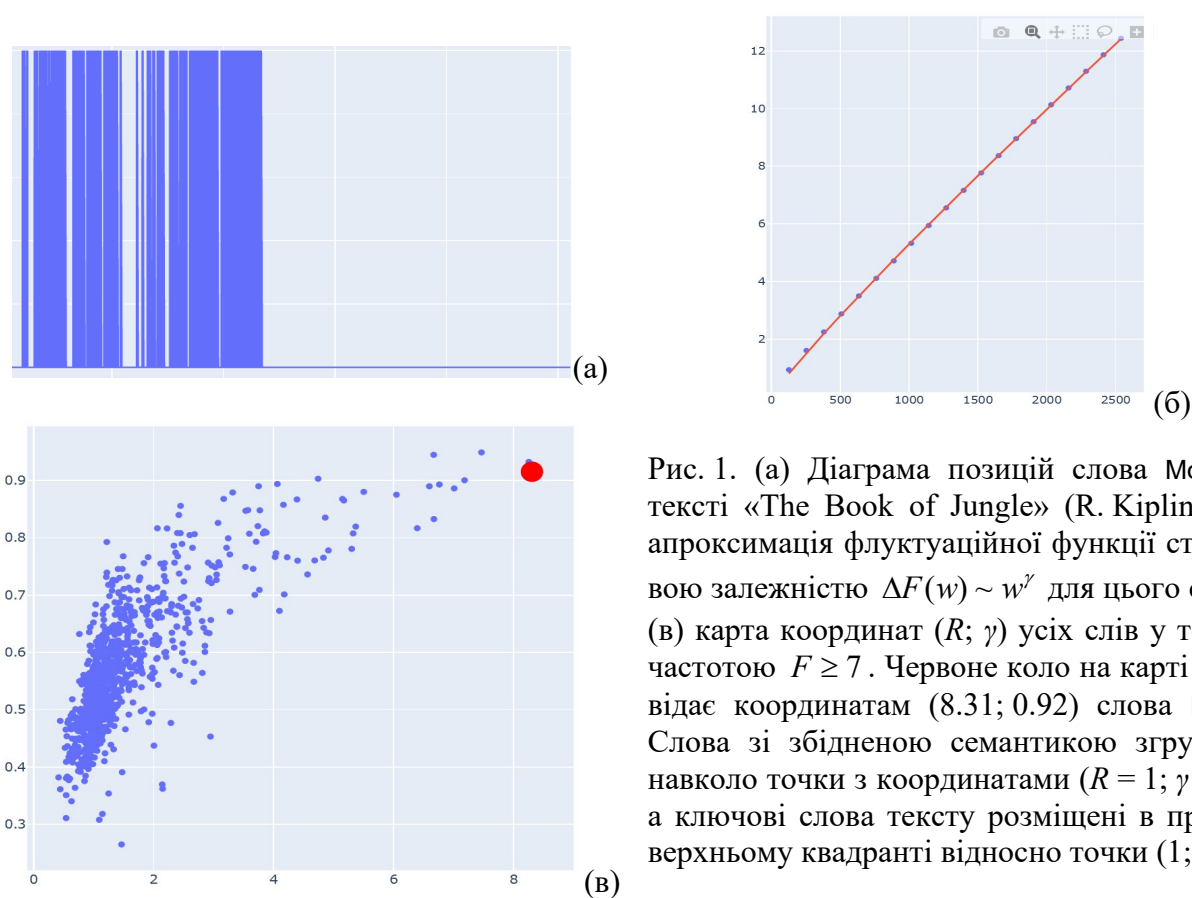


Рис. 1. (а) Діаграма позицій слова Mowgli у тексті «The Book of Jungle» (R. Kipling), (б) апроксимація флуктуаційної функції степеневною залежністю $\Delta F(w) \sim w^\gamma$ для цього слова і (в) карта координат $(R; \gamma)$ усіх слів у тексті з частотою $F \geq 7$. Червоне коло на карті відповідає координатам $(8.31; 0.92)$ слова Mowgli. Слова зі збідненою семантикою згруповані навколо точки з координатами $(R = 1; \gamma = 0.5)$, а ключові слова тексту розміщені в правому верхньому квадранті відносно точки $(1; 0.5)$.

На Рис. 1 проілюстровано явища кластеризації та ДСК для слова Mowgli, а також показано координати $(R; \gamma)$ усіх слів у тексті, що пройшли частотний фільтр $F = 7$. Із Рис. 1в видно, що ключові слова на карті розміщені в правому верхньому куті, де обидва параметри R і γ найвищі. Отже, ДСК слугують якісним критерієм релевантності слів у текстах, а параметр ДСК γ можна використовувати для визначення ключових слів. Ще вищою ефективністю володітиме нормований добуток на зразок $(R\gamma)/(R_{\max}\gamma_{\max})$.

Бібліографія

1. J. P. Herrera, P. A. Pury, Eur. Phys. J. B, **63**, 135 (2008).
2. О. Кушнір, А. Волоско, Л. Іваніцький, С. Рихлюк, Електрон. та інф. технол., **6**, 155 (2016).
3. K.-I. Goh, A.-L. Barabási, Europhys. Lett., **81**, 48002 (2008).
4. E. G. Altmann, G. Cristadoro, M. D. Esposti, Proc. Natl. Acad. Sci., **109**, 11582 (2012).