ВИКОРИСТАННЯ LSA-АЛГОРИТМУ ДЛЯ ПОПЕРЕДНЬОГО АНАЛІЗУ ТЕКСТІВ

Дудник О.В., Євсіна Н.О.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Аналіз текстів можна віднести до допоміжних завдань, які вирішуються при побудові складних систем управління і систем підтримки прийняття рішень. Серед них кластеризація, вибракування штучно синтезованих текстів тощо. Для цього широко застосовуються алгоритми латентно-семантичного аналізу (LSA). У [1] запропоновано LSA-алгоритм, що використовує сингулярне розбиття попередньо перетвореного корпусу текстів. Запропонована робота присвячена продовженню дослідження та удосконаленню модифікованого алгоритму, що було розпочато в [2].

З 15 текстів, кожен завбільшки за 1000 слів, після попередньої обробки, була сформована матриця розмірністю 4277 х 15, що містить вагу кожного слова у відповідному тексті. Здійснивши її сингулярний розклад, отримали три матриці, зміст яких інтерпретується так: теми текстів (матриця розкладання), слова і теми, тексти і теми. Матриця текстів і тем має розмірність 15 х 15, тобто. передбачається, що кожен текст присвячений одній основній темі, але містить додаткові теми, що характеризуються меншим ваговим значенням. Це припущення підтверджується результатом розкладання: у кожному стовпці матриці, що відповідно належить своєму текстові, міститься своє максимальне значення, що також мусить бути максимумом у відповідному рядку до кожної теми.

Якщо прийняти за осі координатної площини дві теми і розмістити на цій площині тексти відповідно до їх тематичних ваг, то більшість текстів сформує хмару поблизу початку координат, позаяк це тексти, в яких наведені теми представлені слабкіше. Два тексти, що мають максимальні ваги для однієї з вибраних тем, будуть відсунені від початку координат, маючи зрушення вздовж осі пануючої теми. Також буде помітно 2 – 3 тексти, що дистанціюються від «нульової» хмари, одночасно їх вагові значення суттєво менші за максимум – це тексти, в яких обрані теми мають другорядне значення. Для аналізу слід обирати теми, що містять найбільший сингулярний номер.

Дослідження було проведено з використанням середовища MATLAB, деякі модулі було втілено мовою C++. Подальші дослідження спрямовані на удосконалення попередньої обробки текстів.

Література:

1. Алгоритм LSA для поиска похожих документов. // [Электронный ресурс] –URL: https://netpeak.net/ru/blog/algoritm-lsa-dlya-poiska-pohozhih-dokumentov/

2. Дудник А.В. Модуль предварительного анализа текстов / А.В. Дудник, Н.А. Евсина, Е.В. Клевцова / Актуальні проблеми автоматики та приладобудування: матеріали III Міжнародної науково-технічної конференції 3 – 4 грудня 2020 – Харків: ФОП Панов А.М., 2020, с. 13 – 14.