Міністерство освіти і науки України
Сумський державний університет
Шосткинський інститут Сумського державного університету
Фармацевтична компанія «Фармак»
Управління освіти Шосткинської міської ради
Виконавчий комітет Шосткинської міської ради

ОСВІТА, НАУКА ТА ВИРОБНИЦТВО: РОЗВИТОК І ПЕРСПЕКТИВИ

МАТЕРІАЛИ

ІІ Всеукраїнської науково-методичної конференції,

(Шостка, 20 квітня 2017 року)

Суми
Сумський державний університет
2017

УДК УДК 519.237.8:519.254:004.9

МАТЕМАТИЧНЕ ТА ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ КЛАСТЕРИЗАЦІЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ

М.Г. Сидорова, І.С. Коробіхіна
Дніпровський національний університет імені Олеся Гончара
пр. Гагаріна 72, м. Дніпро, 49050
korobihina.inna@gmail.com

Останнім часом кількість інформації значно збільшилася і продовжує рости. При цьому отримання доступу до необхідних даних стає все більш проблематичним. Всі ці дані потрібно упорядковувати для зручності зберігання та швидкості і простоти знаходження потрібної інформації. Певним чином у цьому може допомогти кластеризація текстових документів – розподілення вибірки текстів на групи (кластери). До кластеру входять документи, які схожі між собою за змістом. При цьому кожна пара кластерів повинна містити максимально різні дані.

Проблема аналізу текстів з їх подальшою кластеризацією є досить актуальною, оскільки механізми порівняння стилів текстів є корисними в різних предметних галузях.

Робота присвячена розробці математичних моделей та створенню на їх основі програмного забезпечення кластерного аналізу текстових документів. Задана множина текстових файлів $X = \left\{ x_i ; i = \overline{1, N} \right\}$. Необхідно розподілити документи на групи $G = \{ g_1, g_2, \ldots, g_K \}$, $\bigcup_{i=1}^{K} g_i = X$, $g_i \cap g_j = \varnothing$, $i, j = \overline{1, K}, i \neq j$ за схожістю їх змісту.

Для виявлення груп семантично схожих текстів серед заданої фіксованої множини документів застосовувалися латентно-семантичний аналіз, метод K-means та гібридний метод FastDBSCAN. Гібридні підходи є досить перспективними, оскільки дозволяють досягти високої точності та швидкості аналізу текстових документів.

Перед застосування методів кластеризації було проведено попередню обробку текстової інформації, що складається з таких етапів: завантаження текстових документів, побудова матриці термів, виключення стоп-слів (слів, які є загальновживаними і не відносяться до якоїсь конкретної тематики, наприклад, займенники, сполучники, прийменники і т.п.), стеммінг, побудова матриці терми-на-документи, яка і є вхідними даними для методів кластеризації.

Для вирішення задачі стеммінгу було обрано метод Портера, який для знаходження основи слова не використовує бази основ слів, а працює, послідовно застосовуючи ряд правил відсікання закінчень і суфіксів.

Створене програмне забезпечення зручне у використанні та детально протестоване. Здійснено апробацію на різноманітних наборах текстових документів. Для оцінювання якості отримуваних результатів реалізовано зовнішні критерії якості: Ренда, Жакарда та Фолка–Меллоу. У якості мір подібності було обрано евклідову та косинусну метрики. При проведенні апробації можна було наглядно впевнитися в залежності точності кластеризації від обраної розмірності факторного простору (параметр $\mathbf{k}$). Це підтверджує твердження, що цей параметр необхідно підбирати емпірично. Відносно методу K-means, можна впевнитися у його недостатній точності та наглядно побачити, що основним недоліком алгоритму є дуже велика чутливість до обрання початкових центрів кластерів. Чим точніше будуть обрані початкові центри, тим точніше буде проведена кластеризація.