Бєгун А.В., Білошицький О.В. **УДК 004.91+004.946**

КВАЗІДИНАМІЧНЕ МОДЕЛЮВАННЯ АНАЛІЗУ ВІРТУАЛЬНИХ ТЕКСТІВ

Вступ. Ринок методів аналізу текстів відображає багатоплановість задач аналізу в різноманітних сферах діяльності людини. Огляд таких методів наводиться в спеціальних довідниках, що містять скорочений опис їх призначення, вимоги до технічних характеристик, відомості про додаткові сервісні можливості, ціни та інше.

Разом з тим необхідно відмітити, що значна частина такої інформації швидко старіє. Це зв'язано з умовами жорсткої конкуренції, де відбувається процес консолідації і на якому пропонується найкращий продукт. Одним із сучасних підходів до аналізу текстів являється використання методу кластеризації текстів – **kmeans.**

Основні результати. На відміну від ієрархічної кластеризації [5], в моделі kmeans задається кількість кластерів, які ми хочемо отримати. Формування такої моделі полягає в наступному:

1) вибирається в *n*-вимірному векторному просторі випадковим чином кількість початкових центрів *k* (середніх, *means*);

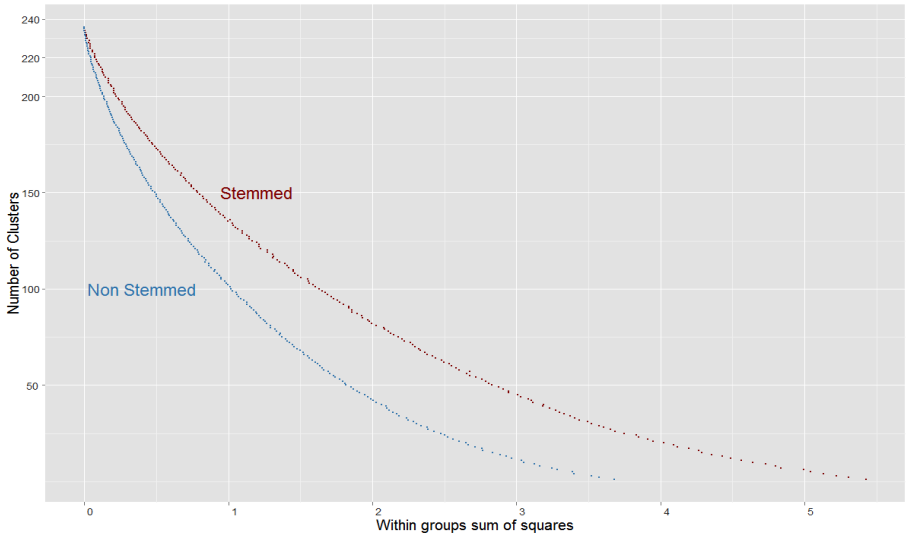
2) створюється *k* кластерів шляхом асоціації кожного спостереження з найближчим *середнім.* Такі кластери представляють собою діаграму Вороного, що сгенерована *середніми* значеннями[1];

3) центроїди (центри тяжіння) кожного з *k* кластерів становляться новими *середніми*;

4) кроки 2) і 3) повторюються до тих пір, поки не буде досягнуто повного сходження, тобто, поки остаточні кластери не будуть створені (умова повного сходження досягається тоді, коли сума квадратів відстаней між елементами спостережень та центроїдом на наступній ітерації не зменшується).

Встановлено, що одним з найбільш важливих кроків при кластеризації за допомогою даної моделі є вибір кількості центрів (кластерів). А тому виникає потреба в оцінюванні похибки для різноманітної кількості кластерів як суми квадратів відхилень (рис. 1).

Як бачимо з результатів оцінювання похибки – текст без стемінгу має меншу похибку для різних варіантів кластерів. Тобто, в першому наближенні кластеризація нестемінгованого тексту дає кращі результати поки кількість кластерів не досягне 236 (кількість спостережень). Це дійсно так, оскільки нестемінгований текст матиме меншу похибку, проте рівень якості таких кластерів буде набагато меншим ніж рівень кластеризації стемінгованого тексту.



$$\lim_{WSS \to 0} \left[N_{ns} - N_s \right] = 0$$

Рис. 1. Оцінювання похибки певної кількості кластерів.

Тут

- *WSS* (within groups sum of squares) – сума квадратів відхилень кластеру;
- N_{ns} – кількість кластерів для тексту без стемінгу;
- N_s – кількість кластерів для тексту після стемінгу.

Так, виходячи з попередніх результатів, виконаємо кластеризацію для 20 кластерів. Вхідними параметрами моделі будуть наступні:

>dtm.clust<-kmeans(x=dtm.k, centers=20, iter.max=40, nstart=10, algorithm="Hartigan-Wong"),

де

- *x=dtm.k* – терм-матриця DocumentTermMatrix, що нормалізована методом TF-IDF з 349 термінами, відфільтрована від термінів з низькою частотою входжень з коефіцієнтом спарингу 0.9[1];
- *centers* – кількість початкових центрів (кластерів);
- *iter.max* – максимальна кількість ітерацій, що виконуються до повного зходження. На останній ітерації визначаються кінцеві кластери;
- *nstart* – кількість стартів моделі, тобто, кількість проходжень усіх ітерацій моделі. В кінцевому підсумку будуть обрані кластери, які мають найменшу сумарну похибку в межах певного старту;
- *algorithm* – визначає алгоритм моделі kmeans. Алгоритм Хартігана-Вонга є найбільш поширеним при виконанні кластеризації kmeans.

Дійсно, неможливо виконати кластеризацію із занадто малою кількістю термінів (наприклад, 75 термінів при коефіцієнті спарингу 0.8), оскільки 75 термінів для 237 документів є досить малою кількістю для кластеризації. З іншого боку, неможливо виконати якісний кластерний аналіз при досить великій кількості спарсових термінів (наприклад, 4105 термінів з коефіцієнтом спарсингу 0.998), оскільки через велику кількість спарсових термінів загальна похибка буде занадто високою і сенс кластеризації буде втрачено [2]. А тому необхідно завжди ретельно оцінювати і знаходити компроміс між кількістю термінів, кількістю спарсових термінів і кількістю спостережень (об'єктів).

В результаті проведеної кластеризації отримано кластери наступних розмірів:

>dtm.clust$size

[1] 41 **21 4** 1 1 **5** 1 **7 12 5** 98 **2 3 7 10** 1 **4 2 1 11**

Відповідні похибки для отриманих кластерів становлять:

>dtm.clust$withinss

[1] 0.75166171 **0.37998302 0.08702162** 0.00000000 0.00000000 **0.10884947** 0.00000000 **0.21350480 0.22052166**

[10] **0.07426058** 1.35245927 **0.03003547 0.05145358 0.12662083 0.25722734** 0.00000000 **0.08037547 0.02691182**

[19] 0.00000000 **0.22561816**

Як бачимо, кластери одиничного розміру мають нульову похибку, і не є цікавими для подальшого аналізу. В той же час, отримані кластери з великим значенням похибки як раз ілюструють проблему аналізу в умовах обмеженості спостережень. В умовах даного дослідження найбільш репрезентативними є ті кластери, що мають відносно невелике значення похибки (виділені напівжирним).

Візуалізація кластерів.

Для графічної ілюстрації кластерів необхідно виконати трансформацію кластерів з *n*-мірного простору (де *n*=349 – кількість елементів вектору) в двомірний. Для цього виконаємо розрахунок відстаней

[1] Відфільтровані терміни, що мають нульове значення входжень в 90% документів корпусу.

Евклідовим методом (середньоквадратичні відхилення). В результаті трансформації отримаємо діаграму візуалізації кластерів (рис.2).

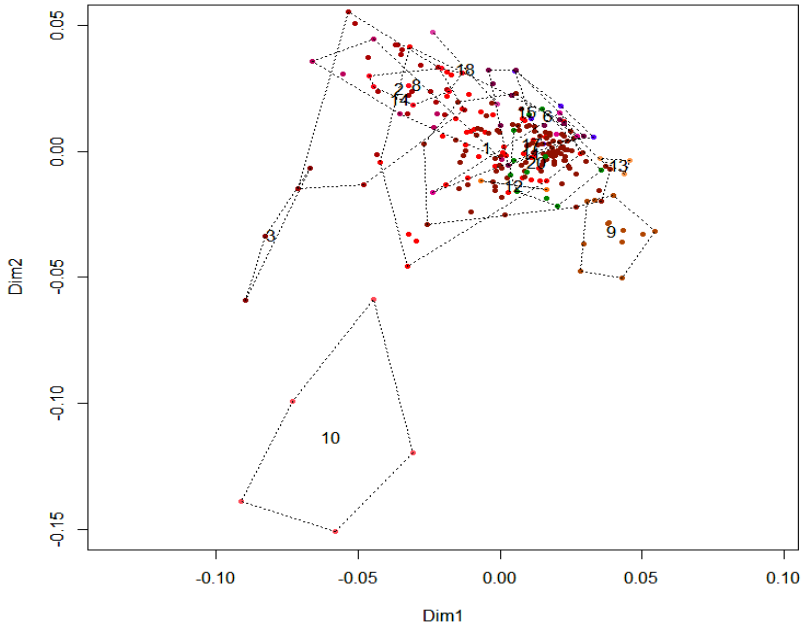


Рис. 2. Візуалізація кластерів.

Таким чином, виконано кластеризацію терм-матриці з 349 термінів методом kmeans. Дана кластеризація виконана на основі розрахунку центроідів (*середніх*), при цьому центроіді не належать множині елементів ($\{c \notin d\}$).

Як було показано раніше, для оцінки похибки розраховуються середньоквадратичні відхилення, де похибка зменшується пропорційно збільшенню кількості кластерів. Після проведення стемінгу, терм-матриця *DTM* має більше значення похибки та більшу якість кластерів.

Таким чином, модель *kmeans* є досить ефективною при кластеризації в умовах коли кількість елементів вектору (349 термінів) більша за загальну кількість спостережень (237).

Кластеризація методом kmedoids.

Аналогічно до попереднього методу *kmeans,* виконаємо кластеризацію тієї ж самої терм-матриці методом *kmedoids.* Основна відмінність в даному випадку полягатиме в тому, що початкові центри будуть обрані із множини об'єктів (кількість спостережень – 237) [3]. Як і в попередньому випадку, кожний елемент множини об'єктів представлений у вигляді вектора і складається із 349 елементів.

Виконаємо аналіз доцільної кількості кластерів для певного часу (рис. 3).

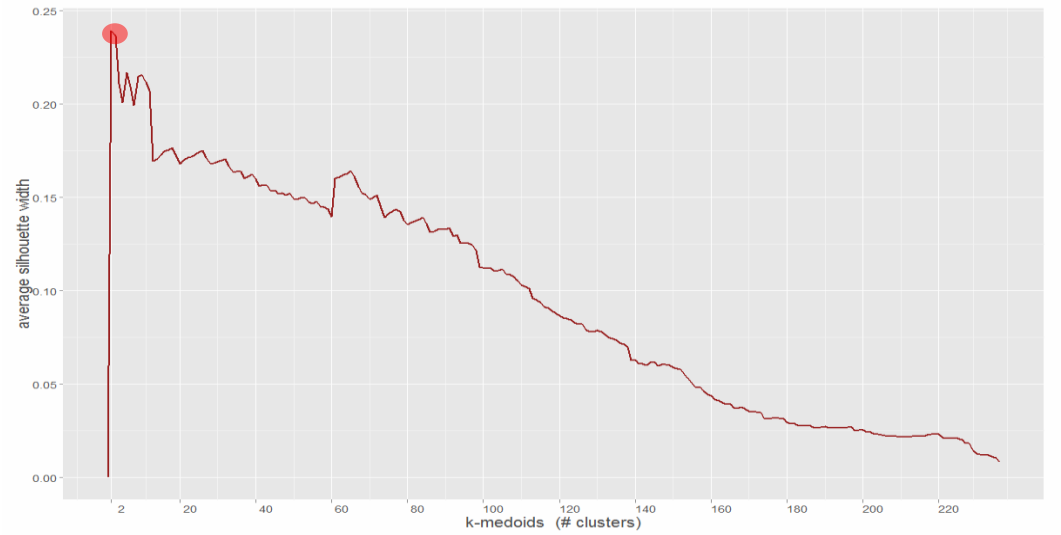


Рис. 3. Аналіз кількості кластерів методом *kmedoids.*

Як бачимо, найкраще значення кількості кластерів (при найменшій сумарній похибці) становить 2 кластери. Цей факт досить суттєво відрізняє існуючумодель від попередньої, і говорить про її чуттєвість до початково обраних центрів. При цьому якість кластеризації погіршується при збільшенні кількості кластерів до 60, і далі від 75 до 237.

З метою репрезентативності аналізу оберемо кількість кластерів 20, що має більшу сумарну похибку, проте в умовах обмеженості спостережень це є вимушеною необхідністю.

В результаті отримано наступні кластери

	size	max_diss	av_diss	diameter	separation
[1,]	178	0.2239088	0.13819608	0.3092605	0.08002509
[2,]	3	0.2214967	0.07383222	0.2214967	0.17892849
[3,]	4	0.1998083	0.14066632	0.2059998	0.17462258
[4,]	36	0.2231504	0.16221406	0.3089206	0.08002509
[5,]	1	0.0000000	0.00000000	0.0000000	0.25646612
[6,]	1	0.0000000	0.00000000	0.0000000	0.23393322
[7,]	1	0.0000000	0.00000000	0.0000000	0.24217552
[8,]	1	0.0000000	0.00000000	0.0000000	0.21947324
[9,]	1	0.0000000	0.00000000	0.0000000	0.23903947
[10,]	1	0.0000000	0.00000000	0.0000000	0.25999796
[11,]	1	0.0000000	0.00000000	0.0000000	0.25263192
[12,]	1	0.0000000	0.00000000	0.0000000	0.23012385
[13,]	1	0.0000000	0.00000000	0.0000000	0.27107264
[14,]	1	0.0000000	0.00000000	0.0000000	0.25148739
[15,]	1	0.0000000	0.00000000	0.0000000	0.22102618
[16,]	1	0.0000000	0.00000000	0.0000000	0.24236543
[17,]	1	0.0000000	0.00000000	0.0000000	0.23445565
[18,]	1	0.0000000	0.00000000	0.0000000	0.21347163
[19,]	1	0.0000000	0.00000000	0.0000000	0.22173467
[20,]	1	0.0000000	0.00000000	0.0000000	0.20469086

Тут

- *Size* – розмір кластеру (кількість термінів в кластері);
- *Max_diss, av_diss* – відповідно максимальна та середня відстань між елементами кластеру та *модоїдом* кластера (*центром*);
- *Diameter* – максимальна відстань між двома елементами кластеру;
- *Separation* – мінімальна відстань між елементом кластеру та елементом іншого кластеру.

Отже, отримано 4 кластери. Виконаємо їх візуальну інтерпретацію шляхом трансформації кластерів з *n*-мірного простору (рис. 4) (де *n=237* – кількість елементів вектору об'єкта) аналогічно до попередніх результатів. Маємо

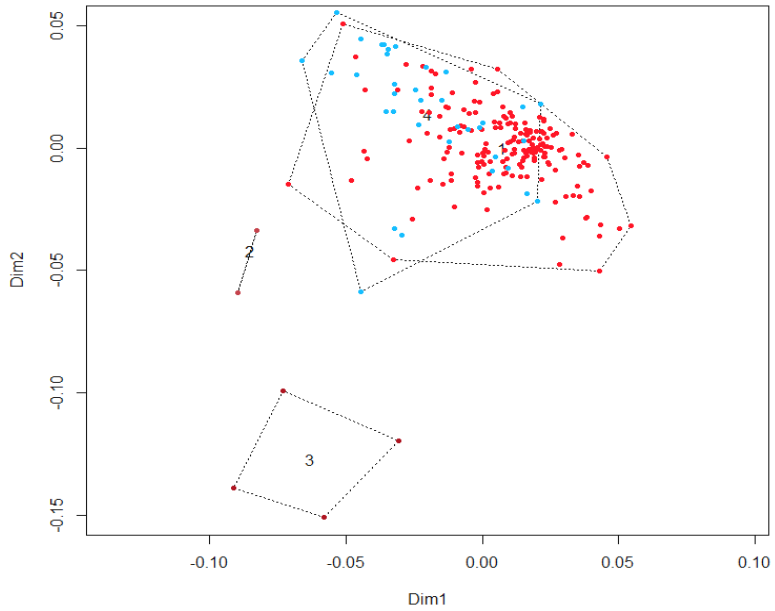


Рис. 4. Трансформація кластерів з *n*-мірного простору.

Таким чином, кластеризація за допомогою моделі *kmedoids* грунтується на розрахунку центроідів (*медоід*). Кількість оцінених при цьому кластерів є очікуваною, оскільки початкові центри кластерів було обрано із множини елементів (спостережень), $\{m \in d\}$. Похибка кластеризації при цьому збільшується пропорційно збільшенню кількості кластерів (на відміну від *kmeans*). Дана модель є досить схожою до попередньої з точки зору якості результатів, і є досить ефективною при аналізі в умовах обмеженості спостережень (тобто, в умовах, коли кількість термінів терм-матриці є більшою за кількість елементів матриці (спостережень)).

Після розпізнавання кластерів отримаємо наступні результати:

Загальна кількість розпізнаних постів – 91 із 236 (~40%).

Висновки. За результатами проведеного дослідження можна стверджувати, що для проведення якісної кластеризації текстів необхідно, щоб кількість елементів вектору даних була значно меншою за кількість спостережень (245). Відповідно до отриманих векторів, деякі кластери можуть містити «неподібні» пости (див. середньоквадратичні відхилення). При цьому варто виконати кластеризацію методом *kmeans* для кращих результатів в умовах, коли необхідно аналізувати великі **неподібні** тексти із малою кількістю спостережень. В іншому випадку *kmedoids* зазвичай дає кращі результати (рис. 5).

Таким чином, для отримання більш якісних результатів необхідно акцентувати дослідження на більш подібних текстах (наприклад, виконати їх первісну категоризацію за темами, часовим інтервалом тощо).

В якості подальшого розвитку даного дослідження досить важливо також проводити **семантичний** аналіз текстів [4], який дасть змогу не лише кластеризувати тексти, але і визначити характер того чи іншого кластера (тексту), наприклад, негативний/позитивний/нейтральний текст тощо.

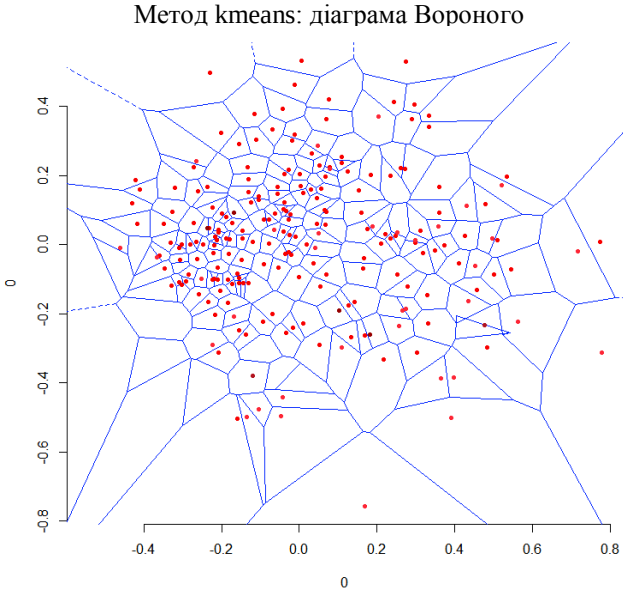


Рис. 5. Ілюстрація кластерів (heatmap) для kmeans.

Джерела та література:

1. Voronoi diagram : [Електронний ресурс]. – Режим доступу : http://en.wikipedia.org/wiki/Voronoi_diagram
2. TextMining with R : [Електронний ресурс]. – Режим доступу : http://www.slideshare.net/whitish/textmining-with-r
3. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications / G. Miner, J. Elder, T. Hill, R. Nisbet, D. Delen, A. Fast. – Elsevier Academic Press, 2012.
4. Indurkhya N. Handbook of Natural Language Processing / N. Indurkhya, F. Damerau. – 2nd Edition. – Boca Raton, FL : CRC Press, 2010.
5. Бєгун А. В. Ієрархічна кластерізація текстів в умовах обмеженості спостережень / А. В. Бєгун, О. В. Білошицький // Культура народов Причорноморья. – 2012. – № 233. – С. 15-19.