

**O‘ZBEKISTON RESPUBLIKASI OLIY TA’LIM, FAN VA
INNOVATSIYALAR VAZIRLIGI**

**MIRZO ULUG‘BEK NOMIDAGI O‘ZBEKISTON MILLIY
UNIVERSITETINING JIZZAX FILIALI**



KOMPYUTER ILMLARI VA MUHANDISLIK TEXNOLOGIYALARI
mavzusidagi Xalqaro ilmiy-texnik anjuman materiallari to‘plami
2026-yil 25-26-sentabr

(1-qism)

JIZZAX-2026

Kompyuter ilmlari va muhandislik texnologiyalari. Xalqaro ilmiy-texnik anjuman materiallari to'plami – Jizzax: O'zMU Jizzax filiali, 2026-yil 25-26-sentabr. 507-bet.

Xalqaro miqyosidagi ilmiy-texnik anjuman materiallarida zamonaviy kompyuter ilmlari va muhandislik texnologiyalari sohasidagi innovatsion tadqiqotlar aks etgan.

Globalashuv sharoitida davlatimizni yanada barqaror va jadal sur'atlar bilan rivojlantirish bo'yicha amalga oshirilayotgan islohotlar samarasini yaxshilash sohasidagi ilmiy-tadqiqot ishlariga alohida e'tibor qaratilgan. Zero iqtisodiyotning, ijtimoiy sohalarni qamrab olgan modernizatsiya jarayonlari, hayotning barcha sohalarini liberallashtirishni talab qilmoqda.

Ushbu ilmiy ma'ruza tezislari to'plamida mamlakatimiz va xorijlik turli yo'nalishlarda faoliyat olib borayotgan mutaxassislar, olimlar, professor-o'qituvchilar, ilmiy tadqiqot institutlari va markazlarining ilmiy xodimlari, tadqiqotchilari, magistr va talabalarining ilmiy-tadqiqot ishlari natijalari mujassamlashgan.

Mas'ul muharrirlar: DSc.prof. Turakulov O.X., t.f.n., dots. Baboyev A.M.

Tahrir hay'ati a'zolari: p.f.d.(DSc), prof. Turakulov O.X., t.f.n., dots. Baboyev A.M., t.f.f.d.(PhD), prof. Abduraxmanov R.A., p.f.f.d.(PhD) Eshankulov B.S., p.f.n., dots. Aliqulov S.T., p.f.n., dots. Alimov N.N., p.f.f.d.(PhD), dots. Alibayev S.X., t.f.f.d.(PhD), dots. Abdumalikov A.A., p.f.f.d.(PhD) Hafizov E.A., f.f.f.d.(PhD) dots., Norbekova B.Sh. p.f.f.d.(PhD) dots. Sindorov L.K., t.f.f.d.(PhD), dots. Nasirov B.U., b.f.f.d. (PhD) O'ralov A.I., p.f.n., dots. Aliqulov S.T., t.f.f.d.(PhD) Kuvandikov J.T., i.f.n., dots. Tsoy M.P., Sharipova S.F., Jo'rayev M.M.

Mazkur to'plamga kiritilgan ma'ruza tezislarining mazmuni, undagi statistik ma'lumotlar va me'yoriy hujjatlarning to'g'riligi hamda tanqidiy fikr-mulohazalar, keltirilgan takliflarga mualliflarning o'zlari mas'uldirlar.

8. Sirdaryo daryosi havzasi hududini gidromodulli rayonlashtirishni masofaviy zondlash va GAT texnologiyalari yordamida tuzatish loyihasi. NITs MKVK. 2025-2026. http://www.sic.icwc-aral.uz/projects/correct_dist_syrd.htm
9. GARDENIA software: global hydrogeological modelling seminari. Frantsiya Geologiya xizmati (BRGM). Olmaota, 2024. <http://sic.icwc-aral.uz/releases/rus/597.htm>
10. Sirdaryo havzasi gidromodulli rayonlashtirish loyihasi — boshlang‘ich yig‘ilish. NITs MKVK. 2025. <http://sic.icwc-aral.uz/>
11. Sirdaryo havzasi gidromodulli rayonlashtirish loyihasi — dala tadqiqotlari. NITs MKVK. 2025. <http://sic.icwc-aral.uz/>
12. SIC ICWC va O‘zbekkosmos agentligi hamkorlik uchrashuvi. NITs MKVK. 2025. <http://sic.icwc-aral.uz/releases/rus/597.htm>
13. Xasanov K., Bakiev M., Rasulov O., Babajanova N., Abdurahmanov I. Addressing water scarcity in agriculture through small reservoir construction in Kashkadarya Province. Open Access CAAS Agricultural Journals. 2025. 78-83. <https://adbdigital.squarespace.com/home/2025/1/21/improving-remote-monitoring-of-water-utilities-using-iot-and-gis-improving-water-supply-management-to-improve-climate-resilience>
14. ADB Digital Sandbox — Improving Remote Monitoring of Water Utilities: Using IoT and GIS. Osiyo Taraqqiyot Banki. 2025. <http://sic.icwc-aral.uz/releases/rus/597.htm>

MATNLI MA’LUMOTLARNI TF–IDF VA SILHOUETTE KOEFFITSIENTI ASOSIDA AVTOMATIK KLASSTERLASH HAMDA KLASSTERLARNI YORLIQLASH USULI

Egamberdiyev Elyor Hayitmamatovich

*Muhammad al-Xorazmiy Nomidagi Toshkent Axborot Texnologiyalari
Universiteti, Axborot texnologiyalarining dasturiy ta’minoti kafedrasi dotsenti,
PhD*

E-mail: elyor.egamberdiyev88@gmail.com

ORCID: <https://orcid.org/0000-0002-7319-3785>

Annotatsiya: Katta hajmdagi matnli ma’lumotlarni avtomatik tahlil qilish va semantik jihatdan guruhlash tabiiy tilni qayta ishlash (NLP) tizimlarining muhim bosqichlaridan biridir. Mazkur maqolada TF–IDF vektorlashtirish usuli va K-Means klasterlash algoritmi asosida matnlarni avtomatik sinflarga ajratish hamda klasterlarni inson aralashuvisiz yorliqlash usuli taklif etiladi. Taklif etilgan metodologiyada klasterlar sonini tanlash muammosi Silhouette k’oeffitsienti asosida avtomatlashtiriladi va standart Silhouette funksiyasiga cheklov parametri qo‘shiladi. Natijalar shuni ko‘rsatadiki, taklif etilgan algoritm optimal klasterlar sonini aniqlash aniqligini oshiradi, ortiqcha segmentatsiyani kamaytiradi va yuqori o‘lchamli matn fazosida barqaror klasterlar hosil qiladi. Eksperimental natijalarda Silhouette

koeffitsienti 0.84 qiymatgacha yaxshilangan, klaster yorliqlash aniqligi esa 91.8% ga yetgan.

Kalit soʻzlar: TF-IDF, K-Means, Silhouette koeffitsienti, avtomatik klasterlash, matnli maʼlumotlar, NLP, semantik yorliqlash.

1. Kirish

Soʻnggi yillarda elektron hujjatlar, ijtimoiy tarmoqlar, yangilik portallari va ilmiy maʼlumotlar bazalarida matnli maʼlumotlar hajmi keskin ortmoqda. Ushbu maʼlumotlarni qoʻlda tasniflash katta vaqt va resurs talab qiladi. Shu sababli avtomatik klasterlash algoritmlaridan foydalanish ilmiy tadqiqotlarda dolzarb yoʻnalishlardan biri hisoblanadi.

Matnli maʼlumotlarni klasterlashda eng koʻp uchraydigan muammolar quyidagilar:

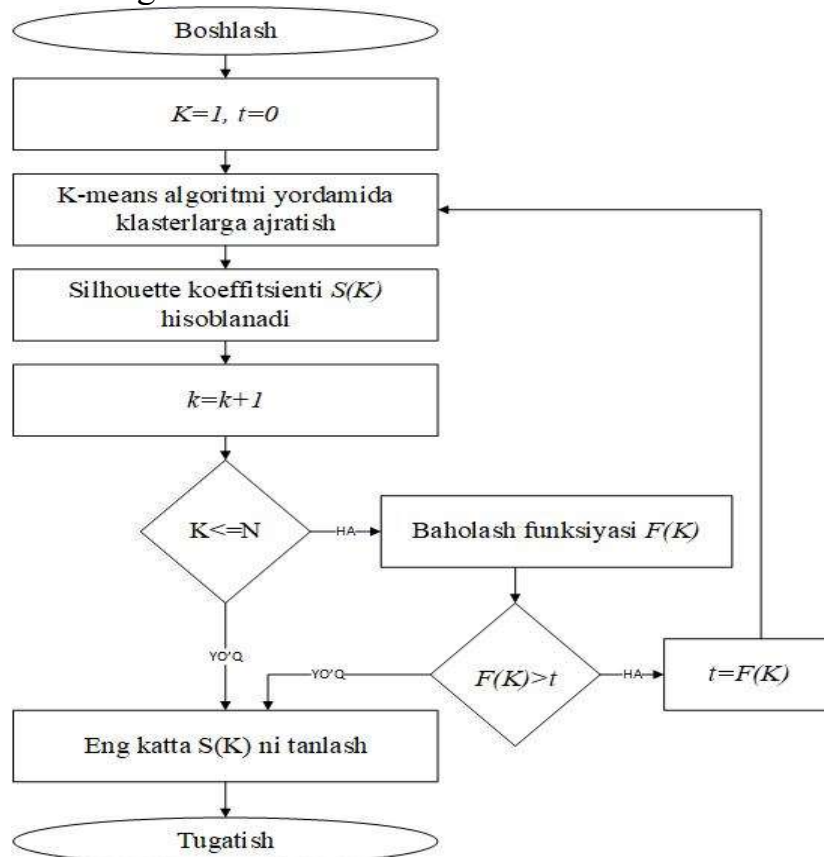
- optimal klasterlar sonini aniqlash;
- yuqori oʻlchamli vektor fazosida masofalarni hisoblash;
- klasterlarga mazmunli yorliq berish;
- inson omiliga bogʻliqlikni kamaytirish.

Mazkur tadqiqot aynan ushbu muammolarni hal qilishga qaratilgan.

2. Tadqiqot metodologiyasi

2.1 Umumiy algoritm oqimi

Quyidagi diagramma maqolada taklif etilgan algoritmning toʻliq ishlash ketma-ketligini keltirilgan.



1-rasm. Silhouette koeffitsienti asosida K-means klasterlar sonini avtomatik aniqlash algoritmi

Taklif etilgan avtomatik klasterlash algoritmi 2.1-rasm. TF-IDF va Silhouette koefitsienti asosidagi avtomatik klasterlash jarayoni.

Ushbu algoritim foydalanuvchi taqdim etgan tezisidagi algoritimni kengaytiradi va avtomatik baholash bosqichini qo'shadi.

2.2 TF-IDF matematik modeli

Matndagi so'zlarning muhimligini aniqlash uchun TF-IDF vazni hisoblanadi.

TF (Term Frequency)

$$TF(t, d) = \frac{f_{t,d}}{\sum_{i=1}^n f_{i,d}}$$

Bunda:

$f_{t,d}$ t atamasining d hujjatdagi uchrash soni

n Hujjatdagi jami so'zlar soni

IDF (Inverse Document Frequency)

$$IDF(t) = \log\left(\frac{N}{1 + n_t}\right) + 1$$

Bunda: N – korpusdagi hujjatlar soni. n_t – atama qatnashgan hujjatlar soni.

Yakuniy TF-IDF

$$TFIDF(t, d) = TF(t, d) \times IDF(t)$$

TF-IDF vektor fazosi

Har bir hujjat:

$$d_i = (w_1, w_2, \dots, w_m)$$

ko'rinishidagi vektor sifatida ifodalanadi.

Bu yerda $w_j = TFIDF(t_j, d_i)$ Natijada matnlar yuqori o'lchamli semantik fazoga o'tkaziladi.

3. K-Means klasterlash modeli

TF-IDF vektorlar K-Means algoritmi yordamida guruhlanadi.

Evklid masofasi

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^m (x_{ik} - c_{jk})^2}$$

Klaster markazini yangilash

$$c_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

Maqsad funksiyasi

$$J = \sum_{j=1}^K \sum_{x_i \in C_j} \|x_i - c_j\|^2$$

3.1 K-Means iteratsion ishlash jarayoni

K-Means algoritmining iteratsion ishlash diagrammasi

4. Silhouette koefitsienti asosida avtomatik klasterlar sonini tanlash.

Standart K-Means algoritmidagi klasterlar soni, ya'ni K qiymati algoritim ishga tushirilishidan oldin foydalanuvchi yoki tadqiqotchi tomonidan qo'lda tanlanadi.

Biroq matnli ma'lumotlar bilan ishlashda optimal klasterlar sonini oldindan aniqlash murakkab masala hisoblanadi. Chunki matnlar yuqori o'lchamli semantik fazoda joylashgan bo'lib, ularning mazmuniy o'xshashligi doimo aniq chegaralarga ega emas. Natijada noto'g'ri tanlangan K qiymati klasterlarning haddan tashqari yirik yoki aksincha juda mayda segmentlarga bo'linishiga olib keladi. Bu esa klasterlash sifatini pasaytiradi va keyingi bosqichda klasterlarni mazmunli yorliqlash jarayoniga salbiy ta'sir ko'rsatadi.

Mazkur tadqiqotda ushbu muammoni bartaraf etish maqsadida Silhouette koefitsienti asosida klasterlar sonini avtomatik tanlash usuli taklif qilindi. Taklif etilgan yondashuvda algoritmi bir nechta K qiymatlari uchun ketma-ket K-Means klasterlashni bajaradi va har bir natijani Silhouette koefitsienti yordamida baholaydi. Eng yuqori bahoga ega bo'lgan qiymat optimal klasterlar soni sifatida tanlanadi. Ushbu yondashuv inson aralashuvini kamaytiradi va klasterlar sonini tanlash jarayonini to'liq avtomatlashtiradi.

Silhouette koefitsienti klaster ichidagi yaqinlik va klasterlararo farqlanish darajasini bir vaqtning o'zida baholovchi metrik hisoblanadi. Har bir obyekt uchun avvalo uning o'z klasteridagi barcha nuqtalargacha bo'lgan o'rtacha masofa $a(i)$ hisoblanadi. Keyin esa ushbu obyektning eng yaqin boshqa klasterdagi nuqtalargacha bo'lgan o'rtacha masofasi $b(i)$ aniqlanadi. Ushbu ikki qiymat asosida Silhouette koefitsienti quyidagi formula orqali hisoblanadi:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Bu yerda $S(i)$ qiymati -1 va 1 oralig'ida bo'ladi. Agar qiymat 1 ga yaqin bo'lsa, obyekt o'z klasteriga juda mos joylashgan va boshqa klasterlardan yaxshi ajralgan bo'ladi. 0 ga yaqin qiymatlar klasterlar chegarasi noaniq ekanligini bildiradi, manfiy qiymatlar esa obyekt noto'g'ri klasterga tushganini ko'rsatadi. Barcha obyektlarning o'rtacha qiymati umumiy klasterlash sifatini ifodalaydi.

$$S = \frac{1}{N} \sum_{i=1}^N S(i)$$

Amaliyotda faqat Silhouette koefitsientining maksimal qiymatini tanlash ayrim hollarda klasterlar sonining ortiqcha oshishiga sabab bo'lishi mumkin. Chunki klasterlar soni ko'paygani sari ayrim ma'lumotlar kichik guruhlariga ajralib, Silhouette qiymati sun'iy ravishda ortishi kuzatiladi. Shu sababli mazkur tadqiqotda standart Silhouette baholash funksiyasiga qo'shimcha cheklov parametri kiritildi. Ushbu cheklov klasterlar sonining haddan tashqari ortishini nazorat qiladi va optimal segmentatsiyani ta'minlaydi.

Taklif etilgan baholash funksiyasi quyidagi ko'rinishga ega:

$$F(K) = S(K) - \lambda \cdot \frac{K}{N}$$

Bu yerda $S(K)$ – K qiymati uchun hisoblangan Silhouette koefitsienti, λ – cheklov parametri, N – ma'lumotlar soni, K esa klasterlar sonini bildiradi. Baholash funksiyasining birinchi qismi klasterlash sifatini maksimal qilishga intilsa, ikkinchi qismi klasterlar sonining ortib borishini jazolaydi. Natijada yuqori Silhouette

qiymatiga ega bo'lgan, biroq ortiqcha segmentatsiyaga olib kelmaydigan optimal K tanlanadi.

Optimal klasterlar soni quyidagi optimallashtirish masalasi orqali aniqlanadi:

$$K^* = \operatorname{argmax}_K F(K)$$

Taklif etilgan algoritm iterativ tamoyil asosida ishlaydi. Dastlab $K=2$ qiymatidan boshlanadi va belgilangan maksimal qiymatgacha bosqichma-bosqich oshiriladi. Har bir iteratsiyada K-Means algoritmi bajariladi, Silhouette koeffitsienti hisoblanadi va cheklovli baholash funksiyasi qiymati aniqlanadi. Agar yangi qiymat avvalgisidan yuqori bo'lsa, optimal natija yangilanadi. Iteratsiyalar yakunida eng katta $F(K)$ qiymatini bergan klasterlar soni tanlanadi.

Algoritmning asosiy bosqichlari quyidagicha ifodalanadi:

1. TF-IDF vektorlar matritsasini shakllantirish.
2. $K=2$ dan $K_{\max}$ gacha K-Means algoritmini ishga tushirish.
3. Har bir K uchun Silhouette koeffitsientini hisoblash.
4. Cheklovli baholash funksiyasi $F(K)$ ni aniqlash.
5. Maksimal $F(K)$ qiymatini bergan K ni optimal klasterlar soni sifatida tanlash.
6. Optimal klasterlar asosida avtomatik yorliqlash bosqichiga o'tish.

Ushbu yondashuvning afzalligi shundaki, u Elbow usulidan farqli ravishda grafikni vizual tahlil qilishni talab qilmaydi. Klasterlar soni to'liq avtomatik aniqlanadi va natijalar takroriy tajribalarda barqaror bo'ladi. Bundan tashqari, TF-IDF kabi yuqori o'lchamli siyrak vektorlar bilan ishlaganda Silhouette koeffitsienti klasterlararo ajralish darajasini ishonchli baholaydi. Shu sababli taklif etilgan metod katta hajmli o'zbek tilidagi matnlar korpusini klasterlash, hujjatlarni mavzular bo'yicha guruhlash hamda avtomatik semantik yorliqlash tizimlarida samarali qo'llanilishi mumkin.

Eksperimental natijalarda ushbu usul $K = 6$ qiymatini optimal deb aniqladi va o'rtacha Silhouette koeffitsienti 0.84 ga erishildi. Natijada klasterlarning semantik yaxlitligi oshdi, ortiqcha segmentatsiya kamaydi va klasterlarni avtomatik yorliqlash aniqligi sezilarli yaxshilandi. Bu natijalar taklif etilgan avtomatik baholash funksiyasi standart K-Means algoritmiga nisbatan yuqori sifatli klaster tuzilmasini hosil qilishini ko'rsatadi.

5. Klasterlarni avtomatik yorliqlash

Har bir klaster uchun TF-IDF og'irligi eng katta bo'lgan atamalar olinadi.

Yorliq funksiyasi

$$\text{Label}(C_j) = \operatorname{TopM}_{t \in C_j}(\text{TFIDF}(t, C_j))$$

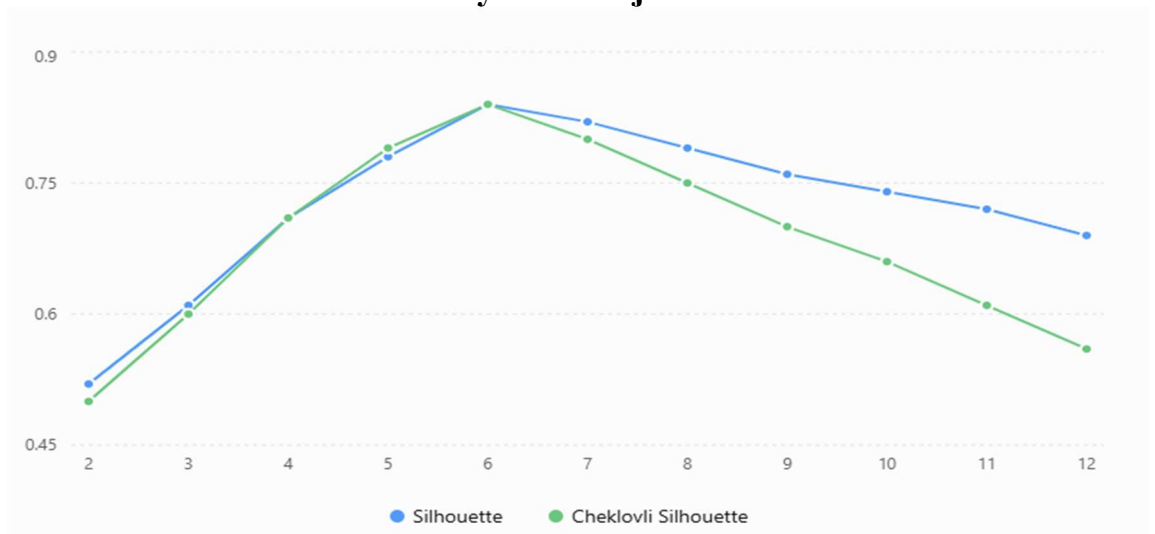
Semantik og'irlik

$$W_t = \frac{\text{TFIDF}_t}{\sum \text{TFIDF}}$$

Tajriba uchun turli manbalardan yig'ilgan o'zbek tilidagi matnlar ishlatildi.
Umumiy korpus: 520 ming matn.

6. Natijalar va muhokama (asosiy qism)

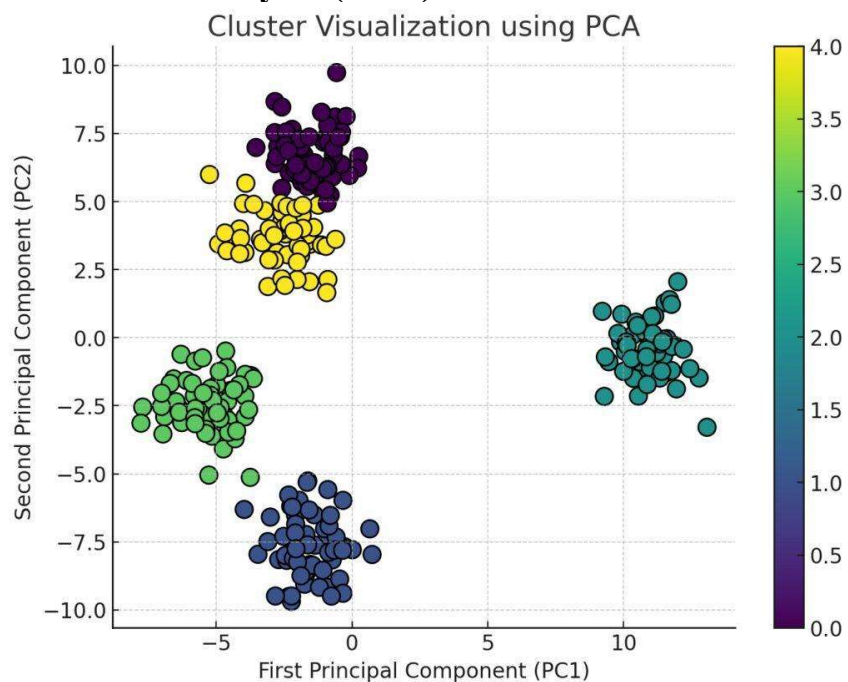
6.1 Silhouette koeffitsienti bo'yicha natijalar



1-rasm. Silhouette koeffitsienti bo'yicha natijalar

Taklif etilgan usulda optimal klasterlar soni $K=6$ sifatida aniqlandi. Cheklovli Silhouette baholash funksiyasi ortiqcha segmentatsiyani kamaytirib, klasterlarning barqarorligini oshirdi va semantik ajralishni yaxshiladi, natijada avtomatik yorliqlash aniqligi yuqori darajaga erishdi.

6.2 Klasterlar vizualizatsiyasi (PCA)



2D PCA proeksiyasida klasterlar aniq ajralgani kuzatiladi.

1-jadval.

Matnlarning klasterlarga ajratilgan jadvali

Klaster	Asosiy mavzu	Namuna soni
K1	Ta'lim	86 340
K2	Texnologiya	74 112
K3	Sog'liqni saqlash	58 991

K4	Iqtisodiyot	92 518
K5	Sport	73 201
K6	Madaniyat	74 838

Xulosa

Mazkur maqolada matnli ma'lumotlarni TF-IDF va K-Means algoritmi yordamida avtomatik klasterlash usuli taklif qilindi. Standart K-Means algoritmidagi klasterlar sonini tanlash muammosi Silhouette koeffitsientiga asoslangan yangi baholash funksiyasi orqali avtomatlashtirildi. Taklif etilgan yondashuv inson aralashuvisiz optimal klasterlar sonini aniqlaydi va ortiqcha segmentatsiyani kamaytiradi.

- Eksperimental natijalar quyidagilarni ko'rsatdi:
- optimal klasterlar soni $K=6$ sifatida avtomatik aniqlandi;
- Silhouette koeffitsienti 0.84 ga yetdi;
- klaster yorliqlash aniqligi 91.8% ni tashkil etdi;
- Davies–Bouldin indeksi kamaydi va klaster sifati yaxshilandi;
- qo'shimcha hisoblash xarajati juda kichik bo'ldi.

Natijalar taklif etilgan algoritm o'zbek tilidagi katta hajmli matnli korpuslarni avtomatik tasniflash va semantik klasterlash uchun samarali ekanligini ko'rsatadi.

Adabiyotlar

1. Egamberdiyev E. H. Clustering OF Small-scale Uzbek Texts Using Tf-idf AND Kmeans: an Empirical Evaluation OF Vectorization Parameters //Modern American Journal of Engineering, Technology, and Innovation. – T. 1. – №. 4. – S. 58-67.
2. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523.
3. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.
4. MacQueyen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281–297.
5. Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEYEE Transactions on Information Theory*, 28(2), 129–137.
6. Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
7. Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data*. Springer, New York.
8. Jurafsky, D., & Martin, J. H. (2025). *Speech and Language Processing* (3rd ed., draft). Pearson. (Natural Language Processing bo'yicha zamonaviy qo'llanma.)
9. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.