MULTITIPLI MA'LUMOTLARGA INTELLEKTUAL ISHLOV BERISH.

Mo'minov Bahodir Boltayevich[1], Egamberdiyev Elyor Hayitmamatovich[2].

[1]Toshkent davlat iqtisodiyot universiteti, t.f.d. professor
[2]Toshkent axborot texnologiyalari universiteti, doktarant

KALIT SO'ZLAR

Multitipli ma'lumotlar, Discrete Fourier Transform, Principal Component Analysis, K-Means Clustering, Tokenizatsiya, Stem-ming, Lemmatization, Word2Vec, TF-IDF, NLTK, spaCy.

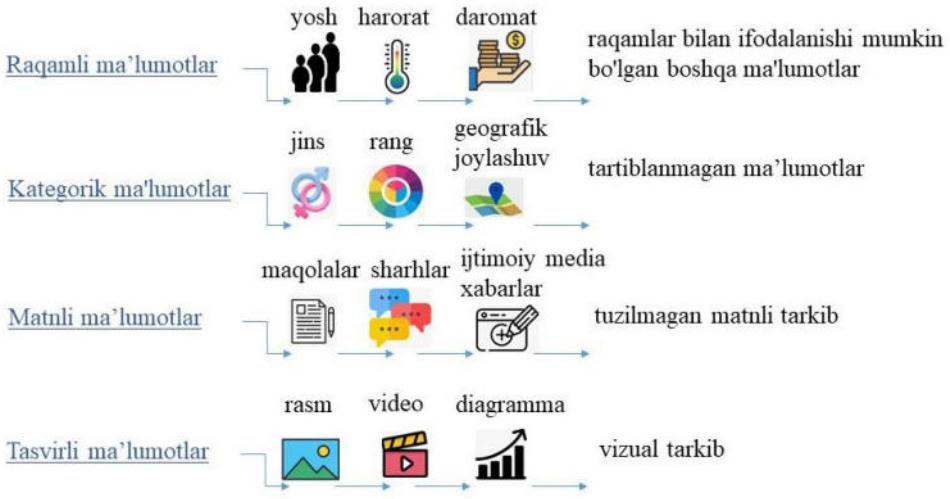
ANNOTATSIYA

Bigdatalarning paydo bo'lishi raqamli, toifali, matnli va tasvirli kabi turli xil ma'lumotlar turlarini o'z ichiga olgan davrni boshladi. Bunday multitipli ma'lumotlardan samarali foydalanish turli sohalarda, jumladan, mashinani o'rganish, data science va sun'iy intellektda hal qiluvchi vazifaga aylandi. Ushbu maqola multitipli ma'lumotlarni qayta ishlashni o'rganish, tahlil qilish va qaror qabul qilishni yaxshilash uchun ma'lumotlar manbalarini integratsiyalash usullarini, qiyinchiliklarini va imkoniyatlarini muhokama qilishdan iborat.

I.Kirish:

Turli xil ma'lumotlar turlarining yaqinlashishi ma'lumotlarni qayta ishlashga yaxlit yondashuvni talab qiladi, har bir turdagi o'ziga xos xususiyatlardan foydalanish uchun maxsus usullar birlashtirishi kerak. Multitipli ma'lumotlarni qayta ishlash chuqurroq tushunchalarni ochish, bashoratli modellashtirishni yaxshilash va murakkab hodisalarni yanada to'liqroq tushunish imkonini beradi.

Multitipli ma'lumotlar turli xil axborot manbalarini o'z ichiga oladi, ularning har biri o'ziga xos formati, tuzilishi va o'ziga xos qiyinchiliklari bilan ajralib turadi. Ushbu ma'lumotlar turlarining yagona ma'lumotlar to'plamida birlashishi Multitipli ma'lumotlarni qayta ishlash uchun asos bo'ladi. Multitipli ma'lumotlarning ba'zi keng tarqalgan turlari:



1-rasm. Multitipli ma'lumotlar.

Multitipli ma'lumotlarning xilma-xilligi raqamli, toifali, matnli va tasvirli ma'lumotlarni o'z ichiga oladi, ularning har biri tahlil qilish uchun alohida qiyinchiliklar va imkoniyatlarni taqdim etadi. Raqamli ma'lumotlar ko'pincha miqdoriy tahlilning asosini tashkil qiladi, kategorik ma'lumotlar esa kodlash va taqdim etish bilan bog'liq muammolarni keltirib chiqaradi. Matn va

© *Mo'minov Bahodir Boltayevich, Egamberdiyev Elyor Hayitmamatovich.*

tasvir ma'lumotlari esa, tuzilmagan ma'lumotni mazmunli xususiyatlarga aylantirish uchun maxsus usullarni talab qiladi.

Multitipli ma'lumotlarni qayta ishlashning muhim bosqichi har bir ma'lumot turiga moslashtirilgan texnikani o'z ichiga olgan dastlabki ishlov berishdir. Masalan raqamli qiymatlarga ishlov berish uchun quyidagi usullar mavjud.

Filtrlash usullari: null qiymatlarni olib tashlash yoki qiymatlardan ma'lum komponentlarni olish uchun ishlatiladi [1-2]. Filtrlash formulasi,

$$y[n] = \frac{1}{M}\sum_{i=0}^{M-1} x[n-i] \qquad (1)$$

Bu yerda M to'plamda elementlar soni, x[n] to'plam elementlari va y[n] mos filterlangan qiymat.

Discrete Fourier Transform (DFT): diskret signalning chastota komponentlarini tahlil qilish uchun ishlatiladi[3-4]. DFT uchun formula quyidagicha berilgan:

$$X[k] = \sum_{n=0}^{N-1} x[n] * e^{-j\frac{2\pi}{N}kn} \qquad (2)$$

Bu yerda x[n] kirish signali, n signal uzunligi, k chastota indeksi.

Principal Component Analysis (PCA): bu ma'lumotlarni yangi koordinatalar tizimiga aylantirish uchun ishlatiladigan o'lchamlarni kamaytirish texnikasi bo'lib, u farqni maksimal darajada oshiradi[5]. Asosiy komponentlarni topish formulasi o'z vektor tahlilini o'z ichiga oladi:

$$v_i = \frac{1}{\sqrt{\tau_i}} A^T x_i \qquad (3)$$

Bu yerda v_i asosiy komponent, τ_i xos qiymat, A ma'lumotlar to'plami, x_i A to'plamdagi *i*-chi ustun.

K-Means Clustering: bu nazoratsiz klasterlash algoritmi bo'lib, ma'lumotlarni quyidagilarga k klasterlarga ajratadi[6]. Algoritm kvadrat masofalarning klaster ichidagi yig'indisini minimallashtirishni o'z ichiga oladi:

$$J = \sum_{i=1}^{k}\sum_{x\in C_i} ||x-\mu_i||^2 \qquad (4)$$

Bu yerda C_i i-chi klaster, *x* ma'lumotlar nuqtasi, va μ_i C_i kalsterning o'rtachasi.

Kategorik ma'lumotlarni qayta ishlash kategorik o'zgaruvchilarni tahlil qilish va modellashtirish uchun samarali ishlatilishi mumkin bo'lgan raqamli formatga aylantirishni o'z ichiga oladi[7]. Keng tarqalgan usullardan biri yorliqlarni kodlashdir. Yorliqlarni kodlash formulasi:

$$Label(category_i) = i \qquad (5)$$

Bu yerda $Label(category_i)$ tayinlangan *i*-chi raqamli qiymat.

Misol uchun, "Erkak" va "Ayol" toifalari bilan "Gender" toifali o'zgaruvchini ko'rib chiqamiz. Yorliq kodlashdan so'ng, o'zgaruvchi quyidagicha o'zgartirilishi mumkin:

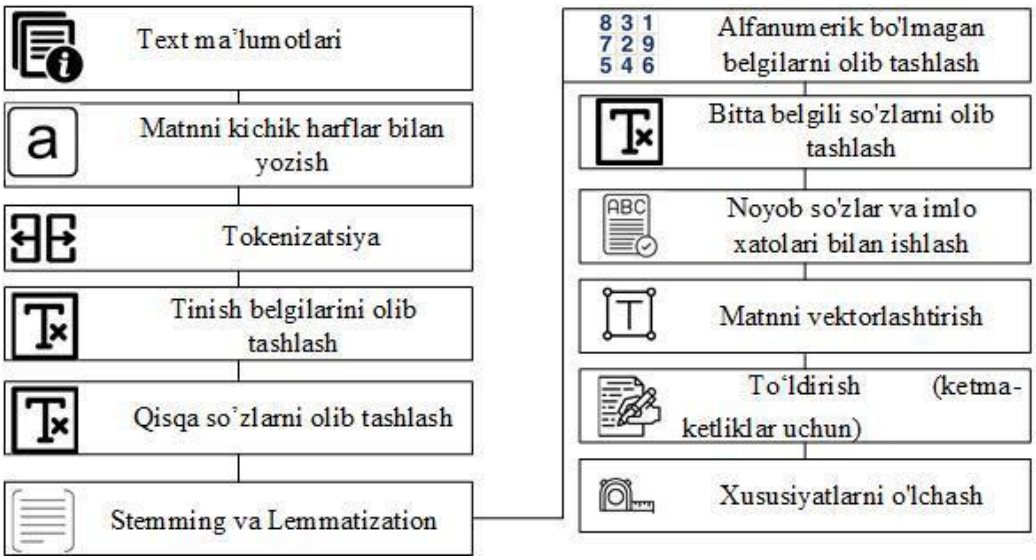
Jins	*Yorliq kodlangan jins*
Erkak	*1*
Ayol	*0*
Erkak	*1*

Ushbu tasvirda har bir toifa raqamli belgi bilan almashtiriladi.

Yorliq kodlash oddiy va samarali, lekin uning cheklovlari bor. Algoritmlar raqamli belgilarni tartibli qiymatlar sifatida izohlashi va barcha toifali o‘zgaruvchilar uchun to‘g‘ri kelmasligi mumkin bo‘lgan mazmunli tartibni qabul qilishi mumkin. Bu noto'g'ri talqin va natijalarga olib kelishi mumkin.

Matn ma'lumotlarini qayta ishlash matn ma'lumotlarini tahlil qilish, tabiiy tilni qayta ishlash (NLP) va mashinani o'rganish vazifalari uchun tayyorlashda muhim qadamdir[8]. To'liqmas matn ma'lumotlari ko'pincha tartibsizliklar va modellarning ishlashiga xalaqit beradigan keraksiz ma'lumotlarni o'z ichiga oladi[9]. Matn ma'lumotlarini qayta ishlash bosqichlarining umumiy sxemasi:

© *Mo'minov Bahodir Boltayevich, Egamberdiyev Elyor Hayitmamatovich.*



2-rasm. Matn ma'lumotlarini qayta ishlash bosqichlari.

Matnni kichik harflar bilan yozish: Barcha matnni kichik harflarga aylantirish. Bu turli bosh harflar bilan yozilgan so'zlar bir xil bo'lishini ta'minlaydi.

Tokenizatsiya: Tokenizatsiya matnni alohida so'zlarga yoki tokenlarga bo'lishni o'z ichiga oladi. Ushbu qadam matnning tuzilgan tasvirini yaratish uchun zarurdir.

Tinish belgilarini olib tashlash: Matndan tinish belgilarini (masalan, vergul, nuqta, savol belgilari) olib tashlang, chunki ular odatda matnning asosiy ma'nosiga hissa qo'shmaydi deb hisoblanishi mumkin.

Bazi so'zlarini olib tashlash: Bazi so'zlar matnda tez-tez uchraydi, lekin ko'pincha kam semantik ma'noga ega bo'lgan "va", "yoki" va hokazo kabi keng tarqalgan so'zlardir. Ularni olib tashlash ma'lumotlarni qayta ishlashga va samaradorlikni oshirishga yordam beradi.

Stemming va Lemmatization: Stemming prefiks va qo'shimchalarni olib tashlash orqali so'zlarni ildiz shakliga keltiradi (masalan, "yugurish" "yugur" ga aylanadi). Lemmatizatsiya esa so'zlarni asos yoki lug'at shakliga keltiradi (masalan, "yaxshiroq" "yaxshi" bo'ladi). Ikkala usul ham so'zlarning xilma-xilligini ularning asosiy ma'nosiga moslashtirishga yordam beradi.

Alfanumerik bo'lmagan belgilarni olib tashlash: Matnning ma'nosiga hissa qo'shmasligi mumkin bo'lgan maxsus belgilar, raqamlar va boshqa harf-raqamli bo'lmagan belgilarni olib tashlanadi.

Bitta belgili so'zlarni olib tashlash: Ko'pincha ko'p ma'lumot bermaydigan juda qisqa so'zlarni olib tashlanadi.

Noyob so'zlar va imlo xatolari bilan ishlash: Juda kam uchraydigan so'zlarni olib tashlash yoki almashtirish, shuningdek, imloni tekshirish yoki oldindan tuzilgan kutubxonalardan foydalanish kabi keng tarqalgan imlo xatolarini tuzatishni ko'rib chiqiladi.

Matnni vektorlashtirish: Qayta ishlangan matnni mashina o'rganish algoritmlari uchun kirish sifatida ishlatilishi mumkin bo'lgan raqamli vektorlarga aylantiriladi. Keng tarqalgan usullarga TF-IDF (Term Frequency-Inverse Document Frequency) va so'zlarni kiritish (masalan, Word2Vec, GloVe) kiradi.

To'ldirish (ketma-ketliklar uchun): Agar siz matn ketma-ketligi (masalan, jumlalar) bilan ishlayotgan bo'lsangiz, qisqaroq ketma-ketliklarni nol yoki boshqa belgilar bilan to'ldirish orqali ular teng uzunlikda ekanligiga ishonch hosil qilinadi.

Xususiyatlarni o'lchash (ixtiyoriy): Amaldagi algoritmga qarab, har bir xususiyat tahlilga teng hissa qo'shishini ta'minlash uchun vektorlashtirilgan matn xususiyatlarini masshtablash zarur bo'lishi mumkin.

© Mo'minov Bahodir Boltayevich, Egamberdiyev Elyor Hayitmamatovich.

Matnga oldindan ishlov berish har doim ham hammaga mos keladigan yondashuv emas; u matnning tabiatiga va siz ishlayotgan aniq vazifaga qarab farq qilishi mumkin. NLTK (Natural Language Toolkit) va spaCy kabi kutubxonalar ushbu dastlabki ishlov berish bosqichlarini soddalashtirilgan tarzda osonlashtirish uchun oldindan yaratilgan vositalarni taklif qiladi[10].

Qiyinchiliklar va kelajak yo'nalishlari: Multitipli ma'lumotlarni qayta ishlash xususiyatlarni tanlash, o'lchamlarni kamaytirish, modelni izohlash va hisoblash samaradorligi bilan bog'liq muammolarni keltirib chiqaradi. Integratsiya jarayoni alohida modellarning kuchli tomonlarini saqlab turishini ta'minlash va ularning zaif tomonlarini yumshatish uchun ehtiyotkorlik bilan ko'rib chiqishni talab qiladi. Kelajakdagi tadqiqotlar turli xil ma'lumotlar turlari o'rtasidagi o'ziga xos munosabatlardan foydalanadigan yanada murakkab integratsiya usullarini ishlab chiqishga qaratilishi kerak.

II.Xulosa.

Multitipli ma'lumotlarni qayta ishlash zamonaviy ma'lumotlarni tahlil qilishning muhim jihati bo'lib, yashirin algoritmlarni ochish va qaror qabul qilishni yaxshilash potentsialini taklif qiladi[11]. Ushbu maqola turli xil ma'lumotlar turlarini integratsiyalash bilan bog'liq texnikalar, qiyinchiliklar va imkoniyatlarni yoritib berdi va murakkab tizimlar haqida yanada kengroq tushuncha yaratish uchun har bir turdagi kuchli tomonlardan foydalanadigan yaxlit yondashuv zarurligini ta'kidladi.

Xulosa qilib aytadigan bo'lsak, Multitipli ma'lumotlarni qayta ishlash dinamik soha bo'lib, u turli xil ma'lumotlar turlari va analitik usullar o'rtasidagi bo'shliqni yopadi. Bu tadqiqotchilar va amaliyotchilar uchun raqamli, toifali, matnli va tasvirli ma'lumotlarning birlashuvidan chuqurroq tushunchalarni ochish uchun qiziqarli yo'lni taqdim etadi. Ma'lumotlarning murakkabligi ortib borar ekan, Multitipli ma'lumotlarni qayta ishlash turli sohalarda innovatsiyalar va yutuqlarni qo'llab-quvvatlovchi ma'lumotlar fanining asboblar to'plamida muhim mahorat bo'lib qoladi.

III.Adabiyotlar.

1. Regalia P. Adaptive IIR filtering in signal processing and control. – Routledge, 2018.
2. Aström K. J., Goodwin G. C., Kumar P. R. (ed.). Adaptive control, filtering, and signal processing. – Springer Science & Business Media, 2012. – T. 74.
3. Weiqi Dai, Chunkai Dai, Kim-Kwang Raymond Choo, Changze Cui, Deiqing Zou, Hai Jin. SDTE: A secure blockchain-based data trading ecosystem. IEEE Transactions on Information Forensics and Security, 15: 725-737, 2020.
4. Wang Z. Fast algorithms for the discrete W transform and for the discrete Fourier transform //IEEE Transactions on Acoustics, Speech, and Signal Processing. – 1984. – T. 32. – №. 4. – C. 803-816.
5. Kurita T. Principal component analysis (PCA) //Computer Vision: A Reference Guide. – 2019. – C. 1-4.
6. Hartigan J. A., Wong M. A. Algorithm AS 136: A k-means clustering algorithm //Journal of the royal statistical society. series c (applied statistics). – 1979. – T. 28. – №. 1. – C. 100-108.
7. Parygin D. S. et al. Categorical data processing for real estate objects valuation using statistical analysis //Journal of Physics: Conference Series. – IOP Publishing, 2018. – T. 1015. – №. 3. – C. 032102.
8. El-Metwally S. et al. Next-generation sequence assembly: four stages of data processing and computational challenges //PLoS computational biology. – 2013. – T. 9. – №. 12. – C. e1003345.
9. Ye M. et al. Specific person retrieval via incomplete text description //Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. – 2015. – C. 547-550.
10. Bird S., Klein E., Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. – " O'Reilly Media, Inc.", 2009.
11. Mo'minov B., Egamberdiyev E. A NEURO-FUZZY MODEL FOR PREDICTING SUCCESSFUL PASSING OF ENTRANCE EXAMS OF APPLICANTS TO HIGHER EDUCATION //Central Asian Journal of Education and Computer Sciences (CAJECS). – 2023. – T. 2. – №. 2. – C. 6-13.

© *Mo'minov Bahodir Boltayevich, Egamberdiyev Elyor Hayitmamatovich.*