SO'ZLARNI VEKTORLI IFODALASH TEXNOLOGIYASI

[1]Akmuradov Baxtiyor Uralovich, [2]Axmedova Xusniya Xusanovna

[1] Muhammad al-Xorazmiy nomidagi Toshkent Axborot Texnologiyalari Universiteti dotsenti, PhD.,b.u.akmuradov@gmail.com

[2]Muhammad al-Xorazmiy nomidagi Toshkent Axborot Texnologiyalari Universiteti o'qituvchisi, h.ahmedova86@mail.ru

https://doi.org/10.5281/zenodo.7854470

Annotatsiya. *Tabbiy tilni qayta ishlash texnologiyalarini ishlab chiqish va samarali foydalanishni tashkil etish bugungi kunning dolzarb maslalaridan biridir. Ushbu ishda elektron matndagi so'zlarni qidirish va ma'no jihatidan tahlil qilishni tashkil etishda qo'llaniladigan samarali texnologiyalardan biri bo'lgan so'zni vektorli ifodalash texnologiyalari haqida so'z yuritilgan. Xususan, Word2vec texnologiyasi bo'yicha so'zni vektor qiymatlarini aniqlash jarayoni tushuntirilgan.*

Kalit so'zlar: *matn, so'z, texnologiya, vektor, Word2vec, Ochiai koeffitsiyenti, GloVe.*

Hisoblash mashinalari yordamida matnni elektron tarzda saqlash, o'zgartirish va uzatishni amalga oshiruvchi ko'plab vositalar mavjud. Zamonaviy sun'iy intellekt texnologiyalarining, NLP (Natural Language Processing) texnologiyalarining rivojlanishi bilan tabiiy tilga ishlov berish vositalari soni va sifati ham keskin o'sdi[1]. Bunda matn tahlili, so'zlarni tanib olish, guruhlash va ma'no jihatidan tahlil qilish hamda ranjirlashni amalga oshirish uchun so'zni kompyuter tushunadigan tilga o'girish talab etiladi[3]. So'zning barcha xususiyatlarini inobatga olgan holda raqamli ko'rinishning samarali usullaridan biri bu vektorli ifodalashdir. So'zlarni vektorli ifodalashni hozirda keng qo'llanilayotgan texnologiyalardan biri misolida tushuntirishga harakat qilamiz.

Word2vec – so'zlarni vektorli ifodalashning samarali texnologiyalaridan biri bo'lib, u 2013-yilda ishlab chiqilgan. So'zlar bilan ishlashdan tashqari, uning ba'zi kontsepsiyalari tavsiya berish mexanizmlarini ishlab chiqishda va ma'lumotlar ma'nosini ifodalashda, hatto tijoriy, lingvistik bo'lmagan vazifalarda ham samarali qo'llanilib kelinmoqda[6]. Xususan, Airbnb, Alibaba, Spotify va Anghami kabi kompaniyalar ushbu texnologiyadan o'zlarining tavsiya tizimlarida foydalanganlar[2].

Ob'ektlarni vektor ko'rinishida tasvirlashni sodda tilde tushuntirish uchun quyidagi misolni ko'rib chiqamiz. Masalan: bir insonni tasvirlash uchun uning bir nechta xususiyatlarini aytish mumkin. Xususiyatlar sonining ortib borishi bilan tasvirlanayotgan shaxs haqida aniqroq tasovvur uyg'ona boshlaydi. Misol uchun biz tanlagan shaxsni tasvirlash uchun 0 dan 100 gacha bo'lgan shkalada 5 ta xususiyatlarini baholash kerak bo'lsin:

1-jadval.

Xususiyatlar jadvali

№	Xususiyat	Baho (0 da 100 gacha oraliqda)
1.	Hissiyotga beriluvchanlik	38
2.	Kuchli/kuchsiz	76
3.	Sog'lom/xasta	82
4.	Yosh/qari	42
5.	Baland/past	80

Jadvaldan ko’rinib turibtiki tasvirlanayotgan shaxsning hissiyotga beriluvchanlik darajasi 38 ga teng va hk. Hisoblash mashinasida amalga oshiriladigan hisob kitoblarni soddalashtirish uchun qiymatlarni -1 da +1 oralig’idagi shkalaga o’tkazsak quyidagi qiymatlarga ega bo’lamiz:

Shaxs	0,4	,8	,8	0,4	,8

Tanlangan shaxga o’xshash bo’lgan shaxsni topish uchun yana 2 ta shaxsning 2 turdagi xususiyatlari berilgan bo’lsin.

Shaxs-1	-0,3	0,2	0,3	-0,4	0,9
Shaxs-2	-0,5	-0,4	-0,2	0,7	-0,1

Berilgan xususiyatlarga ega ikkita shaxsning qaysi biri tanlangan shaxsga ko’proq o’xshashligini aniqlash uchun ularni vektor ko’rinishida ifodalash va taqqoslash mumkin. Quyidagi 1-rasmda ularning vektor ifodasi keltirilgan.

1-rasm. Shaxslar xususiyatlarining vektorli ifodasi

Keltirilgan vektorli ifodadan qaysi bir shaxs tanlangan shaxsga ko’proq o’xshashligi haqida fikr yuritish mumkin. Biroq, hisoblash mashinalarida vektorlar orasidagi o’xshashlik odatda Ochiai koeffitsienti (geometrik koeffitsient) bilan aniqlanadi. Berilgan ikki shaxsning tanlangan shaxsga o’xshashlik darajasini Ochiai koeffitsienti bo’yicha hisoblasak quyidagi natijaga ega bo’lamiz.

O‘xshashlik (	-0,4	0,8	,	-0.3	0,2	)	=	0,87
O‘xshashlik (	-0,4	0,8	,	-0.5	-0,4	)	=	-0,20

Natijadan ko’rish mumkinki birinchi shaxs tanlangan shaxsga ko’proq o’xshashlikka ega. Biroq, ko’rilgan misolda faqat 2 ta xususiyat bo’yicha vektor qiymatlarining o’xshashligi hisoblandi. Shaxsning taqqoslash mumkin bo’lgan parametrlari soni ortib borishi bilan vektor qiymatlarining o’xshashligi yanada aniqroq ifodalash mumkin bo’ladi.

Ochiai koeffitsiyenti yapon biologi Akira Ochiai tomonidan 1957 yilda taklif qilingan ikkilik o'xshashlik o'lchovi bo’lib, keyinchalik biologiyadan tashqarida umumlashtirilgan va turli ilovalarda qo’llanila boshlandi. Standart ta'rifda ikkita ixtiyoriy A va B to'plamlari uchun koeffitsiyent quyidagicha hisoblanadi:

$$K = \frac{|A \cap B|}{\sqrt{|A| \cdot |B|}} \qquad (1)$$

bu yerda, |A|- A to‘plamning (мощности) kardinalligi. Ko’p hollarda Ochiai o‘lchovi quyidagi shaklda qo’llaniladi:

$$K = \frac{|A \cap B|^2}{|A| \cdot |B|} \qquad (2)$$

Ta'riflovchi to'plamlar uchun Ochiai koeffitsienti quyidagicha hisoblanadi:

$$K = \frac{\sum_{i=1}^{r} \min(A_i, B_i)}{\sqrt{\sum_{i-1}^{r}(A_i) \cdot \sum_{i-1}^{r}(B_i)}} \quad (3)$$

Bunday o'lchov ekologiyada ko'plik namunalar uchun keng qo'llaniladi. Agar ob'ektlar turlarning paydo bo'lishi bilan solishtirilsa, ya'ni P (A) ehtimolliklari - turlar uchraydigan namunalarning nisbiy soni hisobga olinsa, u holda koeffitsient hodisalarning muvofiqligi bo'yicha hisoblanadi:

$$K_{0,0} = \frac{P(\mathrm{A} \cap \mathrm{B})}{\sqrt{P(A) \cdot P(B)}} \quad (4)$$

Bundan quyidagi xulosalarni chiqarishimiz mumkin:

Nafaqat odamlar balki boshqa obyektlarni raqamli vektorlar sifatida ifodalash mumkin. Vektorlar o'rtasida o'xshashlikni osongina hisoblash mumkin.

Ko'rib chiqilgan xususiyatlarni so'zlarga qo'llash orqali ma'no jihatidan bir biriga yaqin so'zlarni qidirish va ajratib olish mumkin[5]. Buning uchun so'zlarni ma'no jihatida to'liq ifodalangan holda vektor ko'rinishiga o'tkazish talab etiladi. Hozirgi kunda so'zlarni vektor ko'rinishida ifodalovchi bir qator texnologiyalar va dasturlash kutubxonalari mavjud. Ularga Word2vec, GloVe, CBOW, SkipGram, FastTextni misol qilish mumkin. Bunday texnologiyalar hisoblash mashinalarida so'zlarni ma'no jihatidan barcha xususiyatlarini ifodalovchi qiymatlarni umumlashtiruvchi vektor ko'rinishida ifodalash uchun qo'llaniladi.

Xulosa sifatida shuni aytish mumkinki, so'zning vektor shaklini hosil qilish uchun uning shakl va ma'no jihatidan kuzatilishi mumkin bo'lgan barcha xususiyatlari inobatga olishi zarur. Shunda, vektor qiymatlari o'zaro yaqin bo'lgan so'zlar ma'no jihatidan ham bir biriga yaqin so'zlar bo'ladi. Hozirda ko'plab tizimlarda qidirish, guruhlash va bashorat qilish algoritmlarini ishlab chiqishda ushbu yondashuv keng qo'llanilmoqda[4]. O'zbek tilidagi so'zlarning ham shunday vektor korpusini ishlab chiqish kompyuter lingvistikasi va Tabbiy tilni qayta ishlash sohasidagi ko'plab muammolarga yechim topish imkoniyatini yaratadi.

REFERENCES

1. A. M. Цитульский, A. B. Иванников, И. С. Рогол "Интеллектульный анализ текста", StudNet 2020. №6. С. 476-483.
2. U.Khamdamov, M. Mukhiddinov, O. Djuraev, A. Mukhamedaminov "A novel method for extracting text from natural scene images and TTS", "European science review" scientific journal, Premier publishing, Vienna – 2018. № 11-12. Vol. 1. -P. 30-33
3. Makhmudov, F., Mukhiddinov, M., Abdusalomov, A., Avazov, K., Khamdamov, U., & Cho, Y. I. (2020). Improvement of the end-to-end scene text recognition method for "text-to-speech" conversion. International Journal of Wavelets, Multiresolution and Information Processing, 18(06), 2050052.
4. J. Elov, U. Khamdamov, A. Abdullayev, I. Narzullayev and D. Sultanov, "Development of a database of higher education process management information system based on the relational model," 2021 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2021, pp. 01-05
5. https://nlp.stanford.edu/projects/glove
6. https://towardsdatascience.com/word2vec-explained-49c52b4ccb71