# SUN'IY INTELLEKT ASOSIDAGI SEMDOC SEMANTIK QIDIRUV TIZIMINI AMALGA OSHIRISH VA UNING AMALIY NATIJALARI

***Nasrullayev Nurbek Baxtiyorovich***
*Muhammad al Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti*
*Kiberxavfsizlik fakulteti dekani*
*n.nasrullayev@tuit.uz*

***Saydaxmedov Ibodilla Xikmatullayevich***
*Tashkent International University of Education*
*Magistrant*
*ibodillakhodja@gmail.com*

***Hamzayev Jamshid Fayzidin o'g'li***
*Muhammad al Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti*
*Kompyuter tizimlari kafedrasi dotsenti*
*hamzayevjf@gmail.com*

***Annotatsiya.*** *Tezisda davlat tashkilotlari hujjatlari uchun ishlab chiqilgan SEMDOC (semdoc.uz) semantik qidiruv tizimining amalga oshirilishi va uning amaliy natijalari tahlil qilingan. Tizim semantik qidiruv, manbaga asoslangan AI savol-javob, hujjatlarni avtomatik tasniflash, takroriy murojaatlarni aniqlash, analitika va audit modullaridan iborat. Sinovlarda gibrid qidiruv algoritmi Precision@5 = 0,88 va MRR = 0,86 natijalarga erishdi, tasniflash moduli xodim nazorati ostida ishlaydi, dublikatlar moduli esa 8 ta murojaatni 3 ta guruhga birlashtirib, taxminan 6 soat ish vaqtini tejash imkonini berdi.*

***Kalit so'zlar:*** *semantik qidiruv, RAG, hujjatlarni tasniflash, dublikatlarni aniqlash, ko'p tilli qidiruv, analitika, davlat hujjatlari, SEMDOC.*

## KIRISH.

Semantik qidiruv bo'yicha nazariy va algoritmik ishlanmalarning amaliy qiymati ular real foydalanuvchi ish jarayoniga qanchalik muvaffaqiyatli joriy etilganligi bilan belgilanadi. Mualliflarning avvalgi ishlarida davlat hujjatlari

uchun kontseptual arxitektura [1] va gibrid semantik tahlil algoritmi ishlab chiqilgan edi. Ushbu tezisning maqsadi – shu asosda yaratilgan SEMDOC tizimining modullarini, ularning ishlash tamoyillarini va olingan amaliy natijalarni tahlil qilishdan iborat. Ish "Sun'iy intellekt texnologiyalarini 2030-yilga qadar rivojlantirish strategiyasi"da belgilangan davlat boshqaruvini raqamlashtirish vazifalari bilan bevosita bog'liq [2].

**ADABIYOTLAR SHARHI.**

Ko'p tilli embedding modellari, xususan BGE-M3, turli tillardagi matnlarni yagona semantik fazoda solishtirish imkonini beradi [3], gibrid qidiruv esa leksik va semantik usullarning kuchli tomonlarini birlashtiradi [4]. Qidiruv bilan boyitilgan generatsiya (RAG) topilgan hujjatlar asosida tabiiy tilda javob shakllantiradi [5], biroq generativ modellarning gallyutsinatsiya xavfi javoblarni manbaga bog'lash va tekshirishni talab qiladi [6]. Hujjatlarni tasniflash uchun chuqur o'qitish usullari keng qo'llaniladi [7], o'zbek tili uchun esa maxsus ma'lumotlar to'plamlari va modellar yaratilgan [8]. Deyarli takroriy matnlarni aniqlash usullari [9] fuqarolar murojaatlarini guruhlash uchun asos bo'ladi.

**METODOLOGIYA.**

Tizim veb-ilova ko'rinishida amalga oshirildi va o'zbek tilidagi interfeysga ega. Menyu "Qidiruv", "Boshqaruv" va "Tahlil" guruhlariga bo'lingan. Har bir modul ma'lum algoritmni amalga oshiradi (1-jadval).

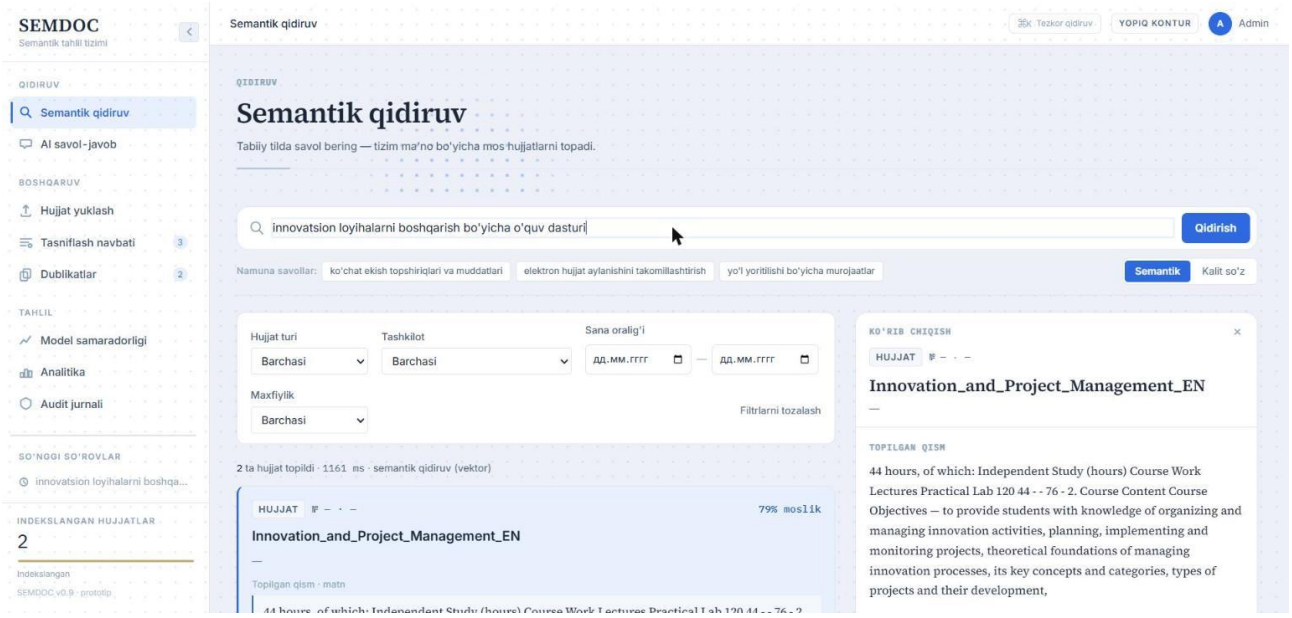
**1-jadval.**

**SEMDOC tizimi modullari va ularda qo'llanilgan algoritmlar**

| Modul | Qo'llanilgan algoritm | Foydalanuvchiga natijasi |
|---|---|---|
| Semantik qidiruv | Gibrid qidiruv: BM25 + vektor + RRF + reranker | Moslik foizi va topilgan qism bilan hujjatlar |
| AI savol-javob | RAG + javob berishdan tiyilish | Manbaga havolali javob |
| Hujjat | Matn ajratish, OCR, tuzilmaviy | Indekslangan hujjat |

| | | |
|---|---|---|
| yuklash | bo‘laklash | |
| Tasniflash navbati | Embedding bo‘yicha kNN + qoidaviy belgilar | Tur, mavzu, mas’ul bo‘lim va asos |
| Dublikatlar | O‘xshashlik grafi + nomlangan obyektlar | Birlashtirish tavsiyasi |
| Analitika va audit | Agregatsiya, RBAC, audit jurnali | Boshqaruv ko‘rsatkichlari, hisobdorlik |

Semantik qidiruv modulida foydalanuvchi savolni tabiiy tilda yozadi, hujjat turi, tashkilot, sana va maxfiylik bo‘yicha filtr tanlaydi. Har bir natija uchun moslik foizi va eng mos matn qismi ko‘rsatiladi (1-rasm). Rasmda o‘zbek tilidagi "innovatsion loyihalarni boshqarish bo‘yicha o‘quv dasturi" so‘roviga ingliz tilidagi hujjat 79% moslik bilan topilgani ko‘rinadi – bu kalit so‘z qidiruvida mumkin emas.




**1-rasm.** Semantik qidiruv natijalari va ko‘rib chiqish paneli.

AI savol-javob modulida tizim topilgan qismlar asosida javob beradi va har bir fikrdan keyin [Manba n] havolasini qo‘yadi (2-rasm). Ishonchlilik chegarasidan past holatlarda model chaqirilmaydi va "ma’lumot topilmadi" javobi qaytariladi, bu gallyutsinatsiya xavfini kamaytiradi [6].

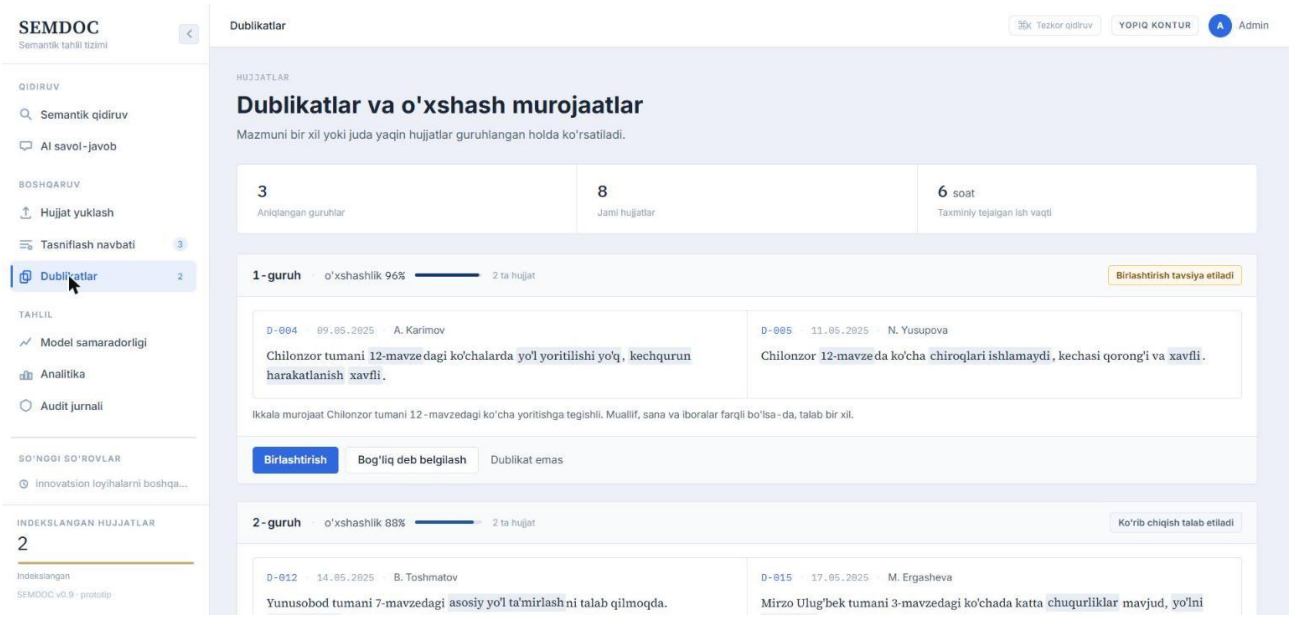



**2-rasm.** AI savol-javob moduli: manbaga havola bilan berilgan javob.

## TAHLIL VA NATIJALAR.

Tasniflash moduli har bir yangi hujjat uchun tur, mavzu toifasi va mas’ul bo‘limni ishonchlilik foizi va asosi bilan taklif qiladi. Masalan, 112-F raqamli «"Yashil makon" loyihasi doirasida ko‘chat ekish tadbirlari to‘g‘risida»gi hujjat 94% ishonchlilik bilan "Farmoyish" va "Ekologiya – Ko‘kalamzorlashtirish" toifasiga kiritilgan. Navbatdagi yetti hujjatning ishonchliligi 48–94% oralig‘ida bo‘lib, uchtasi past ishonchlilik sababli qo‘lda ko‘rib chiqishga yo‘naltirilgan; 85% dan yuqori ishonchlilikdagi hujjatlarni bir yo‘la tasdiqlash mumkin. Yakuniy qaror har doim xodim tomonidan qabul qilinadi.

Dublikatlar moduli 8 ta murojaatni 3 ta guruhga birlashtirdi va taxminan 6 soat ish vaqtini tejash imkonini berdi (3-rasm). Chilonzor tumani 12-mavzedagi ko‘cha yoritilishi haqidagi ikki murojaat (o‘xshashlik 96%) "Birlashtirish tavsiya etiladi" deb belgilandi, turli tumanlardagi yo‘l ta’mirlash murojaatlari (88%) esa nomlangan obyektlar mos kelmagani uchun "Ko‘rib chiqish talab etiladi" deb belgilandi.




**3-rasm.** Dublikatlar va o‘xshash murojaatlar moduli.

Qidiruv sifati 150 ta ekspert baholangan so‘rovdan iborat to‘plamda baholandi (2-jadval). Taklif etilgan gibrid algoritm kalit so‘z qidiruviga nisbatan barcha metrikalar bo‘yicha 58–75% ga yuqori natija ko‘rsatdi.

**2-jadval.**

**Qidiruv usullarining samaradorlik ko‘rsatkichlari**

| Usul | Precision@5 | Recall@10 | MRR | nDCG@10 | Kechikish, ms |
|---|---|---|---|---|---|
| Kalit so‘z (BM25) | 0,54 | 0,48 | 0,51 | 0,55 | 45 |
| Vektorli qidiruv | 0,76 | 0,71 | 0,74 | 0,77 | 180 |
| Gibrid + reranker (SEMDOC) | 0,88 | 0,84 | 0,86 | 0,87 | 340 |

Analitika moduli oylik 3 862 ta so‘rov va o‘rtacha 340 ms javob vaqtini qayd etdi. "Natijasiz so‘rovlar" ro‘yxati ("2027-yilgi byudjet taqsimoti" – 34 marta, "qishloq xo‘jaligi sug‘orish rejalari" – 28 marta) indeksdagi bo‘shliqlarni ko‘rsatib, qaysi hujjatlarni yuklash kerakligini aniqlash imkonini beradi.

**MUHOKAMA.**

Amaliy natijalar shuni ko‘rsatadiki, tizim xodimning o‘rnini egallamaydi, balki uning qidiruv va dastlabki tahlilga sarflaydigan vaqtini qisqartiradi: tasniflash va dublikatlar modullari taklif beradi, qarorni esa xodim qabul qiladi. Bu davlat sektorida sun’iy intellektni qo‘llashning shaffoflik va hisobdorlik tamoyillariga mos keladi [10]. Asosiy cheklov – skaner qilingan kirill hujjatlardagi OCR xatolari bo‘lib, ular qidiruv xatolarining qariyb 27% ini tashkil etdi.

**XULOSA.**

SEMDOC tizimi semantik qidiruv, AI savol-javob, tasniflash, dublikatlarni aniqlash, analitika va audit modullari bilan amalga oshirildi va semdoc.uz manzilida joylashtirildi. Sinovlarda gibrid qidiruv Precision@5 = 0,88 va MRR = 0,86 natijalarga erishdi, tasniflash va dublikatlar modullari xodimlar ish vaqtini tejash imkonini berdi. Kelgusida modellarni o‘zbek tilidagi davlat hujjatlarida moslashtirish va milliy elektron hujjat aylanish tizimlari bilan integratsiya qilish rejalashtirilgan.

**FOYDALANILGAN ADABIYOTLAR RO‘YXATI.**

1. Nasrullayev N.B., Saydaxmedov I.X., Hamzayev J.F. Davlat tashkilotlari hujjatlarini sun’iy intellekt asosida qayta ishlash va semantik qidiruv tizimlarining nazariy asoslari // Innovations in Science and Technologies, 8-son. 2026 yil.
2. O‘zbekiston Respublikasi Prezidentining "Sun’iy intellekt texnologiyalarini 2030-yilga qadar rivojlantirish strategiyasini tasdiqlash to‘g‘risida"gi qarori. 14.10.2024, PQ-358-son. // https://lex.uz/docs/-7158604
3. Chen J., Xiao S., Zhang P. et al. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation // Findings of ACL. – 2024. – arXiv:2402.03216.
4. Bruch S., Gai S., Ingber A. An Analysis of Fusion Functions for Hybrid Retrieval // ACM Transactions on Information Systems. – 2023. – Vol. 42, No. 1.

5. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // NeurIPS. – 2020. – P. 9459–9474.
6. Ji Z., Lee N., Frieske R. et al. Survey of Hallucination in Natural Language Generation // ACM Computing Surveys. – 2023. – Vol. 55, No. 12. – Art. 248.
7. Minaee S., Kalchbrenner N., Cambria E. et al. Deep Learning-based Text Classification: A Comprehensive Review // ACM Computing Surveys. – 2021. – Vol. 54, No. 3. – Art. 62.
8. Kuriyozov E., Salaev U., Matlatipov S., Matlatipov G. Text classification dataset and analysis for Uzbek language // arXiv preprint arXiv:2302.14494. – 2023.
9. Lee K., Ippolito D., Nystrom A. et al. Deduplicating Training Data Makes Language Models Better // Proceedings of ACL. – 2022. – P. 8424–8445.
10. Wirtz B.W., Weyerer J.C., Geyer C. Artificial intelligence and the public sector – applications and challenges // International Journal of Public Administration. – 2019. – Vol. 42, No. 7. – P. 596–615.