O'ZBEKCHA TA'LIM MATNLARIDA DALIL QIDIRISHGA SAHIFA SARLAVHASINING TA'SIRI

Erkayeva Laylo Odilbek qizi
Urganch davlat universiteti bitiruvchisi
"Kompyuter injiniringi" bo'yicha magistr
ORCID: 0009-0001-5380-5721
layloerkayeva0@gmail.com

Annotatsiya

RAG tizimida savolga javob tuzishdan oldin unga mos dalil parchasini topish kerak. O'zbekcha ta'lim sahifalaridagi qisqa parchalar sarlavhasiz qaysi mavzuga oidligini har doim ham bildirmaydi. Ushbu tezis sahifa sarlavhasini parchaga qo'shish dalil qidirish natijasiga qanday ta'sir qilishini o'rganadi. Sakkizta rasmiy sahifadan olingan 111 parcha va 50 savolda uchta TF-IDF usuli sarlavhali hamda sarlavhasiz holatda solishtirildi. Belgili TF-IDFda to'g'ri dalilning dastlabki uch natijadan topilishi 41/50 dan 45/50 ga oshdi, birinchi natijadan topilishi esa 35/50 dan 34/50 ga kamaydi. Demak, sarlavhaning foydasi qidiruv usuli va nechta natija ko'rib chiqilishiga bog'liq. Xulosa ushbu kichik to'plamga tegishli.

Kalit so'zlar: o'zbek tili, ta'lim hujjatlari, RAG, dalil qidirish, sarlavha, TF-IDF.

Abstract

Before a RAG system can answer a question, it must find a relevant evidence passage. Short passages from Uzbek education pages may not clearly indicate their topic without the page title. This thesis examines how adding that title affects evidence retrieval. Three TF-IDF methods were compared with and without titles using 111 passages from eight official pages and 50 questions. With character-level TF-IDF, retrieval of the designated passage among the first three results increased from 41/50 to 45/50, while retrieval in the first position decreased from 35/50 to 34/50. The usefulness of a title therefore depends on the retrieval method and how many results are considered. The findings are limited to this small collection.

Keywords: Uzbek, education documents, RAG, evidence retrieval, page title, TF-IDF.

Аннотация

Прежде чем система RAG ответит на вопрос, ей необходимо найти подходящий фрагмент с подтверждающими сведениями. Короткие фрагменты узбекских образовательных страниц без заголовка не всегда раскрывают свою тему. В тезисе изучено, как добавление заголовка страницы влияет на поиск таких фрагментов. На 111 фрагментах из восьми официальных страниц и 50 вопросах сопоставлены три метода TF-IDF с заголовками и без них. При использовании символьного TF-IDF число вопросов, для которых заданный фрагмент найден среди первых трёх результатов, выросло с 41/50 до 45/50, однако на первом месте — снизилось с 35/50 до 34/50. Следовательно, польза заголовка зависит от метода поиска и числа рассматриваемых результатов. Выводы ограничены небольшим исследованным корпусом.

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

Ключевые слова: узбекский язык, образовательные тексты, RAG, поиск фрагментов, заголовок страницы, TF-IDF.

KIRISH

Hujjatga tayangan savol-javob tizimida javob yozishdan avval savolga mos dalil parchasi qidiriladi. RAG yondashuvida bu bosqich tashqi manbadagi ma'lumotni javob jarayoniga olib kiradi [1]. Ammo «harbiy xizmatni o'tash uchun» kabi qisqa ro'yxat bandi alohida olinganda qaysi mavzuga tegishli ekani noaniq bo'lishi mumkin. Sahifa sarlavhasini parchaga qo'shish shu kontekstni to'ldirishi ehtimol. Wang va hammualliflar hujjat konteksti, jumladan sarlavha, parcha qidiruviga ta'sir qilishi mumkinligini ko'rsatgan; ularning baholashida ham barcha savollar uchun bir xil yaxshilanish kuzatilmagan [2].

Ushbu tezisning maqsadi — o'zbek tilidagi kichik ta'lim matnlari to'plamida «sarlavha + parcha» va «faqat parcha» ko'rinishlarining qidiruv natijalarini bir xil savollarda solishtirish. Tadqiqot savoli: sarlavha qo'shilganda to'g'ri dalil birinchi yoki dastlabki uchta qidiruv natijasida ko'proq uchraydimi? Baholash faqat dalil qidirishga tegishli; javob yaratadigan sun'iy intellekt ishlatilmaydi.

MATERIALLAR VA METODLAR

Materiallar: 2026-yil 24-sentabr holatida gov.uz portalidan olingan ta'lim mavzusidagi sakkiz ochiq sahifa [4–11]. Ulardagi 111 qisqa matn parchasi va har biriga bittadan javobni o'z ichiga oluvchi parcha belgilangan 50 savol ishlatildi. Savollarning 25 tasi manbadagi ifodaga yaqin, 25 tasi qayta ifodalangan. Savollar manbalar va yordamchi raqamli vosita ko'magida tuzilgan; ular real foydalanuvchilardan yig'ilmagan. Ba'zi juda qisqa bandlar sarlavhasiz holatda to'liq mavzuni ifodalamaydi. Shu holat sinovning o'zi uchun muhim, lekin natijaning boshqa korpuslarga tatbiqini cheklaydi.

Har parcha ikki xil ko'rinishda tayyorlandi: A) parchaning o'z matni; B) shu matn oldidan sahifa sarlavhasi va nuqta. Savol matni hamda to'g'ri dalil belgilari har ikkala shartda aynan bir xil saqlandi. scikit-learn 1.8.0 kutubxonasidagi TfidfVectorizer [3] yordamida uch variant hisoblandi: 1–2 so'zli n-grammlar, butun satr bo'ylab 3–5 belgili n-grammlar va so'z chegarasi ichidagi 3–5 belgili n-grammlar. Har bir shart uchun TF-IDF matritsasi o'sha shartdagi 111 parchada qayta hisoblandi; savollar faqat tayyor vektor fazoga o'tkazildi. Matnlar kichik harfga aylantirildi, minimal uchrash soni 1 bo'ldi; o'zaklash yoki tayyor neyron model qo'llanmadi.

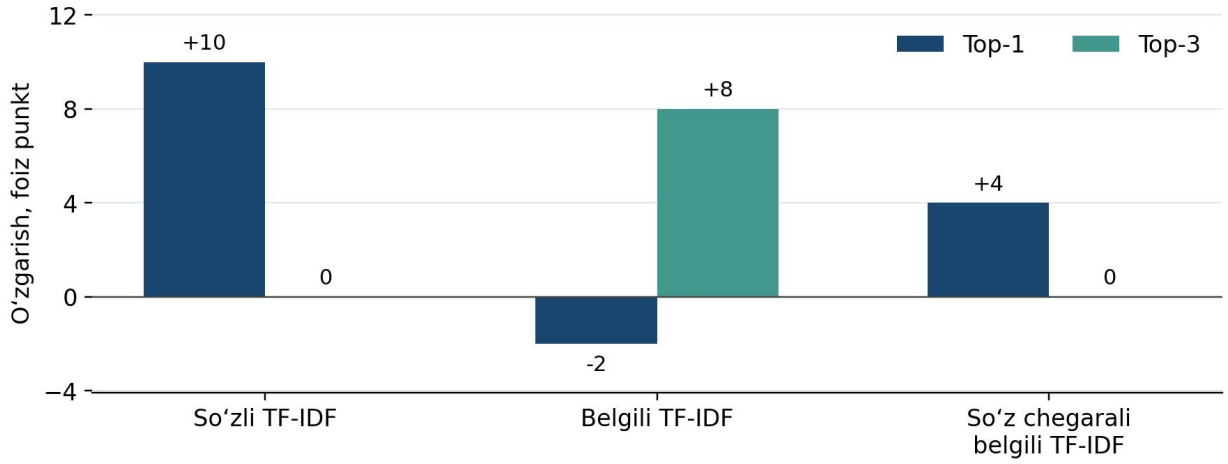
Normallashgan savol va parcha vektorlarining skalyar ko'paytmasi bo'yicha natijalar tartiblandi; tengliklar dastlabki parcha tartibi bilan hal qilindi. Top-1 — to'g'ri parcha birinchi o'rindagi savollar ulushi; Top-3 — dastlabki uch o'rindagi ulush; MRR@3 — to'g'ri dalil o'rnining teskari qiymatlari o'rtachasi (uchlikdan tashqari uchun 0). Har bir ko'rsatkichning maxraji 50 savol. Bir xil 50 savol ikkala shartda ishlatilgani juft qiyoslashga imkon beradi, ammo mustaqil yakuniy test o'rnini bosa olmaydi.

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

NATIJALAR

1-jadval. Bir xil 50 savolda sarlavhali va sarlavhasiz qidiruv natijalari.

Usul	Parcha ko'rinishi	Top-1	Top-3	MRR@3
So'zli TF-IDF (1–2)	Sarlavhasiz	25/50 (50%)	36/50 (72%)	0,600
So'zli TF-IDF (1–2)	Sarlavhali	30/50 (60%)	36/50 (72%)	0,657
Belgili TF-IDF (3–5)	Sarlavhasiz	35/50 (70%)	41/50 (82%)	0,760
Belgili TF-IDF (3–5)	Sarlavhali	34/50 (68%)	45/50 (90%)	0,767
So'z chegarali belgili (3–5)	Sarlavhasiz	34/50 (68%)	43/50 (86%)	0,760
So'z chegarali belgili (3–5)	Sarlavhali	36/50 (72%)	43/50 (86%)	0,783



1-rasm. Sarlavha qo'shilganda Top-1 va Top-3 o'zgarishi (foiz punkt).

Butun satrli belgili variantda sarlavha qo'shilishi bilan Top-3 to'g'ri savollar soni 41 tadan 45 taga oshdi (+8 foiz punkt). Birinchi o'rinda esa to'g'ri parcha 35 emas, 34 savolda topildi (−2 foiz punkt). So'zli variantda Top-3 o'zgarmadi, biroq Top-1 25 tadan 30 taga chiqdi (+10 foiz punkt). So'z chegarali belgili variantda Top-1 34 tadan 36 taga oshdi (+4 foiz punkt), Top-3 esa 43 holida qoldi (1-jadval va 1-rasm).

Butun satrli belgili variantda Q001 va Q047 kabi qayta ifodalangan, Q002 va Q016 kabi bevosita savollar sarlavha qo'shilgach dastlabki uch natijaga kirdi; ushbu variantda Top-3 dan chiqib qolgan savol kuzatilmadi. Masalan, Q001 ning to'g'ri parchasi faqat «oilasining betob a'zosini ... parvarish qilish uchun» mazmunidan iborat, sarlavhasi esa «Akademik ta'til». Bu holat sarlavha mavzu signalini qo'shgan bo'lishi mumkinligini ko'rsatadi, lekin aynan qaysi token o'zgarishni keltirib chiqargani alohida nazorat sinovida ajratilmagan. So'z chegarali variantda Q047 yaxshilandi, Q035 yomonlashdi: Top-3 bo'yicha umumiy son shu sabab o'zgarmadi.

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

MUHOKAMA

Sarlavha qisqa ro'yxat bandining mavzusini aniqlashtira oladi, lekin har bir parchaga takroran qo'shilgan so'zlar TF-IDF vaznlarini ham o'zgartiradi. Bu tajribada belgili usulning Top-3 bo'yicha foydasi Top-1 yaxshilanishi bilan birga kelmadi. Demak, qidiruv tizimi foydalanuvchiga bitta parcha ko'rsatadimi yoki uchta parcha yuboradimi — mezonni tanlashda hal qiluvchi savol. Natijalar boshqa tillar va kattaroq korpuslar uchun umumiy qoida emas.

Cheklovlar: sakkiz sahifa va 50 tuzilgan savol; har savolga faqat bitta to'g'ri parcha belgilangan; belgilashni mustaqil ekspert takrorlamagan; avvalgi usullar shu kichik savollar to'plamida ko'rilgan; alohida yakuniy test yo'q. Natijalar tavsifiydir, statistik ahamiyat yoki sababiy ta'sir da'vo qilinmaydi. Javob yaratish va tayyor javobning manbaga sodiqligi o'lchanmadi. Keyingi bosqichda real foydalanuvchi savollari, mustaqil annotatorlar hamda boshqa toifadagi o'zbekcha hujjatlar bilan sinov o'tkazish kerak.

XULOSA

Sarlavha qo'shish ushbu to'plamda butun satrli belgili TF-IDFning Top-3 natijasini 82% dan 90% ga ko'tardi, ammo uning Top-1 natijasini 70% dan 68% ga tushirdi. So'zli TF-IDFda Top-1 50% dan 60% ga ko'tarildi, Top-3 72% holida qoldi. Shuning uchun qisqa o'zbekcha ta'lim parchalarini qidirishda sarlavhani qo'shish yoki qo'shmaslikni qidiruv usuli va foydalaniladigan natijalar soni bo'yicha alohida tekshirish tavsiya etiladi. Bu kichik qiyoslash yakuniy RAG javoblarining sifatini baholamaydi.

FOYDALANILGAN ADABIYOTLAR

1. Lewis P. va boshq. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 2020. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
2. Wang K., Reimers N., Gurevych I. DAPR: A Benchmark on Document-Aware Passage Retrieval. Proceedings of ACL, 2024, 4313–4330. https://doi.org/10.18653/v1/2024.acl-long.236
3. scikit-learn. TfidfVectorizer texnik hujjatlari. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
4. O'zbekiston Respublikasi Hukumat portali. Akademik ta'til. https://gov.uz/oz/advice/65/document/1755
5. O'zbekiston Respublikasi Hukumat portali. Talabalarga ijara to'lovining bir qismini qoplab berish tartibi. https://gov.uz/oz/advice/64/document/2562
6. O'zbekiston Respublikasi Hukumat portali. Davlat OTMlarining talabalar turar joyiga joylashish uchun ariza yuborish. https://gov.uz/oz/advice/65/document/3106
7. O'zbekiston Respublikasi Hukumat portali. Ta'lim krediti. https://gov.uz/oz/advice/65/document/3248
8. O'zbekiston Respublikasi Hukumat portali. Talabalar bilimini baholash. https://gov.uz/oz/advice/65/document/1758
9. O'zbekiston Respublikasi Hukumat portali. Ta'lim grantida tahsil olgan talabalarning 3 yil ishlab berish majburiyati bekor qilindi. https://gov.uz/oz/advice/65/document/3250
10. O'zbekiston Respublikasi Hukumat portali. O'qishni ko'chirishga oid muhim 7 savolga javob. https://gov.uz/oz/advice/65/document/4173

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.