ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O'ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 3 (6)

www.compling.tsuull.uz

MUNDARIJA

ISSN 2181-922X 78-90

O‘ZBEK TILIDAGI MATNLI HUJJATLARDA IZLASHNI AMALGA OSHIRISHNI TAKOMILLASHTIRISH

Ruhillo Alayev[1]
Gulshaxnoz Maxmudjonova[2]

Annotatsiya. Ushbu maqolada o‘zbek tilidagi matnli hujjatlarda izlashni amalga oshirishda TF-IDF usuli bilan takomillashtirish bosqichlari keltirilgan. O‘zbek tilida izlash natijalarini yaxshilash uchun stemming jarayoni, stemming uchun tanlangan so‘z turkumlari haqida to‘xtalinadi. Natijalar va unga sarflangan vaqt keltiriladi. Shuningdek, stemming jarayoni uchun analitik usul tatbiq qilingan.

Kalit so‘zlar: *stemming, analitik, so‘z turkumlari, hujjat, tokenizatsiya, vektorlar, kosinus.*

Kirish

Axborot qidirish (IR) – foydalanuvchi so‘rovi yoki so‘roviga javoban ma’lumotlar, hujjatlar yoki ma’lumotlarning katta ombori, ya’ni ma’lumotlar bazasidan kerakli ma’lumot olish jarayoni [Korfhage, 1997. 88]. Axborotni qidirishning asosiy maqsadi foydalanuvchining qidiruv so‘rovi asosida eng dolzarb va foydali ma’lumotlarni olish va taqdim etishdir. Ma’lumotni qidirish jarayoni foydalanuvchi so‘rovi asosida tegishli ma’lumotlarni samarali topish va taqdim etish uchun bir necha bosqichlarni o‘z ichiga oladi. Oddiy ma’lumot olish jarayonining umumiy ko‘rinishlari quyidagilar*: Foydalanuvchi so‘rovi*: Jarayon foydalanuvchi kalit so‘zlar, iboralar va savol ko‘rinishida bo‘lishi mumkin bo‘lgan so‘rov yuborganida boshlanadi. Masalan, foydalanuvchi “texnologiyalarning so‘nggi yutuqlari”ni qidirishi mumkin. *So‘rovlar tahlili:* Axborot qidirish tizimi foydalanuvchi so‘rovini uning maqsadini tushunish va asosiy atamalarni chiqarish uchun tahlil qiladi. Bu tokenizatsiya (so‘rovni alohida atamalarga

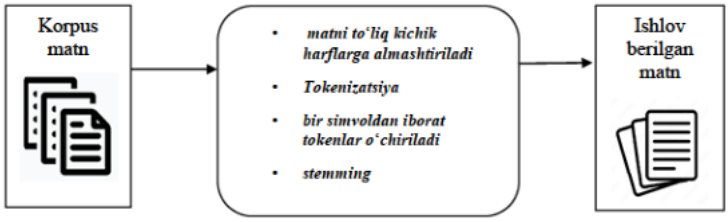
[1]*Alayev Ruhillo Habibovich* – O‘zbekiston Milliy universiteti dotsenti.
E-pochta: alayev_r@nuu.uz
ORCID: 0000-0003-3757-7711
[2]*Maxmudjonova Gulshaxnoz Ulug‘bek qizi* – ToshDO‘TAU Kompyuter lingvistikasi va raqamli texnologiyalar fakulteti magistranti.
E-pochta: gulshaxnozmahmudjonova@gmail.com
ORCID: 0009-0002-8536-0680

bo‘lish), nomuhim so‘zlarni olib tashlash va stemming kabi vazifalarni o‘z ichiga olishi mumkin*. Indekslash:* Tizim tegishli hujjatlarni samarali joylashtirish uchun indeksdan foydalanadi. Indeks – atamalarni (so‘zlarni yoki iboralarni) ular paydo bo‘lgan hujjatlar yoki yozuvlar bilan taqqoslaydigan ma’lumotlar tuzilmasi. Indeks qidiruv maydonini toraytirib, qidirish jarayonini tezlashtirishga yordam beradi [https://www.elastic.co/what-is/information-retrieval]. Axborot-qidiruv tizimlari doimiy ravishda foydalanuvchilarning fikr-mulohazalari va tizim yangilanishlari bilan aniqroq va mos keladigan qidiruv natijalarini taqdim etish orqali takomillashtirilib boradi.

Asosiy qism

Korpus – istalgan tabiiy (real) tildagi elektron shaklda saqlanadigan yozma yoki og‘zaki, kompyuterlashtirilgan qidiruv tizimiga joylashtirilgan matnlar yig‘indisi [Zaxarov, Mengliyev, Xamroyeva, 2021. 25].

Korpusda ma’lumot olish, matnlar yoki hujjatlar to‘plamidan tegishli ma’lumotlarni qidirish va olish jarayonini anglatadi. Korpus matnlarning har qanday tuzilgan yoki tuzilmagan ma’lumotlar to‘plami bo‘lishi mumkin, masalan, kitoblar, maqolalar, veb-sahifalar, tadqiqot ishlari yoki boshqa matnli ma’lumotlar to‘plami. Korpusda ma’lumot olishning maqsadi foydalanuvchining so‘roviga yoki ushbu korpusdagi ma’lumotlarga bo‘lgan ehtiyojga mos keladigan hujjatlar topish va taqdim etishdir. Foydalanuvchining berilgan so‘roviga ko‘proq mos keladigan hujjatlarni aniqlab olish uchun, dastavval hujjat matnlari oldindan qayta ishlanadi. NLPda matnga oldindan ishlov berish (text preprocessing) jarayoni hujjatni kompyuter dasturlari tushunadigan, tahlil qiladigan va formatlaydigan holatga keltirishni anglatadi, ya’ni kompyuter tizimlari uchun matnni oldindan tayyorlash va matn ustida bajariladigan maqsadlar uchun ma’lumotlarni ochishni o‘z ichiga oladi. Shuningdek, bu jarayon axborotni qidirish tizimiga yangi hujjat qo‘shishdir [Xusainova, 2022. 155]. Har bir qidiruv tizimlari uni hujjat sifatida osongina tushunishi va algoritmlari yordamida qayta ishlash uchun turli xil dastlabki ishlov berish bosqichlaridan o‘tishi kerak [Elov, 2022. 43] va bu bosqichlar ko‘p bo‘lib (lemmatizatsiya, normalizatsiya, stemming, tokenizatsiya kabi), hujjat ustida bajariladigan maqsadlarga qarab belgilab olinadi. [1-rasm]



1-rasm. Korpusdagi matnga boshlang'ich ishlov berish

Matnli hujjatlarda qidirishni amalga oshirishda TF-IDF usulini qo'llash uchun matnga oldindan ishlov berish uchun [1-rasm] yuqoridagi bosqichlarni tanlab oldik.

Hujjat matni to'liq kichik harflarga almashtiriladi. Bu qadam matnni oldindan qayta ishlashning eng oddiy va samarali bosqichidir. Matn odatda qisqartma so'zlardan yoki barcha so'zlari bosh harflardan iborat bo'lishi mumkin. Misol uchun matn boshida kelgan "Ona" so'zi bosh harfda yozilib, matn ichida esa kichik harf bilan yozilgan bo'lsa bu so'zlar kompyuter tomonidan ikki xil so'z sifatida qabul qilinadi va so'zlarni joylashtirishning keyingi bosqichlarida ikki xil so'z vektorlari hosil bo'ladi. Shunday qilib, barcha so'zlarni kichik harflar bilan yozish matnni qayta ishlashda eng yaxshi amaliyot bo'ladi.

Hujjat matni tokenizatsiya qilinadi. Tokenizatsiya – bu tabiiy tilni qayta ishlashning (NLP) asosiy bosqichi bo'lib, u matnni kichikroq birliklarga yoki tokenlarga bo'lishni o'z ichiga oladi. Ushbu tokenlar ko'pincha so'zlar, iboralar va individual belgilarni o'z ichiga oladi. Tokenizatsiya – berilgan matndagi gaplarni eng kichik o'lchov birligi hisoblangan token deb nomlanuvchi elementlarga ajratish jarayoni. Gapdagi tinish belgilari, so'zlar va raqamlar token sifatida aniqlanishi mumkin. Tokenlarni aniqlash orqali matndagi so'zlarning uchrash chastotasini topish mumkin. Ushbu qadam keyingi ishlov berish bosqichlarida keraksiz so'zlarni filtrlashga yordam beradi.

Hujjat matnidagi bir simvoldan iborat tokenlar o'chiriladi. Tokenlarni filtrlash tokenizatsiya jarayonida bir belgili tokenlarni matndan to'g'ridan-to'g'ri olib tashlashni o'z ichiga oladi. Ushbu usul odatda ma'lumot bermaydigan belgilarni yo'q qilish uchun ishlatiladi. Bu orqali matnni tahlil qilish vazifalari samaradorligini oshirish mumkin.

Hujjat matni ustida stemming amalga oshiriladi. Stemming, ma'lumot izlash va tabiiy tilni tushunish sohasida agglyutinativ tillar uchun yuqori natijalarni olishda keng foydalanilmoqda [Kışla, Karaoğlan, 2016. 402]. Ushbu qadamda stemming jarayonida so'z yasovchi qo'shimchalar olib tashlanmaydi. Faqat shakl (lug'aviy va sintaktik shakl) yasovchi qo'shimchalar olib tashlanadi. Keng ma'no-

da, stemming algoritmlarini uch guruhga bo‘lish mumkin: analitik usullar, statistik usullar va gibrid usullar [Xusainova, 2023. 70]. Morfologik tahlil yoki grammatik xususiyatlarni o‘z ichiga olgan tadqiqotlar analitik deb tasniflanadi. Shuningdek, statistika/ehtimolga asoslangan usullar statistik deb ataladi. Morfologik jihatdan so‘zlar o‘zakka qo‘shimchalar qo‘shish orqali hosil qilinadi. Bu jarayonda so‘zda fonetik o‘zgarishlar (phonetic harmony) yuzaga kelishi mumkin va bu bevosita matnda o‘z aksini topadi. Affikslar so‘z yasovchi (derivational suffixes) va shakl yasovchi (inflectional suffixes) turga ajratiladi [Hojiyev, 2005. 12]. Ushbu guruhlarning har birida so‘z variantlarining o‘zaklarini topishning odatiy usuli mavjuddir. Biz stemming jarayonida analitik usulni tatbiq qilgan holda so‘z yasovchi affiksni saqlab, shakl yasovchini olib tashlaymiz. Buni 1-jadvalda ko‘rish mumkin.

So‘z shakli	Stem
qishlog‘imiz	qishlog‘
adabiyotshunoslarni	adabiyotshunos
badavlatning	badavlat
yuragimga	yurag
singlisi	singl
mavzuyim	mavzu
deyildi	de

1-jadval. So‘z shakli stemmingi misollari

Umumiy hisobda, stemming uchun jami 6 ta so‘z turkumidan, tez-tez foydalaniladigan 10 tadan so‘z tanlab olindi. Songa qo‘shiladigan qo‘shimchalar oz bo‘lgani sababi faqat 5 dona so‘z olindi. Ot va fe’l so‘z turkumlaridan esa 10 tadan ko‘p so‘zlar olindi. Ularni 2-jadvalda ko‘rish mumkin

	So‘z turkumlari					
N	**Sifat**	**Ot**	**Ravish**	**Fe’l**	**Son**	**Olmosh**
1	chiroyli	fikr	erta	qilmoq	ikki	barcha
2	sovuq	kitob	oldin	bilmoq	besh	men
3	sariq	davlat	ichkari	yozmoq	to‘rt	o‘sha
4	yaxshi	narsa	yaxshi	boshlamoq	olti	biz
5	katta	katta	katta	bormoq	sakkiz	qaysi
6	yomon	oila	yomon	aytmoq		kim

7	kichik	mavzu	bugun	bermoq		sen
8	baland	so‘z	ko‘p	ketmoq		hamma
9	uzun	maktab	avval	olmoq		bu
10	uzoq	ona	hozir	qolmoq		u

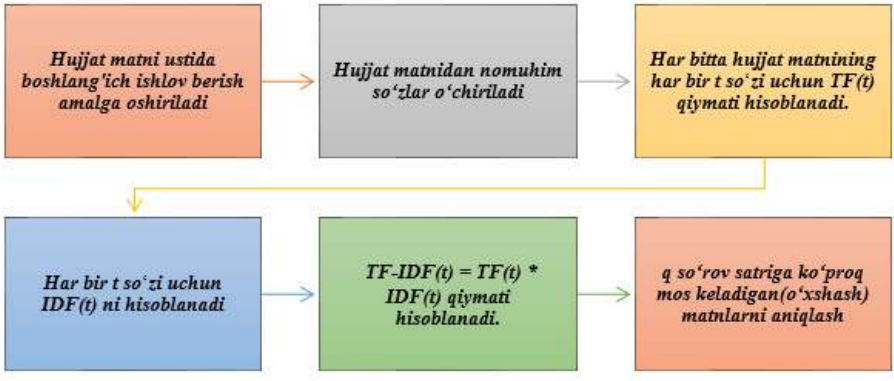
2-jadval. Stem uchun olingan so‘z turkumlari

Ushbu Stemming variantidan foydalanishning 2 ta asosiy sababi mavjud:

a) Agar stemmingning o‘rniga lemmatizatsiya usulidan foydalanilsa, tovush o‘zgarishiga uchragan so‘zlarda lemma satri so‘z satri tarkibida mavjud bo‘lmasligi mumkin. Ya’ni, lemmatizatsiya – so‘zning lug‘atda mavjud bo‘lgan shaklini (leksemani) aniqlash jarayoni hisoblanib, stemmingdan farq qiladi. Ushbu bosqichda bir nechta qo‘shimcha amallar bajariladi. Lemmatizatsiya orqali so‘zning faqat to‘g‘ri lug‘atdagi shakli aniqlanadi. Bu esa so‘z satrini matn satri ichida mavjud ekanligini tekshirish jarayonini murakkablashtiradi. Masalan, “Bu bizning qishlog‘imiz bo‘ladi” gapida “qishlog‘imiz” so‘zining lemmasi “qishloq” satri hisoblanib, “qishloq” satrini “Bu bizning qishlog‘imiz bo‘ladi” satri ichidan oddiy satr qidirish funksiyalari orqali aniqlay olmaydi. Sababi yuqoridagi gapda qishloq so‘zi tovush o‘zgarish hodisasiga uchragan va so‘z satrida *qishlog‘* bo‘lib keladi. Lemmatizatsiyada faqat qishloq so‘zi qidiriladi. Tovush o‘zgarishiga uchragan so‘zlarda bu ko‘p vaqtni va bir nechta qadamni talab qiladi.

b) Agar barcha so‘z yasovchi qo‘shimchalarni olib tashlashga asoslangan “stemming” variantidan foydalanilsa, so‘zning dastlabki ma’nosi yo‘qoladi. Masalan, “adabiyotshunoslarni” so‘zining stemi “adab”, “adabiyot”, “adabiyotshunos” va boshqa so‘zlarning stemi bilan bir xil. Bundan kelib chiqadiki, “adabiyotshunos” qidirilayotgan bo‘lsa, “adab”, “adabiyot”, “adabiyotshunos” so‘zlar qatnashgan matnlar ham qidiruv natijasida paydo bo‘ladi. Shunday ekan stemmer sintaktik va lug‘aviy shakl yasovchi qo‘shimchalarni olib tashlashga qaratilgan tovush o‘zgarishiga uchragan o‘zak so‘zlarni o‘z holicha qoldiradi.

Matnli hujjatlarda qidirishni amalga oshirishda TF-IDF usulini qo‘llash uchun matnga oldindan ishlov bergandan so‘ng keyingi qadamlarga o‘tiladi. Quyidagi jadvalda ketma-ketlikda tushuntirilgan[3-jadval]:



3-jadval. Matnli hujjatlarda qidirishni amalga oshirish jarayoni

1-qadam. Hujjat matni ustida boshlang‘ich ishlov berish amalga oshirildi (1-rasm) .

2-qadam. Hujjat matnidan nomuhim so‘zlar o‘chiriladi. Nomuhim so‘zlar hujjatning katta qismini tashkil etuvchi matn tarkibidagi so‘zlar bo‘lib, ularning umumiy xususiyati shundaki, hujjatda muhim ma’lumotga ega emas; faqat grammatika tufayli ishlatiladi. Hujjatdagi matndan nomuhim so‘zlar tushirib qoldirilsa, matn mazmuni o‘zgarmaydi [Maxmudjonova, 2023. 205].

3-qadam. Har bitta hujjat matnining har bir t so‘zi uchun TF(t) qiymati hisoblanadi.

$TF(t) = N(t,d) / N(d)$,

bunda, N – matnlar soni, N(t,d) – d matndagi t so‘zning uchrashi soni,

N(d) – d matndagi so‘zlar soni.

4-qadam. Har bir t so‘zi uchun IDF(t) ni hisoblanadi.

$IDF(t) = \log(N/DF(t))$ (1)

bunda, DF(t) – t so‘z uchraydigan hujjatlar soni, t so‘z bironta hujjatda ham uchramasa, unda DF(t) = 0 bo‘ladi, natijada (1) formulada nolga bo‘lish holati yuzaga keladi. Shuning uchun quyidagi ko‘rinishdagi formuladan foydalaniladi:

$IDF(t) = \log(N/DF(t)+1)$

5-qadam. TF-IDF(t) = TF(t) * IDF(t) qiymati hisoblanadi. TF-IDF usuli birinchi bo‘lib hujjatlarni qidirish va ma’lumot olish algoritmlarida qo‘llanilgan bo‘lib, foydalanuvchi so‘roviga mos tarzda axborot tizimi ma’lumotlar bazasidan eng kerakli hujjatlarni aniqlab bergan. TF-IDF usuli ikkita statistik ko‘rsatkichni o‘zaro ko‘paytirish orqali aniqlanadi:

– so‘zlar chastotasi (Term Frequency, TF): so‘zning hujjat-

da necha marta uchrashi;

– hujjatning teskari chastotasi (Inverse Document Frequency, IDF): hujjatlar to‘plamidagi so‘zning teskari chastotasi. Ushbu chastota orqali aniqlangan qiymatlarga ko‘ra unikal so‘zlar yuqori ballga, ko‘p qo‘llaniladigan so‘zlar past ballga ega bo‘ladi [Elov, Xusainova, Xudayberganov, 2023. 1775].

– so‘zlar chastotasi (Term Frequency, TF): so‘zning hujjatda necha marta uchrashi;

– hujjatning teskari chastotasi (Inverse Document Frequency, IDF).

6-qadam. Endi q so‘rov satriga ko‘proq mos keladigan (o‘xshash) matnlarni aniqlash kerak. Buning uchun quyidagi 2 ta usuldan foydalanish mumkin.

1) “Muvofiqik ko‘rsatkichi” asosida tartiblash. Bu usulda har bir hujjatning q so‘rovga “muvofiqlik ko‘rsatkichi” hisoblanadi. “Muvofiqlik ko‘rsatkichi” – o‘xshashlikni hisoblashning eng oddiy usuli, bu usulda har bir hujjat uchun so‘rov satrida mavjud bo‘lgan token stemlarining TF-IDF qiymatlari qo‘shiladi. Misol uchun, “qizil mashinada” so‘rovi so‘zlari uchun “qizil” va “mashina” stemlari aniqlanadi. So‘ng har bir hujjatda ushbu stemlar mavjudligi tekshiriladi, agar stemlar hujjatda mavjud bo‘lsa, u holda stemning ushbu hujjatdagi TF-IDF qiymati hujjatning “muvofiqlik ko‘rsatkichi”ga qo‘shiladi. Shu tarzda, eng yuqori “muvofiqlik ko‘rsatkichi”ga ega K ta hujjat saralab olinadi.

2) Kosinus o‘xshashligidan foydalangan holda tartiblash. 1-usul hujjatlarning “muvofiqlik ko‘rsatkichi”ni aniqlashni taqdim etsa-da, lekin u uzun so‘rovlar uchun to‘g‘ri kelmaydi, ularni to‘g‘ri tartiblay olmaydi. Kosinus o‘xshashlik barcha hujjatlarni TF-IDF tokenlari vektorlari sifatida ifodalaydi va kosinus fazodagi o‘xshashlikni(vektorlar orasidagi burchak) o‘lchaydi. Ba’zi hollarda so‘rov uzunligi bir necha marta kichik bo‘lishi, lekin u hujjatga ko‘proq o‘xshashlikka ega bo‘lishi mumkin. Bunday hollarda kosinus o‘xshashligi so‘rov va hujjat o‘rtasidagi o‘xshashlikni aniqlash uchun eng yaxshi usuldir. Misol uchun [Kristian, Filip, 2016. 672] Kosinus masofasi bo‘yicha original va tarjima qilingan hujjatlarni solishtirishda TF/IDF sxemalarini taqqoslangan va bu vaziyatda past chastotali so‘zlarni kiritish bilan bir qatorda atama-chastota (TF) vaznini to‘g‘ri tanlash aniqlikni amalga oshirishda hal qiluvchi ahamiyatga ega ekanligi aniqlangan. Uzun so‘rovlar uchun kosinus o‘xshashligidan foydalanish maqsadga muvofiq. Kosinus o‘xshashlik barcha hujjatlarni TF-IDF tokenlari vektorlari sifatida ifodalaydi va kosinus

fazodagi o‘xshashlikni (vektorlar orasidagi burchak) o‘lchashga yordam beradi.

Natijalar tahlili: 1,5 gb hajmdagi 8.000.000 ta o‘zbek tili lotin yozuvidagi matnlar ustida berilgan qidiruv so‘rovi satri bo‘yicha qidiruv tajribalari amalga oshirildi. Izlash so‘rovlari SQL Server 2019 ma’lumotlar bazasida amalga oshirildi. Bunda izlash uchun n-gramlardan foydalanildi. N-grammni moslashtirish usullari eng keng tarqalgan usullardan biridir [Elov, Xudayberganov, Xusainova, 2023. 588]. N-gramm – N ta token (so‘z) lardan iborat ketma-ketligidir. Matndagi Ngrammlar ko‘p so‘zli iboralar yoki leksik birliklar sifatida aniqlanadi. Quyidagi so‘z birikmalari mos ravishda 2- va 3-grammni ifodalaydi: “Amir Temur”, “Katta Buxoro kanali”. Ko‘p hollarda matndagi alohida so‘zlarni (tokenlarni) tahlil qilishdan ko‘ra N-grammlarni tahlil qilish samarali natijalarni qaytaradi [Ruambo, Nikolas, 2019. 88]. Tahlil uchun 5 ta 1-gramm, ya’ni 1 so‘zdan, 5 ta 2-gramm – 2 so‘zdan, 5 ta 3-gramm – 3 so‘zdan [4-jadval] iborat qidiruv so‘rovlari uchun 4 xil usulda tajriba o‘tkazildi.

1-grammlar	**2-grammlar**	**3-grammlar**
Qalam	*Shahrimiz aholisi*	*Badiiy kitoblarni o‘qish*
Maqolalar	*Ko‘m-ko‘k osmon*	*Tabiat manzarasini chizish*
Chiroyli	*Qora bulutlar*	*Birlashgan Millatlar tashkiloti*
Qishlog‘imiz	*Issiq nonlar*	*Qishloq aholisining uylari*
Bolalar	*Dengizning qurishi*	*Yosh bolalarning qiziqishlari*

4-jadval. N-grammlar ro‘yxati

Qidiruv so‘rovlari uchun 4 xil usul
1) Ma’lumotlar bazasining “like” operatori orqali izlash amalga oshirildi.
2) SQL Server-ning Full-Text Search komponentidan foydalangan holda izlash amalga oshirildi.
3) Qidiruv so‘rovi so‘zlarining stemi hisoblanmagan holda, hujjatlarning butun matni bo‘yicha izlash amalga oshirildi.
4) Hujjatlar matni stemlangan holda oldindan tayyorlandi. Stemlangan hujjat matnlarining TF/IDF qiymatlari hisoblandi. Qidiruv so‘rovi so‘zlarining stemini hisoblangan holda izlash amalga oshirildi.

5-jadval. Qidiruv so‘rovlari uchun usul

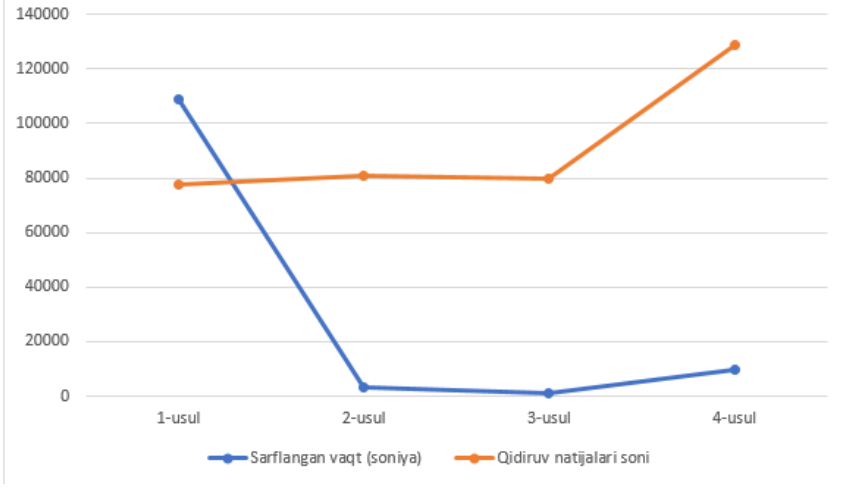
Qidiruv so‘rovlari uchun jami 15 ta so‘z olinib 4 xil usulda tajriba o‘tkazildi. Har bir qidiruv so‘rovlari bo‘yicha qidiruv natijalari soni quyidagi 6-jadvalda keltirildi.

Qidiruv so‘rovlari	**1-usul**	**2-usul**	**3-usul**	**4-usul**

qalam	1595	1640	1570	1868
maqolalar	3548	3847	3839	22434
chiroyli	9002	9148	9132	9240
qishlog‘imiz	225	237	237	19702
bolalar	60668	62742	61945	63713
shahrimiz aholisi	46	47	47	1821
ko‘m-ko‘k osmon	8	8	14	4871
qora bulutlar	264	271	260	300
issiq nonlar	13	13	12	165
dengizning qurishi	13	14	14	34
badiiy kitoblarni o‘qish	3	3	3	25
tabiat manzarasini chizish	0	0	0	1
Birlashgan Millatlar tashkiloti	2058	2832	2846	4811
qishloq aholisining uylari	0	0	0	13
yosh bolalarning qiziqishlari	0	0	0	2

6-jadval. Qidiruv so‘rovlari bo‘yicha qidiruv natijalari

Har bir usul bo‘yicha barcha qidiruv so‘rovlarini amalga oshirish uchun sarflangan vaqt va jami qidiruv natijalari soni yig‘indisi



1-diagrammada keltirilgan.

1-diagramma. Usullar bo‘yicha sarflangan vaqt va jami qidiruv natijalari soni

Xulosa

Sinov natijalaridan ko‘rish mumkinki, qidiruv so‘zlarining stemini hisoblangan holda izlash ko‘proq qidiruv natijalarini (matn-

larni) topish imkonini bergan. Har qanday qidiruv jarayonida stemlash yaxshi natija beradi va tez ishlash samaradorligini oshiradi. Bu qadamda qidiruv so‘roviga mos hujjatlar aniqlandi.

Keyingi qadamda yuqoridagi qadam natijalari (hujjat matnlari) “Kosinus o‘xshashligi”dan foydalangan holda tartiblash amalga oshirildi.

Foydalanilgan adabiyotlar

Elov B., Xusainova Z., Xudayberganov N., 2023. “O‘zbek tili korpusi matnlari uchun TF-IDF statistik ko‘rsatkichni hisoblash”. Journal of Science and Innovative Development.

Elov B., 2022. “Tabiiy tilni qayta ishlash (NLP) da spacy modulidan foydalanish”. Journal of Science and Innovative Development.

Elov B., Xudayberganov N., Xusainova Z., 2023. “Morfologik teg va n-grammlar vositasida soxta yangiliklarni tasniflash”. EasyChair Preprint. https://www.elastic.co/what-is/information-retrieval

Kışla T., Karaoğlan B. 2016. “A hybrid Statistical Approach to Stemming in Turkish: An Agglutinative Language”. Anadolu University Journal of Science and Technology A Applied Sciences and Engineering. https://doi.org/10.18038/btda.31812

Kristian B., Filip K., 2016. “Quick and Reliable Document Alignment via TF/IDF-weighted Cosine Distance”. Proceedings of the First Conference on Machine Translation.

Maxmudjonova G., 2023. “Nomuhim so‘zlar tushunchasi va uning ahamiyati”. Kompyuter lingvistikasi: muammolar, yechim, istiqbollar Xalqaro ilmiy-amaliy konferensiya materiallari, 204 – 211. Toshkent: ToshDO‘TAU nashriyoti, http://compling.navoiy-uni.uz

Robert R. Korfhage., 1997. Information Storage and Retrieval. Wiley Computer Pub

Ruambo F A., Nikolas M R., 2019. “Towards enhancing information retrieval systems: A brief survey of strategies and challenges,” International Congress on Ultra Modern Telecommunications and Control Systems and Workshops.

Xusainova Z., 2022. “NLP: Tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash”. O‘zbek amaliy filologiyasi istiqbollari mavzusidagi Respublika ilmiy-amaliy konferensiyasi materiallari, 154 – 163. Toshkent: ToshDO‘TAU nashriyoti.

Xusainova Z., 2023. “O‘zbek tilida stemmingni amalga oshirishning gibrid statistik yondashuvi”. Kompyuter lingvistikasi: muammolar, yechim, istiqbollar Xalqaro ilmiy-amaliy konferensiya materiallari, 69 – 75. Toshkent: ToshDO‘TAU nashriyoti, http://compling.navoiy-uni.uz

Zaxarov V., Mengliyev B., Xamroyeva Sh. 2021. Korpus lingvistikasi O‘quv qo‘llanma.

Ҳожиев А. 2005. Ўзбек тилида сўз ясалиши.

IMPROVING SEARCH IN TEXT DOCUMENTS IN THE UZBEK LANGUAGE

Ruhillo Alayev[1]
Gulshaxnoz Maxmudjonova [2]

Abstract. This article outlines the improvement of the search in Uzbek text documents using the TF-IDF method. The process of stemming to improve search results in the Uzbek language, the part of speech selected for stemming will be discussed. The results and the time spent on it are given. Also, an analytical method for the stemming process was applied.

Keywords: stemming, analytics, part of speech, document, tokenization, vectors, cosine similarity.

References

Elov B., Xusainova Z., Xudayberganov N., 2023. “O‘zbek tili korpusi matnlari uchun TF-IDF statistik ko‘rsatkichni hisoblash”. Journal of Science and Innovative Development.

Elov B., 2022. “Tabiiy tilni qayta ishlash (NLP) da spacy modulidan foydalanish”. Journal of Science and Innovative Development.

Elov B., Xudayberganov N., Xusainova Z., 2023. “Morfologik teg va n-grammlar vositasida soxta yangiliklarni tasniflash”. EasyChair Preprint. https://www.elastic.co/what-is/information-retrieval

Kışla T., Karaoğlan B. 2016. “A hybrid Statistical Approach to Stemming in Turkish: An Agglutinative Language”. Anadolu University Journal of Science and Technology A Applied Sciences and Engineering. https://doi.org/10.18038/btda.31812

Kristian B., Filip K., 2016. “Quick and Reliable Document Alignment via TF/IDF-weighted Cosine Distance”. Proceedings of the First Conference on Machine Translation.

Maxmudjonova G., 2023. “Nomuhim so‘zlar tushunchasi va uning

[1]*Alayev Ruhillo Habibovich* – Associate Professor, National University of Uzbekistan.
E-mail: alayev_r@nuu.uz
ORCID: 0000-0003-3757-7711
[2]*Maxmudjonova Gulshaxnoz Ulug‘bek qizi* – Master of degree, Alisher Navo’i Tashkent State University of Uzbek Language and Literature.
E-mail: *gulshaxnozmahmudjonova@gmail.com*
ORCID: 0009-0002-8536-0680

ahamiyati". Kompyuter lingvistikasi: muammolar, yechim, istiqbollar Xalqaro ilmiy-amaliy konferensiya materiallari, 204 – 211. Toshkent: ToshDO'TAU nashriyoti, http://compling.navoiy-uni.uz

Robert R. Korfhage., 1997. Information Storage and Retrieval. Wiley Computer Pub

Ruambo F A., Nikolas M R., 2019. "Towards enhancing information retrieval systems: A brief survey of strategies and challenges," International Congress on Ultra Modern Telecommunications and Control Systems and Workshops.

Xusainova Z., 2022. "NLP: Tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash". O'zbek amaliy filologiyasi istiqbollari mavzusidagi Respublika ilmiy-amaliy konferensiyasi materiallari, 154 – 163. Toshkent: ToshDO'TAU nashriyoti.

Xusainova Z., 2023. "O'zbek tilida stemmingni amalga oshirishning gibrid statistik yondashuvi". Kompyuter lingvistikasi: muammolar, yechim, istiqbollar Xalqaro ilmiy-amaliy konferensiya materiallari, 69 – 75. Toshkent: ToshDO'TAU nashriyoti, http://compling.navoiy-uni.uz

Zaxarov V., Mengliyev B., Xamroyeva Sh. 2021. Korpus lingvistikasi O'quv qo'llanma.

Hojiyev A., 2005. O'zbek tilida so'z yasalishi. – Toshkent.