ISSN 2181-922X

UZBEKISTAN O‘ZBEKISTON

LANGUAGE & CULTURE

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 2 (6)

www.compling.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 2 (6)

compling.tsuull.uz

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

Jurnal haqida ma’lumot

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi – Oliy attestatsiya komissiyasi ilmiy nashrlar ro‘yxatidagi “O‘zbekiston: til va madaniyat” akademik jurnalining ilovasi hisoblanib, unda professor-o‘qituvchilar, doktorantlar, stajor-tadqiqotchilar, mustaqil izlanuvchilar, magistrantlarning kompyuter lingvistikasi, jumladan, tabiiy tilga ishlov berish (NLP), o‘zbek tilining formal grammatikasi, korpus lingvistikasi, mashina tarjimasi, nutqni qayta ishlash tizimlari, intellektual tizimlar, kompyuter leksikografiyasi hamda lingvistik ontologiyalar kabi sohalarga oid tadqiqotlari nashr qilinadi.

Jurnal ilovasi bir yilda to‘rt marta chop etiladi.

O‘zbek, turk, rus va ingliz tillarida yozilgan maqolalar qabul qilinadi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e‘lon qilinadi.

Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi 2023-yildan chiqa boshlagan.

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti. O‘zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103-uy.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

Information about the magazine

"Uzbekistan: language and culture. "Computer Linguistics" series is an appendix of the academic journal "Uzbekistan: Language and Culture" in the list of scientific publications of the Higher Attestation Commission, in which computer linguistics, including natural language processing (NLP) of professors-teachers, doctoral students, intern-researchers, independent researchers, master's students, researches related to formal grammar of the Uzbek language, corpus linguistics, machine translation, speech processing systems, intelligent systems, computer lexicography and linguistic ontologies are published.

The magazine supplement is published four times a year.

Articles written in Uzbek, Turkish, Russian and English languages are accepted.

The journal also publishes book reviews, literature reviews, conference reports, and research project results.

The opinion of the authors may differ from the editorial point of view.

"Uzbekistan: language and culture. "Computer Linguistics" series has been published since 2023.

Tashkent State University of Uzbek Language and Literature named after Alisher Navoi. Yusuf Khos Hajib street, 103, Yakkasaray district, Tashkent, Uzbekistan.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

MUNDARIJA

CONTENT

ISSN 2181-922X 20-37

O‘ZBEK TILI MILLIY KORPUSI QIDIRUV TIZIMINI OPTIMALLASHTIRISHDA LEMMATIZATSIYADAN FOYDALANISH

Zilola Xusainova[1]

Annotatsiya

Lemmatizatsiya – lemma(leksema)ning ma’noli shaklini aniqlash va uni flektiv/hosila shakllari o‘rnida ishlatishni o‘z ichiga olgan tabiiy tilni qayta ishlash usuli. Ushbu maqolada o‘zbek tili milliy korpusi matnlarini lemmalash va uning qidiruv tizimi natijalarini optimallashtirishga (SEO, Search Engine optimization) qo‘llash masalasi ko‘rib chiqilgan. Axborot qidiruv tizimlari algoritmlari doimiy rivojlanishda bo‘lib, inson tilini tushunish va talqin qilish qobiliyatini yaxshilashda davom etmoqda. *Lemmatizatsiya* kabi tabiiy tilni qayta ishlash usullaridan foydalanish til korpusidagi qidiruv tizimi natijasini optimallashtirish uchun tobora muhim ahamiyat kasb etmoqda.

Kalit so‘zlar: *lemmatizatsiya, lemma, POS, SEO, qidiruv tizimlari, o‘zbek tili milliy korpusi.*

KIRISH

Tabiiy tilni qayta ishlash (Natural Language Processing, NLP) dunyodagi eng tez rivojlanayotgan sohalardan biridir. NLP – bu sun’iy intellekt sohasi bo‘lib, unda inson tilidagi ma’lumotlarni kompyuterlar vositasida intellektual tarzda tahlil qilish va tushunish amalga oshiriladi. Bugungi kunda NLP algoritmlaridan *elektron pochtada spamni aniqlash* va *qidiruv tizimlarida foydalanuvchi so‘rovlarini optimallashtirish* kabi masalalarni hal qilishda qo‘llanilmoqda. NLP - mashinalarga insonlar bilan *tabiiy tilda muloqot qilishda*, *matnni intellektual tahlil qilish, nutqni tushunish* va *uni to‘g‘ri formatda talqin qilish* kabi vazifalarni bajarishda yordam beradi [Elov, Hamroyeva, Xusainova, 2022: 19-26]. Zamonaviy mashinalar katta hajmdagi ma’lumotlar (BigData)ni samarali va tez tahlil qilish imkoniyatiga ega [Plale, 2013]. Axborot tizimlaridan muntazam foydalanish jarayonida juda ko‘p yangi strukturalanmagan ma’lumotlar hosil qilina-

[1] *Xusainova Zilola Yuldashevna* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti tayanch doktoranti.
E-pochta: xusainovazilola@navoiy-uni.uz
ORCID: 0000-0003-4357-7515

di. Strukturalanmagan ma’lumotlarni NLP vositalari orqali qayta ishlash murakkab va dolzarb amal bo‘lib, NLP usullari vositasida matndan kerakli ma’lumotlarni aniq olish uchun juda muhimdir. NLPda matnni qayta ishlash quyidagi boshqichlarda bajariladi [Elov, Khamroeva, Xusainova, 2023: 183]:

– *morfologik tahlil – so‘zning birlamchi ma’nosini aniqlash bosqichi;*

– *leksik tahlil – alohida so‘zlarning ma’nosini izohlash;*

– *sintaksis tahlil – gaplarning grammatik ma’nosini aniqlash;*

– *semantik tahlil – ma’nolilikni tekshirish;*

– *ochiqlash integratsiyasi – matnning yaxlit xususiyatlariga e’tibor qaratish va jumlalar o‘rtasidagi aloqalarni o‘rnatish;*

– *pragmatik tahlil – gapning turli holatdagi ma’nosi.*

Yuqorida keltirilgan matnni qayta ishlash bosqichlarning aksariyatida *tokenizatsiya, stemming, lemmatizatsiya* jarayonlari kabi boshlang‘ich amallar bajariladi [Saloot, Pham, 2021; Vajjala, Majumder, Gupta, Surana, 2020: 455]. O‘zbek tili leksik birliklarini tokenlash va stemlash jarayoni B.B.Elov, Sh.M.Hamroyeva, R.H.Alayev va Z.Y.Xusainovalarning ilmiy tadqiqot ishlarida batafsil keltirilgan.

Lemmatizatsiya so‘zning lemma (leksema) shaklini aniqlash va uni flektiv/hosila shakllari o‘rnida ishlatishni o‘z ichiga oladi. Lemmatizatsiya jarayoni orqali til korpusi matnlari mazmunini *tahlil qilish aniqligini oshirish* mumkin. Shuningdek, lemmatizatsiya jarayoni orqali qidiruv tizimlari algoritmlari korpus matnlari mazmunini tushunishi, uni tartiblash jarayonini amalga oshirishga yordam beradi [Jabeen, 2018; Balakrishnan, Ethel, 2014; Khyani, 2021].

Ushbu maqolada NLPda lemmatizatsiya jarayoni va uning SEOdagi roli ko‘rib chiqiladi. Lemmatizatsiyani amalga oshirish bosqichlari, metodlari va algoritmlarining qanday ishlashini, uni korpus matnlariga qo‘llash usullari keltiriladi. Shuningdek, lemmatizatsiya jarayonida yuzaga kelishi mumkin bo‘lgan xatoliklar, ularni bartaraf qilish usullari, SEOni yaxshilash uchun lemmatizatsiyaga alternativ yondashuvlar muhokama qilinadi [Baklanov, Bezkorovainyi, Kolesnyk, 2022; marketbrew.ai].

Lemmatizatsiya va stemming

Lemmatizatsiya va stemming – tabiiy tilni qayta ishlash (NLP)da tahlil murakkabligini kamaytirish maqsadida ishlatiladigan ikkita usul. Bu ikkala usul ham so‘zlarni asosiy shakllariga qisqartirishdan iborat, ammo ular bu amalni bajarish maqsadi va

unga erishish usullari bilan farqlanadi [Elov, Khamroeva, Xusainova, 2023: 185; Elov, Hamroyeva, Abdullayeva, Xusainova, Xudayberganov, 2023: 46].

Lemmatizatsiya – so‘zning asosiy shakli bo‘lgan **lemmasi(leksema)ga** qisqarish jarayoni. Bu asosiy shakl ko‘pincha so‘zning **lug‘atdagi shakli** deb ham ataladi. Masalan, *“bajargan”* so‘zining lemmasi *“bajarmoq”*, *“sening”* so‘zining lemmasi esa “*sen*” hisoblanadi. Lemmatizatsiya so‘zning *kontekstini*, jumladan, *POS tegini, zamon* va *turkumi*ni hisobga oladi.

1-jadval. Ingliz, rus, turk va o‘zbek tilidagi so‘zlar lemmasi

o‘zbek		**turk**		**rus**		**ingliz**	
so‘zshakl	**lemma**	**so‘zshakl**	**lemma**	**so‘zshakl**	**lemma**	**so‘zshakl**	**lemma**
izladim	izlamoq	gelmiştim	gelmek	идем	идти	went	go
olmalar	olma	sebzesini	sebze	пятый	пять	spoken	speak
yutug‘imiz	yutuq	Mutluluk	mutlu	телевизору	телевизор	better	good

NLP vazifalarini amalga oshirishda ko‘p hollarda lemmatizatsiya jarayonidan foydalaniladi. Lemmatizatsiya jarayoni stemmingdan farqli ravishda so‘z ma’nosini aniqroq ifodalashni ta’minlaydi [Nunzio, Vezzani, 2018: 2253]. Chunki lemmatizatsiya jarayonida so‘zning konteksti hisobga olinsa, stemmingda hisobga olinmaydi.

Stemming – kontekstni hisobga olmagan holda so‘zning asosiy shakli bo‘lgan o‘zagiga qisqartirish jarayoni. Stemlar ko‘pincha so‘zning o‘zak shakli deb ataladi. Masalan, *“yugurish”* so‘zining o‘zagi *“yugur”*, *“sening”* so‘zining o‘zagi esa *“sen”, sening* so‘zining stemi esa *“se”*dir. *Stemming* – bu NLP masalalarini yechishdagi murakkabliklarni kamaytirishning asosiy omili bo‘lib, so‘zning kontekstini hisobga olmaydi. Bu esa so‘z ma’nosini ifodalashda ko‘plab xatolarga olib kelishi mumkin. Chunki so‘zning stemi har doim ham so‘z ma’nosini to‘g‘ri ifodalamasligi mumkin [Elov, Hamroyeva, Abdullayeva, Xusainova, Xudayberganov, 2023: 52; Xusainova, 2023: 72; Xusainova, 2022: 158]. Mualliflar tomonidan olib borilgan tadqiqot natijasiga ko‘ra o‘zbek tili milliy korpusi matnlarida UzbStemmer algoritmning samaradorligi 95.5%ni tashkil etgan [uznatcorpara.uz/uz/POSTag]. Masalan, *“sotib oldim”* so‘zining stemi 2 ta *“sot”* va *”ol”* bo‘lsa, lemmasi *“sotib olmoq”*dan iborat 1 ta leksik birlikni ifodalaydi.

Bugungi kunda NLPning ko‘plab vazifalarida stemming jarayonidan foydalaniladi. Chunki stemming so‘zlarning murakkabligini kamaytirish uchun oddiyroq va tezroq yondashuv hisoblanadi [Balakrishnan, Ethel, 2014; Khyani, 2021; Pattnaik, Nayak, Patnaik, 2020]. Shuningdek, qidiruv natijalariga so‘zning

o‘zgarishlarini kiritish imkonini taqdim etganligi sababli stemming jarayoni ko‘proq qidiruv tizimlarida qo‘llaniladi.

Til korpusi qidiruv tizimini optimallashtirish

Lemmatizatsiya – qidiruv tizimi samaradorligini oshirish uchun so‘zlarni lemma (asosiy) shakliga qisqartirish jarayoni. Ushbu jarayon, ayniqsa, qidiruv natijalarining aniqligini va korpus matni kontentini tushunish uchun foydali bo‘lishi mumkin. Lemmatizatsiyaning asosiy afzalliklaridan biri shundaki, u *qidiruv tizimlariga matn konteksti va “ma’nosi”ni yaxshiroq tushunishga yordam beradi*. Masalan, “*olma*” so‘zi *harakat*ni bildiruvchi fe’l yoki *meva*ni bildiruvchi ot bo‘lishi mumkin. Ushbu so‘zni lemmatizatsiya qilish orqali qidiruv tizimlari qaysi ma’no yuzaga chiqqanligini aniqroq topishi va tegishli qidiruv natijalarini taqdim etishi mumkin.

Lemmatizatsiyaning yana bir afzalligi shundaki, u qidiruv tizimlariga *turli so‘zlar va tushunchalar o‘rtasidagi munosabatlarni yaxshiroq tushunish* imkonini beradi. Masalan, agar foydalanuvchi qidiruv tizimida *“yugurish uchun poyabzal”* so‘z birikmasini qidirsa, *“yugurish”* lemmasi *“poyabzal”* lemmasi bilan bog‘lanadi. Bu qidiruv tizimiga foydalanuvchining ma’lum turdagi poyabzallarni qidirayotganini tushunish imkonini beradi. Bu esa ma’lumotlar bazasidan so‘rovga ko‘proq mos keladigan qidiruv natijalarini shakllantirishga xizmat qiladi.

Lemmatizatsiya, shuningdek, korpus matnlari tarkibi (mazmuni)ning umumiy o‘qilishi va aniqligini yaxshilaydi. So‘zlarni lemma shakliga qisqartirish orqali qidiruv tizimlarida kontentning ma’nosi va kontekstini tushunishni osonlashtiradi. Bu esa matnning muayyan so‘z birikmalari bo‘yicha yuqori qidiruv reytinglari va korpus matnlariga murojaatlar sonini oshishiga olib kelishi mumkin. Shuningdek, lemmatizatsiya SEO samaradorligini oshirishga ham yordam beradi. So‘zlarni standartlashtirish va ularni asosiy shaklga qisqartirish orqali o‘zbek tili korpus ma’lumotlari, tendensiyalarni kuzatish, tahlil qilishni osonlashtirishi mumkin. Bu SEO mutaxassislariga qidiruv samaradorligini aniq o‘lchash va kelajakdagi strategiyalar haqida to‘g‘ri qaror qabul qilishga yordam beradi.

Ta’kidlash kerakki, SEOda lemmatizatsiyadan tashqari boshqa optimallashtirish usullari ham mavjud [Fisenko, Baliun, Grigorenko, 2022; Andrikopoulos, Sun, Guo, 2017]. Korpus matnlari (veb-sayt) dagi *kalit so‘zlarni o‘rganish*, *sahifani optimallashtirish* va *qidiruv reytinglariga ta’sir ko‘rsatadigan boshqa omillarga e’tibor qaratish* ham muhim. Lemmatizatsiya eng yaxshi natijalarga erishish uchun

keng qamrovli SEO strategiyasining bir qismi sifatida ishlatilishi kerak.

Umuman olganda, lemmatizatsiya qidiruv tizimini optimallashtirish va veb-saytning foydalanuvchi tajribasini yaxshilash uchun muhim vosita sifatida ishlatilishi mumkin. So‘zlarni asosiy (lemma) shakliga qisqartirish, qidiruv natijasining aniqligi, dolzarbligini oshirish orqali *veb-saytga ko‘proq maqsadli trafikni jalb qilish* va *daromadni oshirishga* yordam beradi. Quyidagi 2-jadvalda stemming va lemmatizatsiya jarayonlari farqi keltirilgan.

2-jadval. *Stemming va lemmatizatsiya jarayonlari*

so‘zshakl	stem	lemma	asos
kelajagimiz	kelajag	kelajak	kelajak
o‘quvchilar	o‘quv	o‘quvchi	o‘qi
borib ketdi	bor ket (2ta)	borib ketmoq	bor ket (2ta)
ega bo‘lishdi	ega bo‘l (2ta)	ega bo‘lmoq	ega bo‘l (2ta)
taqillatdi	taqilla	taqillamoq	taq
undami	un	u	u
keldilar	kel	kelmoq	kel
uyda	uy	uy	uy
har birimiz	har bir (2ta)	har bir	har bir (2ta)

Qidiruv tizimlari algoritmlarida lemmatizatsiyadan foydalanish.

Bugungi kunda qidiruv tizimlari lemmatizatsiyani amalga oshirishda foydalaniladigan ko‘plab usullar mavjud bo‘lib, keng tarqalgan usullardan biri **lemma lug‘ati**dan foydalanishdir [marketbrew.ai, Gashkov, Eltsova, 2018: 55]. Ushbu lug‘atda *so‘zlar ro‘yxati* va ularga mos *lemmalar* mavjud, ular ko‘pincha tildagi so‘zning **ildiziga (ing. root)** asoslanadi. Qidiruv tizimi matnda biror so‘zga duch kelganda, u o‘zining *lemma lug‘atidan so‘zni qidiradi* va *so‘zning asosiy shakli sifatida tegishli lemmadan* foydalanadi. Bu esa qidiruv tizimiga so‘zning ma’nosi va kontekstini aniqroq “tushunish”ga, uni tegishli qidiruv so‘rovlariga yaxshiroq moslashtirishga imkon beradi.

Lemmatizatsiya qidiruv tizimlari uchun, ayniqsa, bir so‘zning turli shakllariga ega bo‘lgan ingliz tili kabi flektiv tillar bilan ishlashda foydali. Misol uchun, “go” so‘zi “went”, “gone”, “going” kabi bir nechta shaklga ega bo‘lishi mumkin, ularning barchasi bir-biridan biroz boshqacha ma’nosi bilan farqlanadi [Plisson, Lavrac, Mladenić, 2004; Stanković, Krstev, Obradović, Lazić, Trtovac, 2016]. Lemma lug‘atidan foydalanishdan tashqari, qidiruv tizimlari lemmatizatsiyani amalga oshirish uchun *tabiiy tilni qayta ishlash algoritmlari* yoki *mashinali o‘rganish usullari* kabi boshqa usullardan ham foydalanishi mumkin [Kanerva, Ginter, Salakoski, 2021; Dave, Balani, 2015]. Ushbu

algoritmlar so‘zning lemma lug‘atida ro‘yxatga kiritilmagan bo‘lsa ham matnning konteksti va sintaksisini tahlil qilib, so‘zning asosiy shaklini aniqlashi mumkin.

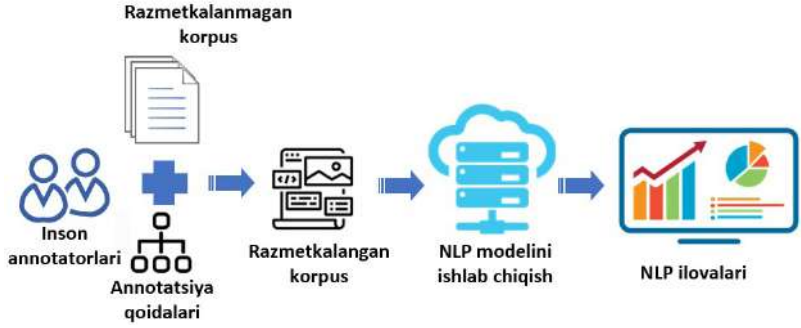
Qidiruv tizimlari uchun lemmatizatsiyaning muhim jihatlaridan biri bu **bir nechta ma’noga ega bo‘lgan so‘zlarni tahlil qilish** imkoniyatidir. Misol uchun, *“suzma”* so‘zi *ish-harakat* yoki *ovqat turiga* ishora qilishi mumkin. Lemmatizatsiya, shuningdek, *hujjatning asosiy mavzulari* va *kalit so‘zlarini aniqlashga* yordam beradi. Bu, ayniqsa, hujjat mazmunini tushunish va uni tegishli qidiruv so‘rovlariga moslashtirish uchun mashinali o‘rganish algoritmlariga asoslangan qidiruv tizimlari uchun juda muhim amaldir. Umuman olganda, lemmatizatsiya qidiruv tizimlari *matn ichidagi so‘zlarning ma’nosi* va *kontekstini yaxshiroq tushunish, qidiruv so‘rovlarini tegishli tarkibga to‘g‘ri moslashtirish* uchun foydalanadigan muhim vositadir.

Lemmatizatsiya orqali veb-sayt mazmunining ravshanligini oshirish. Lemmatizatsiya jarayoni sayt mazmunini tushunishni yaxshilash usullaridan biri bo‘lib, so‘zning shakllari sonini kamaytiradi. Misol uchun, “bormoq” so‘zi bordim, *boruvdim, borgani, bordingiz, boryapsiz, bormoqdasiz* kabi bir necha shaklga ega bo‘lishi mumkin. Mazkur so‘zlarni lemmatizatsiya qilish orqali “bormoq” so‘zi uchragan kontekstning tushunilishi oson bo‘ladi. Shuningdek, lemmatizatsiya so‘zning barcha shakllarini bilmaganlar uchun chalkashlikni kamaytirishga yordam beradi. Bunday muammo, odatda, flektiv tillarda uchraydi.

Lemmatizatsiyaning yana bir afzalligi shundaki, u veb-sayt tarkibining umumiy tuzilishi va kalit so‘zlarni aniqlashga yordam beradi. So‘zlarni lemmaga keltirish orqali sayt mazmunini tahlil qilish yanada izchil va mantiqiy bo‘ladi. Bu bir qancha texnik yoki ilmiy tarkibga ega bo‘lgan veb-saytlar uchun foydali. Chunki bu ma’lumotni mutaxassis bo‘lmaganlar uchun osonroq tushunishga yordam beradi.

Qidiruv mexanizmlari *veb-saytlarni skanerlash* va *indekslash* uchun turli algoritmlardan foydalanadi. Lemmatizatsiya jarayoni bu algoritmlarning aniqligini oshirishga yordam beradi. Lemmatizatsiyaning bir necha xil yondashuvlari mavjud bo‘lib, veb-sayt uchun eng samarali bo‘lganlari veb-saytning o‘ziga xos ehtiyojlari va maqsadlariga bog‘liq bo‘ladi. Korpus matnini lemmatizatsiya qilishni **“qo‘lda”** yoki **dasturiy ta’minot** orqali amalga oshirish mumkin [Plisson, Lavrac, Mladenić, 2004; Stanković, R., Krstev, Obradović, Lazić, Trtovac, 2016]. Ushbu dasturiy vositalar har doim

ham eng aniq natijalarni bermasligi mumkin. Korpus matnlarini qo‘lda lemmatizatsiya qilish ko‘proq vaqt talab qilsa-da, aniqroq natijalarni beradi. Quyidagi rasmda korpus matnlarini razmetkalash bosqichlari keltirilgan:



1-rasm. Korpus matnlarini razmetkalash bosqichlari

Veb-saytda lemmatizatsiyani amalga oshirish

Matn tasnifi, ma’lumotlarni qidirish va *mashina tarjimasi* kabi vazifalarini hal qilishda lemmatizatsiya yordam beradi. Veb-saytda lemmatizatsiyani amalga oshirish orqali qidiruv funksiyasi va matn tahlili samaradorligi oshirish mumkin. Veb-saytda lemmatizatsiyani amalga oshirish uchun turli yondashuvlar mavjud:

Oldindan o‘qitilgan lemmatizatsiya modelidan foydalanish. Veb-saytga integratsiya qilinishi mumkin bo‘lgan, oldindan o‘qitilgan lemmatizatsiya modelidan foydalanishdir [Khalil, Houby, Mohamed, 2021]. Bugungi kunda ko‘plab dunyo tillari uchun **WordNet** kabi leksik ma’lumotlar bazasiga asoslangan **WordNet Lemmatizer** va Stenford universitetida ishlab chiqilgan **Stanford CoreNLP** kabi turli xil ochiq kodli *lemmatizatsiya modellari* mavjud. Ushbu modellar tegishli kutubxonalarni o‘rnatish va model tomonidan taqdim etilgan lemmatizatsiya funksiyasini chaqirish orqali veb-saytga osongina integratsiya qilinishi mumkin.

Maxsus lemmatizatsiya modelini ishlab chiqish. Yana bir yondashuv veb-saytga xos bo‘lgan maxsus lemmatizatsiya modelini yaratishdir [Gamallo, 2020]. Agar veb-sayt mavjud lemmatizatsiya modellari bilan yaxshi ifodalanmagan katta va xilma-xil matnlar to‘plamiga ega bo‘lsa, bunday yondashuvdan foydalanish mumkin. Maxsus lemmatizatsiya modelini yaratish uchun birinchi qadam veb-saytga tegishli bo‘lgan *katta va xilma-xil matnlarni to‘plash*dir. Keyingi qadamda, nomuvofiqliklarni bartaraf etish uchun *matnlar oldindan qayta ishlanadi,* **lemmalar POS tegi belgilanadi** hamda *morfologik tahlil* kabi usullardan foydalaniladi. Olingan lemmalar keyinchalik har qanday so‘z lemmasini bashorat qilish uchun **qaror daraxti** *(decision tree)* yoki **support vector machine (SVM)** kabi

mashinali o‘rganish modelini o‘rgatish uchun ishlatiladi.

Onlayn lemmatizatsiya xizmatidan foydalanish. Veb-saytdagi matnlarni lemmatizatsiya qilish uchun **Lemmatize.io** yoki **TextBlob** kabi onlayn lemmatizatsiya xizmatidan foydalanish mumkin. Ushbu yondashuv nisbatan sodda va minimal sozlashni talab qiladi. Chunki lemmatizatsiya xizmatidan **API** chaqiruvi orqali foydalanish mumkin. Biroq bu yondashuv katta va murakkab veb-saytlar uchun mos kelmasligi mumkin. Bu yondashuv maxsus lemmatizatsiya modeliga nisbatan sekin va kam aniqlikka ega bo‘lishi mumkin.

Tanlangan yondashuvdan qat’iy nazar, veb-saytda lemmatizatsiyani amalga oshirish uchun bir necha qadamlarni bajarish kerak:

1. **Oldindan ishlov berish.** Birinchi qadam matnda mavjud *shovqin (noise)* va *nomuvofiqliklarni olib tashlash* uchun veb-saytdagi matnlarni oldindan qayta ishlashdir. Bu qadamda *matnlarni kichik harflar bilan yozish*, *tinish belgilarini olib tashlash* va *qisqartmalarni ularning to‘liq shakllari bilan almashtirish* kabi amallar bajariladi.

2. **Tokenizatsiya.** Keyingi qadamda matnlar *alohida so‘zlar* yoki *tokenlarga* aylantirildi. Bu amallarni *so‘zlarni segmentatsiyalash* yoki *jumlalarni ajratish* kabi usullar yordamida amalga oshirilishi mumkin.

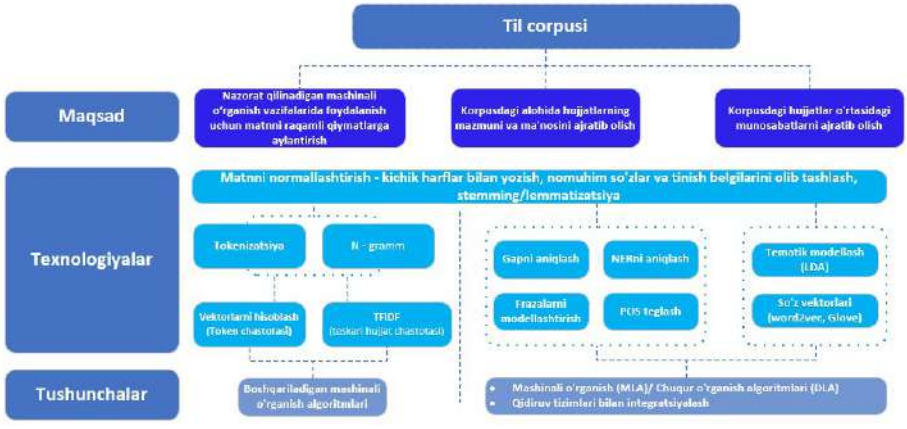
3. **POS teglash.** Uchinchi qadam – tokenlarni *POS teglar* bilan belgilashdan iborat. Buni oldindan o‘rgatilgan *POS tegger* yoki *maxsus POS tegger* yordamida amalga oshirish mumkin.

4. **Lemmatizatsiya.** Nihoyat, lemmatizatsiya jarayoni *lemmatizatsiya modeli* yoki *xizmatidan* foydalangan holda tokenlarga qo‘llaniladi. Bunda lemmani aniqlashda tokenlarning POS teglaridan va matn kontekstidan foydalanadi.

5. **Integratsiya.** Lemmatizatsiya qilingan matnlar veb-sayt talablari va maqsadlariga qarab turli usullar bilan veb-saytga birlashtiriladi. Masalan, lemmatlashtirilgan matnlar asl so‘zlar o‘rniga lemmalarni indekslash orqali qidiruv funksiyalarini yaxshilash uchun ishlatilishi mumkin.

Lemmatizatsiyalangan *matnlar mavzuni modellashtirish* yoki *hissiyotlarni tahlil qilish* kabi matn tahlili vazifalari uchun ham ishlatiladi. Bunda lemma matnning yanada izchil ifodalanishini ta’minlaydi. Veb-saytda lemmatizatsiya qilishda e’tibor qilinishi kerak bo‘lgan bir nechta holatlar mavjud. Lemmatizatsiya modeli yoki xizmatini tanlash muhim ahamiyat kasb etadi, turli model/xizmatlar turli darajadagi aniqlik va samaradorlikni taqdim etadi. Mavjud variantlarni baholash, veb-sayt ehtiyoj va resurslariga eng maqbulini tanlash

muhim. Lemmatizatsiyalangan matnlarni veb-saytga integratsiya qilishda ham turli muammolar yuzaga keladi. Bu asl so‘zlar o‘rniga lemmalarni indekslash uchun asosiy ma’lumotlar bazasi yoki qidiruv algoritmlarini o‘zgartirishni o‘z ichiga olishi mumkin. Integratsiya jarayonida veb-saytning ishlashi yoki ishonchliligiga ta’sir qilmasligini ta’minlash lozim. Nihoyat, lemmatizatsiya modeli yoki xizmatining to‘g‘riligiga ishonch hosil qilish uchun uni muntazam yangilash va qo‘llab-quvvatlash kerak. Bunda model yoki xizmatni yangi matnlar bilan qayta o‘qitish va ishlashga qarab parametrlarni sozlash amalga oshiriladi. Quyidagi rasmda til korpusi arxitekturasi keltirilgan:



2-rasm. Til korpusi arxitekturasi

Lemmatizatsiyaning SEOga salbiy ta’siri

SEO nuqtayi nazaridan lemmatizatsiya qanday qo‘llanilishiga qarab *ham ijobiy, ham salbiy* ta’sir ko‘rsatadi. Ijobiy tomoni shundaki, lemmatizatsiya ishlatiladigan so‘zning turli shakllari sonini kamaytirish orqali mazmunning o‘qilishi va ravshanligini yaxshilashga yordam beradi. Bu qidiruv tizimlariga kontentni tushunishni osonlashtirishi va foydalanuvchilar uchun aniqroq qidiruv natijalarini taqdim etishi mumkin. Bundan tashqari, lemmatizatsiya so‘zning barcha o‘zgarishlariga bir xil munosabatda bo‘lishini ta’minlash orqali kontentning dolzarbligini (faol qo‘llanishini) yaxshilashga yordam beradi. Bu veb-sahifaning reytingini ko‘taradi, chunki qidiruv tizimlari ko‘pincha veb-sahifadagi muayyan kalit so‘zlarning mavjudligi va chastotasidan uning ma’lum bir qidiruv so‘roviga aloqadorligini aniqlashda samara beradi.

Biroq lemmatizatsiya noto‘g‘ri ishlatilganda SEOga salbiy ta’sir ko‘rsatishi mumkin. Lemmatizatsiya so‘zning ma’nosini o‘zgartirib, *chalkashlik* yoki *tushunmovchilikka* olib kelishi ham mumkin.

Misol uchun, “olma” so‘zi harakat ma’nosini anglatuvchi fe’l yoki predmetni bildiruvchi ot bo‘lishi mumkin. Agar lemmatizatsiya jarayonida kontekstdan bexabar holda amalga oshirilsa, u “olma” fe’lini ot turkumiga aylantirishi mumkin, bu esa matnning asl ma’nosini o‘zgartiradi.

Shuningdek, lemmatizatsiya so‘zlardan muhim ma’lumotlarni olib tashlashi mumkin. Misol uchun, “yugurish” so‘zi “kompaniyaning muvaffaqiyatli o‘tishi” iborasi kabi sifatdosh sifatida ham ishlatilishi mumkin. Lemmatizatsiya buni «yugurish» asosiy shakliga aylantiradi, bu esa kampaniya muvaffaqiyatli bo‘lganligi haqidagi muhim ma’lumotlarni yo‘qotadi.

Nihoyat, lemmatizatsiya ortiqcha/keraksiz kalit so‘zlarni hosil bo‘lsihiga olib kelishi mumkin. Bu holda veb-sahifa ma’lum kalit so‘zlardan ortiqcha foydalanish orqali qidiruv natijalarida yuqori o‘rinlarni egallashga harakat qiladi. Misol uchun, agar veb-sahifa “*kitobxon*” kalit so‘zi bo‘yicha tartiblashmoqchi bo‘lsa, u “*kitob*”ning barcha variantlarini “*kitob*”ga aylantirish uchun lemmatizatsiyadan foydalanishi mumkin. Bu esa matnda ortiqcha kalit so‘zlarning hosil bo‘lishiga olib kelishi mumkin. Bu qidiruv tizimlari tomonidan **spam xatti-harakati** sifatida belgilanishi va pastroq reytingga yoki hatto jazoga olib kelishi mumkin.

Umuman olganda, lemmatizatsiya to‘g‘ri ishlatilganda SEOda foydali vosita bo‘lishi mumkin. Ammo uning potentsial ta’siridan xabardor bo‘lish, uni matnning ma’nosi yoki kontekstiga salbiy ta’sir ko‘rsatmaydigan tarzda ishlatish ahamiyatli. Eng yaxshi natijalarni ta’minlash uchun kalit so‘zlarni ortiqcha hosil bo‘lishidan qochish, lemmatizatsiyani kalit so‘zlarni o‘rganish hamda sahifani optimallashtirish kabi boshqa SEO strategiyalari bilan birgalikda ishlatish lozim.

So‘zning konteksti va POS tegining lemmaga ta’siri

So‘zning lemmasi uning asosiy shakli yoki lug‘atdagi shaklini ifodalaydi. O‘zbek tilidagi barcha so‘zlar lug‘atda mavjud lemmalarga shakl yasovchi qo‘shimchalarni qo‘shish orqali hosil qilinadi. **So‘zning konteksti** va **POS tegi** uning lemmasiga sezilarli ta’sir ko‘rsatishi mumkin. Chunki turli kontekstlar va POS teglar ko‘pincha so‘zning turli xil *fleksiyalari* va *hosilalaridan* foydalanishni talab qiladi.

Lemmatizatsiya jarayonidagi xatoliklar

Lemmatizatsiya foydali vosita bo‘lishi mumkin bo‘lsa-da, undan SEO uchun foydalanishda bir necha keng tarqalgan xatolar yuzaga kelishi mumkin. Keng tarqalgan xatolardan biri bu *lemmatizatsiya*

va *stemming* o‘rtasidagi farqni to‘g‘ri tushunmaslikdir. Bu farqni tushunmaslik qidiruv natijalaridagi noaniqliklarga va foydalanuvchi so‘roviga mos bo‘lmagan natijalarning shakllanishiga olib kelishi mumkin.

Lemmatizatsiyadan saytning ba’zi sahifalari yoki bo‘limlariga qo‘llanilsa, nomuvofiq qidiruv natijalariga va foydalanuvchi so‘roviga mos bo‘lmagan chalkashtirishga yuzaga kelishi mumkin. Qidiruv natijalari to‘g‘ri va tegishli bo‘lishini ta’minlash uchun veb-saytda lemmatizatsiyani qo‘llash talab etiladi.

Lemmatizatsiya POS teg va so‘z ishlatilgan kontekstni, so‘ngra tegishli lemmani aniqlaydi. Lemmatizatsiya jarayonida kontekst hisobga olinmasa, natijada aniqlangan lemma noto‘g‘ri bo‘lishi mumkin. Bu noto‘g‘ri qidiruv natijalarini hosil qilishi mumkin.

Lug‘atdagi lemmalar ro‘yxatini muntazam yangilamaslik ham turli xatoliklarga olib keladi. Tilning rivojlanishi, lug‘at tarkibiga yangi so‘zlar qo‘shilishi bilan birga, lemmalar ro‘yxatini yangilash, uning to‘g‘ri va dolzarb bo‘lishini ta’minlash lozim. Agar lemmalar ro‘yxati yangilanmagan bo‘lsa, unda eskirgan/noto‘g‘ri lemmalar bo‘lishi mumkin, bu esa qidiruv natijalarining noto‘g‘ri bo‘lishiga olib keladi.

Lemmatizatsiya qidiruv natijalarini yaxshilashda foydali bo‘lsa-da, unga juda ko‘p ishonmaslik kerak. **Haddan tashqari lemmatizatsiya (over-lemmatizing)** kontekstni va nuansning yo‘qolishiga olib keladi. Qidiruv natijalarini yaxshilash va kontentning sifati, o‘qilishi o‘rtasida muvozanatni saqlash muhim ahamiyat kasb etadi.

Oldini olish kerak bo‘lgan yana bir xato – lemmatizatsiya jarayonini sinovdan o‘tkazmaslik. Lemmatizatsiya jarayonining to‘g‘ri va samarali ishlashini ta’minlash uchun uni muntazam sinab ko‘rish lozim. Bunga lemmatizatsiya qidiruv tajribasini yaxshilash yoki to‘sqinlik qilishini aniqlash uchun qidiruv testlarini o‘tkazish va natijalarni tahlil qilish bilan erishiladi.

Xulosa qilib aytganda, lemmatizatsiya SEOda qidiruv natijalari aniqligi va dolzarbligini oshirishda kuchli vosita bo‘lishi mumkin. Shu bilan birga, *lemmatizatsiya va stemming o‘rtasidagi farqni tushunmaslik, lemmatizatsiyani veb-saytning barcha sahifalariga qo‘llamaslik, kontekstni hisobga olmaslik, lemmatizatsiya ro‘yxatini muntazam yangilamaslik, lemmatizatsiyani haddan tashqari oshirib yuborish* va *lemmatizatsiya jarayonini sinab ko‘rmaslik* kabi keng tarqalgan xatolardan qochish kerak. Ushbu xatolarga yo‘l qo‘ymaslik lemmatizatsiyani foydalanuvchi hamda qidiruv tizimlari uchun

samarali bo‘lishini ta’minlaydi.

Veb-saytda lemmatizatsiya samaradorligini tekshirish

SEO uchun veb-saytda amalga oshirilgan lemmatizatsiya jarayoni samaradorligini tekshirish uchun bir nechta asosiy qadamlarni bajarish lozim.

Kalit so‘zlarni aniqlash. Lemmatizatsiya samaradorligini tekshirishning birinchi bosqichi optimallashtirilgan tegishli *kalit so‘zlarni aniqlashdir*. Bu qadamda maqsadli auditoriya tomonidan ko‘p qidiriladigan va veb-saytga tegishli bo‘lgan atama (termin)lar aniqlanadi.

Kontentni tahlil qilish. Saytdagi kalit so‘zlar aniqlaganidan so‘ng ushbu shartlar uchun qanchalik yaxshi optimallashtirilganligini tahlil qilib, *mavjud kontentni tahlil qilish* kerak. Sayt kontenti tahlili **Google Analytics** yoki **Ahrefs** kabi vositalardan foydalangan holda amalga oshiriladi. Bu qadamda muayyan kalit so‘zlar bo‘yicha veb-saytning *qanchalik yaxshi reytingga ega ekanligi, kontentning optimalligini ko‘rish* qo‘lda ko‘rib chiqishni talab etadi.

Lemmatizatsiyani amalga oshirish. Joriy SEO faoliyatini yaxshi tushungandan so‘ng, keyingi qadamda *lemmatizatsiyani amalga oshirish* lozim. Ingliz tilida so‘zlar uchun bu qadamni **Natural Language Toolkit (NLTK)** yoki **WordNet lemmatizer** kabi vositalar orqali amalga oshirish mumkin. Shuningdek, har bir so‘zning ildiz (root) shaklini aniqlash va lemmatizatsiyani qo‘lda bajarish ham mumkin.

Samaradirlikni tekshirish. Lemmatizatsiya amalga oshirgandan so‘ng, keyingi qadamda veb-sayt ishlashini kuzatish maqsadga muvofiq. Bu amalni **Google Analytics** yoki **Ahrefs** kabi vositalar yordamida bajarish mumkin. Bu ma’lum *kalit so‘zlar bo‘yicha sayt reytingini kuzatish* yoki veb-*saytdagi trafik va ishtirok darajasini kuzatish*dan iborat.

Samaradorlikni taqqoslash. Nihoyat, oxirgi qadamda lemmatizatsiya amalga oshirilgandan oldin va keyingi veb-sayt faoliyatini taqqoslash lozim. Bu lemmatizatsiya jarayoni SEO faoliyatini yaxshilashda qanchalik samarali bo‘lganini ko‘rishga yordam beradi.

Veb-saytda lemmatizatsiya orqali SEO samaradorligini o‘lchash uchun foydalanish mumkin bo‘lgan bir nechta asosiy ko‘rsatkichlar mavjud:

1. Qidiruv tizimidagi reyting. Kuzatiladigan eng muhim ko‘rsatkichlardan biri bu veb-saytning qidiruv tizimidagi aniq kalit so‘zlar bo‘yicha reytingdir. Agar lemmatizatsiya samarali amalga

oshirilgan bo‘lsa, maqsadli kalit so‘zlar uchun reyting yaxshilangani ko‘rinadi.

2. Trafik darajalari. Kuzatiladigan yana bir asosiy ko‘rsatkich – bu veb-saytning trafik darajasi. Agar lemmatizatsiya samarali amalga oshirilgan bo‘lsa, trafikning ko‘payishini qayd etish mumkin.

3. Ishtirok etish ko‘rsatkichlari. Trafik darajasiga qo‘shimcha ravishda, veb-saytga sarflangan vaqt va bir tashrif uchun ko‘rilgan sahifalar soni kabi jalb qilish ko‘rsatkichlarini kuzatish maqsadga muvofiq. Agar lemmatizatsiya samarali amalga oshirilgan bo‘lsa, ushbu ko‘rsatkichlarning o‘sishi kuzatiladi. Chunki tashrif buyuruvchilar sizning kontentingizni osonroq topishi, undagi ma’lumotlardan foydalanishi osonlashadi.

Umuman olganda, lemmatizatsiya veb-saytning SEO faoliyatini o‘stirishda zaruriy vosita hisoblanadi.

Tegishli kalit so‘zlarni aniqlash, mavjud kontentni tahlil qilish, lemmatizatsiyani amalga oshirish, samaradorlikni kuzatish, natijalarni taqqoslash orqali veb-sayting SEOda lemmatizatsiya samaradorligini samarali sinab ko‘rish va qidiruv tizimlari uchun kontentni qanday optimallashtirish haqida to‘g‘ri qaror qabul qilishga olib keladi.

Xulosa

Til korpuslaridagi matnlar soni va korpus hajmi muntazam tarzda oshib borishi tufayli, qidiruv tizimlari orqali korpusda amalga oshirilgan so‘rovni qayta ishlash vazifasi muhim ahamiyat kasb etadi. Til korpusini razmetkalash (teglash/annotatsiyalash/indekslash) foydalanuvchi amalga oshirgan so‘rov natijasini samarali va tez bajarish imkoniyatini taqdim etadi. Ushbu maqolada o‘zbek tili milliy korpusi ma’tnlarini lemmalash va uning qidiruv tizimi natijalarini optimallashtirishga qo‘llash masalasi ko‘rib chiqildi. *Lemmatizatsiya va stemming* so‘zlarni ifodalashni soddalashtirish orqali NLP masalalarini yechishdagi murakkabliklarni kamaytirishda ishlatiladigan usullardir. Ikkala usul ham so‘zni asosiy shakliga qisqartirishni nazarda tutadi. Ammo ular buni amalga oshirish algoritmi va usuli bilan farqlanadi. Lemmatizatsiya so‘z ma’nosini to‘g‘ri ifodalashni ta’minlaydi, biroq bu jarayon murakkab, u sekinroq amalga oshiriladi. Stemming sodda va tezkor yondashuv bo‘lsa-da, so‘z ma’nosini ifodalashda xatolarga olib kelishi mumkin. Maqolada til korpusida lemmatizatsiyani amalga oshirish uchun mavjud yondashuvlar va ularning yutuq hamda kamchiliklari keltiridi. Matnga *oldindan ishlov berish, tokenizatsiya,*

POS teglash, lemmatizatsiya/stemming va integratsiya amallari barcha yondashuvlarda boshlangi`ch qadam bo‘lib xizmat qiladi. Shuningdek, til korpusida lemmatizatsiya jarayonidan noto‘g‘ri qo‘llanilganda kelib chiqadigan xatoliklar va ularni bartaraf qilish usullari keltirildi. Xulosa sifatida, lemmatizatsiya jarayonining til korpusi funksionalligini va tahlil samaradorligini oshirishi mumkin bo‘lgan foydali jarayonligini qayd etish mumkin.

Foydalanilgan adabiyotlar ro‘yxati:

B.B.Elov, Sh.Hamroyeva, Z.Xusainova. *NLP (tabiiy tilga ishlov berish) ning vazifalari va zamonaviy yondashuvlar*. TerDU, Filologik tadqiqotlar: til, adabiyot, ta’lim. 2022, 5-6. 19-26 b.

Plale, B. (2013). *Big data opportunities and challenges for IR, text mining and NLP*. https://doi.org/10.1145/2513549.2514739

Elov B.B., Khamroeva Sh.M., Xusainova Z.Y. *NLP (tabiiy tilga ishlov berish) ning Pipeline konveyeri.* Muhammad al-Xorazmiy avlodlari, № 1 (23), mart 2023, 181-192 b.

Saloot, M. A., & Pham, D. N. (2021). Real-time Text Stream Processing: A Dynamic and Distributed NLP Pipeline. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3459104.3459198

S.Vajjala, B.Majumder, A.Gupta, H.Surana. *Practical Natural Language Processing. A Comprehensive Guide to Building Real-World NLP Systems. 2020. 455 p.*

Hafsa Jabeen. (2018). Stemming and Lemmatization in Python. *Towardsdatascience*.

Balakrishnan, V., & Ethel, L.-Y. (2014). Stemming and Lemmatization: A Comparison of Retrieval Performances. *Lecture Notes on Software Engineering*, *2*(3). https://doi.org/10.7763/lnse.2014.v2.134

Khyani, D., S, S. B., M, N. N., & M, D. B. (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing. *Journal of University of Shanghai for Science and Technology*, *22*(10).

Baklanov, O., Bezkorovainyi, V., & Kolesnyk, L. (2022). Studying cognitive services for websites search engine optimization. *Bulletin of Kharkov National Automobile and Highway University*, *97*. https://doi.org/10.30977/bul.2219-5548.2022.97.0.7

https://marketbrew.ai/a/lemmatization-seo

B.Elov, Sh.Hamroyeva, O.Abdullayeva, Z.Xusainova, N.Xudayberganov. *O‘zbek, turk va uyg‘ur tillarida pos teglash va stemming.* O`zbekiston: til va madaniyat (Kompyuter lingvistikasi), 2023, 1(6). 40-64 b.

di Nunzio, G. M., & Vezzani, F. (2018). A linguistic failure analysis of classification of medical publications: A study on stemming vs lemmatization. *CEUR Workshop Proceedings*, *2253*. https://doi.org/10.4000/books.aaccademia.3327

Z.Xusainova. *O‘zbek tilida stemmingni amalga oshirishning gibrid statistik*

yondashuvi. “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” an’anaviy xalqaro ilmiy-amaliy konferensiya, 2023-yil aprel. 70-76 b.

Z.Xusainova. *NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash.* O‘zbek amaliy filologiyasi istiqbollari. Respublika ilmiy-amaliy konfrensiyasi. 2022-yil oktabr. 154-163 b.

http://uznatcorpara.uz/uz/POSTag

Pattnaik, S., Nayak, A. K., & Patnaik, S. (2020). A Semi-supervised Learning of HMM to Build a POS Tagger for a Low Resourced Language. *Journal of Information and Communication Convergence Engineering*, *18*(4). https://doi.org/10.6109/jicce.2020.18.4.207

Fisenko, T., Baliun, O., & Grigorenko, A. (2022). SEO AS A METHOD OF WEBSITE PROMOTION. *Věda a Perspektivy*, *2 (9)*. https://doi.org/10.52058/2695-1584-2022-2(9)-260-271

Andrikopoulos, P., Sun, J., & Guo, J. (2017). Ownership structure and the choice of SEO issue method in the UK. *International Journal of Managerial Finance*, *13*(4). https://doi.org/10.1108/IJMF-04-2017-0069

Gashkov, A., & Eltsova, M. (2018). Lemmatization with reversed dictionary and fuzzy sets. *SHS Web of Conferences*, *55*. https://doi.org/10.1051/shsconf/20185504007

Plisson, J., Lavrac, N., & Mladenić, Dr. D. (2004). A rule based approach to word lemmatization. *Proceedings of the 7th International Multiconference Information Society (IS’04)*.

Stanković, R., Krstev, C., Obradović, I., Lazić, B., & Trtovac, A. (2016). Rule-based automatic multi-word term extraction and lemmatization. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*.

Kanerva, J., Ginter, F., & Salakoski, T. (2021). Universal Lemmatizer: A sequence-To-sequence model for lemmatizing Universal Dependencies treebanks. *Natural Language Engineering*, *27*(5). https://doi.org/10.1017/S1351324920000224

Dave, R., & Balani, P. (2015). Survey paper of Different Lemmatization Approaches. *International Journal of Research in Advent Technology Science and Technology. Special Issue 1st International Conference on Advent Trends in Engineering, Science and Technology “ICATEST 2015,” Special Issue 1st International Conference on Advent Trends in Engineering, Science and Technology “ICATEST 2015.”*

Khalil, E. A. H., el Houby, E. M. F., & Mohamed, H. K. (2021). Deep learning for emotion analysis in Arabic tweets. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00523-w

Gamallo, P. (2020). The impact of linguistic knowledge in different strategies to learn cross-lingual distributional models. *Frontiers in Artificial Intelligence and Applications*, *325*. https://doi.org/10.3233/FAIA200322

THE USE OF LEMMATIZATION IN OPTIMIZING THE SEARCH ENGINE OF THE NATIONAL CORPUS OF THE UZBEK LANGUAGE

Zilola Khusainova[1]

Abstract.

Lemmatization is a natural language processing technique that involves determining the meaningful form of a lemma (lexeme) and using it instead of inflectional/derivative forms. This article deals with the issue of lemmatization of the texts of the national corpus of the Uzbek language and the application of its results for search engine optimization (SEO, search engine optimization). Search engine algorithms are constantly evolving and improving their ability to understand and interpret human language. The use of natural language processing techniques such as lemmatization is becoming increasingly important for optimizing search engine results in language corpora.

Keywords: *lemmatization, lemma, POS, SEO, search engines, Uzbek national corpus.*

References:

B.B.Elov, Sh.Hamroyeva, Z.Xusainova. *NLP (tabiiy tilga ishlov berish) ning vazifalari va zamonaviy yondashuvlar*. TerDU, Filologik tadqiqotlar: til, adabiyot, ta'lim. 2022, 5-6. 19-26 b.

Plale, B. (2013). *Big data opportunities and challenges for IR, text mining and NLP*. https://doi.org/10.1145/2513549.2514739

Elov B.B., Khamroeva Sh.M., Xusainova Z.Y. *NLP (tabiiy tilga ishlov berish) ning Pipeline konveyeri.* Muhammad al-Xorazmiy avlodlari, № 1 (23), mart 2023, 181-192 b.

Saloot, M. A., & Pham, D. N. (2021). Real-time Text Stream Processing: A Dynamic and Distributed NLP Pipeline. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3459104.3459198

S.Vajjala, B.Majumder, A.Gupta, H.Surana. *Practical Natural Language Processing. A Comprehensive Guide to Building Real-World NLP Systems. 2020. 455 p.*

Hafsa Jabeen. (2018). Stemming and Lemmatization in Python. *Towardsdatascience*.

Balakrishnan, V., & Ethel, L.-Y. (2014). Stemming and Lemmatization: A Comparison of Retrieval Performances. *Lecture Notes on Software Engineering*, *2*(3). https://doi.org/10.7763/lnse.2014.v2.134

Khyani, D., S, S. B., M, N. N., & M, D. B. (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing. *Journal of University of Shanghai for Science and Technology*, *22*(10).

Baklanov, O., Bezkorovainyi, V., & Kolesnyk, L. (2022). Studying cognitive services for websites search engine optimization. *Bulletin of*

[1]*Xusainova Zilola Yuldashevna* – PhD student of Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i.
E-mail: xusainovazilola@navoiy-uni.uz
ORCID: 0000-0003-4357-7515

Kharkov National Automobile and Highway University, *97*. https://doi.org/10.30977/bul.2219-5548.2022.97.0.7

https://marketbrew.ai/a/lemmatization-seo

B.Elov, Sh.Hamroyeva, O.Abdullayeva, Z.Xusainova, N.Xudayberganov. *O‘zbek, turk va uyg‘ur tillarida pos teglash va stemming.* O`zbekiston: til va madaniyat (Kompyuter lingvistikasi), 2023, 1(6). 40-64 b.

di Nunzio, G. M., & Vezzani, F. (2018). A linguistic failure analysis of classification of medical publications: A study on stemming vs lemmatization. *CEUR Workshop Proceedings*, *2253*. https://doi.org/10.4000/books.aaccademia.3327

Z.Xusainova. *O‘zbek tilida stemmingni amalga oshirishning gibrid statistik yondashuvi.* “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” an’anaviy xalqaro ilmiy-amaliy konferensiya, 2023-yil aprel. 70-76 b.

Z.Xusainova. *NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash.* O‘zbek amaliy filologiyasi istiqbollari. Respublika ilmiy-amaliy konfrensiyasi. 2022-yil oktabr. 154-163 b.

http://uznatcorpara.uz/uz/POSTag

Pattnaik, S., Nayak, A. K., & Patnaik, S. (2020). A Semi-supervised Learning of HMM to Build a POS Tagger for a Low Resourced Language. *Journal of Information and Communication Convergence Engineering*, *18*(4). https://doi.org/10.6109/jicce.2020.18.4.207

Fisenko, T., Baliun, O., & Grigorenko, A. (2022). SEO AS A METHOD OF WEBSITE PROMOTION. *Věda a Perspektivy*, *2 (9)*. https://doi.org/10.52058/2695-1584-2022-2(9)-260-271

Andrikopoulos, P., Sun, J., & Guo, J. (2017). Ownership structure and the choice of SEO issue method in the UK. *International Journal of Managerial Finance*, *13*(4). https://doi.org/10.1108/IJMF-04-2017-0069

Gashkov, A., & Eltsova, M. (2018). Lemmatization with reversed dictionary and fuzzy sets. *SHS Web of Conferences*, *55*. https://doi.org/10.1051/shsconf/20185504007

Plisson, J., Lavrac, N., & Mladenić, Dr. D. (2004). A rule based approach to word lemmatization. *Proceedings of the 7th International Multiconference Information Society (IS’04)*.

Stanković, R., Krstev, C., Obradović, I., Lazić, B., & Trtovac, A. (2016). Rule-based automatic multi-word term extraction and lemmatization. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*.

Kanerva, J., Ginter, F., & Salakoski, T. (2021). Universal Lemmatizer: A sequence-To-sequence model for lemmatizing Universal Dependencies treebanks. *Natural Language Engineering*, *27*(5). https://doi.org/10.1017/S1351324920000224

Dave, R., & Balani, P. (2015). Survey paper of Different Lemmatization Approaches. *International Journal of Research in Advent Technology Science and Technology. Special Issue 1st International Conference on Advent Trends in Engineering, Science and*

Technology "ICATEST 2015," Special Issue 1st International Conference on Advent Trends in Engineering, Science and Technology "ICATEST 2015."

Khalil, E. A. H., el Houby, E. M. F., & Mohamed, H. K. (2021). Deep learning for emotion analysis in Arabic tweets. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00523-w

Gamallo, P. (2020). The impact of linguistic knowledge in different strategies to learn cross-lingual distributional models. *Frontiers in Artificial Intelligence and Applications*, *325*. https://doi.org/10.3233/FAIA200322

Jurnal 2017-yil 26-oktyabrda O‘zbekiston Respublikasi Matbuot
va axborot agentligi tomonidan 0936-raqam bilan ro‘yxatdan o‘tgan.

Jurnal O'zbekiston Respublikasi Oliy Attestatsiya Komissiyasi tomonidan
filologiya fanlari bo'yicha falsafa doktori (PhD) va fan doktori (DSc)
dissertatsiyalari asosiy ilmiy natijalari chop etilishi lozim bo'lgan ro'yxatga
kiritilgan (30.10.2021. № 308/6).

Tahririyatga kelgan maqolalar mualliflarga qaytarilmaydi.

Manzil: Toshkent shahri, Yakkasaroy tumani, Yusuf Xos
Hojib ko‘chasi 103-uy.
Telefonlar: +99871 281-45-11, +99871 281-41-93.
Website: compling.tsuull.uz
E-mail: kompling@navoiy-uni.uz

Bosishga 05.09.2023-yilda ruxsat etildi.
Bichimi 70x100 1/16, Ofset bosma. “Cambria” garniturasi.
Shartli b.t. 7,51. Nashr b.t. 7,62.

“O‘zbekiston: til va madaniyat” jurnali tahririyatida
tayyorlandi va sahifalandi.
“YASHNOBOD NASHR” bosmaxonasida chop etildi.
Adadi 300 nusxa. Buyurtma №2.
Bosmaxona manzili: Toshkent shahar Yashnobod tumani,
58-a harbiy shaharcha.