Xusainova Zilola Yuldashevna

O'zbek tili birliklarini tokenlash, stemlash, lemmalashning lingvistik asoslari

Monografiya

O‘ZBEKISTON RESPUBLIKASI
OLIY TA’LIM, FAN VA INNOVATSIYA VAZIRLIGI

ALISHER NAVOIY NOMIDAGI TOSHKENT DAVLAT
O‘ZBEK TILI VA ADABIYOTI UNIVERSITETI

XUSAINOVA ZILOLA YULDASHEVNA

O‘ZBEK TILI BIRLIKLARINI TOKENLASH, STEMLASH, LEMMALASHNING LINGVISTIK ASOSLARI

MONOGRAFIYA

Toshkent – 2025

UO'K: 811.512.133'32
KBK 81.2O'zb
X 93

Xusainova Z.Y.
O'zbek tili birliklarini tokenlash, stemlash, lemmalashning lingvistik asoslari. Monografiya. – Toshkent: Noble Print, 2025. – 148 b.

Ushbu monografiya doktorantlar, magistrantlar, mustaqil izlanuvchilar, bakalavriat bosqichi talabalari uchun mo'ljallangan.

Monografiyadan o'zbek tili birliklarini tokenlash, stemlash va lemmalash jarayonlarini lingvistik hamda dasturiy asosda o'rganish natijalari tilshunoslikning nazariy hamda amaliy yo'nalishlarida qo'llaniladigan samarali vosita sifatida foydalanish mumkin. Shuningdek, ish natijalaridan kompyuter lingvistikasi, tabiiy tilni qayta ishlashga oid o'quv-metodik qo'llanmalar, darsliklar yaratishda, o'zbek tilidagi leksik birliklar, korpus lingvistikasi, avtomatik morfologik tahlil bo'yicha maxsus kurs va seminarlar o'tishda, morfologik va tokenlash asosidagi elektron lug'atlarni boyitishda foydalanish mumkin.

Mas'ul muharrir:
B.B. Elov

Taqrizchilar:
f.f.d., prof. S. Normamatov
t.f.f.d., dotsent R. Alayev

Monografiya Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti Ilmiy-texnikaviy kengashi majlisida muhokama qilinib, nashrga tavsiya etilgan (2025-yil 17-dekabrdagi 6-sonli bayonnoma).

ISBN 978-9910-8466-7-0

© Xusainova Z.Y.
Noble Print, 2025

KIRISH

Jahon kompyuter lingvistikasi amaliyotida tabiiy tilga ishlov berish usullaridan keng foydalanilmoqda. Tabiiy tilga ishlov berish (Natural language prossecing – NLP) loyihalarida doimiy bajariladigan vazifalar mavjud. Ular fundamental xarakterga ega bo'lganligi sababli keng ko'lamda o'rganilgan. Ushbu vazifalarni yaxshi tushunish NLPning turli ilovalarini yaratishga imkon beradi. Tilni modellashtirish, matnni tasniflash, axborot olish, ma'lumot izlash, suhbat agenti, matnni umumlashtirish, savol-javob tizimlari, mashina tarjimasi, tematik modellashtirish NLPning asosiy sohalaridir. Tabiiy tilni qayta ishlash – sun'iy intellekt sohasining mashinalarni inson tilini tushunish va qayta ishlashga qaratilgan yo'nalishi. Tabiiy tilni qayta ishlash mashina uchun oson ish emas, chunki mashina matn bilan emas, raqamlar bilan ishlaydi.

Dunyo tilshunosligida XX asr oxirida avtomatik tarjima sifatini yaxshilash, tilni lingvistik modellashtirish, muayyan tillar doirasida so'zlarni lemmalash nazariyasi, algoritmini yaratish kabi masalalar kun tartibiga qo'yildi. Natijada axborot-qidiruv tizimida faol qo'llanuvchi tokenizator, stemmer, lemmatizator singari ilovalar ishlab chiqildi. Binobarin, til korpuslarini yaratish, ularni lingvistik annotatsiyalash, matnni avtomatik qayta ishlaydigan dasturiy vositalar – morfoanalizator, stemmer, lemmatizator, parser, orfokorrektorlarni ishlab chiqish kompyuter lingvistikasining ilmiy-nazariy muammolariga aylandi.

O'zbek kompyuter lingvistikasi sohasida til korpusi, nutq [illegible], avtomatik tarjima, sun'iy intellekt, o'zbek tilini tushunish va qayta ishlash borasida ilmiy izlanishlarga e'tibor qaratilmoqda. Lekin tabiiy tilga ishlov berishning dastlabki bosqichlarini amalga oshiruvchi [illegible] tokenizator, stemmer, lemmatizatorlarni ishlab chiqish masalasi [illegible] nazariy jihatdan o'rganilmagan. Zero, "...davlat tilining sofligini [illegible] uni boyitib borish va aholining nutq madaniyatini oshirish; davlat [illegible] zamonaviy axborot texnologiyalari va kommunikatsiyalariga faol [illegible] ta'minlash"[1] hozirgi kunda o'zbek kompyuter

[1] [illegible] nufuzi va mavqeyini tubdan oshirish chora-tadbirlari to'g'risida"gi Prezident farmoni. https://lex.uz/docs/-4561730

6-ilova

Statistik usullarga asoslangan stemmerlar

n-gramm stemmeri	
Afzalliklari	**Kamchiliklari**
1) n-gramm va satrlarni o'zaro taqqoshlashga asoslangan; 2) tabiiy tilga bog'liq emas.	1) tahlil juda ko'p vaqt talab qiladi; 2) n-grammlarni saqlash uchun yetarlicha katta hajmdagi xotira talab qilinadi; 3) amalga qo'llash murakkab.
HMM stemmeri	
Afzalliklari	**Kamchiliklari**
1) HMM metodiga asoslangan; 2) tabiiy tilga bog'liq emas.	1) tatbiq qilish murakkab; 2)"over stemming yuzaga kelishi mumkin.
YASS stemmeri	
Afzalliklari	**Kamchiliklari**
1) klasterizatsiya va masofani aniqlashda iyerarxik yondashuvdan foydalaniladi; 2) korpusga asoslangan; 3) tabiiy tilga bog'liq emas.	1) klasterlarni hosil qilishdagi chegara (shart)larni aniqlash murakkab; 2) yetarli tezlikka ega texnika (kompyuter) talab qilinadi; 3) murakkab hisob-kitoblar talab qilinadi.

7-ilova

Fleksiya va so'z yasalishi usullariga asoslangan stemmerlar

Krovetz stemmeri	
Afzalliklari	**Kamchiliklari**
1) tushunarli (sodda); 2) boshqa stemmerlar uchun boshlang'ich stemmer sifatida foydalanish tavsiya qilinadi.	1) katta hajmdagi hujjatlar uchun samarali natija qaytarmaydi; 2) lug'atdan tashqari so'zlarni qayta ishlamaydi; 3) aniqlik ko'rsatkichi past; 4) lug'at avvaldan ishlab chiqilishi lozim.
Xerox stemmeri	
Afzalliklari	**Kamchiliklari**
1) katta hajmdagi hujjatlarni samarali qayta ishlashi mumkin; 2) prefikslarni o'chirish imkoniyati mavjud; 3) stemlash natijasida har doim ma'noga ega so'z hosil qilinadi.	1) lug'atdan tashqari so'zni qayta ishlamaydi; 2) faqat ingliz tiliga moslashtirilgan; 3) tabiiy tilga bog'liq.

MUNDARIJA

KIRISH 3
I BOB. TOKENLASH, STEMLASH VA LEMMALASHNING NAZARIY ASOSLARI 6
§ 1.1. Jahon va o'zbek tilshunosligida tokenlash jarayonining umumiy tavsifi 6
§ 1.2. Stemlash jarayonining umumiy tavsifi 12
§ 1.3. Lemmalash jarayoni: umumiy tavsif va tahlil 33
II BOB. O'ZBEK TILI LEKSIK BIRLIKLARINI STEMLASH VA LEMMALASHNING LINGVISTIK ASOSLARI 43
§ 2.1. O'zbek tili so'zshakllarini stemlash: muammo va yechimlar 43
§ 2.2. O'zbek tilida turli tuzilishli so'zlarni lemmalash usullari 64
§ 2.3. Til korpusi qidiruv tizimini optimallashtirishda lemmalashning ahamiyati 71
III BOB. O'ZBEK TILI LEKSIK BIRLIKLARINI TOKENLASH, STEMLASH VA LEMMALASHNING DASTURIY ASOSLARI 83
§ 3.1. O'zbek tili birliklarini BoW va BPE algoritmlari asosida tokenlash 83
§ 3.2. Affikslarni ajratishga asoslangan UzbStemming algoritmini ishlab chiqish 107
§ 3.3. O'zbek tilidagi so'zlarni lemmalash muammosi va usullari 112
XULOSA 122
FOYDALANILGAN ADABIYOTLAR RO'YXATI 125
ILOVALAR 141

Xusainova Z.Y.

O‘zbek tili birliklarini tokenlash, stemlash, lemmalashning lingvistik asoslari

MONOGRAFIYA

9527

Bosishga ruxsat etildi: 19.12.2025-yil.
Bichimi 60x84 $^{1}/_{16}$, “Times New Roman”
garniturada raqamli bosma usulida bosildi.
Shartli bosma tabog‘i: 9,5. Adadi: 100.

“Noble Print” MChJ bosmaxonasida chop etildi.
100194, Toshkent shahri, Yunusobod-11, 62-uy.