ISSN 2181-922X

UZBEKISTAN O‘ZBEKISTON

LANGUAGE & CULTURE

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 3 (6)

www.compling.tsuull.uz

MUNDARIJA

Mavjuda Alimbekova
Abdurauf Fitrat mualliflik korpusini yaratishning ahamiyati.....6

Madinabonu Qodirova, Shahlo Hamroyeva
Zamonaviy dunyoda mashina tarjimasi tadriji:
tahlillar va natijalar.....22

Noila Matyakubova
"Aligner" dasturiy vositasi uchun o‘zbek-ingliz tilida sifat va uning
darajalarining morfologik tahlili.....41

Mohiyaxon Uzoqova, Mansurbek Narzullayev
Sinonimayzer dasturida RoBERTaForMaskedLM modelidan leksik
sinonimlarni aniqlash uchun foydalanish.....54

Dlafruz Xudoyqulova
O‘zbek-ingliz farmatsevtika terminlari korpusli lingvistik
ta'minotining milliy-madaniy asoslari.....69

Ruhillo Alayev, Gulshaxnoz Maxmudjonova
O‘zbek tilidagi matnli hujjatlarda izlashni amalga
oshirishni takomillashtirish.....78

Sanjarbek Baxodirov
Tabiiy tilni qayta ishlashda matnni tozalash tizimini
ishlab chiqish.....91

Azizaxon Raxmanova
Sun’iy intelekt yordamida o‘zbek va ingliz tili lingvistik asoslarini
o‘qitishning zamonaviy uslublari.....106

ISSN 2181-922X 91-105

TABIIY TILNI QAYTA ISHLASHDA MATNNI TOZALASH TIZIMINI ISHLAB CHIQISH

Sanjarbek Baxodirov [1]

Annotatsiya. Matnni tozalash matnni tahlil qilish sifati va aniqligini oshirish uchun tabiiy tilni qayta ishlashda (NLP) muhim qadamdir. Bu imlo va formatlashdagi nomuvofiqliklarni bartaraf etish orqali matnni kichik harflarga aylantirish bilan birga maxsus belgilar, raqamlar va nomuhim so‘zlar kabi ahamiyatsiz yoki ortiqcha ma’lumotlarni olib tashlashni o‘z ichiga oladi. Matnni tozalash, shuningdek, imlo xatolarni qayta ishlash, so‘zlarni o‘zak shakliga keltirish (lemmatizatsiya) va matnni kodlash muammolarini hal qilish yechimlarini taklif qiladi. Matnni tozalashning maqsadi, matn ma’lumotlarining sentiment analizini tashkil etish, tilni modellashtirish va ma’lumot olish kabi keyingi qayta ishlash va tahlil qilish uchun tayyorlash sanaladi. Ushbu maqolada NLPda matnni tozalashning ahamiyatini, shuningdek, to‘g‘ri tuzilgan matn ma’lumotlariga erishish uchun ishlatiladigan turli xil texnika va vositalar muhokama qilinadi. Shu bilan bir qatorda, matnni tozalashning NLP modellari va ilovalari samaradorligini oshirishdagi ahamiyatini va uning tilni aniqroq va mazmunli tushunish va qayta ishlashni osonlashtirishdagi roli ta’kidlanadi.

Kalit so‘zlar: *normallashtirish, nomuhim so‘zlar, tokenizatsiya, lemmalash, stemlash.*

Kirish

Ma’lumotlar formati har doim ham jadval ko‘rinishda yoki dastur tushuna oladigan ko‘rinishda bo‘lavermaydi. Katta ma’lumotlar davriga kirar ekanmiz, ma’lumotlar juda xilma-xil shaklga ega, jumladan, tasvirlar, matnlar, grafiklar va boshqalar. Format har xil bo‘lganligi sababli, ma’lumotlarni kompyuter o‘qiy oladigan ko‘rinishga keltirib olish muhim bosqich sanaladi. Dasturga kiritilayotgan ma’lumotlarning sifati uning ishlashiga bevosita ta’sir o‘tkazadi. Matn ma’lumotlari konteksida “sifat” degan tushuncha to‘g‘ri tuzilgan, izchil va ahamiyatsiz so‘zlardan holi degan ma’nolarni anglatadi. Xususan, matn terishda xatoliklar, nomuvofiq so‘zlar, turli lahjalar

[1] *Baxodirov Sanjarbek Rahmatali o‘g‘li* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi magistranti.
E-pochta: sanjarbahodirov9901@gmail.com
ORCID: 0009-0008-0132-3510

bilan to‘ldirilgan kitobni o‘qishda qanchalik qiynalamiz. Bunday holatlar turli xil qiyinchiliklarni keltirib chiqarishi mumkin. Xuddi shu jarayon modelning ishlashiga ham ta’sir o‘tkazadi.

1.1. NLPning imkoniyatlari.

Tabiiy tilni qayta ishlash (NLP) tez rivojlanayotgan soha bo‘lib, u mashinalarga inson tilini tushunish va qayta ishlash imkonini beradi. NLP ning muhim jihatlaridan biri bu matnli ma’lumotlarni tozalash va oldindan qayta ishlash vazifasidir.

Matn ma’lumotlarini tozalash matndan ahamiyatsiz ma’lumotlarni olib tashlashni o‘z ichiga oladi. Bu maxsus belgilar, tinish belgilari va raqamlarni olib tashlash, shuningdek, ma’lumotlarni standartlashtirish uchun matnni kichik harflarga aylantirishni o‘z ichiga olishi mumkin. Bundan tashqari, nomuhim so‘zlar (masalan, mustaqil so‘z turkumlaridan sifat darajalari, miqdor-daraja, sabab, maqsad ravishi, olmoshlar, yordamchi so‘zlar, modal, taqlid, undov so‘zlar) ko‘pincha matndan olib tashlanadi, chunki ular NLP vazifalari uchun foydali ma’lumot bermaydi.

1.2. Muammoning o‘rganilganlik darajasi.

Matnni oldindan qayta ishlash yoki matnni tozalash o‘nlab yillar davomida tabiiy tilni qayta ishlash (NLP) va ma’lumotlarni tahlil qilishning muhim bosqichi bo‘lib kelgan. Jarayon nomuvofiq ma’lumotlarni olib tashlash, xatolarni tuzatish va ma’lumotlar formatini standartlashtirishni o‘z ichiga oladi.

Matn tozalash tarixini NLP tadqiqotlari va kompyuter dasturlashning dastlabki kunlaridan kuzatish mumkin. 1950-1960-yillarda tadqiqotchilar matnli ma’lumotlarni avtomatik ravishda qayta ishlash va tahlil qilish imkonini beruvchi dasturlarni ishlab chiqishga kirishdilar. Biroq, bu dastlabki dasturlar tabiiy tilning murakkabligi va o‘zgaruvchanligi bilan cheklangan va matnni tozalash qo‘lda va ko‘p mehnat talab qiladigan jarayon edi. Kompyuter texnologiyalari rivojlangan sari matnni tozalash usullari ham kengayib bordi. 1980 va 1990-yillarda tadqiqotchilar matn ma’lumotlarini avtomatik ravishda tozalash va qayta ishlash uchun yanada ilg‘or algoritmlar va usullar ishlab chiqishni boshladilar. Bu matn ma’lumotlaridagi tinish belgilari, nomuhim so‘zlar va boshqa maxsus belgilarni olib tashlashni tarkibiga qo‘shadi.

2000-yillarda katta ma’lumotlarning o‘sishi va internetda matnli ma’lumotlarning ko‘payishi bilan matnni tozalash ma’lumotlarni tahlil qilishning yanada muhim qismiga aylandi. Tadqiqotchi-

lar va ishlab chiqaruvchilar matnni tozalash uchun yanada murakkab vositalar va kutubxonalarni yaratishni boshladilar, bu jarayon avtomatlashtirilgan va kengaytirilishi mumkin.

Bugungi kunda matnni tozalash NLP va ma'lumotlarni tahlil qilishning asosiy qismi sanaladi. Bu matn ma'lumotlarini tahlil qilish uchun tokenizatsiya, lemmatizatsiya, stemming va lemmalash kabi usullarning kombinatsiyasidan foydalanadi. Matnni tozalash NLP modellari va ma'lumotlarni tahlil qilish natijalarining aniqligi va ishonchliligini ta'minlash uchun juda muhimdir.

Jahon kompyuter tilshunosligi NLP sohasida matnni tozalash borasida bir qator olimlar ilmiy izlanishlar olib borgan. N.Pentapalli, D.Silval, R.Vikkery, R.Chandra, S.Kim, K.Pikes, I.A.Xalid, K.Rastogi, I.Roldos kabi olimlarning tadqiqotlarini ana shunday ishlar sifatida qayd etish mumkin. Tabiiy tilni qayta ishlash (NLP) sohasidagi bir qancha tadqiqotchilar va amaliyotchilar matnni tozalash hamda tegishli mavzular bo'yicha tadqiqotlarga hissa qo'shgan. D.Jurafskiy tilni qayta ishlash va tushunishga e'tibor qaratgan holda NLP bo'yicha tadqiqot olib bordi, bu esa tuzilmagan matn ma'lumotlari va matnni tozalash bilan bog'liq muammolar bilan shug'ullanishni o'z ichiga oladi.

Osiyoda ushbu sohada tadqiqotlar amalga oshirganlardan biri P.Fung NLPda matnni tozalash sohasidagi taniqli tadqiqotchilardan biri sanaladi. P.Fung tabiiy tilni qayta ishlash bo'yicha, jumladan, matnni tozalash va NLP ilovalari uchun dastlabki ishlov berish usullari ustida ish olib bordi. Mazkur tadqiqotchilar matnni tozalash va uning NLP va kompyuter tilshunosligining kengroq kontekstida ahamiyatini o'rgangan ko'plab yetakchi mutaxassislarning bir nechta misolidir. Ularning ishi tabiiy tillarni qayta ishlash ilovalari uchun matn ma'lumotlarini tayyorlash va tozalash bilan bog'liq qiyinchiliklar hamda usullarni tushunishimizga yordam berdi.

Umuman olganda, matnni tozalash tarixi matn ma'lumotlarini qayta ishlash va tahlil qilishning yanada samarali usullarini ishlab chiqish bo'yicha davom etayotgan sa'y-harakatlarni aks ettiradi va u NLP hamda ma'lumotlarni tahlil qilish sohasidagi tadqiqot va ishlanmalarning muhim yo'nalishi bo'lib qolmoqda.

Asosiy qism

Tabiiy tilni qayta ishlash (NLP) da matnni tozalash usullari haqida gap ketganda, matnni tahlil qilish uchun odatda qo'llanadigan bir nechta usullar mavjud. Dastlab, matn tozalash texnikasini Python dasturlash tili asosida ko'rib chiqamiz. Python kuchli va moslashuv-

chan dastur bo‘lib, matnni samarali tozalash uchun turli xil kutubxonalarga murojaat qiladi. Bularga Natural Language Toolkit (NLTK), Regular Expressions (RegEx) va boshqalar kiradi. Ushbu vositalar bizga tinish belgilari va maxsus belgilarni olib tashlashdan tortib, so‘z shakllarini normallashtirishgacha (kichik harflarga aylantirish) bo‘lgan keng ko‘lamli matnni tozalash vazifalarini bajarishga yordam beradi. Maqolaning asosiy qismida ushbu usullarga alohida e’tibor qaratiladi, jumladan:

2.1 Tokenizatsiya

Bu matnni qayta ishlash va tahlil qilishni osonlashtirish uchun so‘zlar yoki jumlalar kabi kichikroq birliklarga ajratishni o‘z ichiga oluvchi usul sanaladi.

Z.Xusainova ma’lumotiga ko‘ra, tokenizatsiya so‘z asosidagi tokenizatsiya (Word Based Tokenization), belgilarga asoslangan tokenizatsiya, so‘z ostilarni aniqlashga asoslangan tokenizatsiya kabi guruhlarga ajratilgan.

So‘z asosidagi tokenizatsiya (Word Based Tokenization). Nomidan ko‘rinib turibdiki, so‘z asosidagi tokenizatsiya usullarida *tinish belgilari*, *bo‘shliqlar*, *chegaralovchilar* va boshqalar bilan ajratilgan so‘zlar token sifatida qabul qilinadi. Ajratish chegarasi qo‘yilgan NLP vazifasiga mos qo‘yiladi va ba’zan qayta ishlanayotgan ma’lumotlarning xususiyatiga bog‘liq bo‘lishi mumkin. Twitter ijtimoiy tarmog‘idagi tvitlarini tokenizatsiya qilish uchun mo‘ljallangan tokenizator yangiliklardan iborat maqolalarini tokenizatsiya qilish jarayonidan biroz farq qiladi. Quyidagi 1-jadvalda, faqat bo‘sh joy belgilariga asoslangan tokenizatsiya jarayoni keltirilgan:

Otam gapni shartta kesdilar: “Bugungi ishni ertaga qo‘yma!”							
Otam	gapni	shartta	kesdilar:	“Bugungi	ishni	ertaga	qo‘yma!”

1-jadval. Bo‘sh joy belgilariga asoslangan tokenizatsiya

Keltiriladigan 2- jadvalda esa, tokenlarni bo‘shliqlar va tinish belgilari asosida ajratib chiqiladi.

Otam gapni shartta kesdilar: “Bugungi ishni ertaga qo‘yma”												
Otam	gapni	shartta	kesdilar	:	“	Bugungi	Ishni	ertaga	qo	‘	yma	“

2-jadval. Bo‘shliqlar, tinish belgilariga asoslangan tokenizatsiya

Lekin bunday tokenlash ham o‘zbek tili qoidalariga to‘g‘ri kelmaydi. Chunki o‘zbek tilidagi tutuq belgisi va apostrofni hisobga olgan holda tokenizatsiya jarayonini tashkillash kerak. Yuqorida keltirilgan misolda **qo/ ‘/ yma/** tarzida qismlarga ajratildi. Lekin **“qo‘yma”** shaklida tokenga ajratsa maqsadga muvofiq bo‘ladi. Xud-

di shu muammoni hal qilish uchun ToshDO‘TAU ilmiy tadqiqotchilari tomonidan tokenizator ishlab chiqildi. Bu semantik jihatdan bir muncha mukammal va tushunishga sodda. Buni 3-jadvalda ko‘rish mumkin:

Otam gapni shartta kesdilar: “Bugungi ishni ertaga qo‘yma”										
Otam	gapni	shartta	Kesdilar	:	“	Bugungi	ishni	ertaga	qo‘yma	“

3-jadval. UzTokenizator yordamida tokenizatsiya

Belgilarga asoslangan tokenizatsiya

Belgilarga asoslangan tokenizatsiya jarayonida har bir belgi alohida bir token sifatida qaraladi. NLP shartiga ko‘ra **UNICODE, ASCII** kabi kodlashtirish usullaridan foydalanish mumkin. (4-jadval)

Otam gapni shartta kesdilar																					
O	t	a	m	g	a	p	n	i	s	h	a	r	t	t	a	k	e	s	d	i	l	a	r

4-jadval. Belgilarga asoslangan tokenizatsiya

Bu usulda tokenlar hech qanday semantik ma’no anglatmaydi. Agar NLP da yuqori samaradorlik masalasi shart qilib qo‘yilmagan bo‘lsa, belgilarga asoslangan tokenizatsiyadan foydalangan ma’qul. Chunki bunday usul orqali katta tezlikda, kam hisoblash orqali tokenlashni amalga oshirish mumkin. [Xusainova, 2022. 73-74]

Tokenlash jarayonini Python dasturlash tili yordamida qilib ko‘ramiz.

Masalan:

1. from nltk.tokenize import word_tokenize
2. text = “Hello, World! @Python #NLP”
3. tokens = word_tokenize(text)
4. print(tokens)
5. Outputs: [‘Hello’, ‚’, ‘World’, ‘!’, ‘@Python’, ‘#NLP’]

2.2. Lemmatizatsiya va Stemming

Bu usullar so‘zlarni asosiy yoki o‘zak shakliga qisqartirishni o‘z ichiga oladi. Bu matnni murakkabligini kamaytirishga va tahlilni aniqligini oshirishga yordam beradi, ammo ular bu amalni bajarish maqsadi va unga erishish usullari bilan farqlanadi.

Lemma yoki **leksema faqat** o‘zak (asos) yoki + so‘z yasovchidan iborat bo‘ladi. Leksema so‘zi yunonchadan olingan bo‘lib, so‘z, ifoda degan ma’nolarni anglatadi.

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, *ishlagan* so‘zining lemmasi *ishlamoq*, *uning* olmoshining lemmasi esa *u* hisoblanadi. Lemmatizatsiya so‘zning kontekstini, POS tegini, zamon va turkumini hisobga oladi.

NLP vazifalarini amalga oshirishda ko‘p hollarda lemmatizatsiya jarayonidan foydalaniladi.

Stem – so‘zshaklning qo‘shimchalarini kesib tashlashdan hosil bo‘luvchi qism bo‘lib, ba‘zi hollarda ma‘no anglatmasligi mumkin. Shuningdek, stem so‘zning morfologik o‘zagi bilan aynan mos bo‘lmasligi yoki mos tushishi mumkin [Xusainova, 2023. 43-44]

Stemming – kontekstni hisobga olmagan holda so‘zning asosiy shakli bo‘lgan o‘zagiga qisqartirish jarayoni. Masalan, *bajarish so‘zining* o‘zagi *bajar, uning* olmoshini stemi esa *u.* Stemming jarayoni NLP masalalarini hal qilishda asosiy rollarda turadi. Lekin jarayonni amalga oshirishda so‘zning kontekstini hisobga olmasligi sababli, turli xil xatoliklar yuzaga kelishi mumkin. Chunki, yuqorida aytilganidek, so‘zning stemi har doim ham so‘z ma’nosini to‘g‘ri ifodalamaydi [Elov, Hamroyeva, Abdullayeva, Xusainova, Xudayberganov, 2023. 41]. Mualliflar tomonidan olib borilgan tadqiqot natijasiga ko‘ra o‘zbek tili milliy korpusi matnlarida UzbStemmer algoritmning samaradorligi 95.5%ni tashkil etgan [Elov, Alayev, Xusainova, 2023. 41]. Masalan, “sotib oldim” so‘zining stemi 2 ta “sot” va “ol” bo‘lsa, lemmasi “sotib olmoq”dan iborat 1 ta leksik birlikni ifodalaydi [Xusainova, 2023. 48].

O‘zbek, Uyg‘ur va Turk tillarida stem va lemmaga qarash turlicha. Uyg‘ur va Turk tillarida “stemming” jarayonida so‘zga qo‘shilgan flektiv qo‘shimchalarni olib tashlash orqali o‘zakkacha qisqartiriladi. O‘zbek tilida esa, so‘zga qo‘shilgan derivatsion va flektiv qo‘shimchalarni kesib tashlashni stemming jarayoni amalga oshiradi. O‘zbek tilida lemma tub ham, yasama ham bo‘lishi mumkin. Misol uchun, *darvoza, darvozabon, chaman, chamanzor.* Demak, o‘zbek tilida lemma lug‘atda mavjud leksemaga to‘g‘ri keladi.

O‘zbek tilida stemming jarayonida barcha qo‘shimchalar: so‘z yasovchi, shakl yasovchi va lug‘aviy shakl yasovchilar olib tashlanadi. Masalan, **sinf**+*dosh+lar+i+miz+ga* tarzida amalga oshiriladi.

Turk tilida esa, bunday emas, so‘z yasovchilar qoldiriladi faqat sintaktik shakl yasovchi qo‘shimchalar va lug‘aviy shakl yasovchilar ajratiladi.

Uyg‘ur tilida ham turk tilidagi kabi sintaktik shakl yasovchilar va lug‘aviy shakl yasovchilar olib tashlanadi va so‘z yasovchilar qoldiriladi.

Z.Xusainova fikriga koʻra, Lemmatizatsiya soʻzshaklni (yoki soʻzni) asosiy shakli – lemmaga aylantiradigan matnni oʻzgartirish jarayoni. Odatda, bu jarayonda lugʻatlar, morfologik tahlil va soʻz turkumlariga ajratish kabi amallardan foydalaniladi. Stemming jarayoni esa, soʻzlarni ularning asos (oʻzak) shakliga qisqartirish usuli (soʻzning kanonik shaklini hosil qilish). Stemming jarayonida, odatda, soʻzlardagi qoʻshimchalarini kesib tashlaydigan evristik amaldan foydalaniladi (5-jadval).

Stemming va lemmatizatsiya oʻrtasidagi farq shundaki, oxirgisi kontekstni oladi va soʻzni lemmaga aylantiradi, stemming esa soʻnggi (baʼzan soʻz boshidagi) bir nechta belgilarni kesib tashlaydi.

Soʻzshakl	Stem	Lemma	Asos
kelajagimiz	Kelajag	kelajak	kelajak
oʻquvchilar	oʻquv	oʻquvchi	oʻqi
borib ketdi	bor ket (2ta)	borib ketmoq	bor ket (2ta)
ega boʻlishdi	ega boʻl (2ta)	ega boʻlmoq	ega boʻl (2ta)
taqillatdi	Taqilla	taqillamoq	Taq
undami	Un	u	u
keldilar	Kel	kelmoq	kel
Uyda	uy	uy	uy

5-jadval. Stemming va lemmatizatsiya jarayonlari

Stemming jarayonida turli xil muammolarga duch kelishimiz mumkin. Ular quyidagilar: turli xil neologizmlarni stemmlash, oʻzak bilan qoʻshimchaning bitta soʻz bilan omonim boʻlib qolishi yoki soʻzning turli tovush oʻzgarish hodisasiga uchrashi.

Xulosa qilib aytganda, lemmatizatsiya va stemming soʻzlarni ifodalashni soddalashtirish orqali NLP masalalarini yechishdagi murakkabliklarni kamaytirishda ishlatiladigan usullardir. Ikkala usul ham soʻzni asosiy shakliga qisqartirishni nazarda tutadi, ammo ular buni amalga oshirish algoritmi va usuli bilan farqlanadi. Lemmatizatsiya soʻz maʼnosini toʻgʻri ifodalashni taʼminlaydi, biroq bu jarayon murakkab, u sekinroq amalga oshiriladi. Stemming sodda va tezkor yondashuv boʻlsa-da, soʻz maʼnosini ifodalashda xatolarga olib kelishi mumkin [Sharma, Kumar, Mansotra, 2016].

2.3. Nomuhim (stopwords) soʻzlarni olib tashlash

Nomuhim soʻzlar barcha tillarda mavjud boʻlib, semantik maʼnoga ega boʻlmagan soʻzlarga aytiladi. Kam maʼnoli maʼlumotlarga ega boʻlgan yoki mustaqil maʼnoga ega boʻlmagan hamda barcha matnlarga xos keng tarqalgan soʻzlar nomuhim soʻzlari deb atala-

di [Madatov, Sharipov, Bekchanov, 2021]. Matn tozalash masalasi maqsad qilinganda nomuhim so‘zlarni olib tashlash kerak bo‘ladi. Chunki u quradigan dastur uchun ahamiyatga ega emas. Bu orqali matn hajmini kamaytirishga erishiladi.

Nomuhim so‘zlarning xususiyatlari:

1. Matnda juda kam qiymatiga ega bo‘ladi.
2. Matnda paydo bo‘lish chastotasi yuqori bo‘ladi
3. Kamdan-kam hollarda so‘rov so‘zlari/qidiruv so‘zlari sifatida ishlatilinadi, ya’ni ma’lumotlarni qidirishda bunday so‘zlardan foydalanilmaydi.
4. Kirish so‘zlar, olmosh, modal, taqlid, undov so‘zlar, yordamchi so‘zlar: bog‘lovchi, ko‘makchi yuklama kabi so‘zlar bo‘lishi mumkin.
5. Matn uchun umumiy so‘zlar bo‘lib, muayyan sohada maxsus ishlatilinmaydi.
6. Tilning qurilishi uchun zarurdir.

Nomuhim so‘zlar ro‘yxati, asosan, mustaqil so‘z turkumlaridan sifat darajalari, miqdor-daraja, sabab, maqsad ravishi, olmoshlar, yordamchi so‘zlar, modal, taqlid, undov so‘zlardan tuzulishi mumkin. Sababi, bunday so‘zlarning matn mazmuniga ahamiyati kam bo‘ladi, ya’ni gapda grammatik ma’no ifodalash uchun foydalaniladi.

O‘zbek tili uchun nomuhim so‘zlarning standart ro‘yxati mavjud emas. Mualliflar tomonidan ishlab chiqilgan o‘zbek tilidagi nomuhim so‘zlarning ba’zilari 6-jadvalda keltiligan.

afsuski	beri	eng	lla	mening	orqada	shak-shubhasiz	ustida
aftidan	bilan	esa	iloyim	Misli	orqaga	shekilli	ustidan
agar	binoan	essiz	ishonamanki	misoli	orqali	shu	ustiga
aksincha	biroq	evaziga	ishqilib	mobaynida	o‘sha	shubhasiz	va
albatta	biroz	faqat	jihatdan	mobodo	ostida	shunchaki	vaqtda
allaqachon	biz	gar	joiz	modomiki	oxir	shunday	xayriyat
alqissa	bizning	garchand	juda	mos	o‘zi	shunga	xo‘sh
ammo	bo‘yi	garchi	jumladan	na, na	o‘zim	shuning uchun	xolos
ammo-lekin	bog‘liq	go‘yo	Kabi	nachora	o‘zimiz	shuningdek	xuddi
aqalli	bois	go‘yoki	kerak	nafaqat	o‘zingiz	singari	xullas
arafasida	boshqa	goh	keyin	nafar	o‘zlari	siz	xususan
aro	bu	chunonchi	keyingi	natijada	pastda	sizniki	ya’ni
aslida	bular	ham	kim	negaki	pastga	sizning	yana
aslo	bundan	hamda	kimdir	nimagaki	payida	tabiiyki	yanada
asosan	bunday	hamma	kimga	o‘rniga	qachonki	tag‘in	yaxshi
asosiy	butun	hammasi	ko‘p	o‘rtasida	qadar	tahminan	yaxshiyam
avvalambor	chamasi	haqiqatda	ko‘plab	o‘sha	qarab	tashqari	yo
avvalgi	chog‘i	haqiqatdan	ko‘proq	o‘z	qarata	to‘g‘risi	yo‘qsa
avvalo	chunki	har	ko‘ra	o‘zi	qaratilgan	toki	yo‘q-yo‘q
axir	chunonchi	har holda	ko‘proq	o‘zim	qayta	tomon	yo‘sinda
aynan	darhaqiqat	har qalay	lekin	o‘zimiz	qaytanga	tufayli	yoki
ayni	darhol	hatto	lozim	o‘zingiz	qisqasi	turli	yonida
ayniqsa	darkor	hech	ma’lum	o‘zini	qo‘yingki	u	yonidan
aytaylik	dastavval	hokazo	mabodo	o‘zlari	quyida	uchun	yoniga
aytgancha	davomida	holda	mana	o‘z-o‘zidan	quyidagi	ular	yo‘q
aytganday	demak	hozir	masalan	ochig‘i	ravishda	ularni	yoxud
ba’zi	deyarli	hozirda	mayli	oid	rostdan	ularning	yuqorida
balki	deylik	ichida	mazkur	olaylik	rosti	unda	yuqoriga
baravarida	doim	ichidan	mazmuni	oldiga	sana	unga	yuzasidan

barcha	doir	ichiga	men	oldin	sari	uni	zero
barchasi	doirasida	ichkarida	meni	orasida	sayin	uning	zeroki
baribir	ehtimol	ichra	menimcha	orasiga	seningcha	ushbu	zotan

6-jadval. Nomuhim so‘zlar ro‘yxati

Nomuhim so‘zlar gap tarkibini grammatik shakllantirishda ishlatilinib, semantik ma’no ahamiyati kam. Bularni gap tarkibidan olib tashlasak ham deyarli ma’no o‘zgarmaydi. Misol sifatida Ashurali Jo‘rayevning “Kichik Vatan” hikoyasidan parchani ko‘rib chiqaylik [Matchonov, 2020]:

O‘shanda *uchinchi sinfda o‘qirdim.* ***Biz*** *oilamiz* ***bilan*** *boshqa qishloqqa ko‘chadigan bo‘ldik. Ko‘chishimizdan bir kun oldin akam* ***bilan*** *otam mollarni haydab, yangi uyimizga ketishdi. Ularga itimiz* ***ham*** *ergashdi.*

Xulosa qilib aytganda, nomuhim so‘zlar – bu hech qanday ma’lumot qiymatiga ega bo‘lmagan va hujjatning katta qismini tashkil etuvchi matn tarkibidagi so‘zlar. Bunday so‘zlardan mashinali o‘qitish modellarida ma’lumotlar hajmini qisqartirish va modellar tez ishlashi uchun foydalanish mumkin. Matnlarni tasniflovchi modellarda nomuhim so‘zlarni olib tashlash yaxshi natija beradi. Biroq his-tuyg‘ularga asoslangan modellarda bunday so‘zlarni olib tashlash notog‘ri xulosaga olib kelishi mumkin

2.4. Maxsus belgilar (special characters) va tinish belgilarini olib tashlash

Maxsus belgilar deyilganda, alifbo va sonlar sistemasida mavjud bo‘lmagan belgilarga aytiladi. Bularga tinish belgilari, belgilar (symbols), bo‘shliq belgilari (yangi qator), boshqaruv belgilari (control characters) va boshqa chop etib bo‘lmaydigan belgilar kiradi.

Maxsus belgilarga misol:

1. Tinish belgilar: ! @ # $ % ^ & * () - _ = + [] {}; : ‘ “ , . < > / ?
2. Belgilar: © ™ ® ° ± §
3. Bo‘shliq belgilari: bo‘sh joy, tab, yangi qator
4. Boshqaruv belgilari: →,←

Tinish belgilarini olib tashlash matnni tozalashda keng tarqalgan vazifa bo‘lib, matnni tabiiy tilni qayta ishlash va tahlil qilish uchun tayyorlashni o‘z ichiga oladi. Bu orqali matnni qayta tayyorlashni osonlashtirish mumkin, ayniqsa hissiyotlarni tahlil qilish, tilni modellashtirishda foydalansa bo‘ladi.

Maxsus belgilar va tinish belgilari ko‘pincha semantik ma’no bermasdan matn ma’nolarida qo‘shimcha xatolarni yuzaga keltirib chiqarishi mumkin. Ularni Pythonning o‘rnatilgan **“string usullari”**(qo‘shtirnoq ichiga olingan Unicode belgilar ketma-ketligi)

yordamida osongina olib tashlashimiz mumkin. Misol tariqasida ko‘rishimiz mumkin:

import re

text = “Hello, World! @Python #NLP”

clean_text = re.sub(r’[^\w\s]’, ‘’, text)

print(clean_text)

Outputs: Hello World Python NLP

Maxsus belgilar va tinish belgilaridan tashqari, raqamlar, belgilar, HTML teglari yoki kulgichlar(emoji) kabi keraksiz belgilarni olib tashlash muhim sanaladi.

Pythonda kodlar yordamida satrdan tinish belgilarini qanday olib tashlashga misol:

import re

text = “Qishki “perevod”ga barcha ham hujjat topshirolmaydimi?. Faqat, 2 ta uzrli sababi borlarga ruxsat!.”

clean_text = re.sub(r’[^\w\s]’,’’, text)

print(clean_text)

Output: Qishki perevodga barcha ham hujjat topshirolmaydimi Faqat 2 ta uzrli sababi borlarga ruxsat

2.5. Normallashtirish (Case normalization)

Harflarni normallashtirish–bu matndagi barcha harflarni katta yoki kichik harfga o‘zgartirish jarayoni. Bu muhim jarayon sanaladi, chunki kompyuter katta va kichik harflarni bir-biridan farqlaydi. Misol uchun gap boshida kelgan ***Sen*** so‘zi bilan gap o‘rtasida kelgan ***sen*** kishilik olmoshi inson uchun bir xil ma’no anglatadi, lekin kompyuter bunga ikki xil so‘z sifatida qaraydi. Shuning uchun, har bir so‘z bir xil holatda bo‘lishi va kompyuter bir xil so‘zni 2 xil token sifatida qayta ishlamasligi uchun so‘zlarimizning holatini normallashtirishimiz zarur. Bu ma’lumotlar izchilligini ta’minlash va tahlilning aniqligini oshirishga yordam beradi.

Ma’lumotlarni qayta ishlash konteksida harflarni normallashtirish turli ilovalar va tizimlarda funksionallik va aniqlikni taminlaydi. Masalan, ma’lumotlar bazasida harflarni normallashtirish foydalanuvchining qidiruv operatsiyalarida va matn tahlilida yordam berishi mumkin. Ushbu jarayonni ijtimoiy media postlaridagi yangiliklarni aniqlash masalasi orqali ko‘rib chiqaylik. Ijtimoiy media matni gazetalarda foydalaniladigan tildan juda farq qiladi. So‘zlar turli yo‘llar bilan yozilishi mumkin, masalan, qisqartirilgan shakllarda, telefon raqami turli formatlarda yozilishi mumkin, ismlar ba’zan kichik harflar bilan yoziladi va hokazo. Bunday ma’lumot-

lar bilan ishlash uchun NLP vositalarini ishlab chiqishda matnning barcha o‘zgarishlarni qamrab oladigan **kanonik shaklini** hosil qilishimiz lozim. Ushbu jarayon matnni normallashtirish deb nomlanadi. Matnni normallashtirishning ba’zi umumiy qadamlari matnni barcha kichik yoki katta harflarga aylantirish, raqamlarni matnga aylantirish (masalan, 9 – “to‘qqiz”), qisqartmalarni kengaytirish va hokazo. Matnni normallashtirishning oddiy usuli Spacy paketida taqdim etilgan.

Harflarni normallashtirish Python, JavaScript yoki SQL kabi turli xil dasturlash tillari va vositalari yordamida amalga oshirilishi mumkin. Ko‘pgina dasturlash tillari harflarni normallashtirish uchun o‘rnatilgan funksiyalar yoki usullarni taqdim etadi, bu ularni kod orqali amalga oshirishni osonlashtiradi. Bundan tashqari, maxsus belgilar, urg‘ular yoki o‘zbek tilida tanish bo‘lmagan belgilar bilan ishlash kabi holatlarni normallashtirish uchun maxsus kutubxonalar va modullar ham mavjud. Ushbu jarayonni Python dasturlash tili asosida ko‘rish mumkin:

Misol uchun:

text = “Hello, World! @Python #NLP”

lowercase_text = text.lower()

print(lowercase_text)

outputs:hello, world! @python #nlp

Umuman olganda, harflarni normallashtirish ma’lumotlarni qayta ishlash va dasturlashning muhim jihati hisoblanadi, chunki u matnni qayta ishlashda va manipulatsiya qilishda aniqlikni ta'minlaydi.

Xulosa

Matnni tozalash tabiiy tilni qayta ishlashning (NLP) muhim bosqichi bo‘lib, matn ma’lumotlarini tahlil qilish va modellashtirish uchun oldindan qayta ishlash va tayyorlashni o‘z ichiga oladi. Ushbu jarayon NLP algoritmlari va modellarining aniqligi va samaradorligini ta’minlash uchun juda muhimdir. Ushbu keng qamrovli xulosada biz matnni tozalash bilan bog‘liq turli komponentlar va usullarni, uning NLPdagi ahamiyatini va ushbu jarayon bilan bog‘liq muammolarni ko‘rib chiqamiz.

Matnni tozalash standartlashtirish va matnli ma’lumotlarni keyingi tahlil qilish uchun tayyorlashga qaratilgan bir qator vazifalarni o‘z ichiga oladi. Ushbu vazifalarga har qanday ahamiyatsiz ma’lumotlarni olib tashlash, matnni kichik harflarga aylantirish, tinish belgilarini olib tashlash va berilgan kontekstda muhim ma’noga

ega bo‘lmagan umumiy so‘zlar bo‘lgan nomuhim so‘zlarini olib tashlash kiradi. Bundan tashqari, tokenizatsiya, stemming yoki lemmatizatsiya hamda imlo xatolari va nomuvofiqliklar bilan ishlash ham matnni tozalash jarayonining bir qismidir. Matnni tozalashning asosiy maqsadlaridan biri matn ma’lumotlarini NLP algoritmlari ularni to‘g‘ri talqin qilish va tahlil qilishini ta’minlash uchun standartlashtirishdir. Ushbu standartlashtirish turli NLP ilovalari, jumladan, hissiyotlarni tahlil qilish, matn tasnifi, ma’lumotlarni qidirish, mashina tarjimasi uchun juda muhimdir. Matn ma’lumotlarining tozalanishi va nomuvofiqliklardan holi bo‘lishini ta’minlash orqali NLP modellari va algoritmlarining ishlashi sezilarli darajada yaxshilanadi.

NLPda matnni tozalashning ahamiyatini oshirib bo‘lmaydi. Toza va standartlashtirilgan matn ma’lumotlari NLP modellarini o‘rgatish va sinab ko‘rish uchun zarurdir. Mos kelmaydigan yoki xato ma’lumotlar noto‘g‘ri tahlil va noto‘g‘ri natijalarga olib kelishi mumkin, bu har qanday NLP dasturining ishonchliligiga ta’sir qiladi. Shuning uchun matnni tozalash NLP modellarining umumiy aniqligi va ishonchliligiga hissa qo‘shadigan NLPni qayta ishlash jarayonida asosiy rol o‘ynaydi.

Biroq, NLPda matnni tozalash bir qator qiyinchiliklarni keltirib chiqaradi. Muhim qiyinchiliklardan biri noto‘g‘ri imlolar, grammatik xatolar va qoidabuzarliklarni o‘z ichiga olishi mumkin bo‘lgan to‘g‘ri tuzilmagan va xilma-xil matn ma’lumotlarini qayta ishlashdir. Bu muammolarni samarali hal qilish uchun imlo tekshiruvi kabi turli usullarni qo‘llashni talab qiladi. Bundan tashqari, turli tillar va shevalar matnni tozalash uchun maxsus yondashuvlarni talab qilishi mumkin, bu jarayonni yanada murakkablashtiradi.

Xulosa qilib aytganda, matnni tozalash NLP modellari va algoritmlarining ishlashi va aniqligini yaxshilash uchun matnli ma’lumotlarni oldindan qayta ishlash va standartlashtirishni o‘z ichiga olgan tabiiy tilni qayta ishlashning muhim tarkibiy qismidir. Matnni tozalash bilan bog‘liq muammolarni hal qilish orqali NLP amaliyotchilari o‘zlarining NLP ilovalarining ishonchliligi va samaradorligini ta’minlashlari mumkin.

Foydalanilgan adabiyotlar

Elov B., Hamroyeva Sh., Xusainova Z. NLPning zamonaviy algoritmlari va konsepsiyalari. – Toshkent: So‘z san’ati xalqaro jurnali, 2023. –52b.

Elov B., Hamroyeva Sh., Xusainova Z., Abdullayeva O., Xudayberganov N. O‘zbek, turk va uyg‘ur tillarida pos teglash va stemming.

– Toshkent: Uzbekistan: til va madaniyat, 2023. – 43-44 b.

Mahmudjonova G. Nomuhim so‘zlar tushunchasi va uning ahamiyati. “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” Xalqaro ilmiy-amaliy konferensiya, 2023.

Madatov X., Sharipov M., Bekchanov Sh. O‘zbek tili matnlaridagi nomuhim so‘zlar. – 2021

Xusainova Z. O‘zbek tili ta’limiy korpusida lemmalash algoritmlari. O‘zbek tili milliy va ta’limiy korpusining nazariy va amaliy masalalari, Respublika ilmiy-amaliy konferensiya to‘plami. Toshkent: ToshDO‘TAU, 05.05.2023. – 48 b.

Xusainova Z. Tokenizatsiya algoritmlari. – Termiz: Filologik tadqiqotlar: til, adabiyot, ta’lim, 2022 / № 5-6. 73-74 b.

Pentapalli N. Text Cleaning in Natural Language Processing (NLP). Analytics Vidhya, 2020, Jun 1. https://medium.com/analytics-vidhya

Rastogi K. Text cleaning methods in NLP. Analytics Vidhya, 2022. https://www.analyticsvidhya.com/

Chandra R. Text Cleaning in Python: Effective Data Cleaning Tutorial, 2023.

Vickery R. Text Cleaning Methods for Natural Language Processing. Towards Data Science. 2020, Feb

DEVELOPMENT OF A SYSTEM FOR CLEANING TEXT IN NATURAL LANGUAGE PROCESSING

Sanjarbek Baxodirov[1]

Abstract. Text cleaning is a crucial step in natural language processing (NLP) to enhance the quality and accuracy of text analysis. This involves eliminating irrelevant or redundant information, such as special characters, numbers, and stopwords, while ensuring consistency in spelling and formatting by converting text to lowercase. Text cleaning also provides solutions for spelling errors, word stemming (lemmatisation), and text encoding. The aim of text cleaning is to prepare text data for further processing and analysis, such as organizing sentiment analysis, language modelling, and data mining. This article discusses the significance of text cleaning in NLP, as well as the different techniques and tools used to achieve well-structured text data. The significance the importance of text cleaning in enhancing the performance of NLP models and applications, as well as its role in enabling more precise and meaningful language comprehension and processing, is also emphasized.

Key words: *normalizing, stopwords, tokenization, lemmatization, stemming.*

References

Elov B., Hamroyeva Sh., Xusainova Z. NLPning zamonaviy algoritmlari va konsepsiyalari. – Toshkent: So‘z san’ati xalqaro jurnali, 2023. –52b.

Elov B., Hamroyeva Sh., Xusainova Z., Abdullayeva O., Xudayberganov N. O‘zbek, turk va uyg‘ur tillarida pos teglash va stemming. – Toshkent: Uzbekistan: til va madaniyat, 2023. – 43-44 b.

Mahmudjonova G. Nomuhim so‘zlar tushunchasi va uning ahamiyati. “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” Xalqaro ilmiy-amaliy konferensiya, 2023.

Madatov X., Sharipov M., Bekchanov Sh. O‘zbek tili matnlaridagi nomuhim so‘zlar. – 2021

Xusainova Z. O‘zbek tili ta’limiy korpusida lemmalash algoritmlari. O‘zbek tili milliy va ta’limiy korpusining nazariy va amaliy masalalari, Respublika ilmiy-amaliy konferensiya to‘plami. Toshkent: ToshDO‘TAU, 05.05.2023. – 48 b.

Xusainova Z. Tokenizatsiya algoritmlari. – Termiz: Filologik tadqiqotlar: til, adabiyot, ta’lim, 2022 / № 5-6. 73-74 b.

[1] *Baxodirov Sanjarbek Rahmatali o‘g‘li* – Master of degree. Alisher Navo’i Tashkent State University of Uzbek Language and Literature.
E-pochta: sanjarbahodirov9901@gmail.com
ORCID: 0009-0008-0132-3510

Pentapalli N. Text Cleaning in Natural Language Processing (NLP). Analytics Vidhya, 2020, Jun 1. https://medium.com/analytics-vidhya

Rastogi K. Text cleaning methods in NLP. Analytics Vidhya, 2022. https://www.analyticsvidhya.com/

Chandra R. Text Cleaning in Python: Effective Data Cleaning Tutorial, 2023.

Vickery R. Text Cleaning Methods for Natural Language Processing. Towards Data Science. 2020, Feb