ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O‘ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 1 (6)

www.compling.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 1 (6)

compling.tsuull.uz

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

Bosh muharrir: **Botir Elov**

Bosh muharrir o‘rinbosari: **Shahlo Hamroyeva**

Mas’ul kotib: **Oqila Abdullayeva**

Tahrir kengashi

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

Alisher Navo'i Tashkent State University of the Uzbek Language and Literature

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

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 29.02.2024-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.

ISSN 2181-922X 21-31

O‘ZBEK-TURK PARALLEL KORPUSI UCHUN MATNLAR TOKENIZATSIYASI MASALASI

Iqbola Xolmonova[1]

Annotatsiya. Ushbu maqolada tokenizatsiya haqida, korpus tuzish uchun tokenizatsiya zarurligining sabablari, o‘zbek-turk parallel matnlari tokenizatsiyasini amalga oshiruvchi dasturlar va ularning mavjud imkoniyatlari yoritilgan. Shu bilan birga, o‘zbek-turk parallel korpusi uchun matn tokenizatsiyasi jarayonida duch kelish mumkin bo‘lgan muammolar va ularning yechimi haqida so‘z boradi. Undan tashqari “Python NLTK yordamida so‘z tokenizatsiyasi”[2] dasturidan foydalanish tartibi haqida ma’lumotlar mavjud.

Kalit so‘zlar: *token, tokenizatsiya jarayoni, parallel korpus, Python NLTK, o‘zbek-turk parallel korpus.*

Tokenizatsiya – berilgan matndagi gaplarni eng kichik o‘lchov birligi hisoblangan token deb nomlanuvchi elementlarga ajratish jarayoni [Rai, 2020. 11]. Gapdagi *tinish belgilari, so‘zlar va raqamlar* **token** sifatida aniqlanishi mumkin. Tokenlarni aniqlash orqali matndagi so‘zlarning uchrash chastotasini topish mumkin. So‘ngra ushbu chastotalar vositasida turli modellarni ishlab chiqishga imkon yaratiladi. Yoki tokenlarni so‘zlar tabiati (guruhi)ga ko‘ra razmetkalashni amalga oshirish mumkin [Xusainova, 2022. 154].

Matnning tokenizatsiyasi matn qismini so‘zlar yoki belgilardan iborat bo‘lgan alohida qismlarga ajaratish jarayonini anglatadi. Ushbu jarayon, odatda matn tasnifi, ma’lumotlarni qidirish va mashina tarjimasi kabi tabiiy tilni qayta ishlash (NLP) vazifalarida qo‘llaniladi. Tokenizatsiya matn ma’lumotlarini yanada boshqariladigan va aniq tuzilgan formatga aylantirish orqali tartibga solish hamda tahlil qilishda yordam beradi. Har bir token matn ichidagi ma’noning diskret birligini ifodalaydi, bu ma’lumotlarni oson boshqarish va tahlil qilish imkonini beradi.

Tokenizatsiya tabiiy tilni qayta ishlashda paragraflar va gaplar ma’nosini osonroq belgilash mumkin bo‘lgan kichikroq

[1]*Xolmonova Iqbola Alisher qizi* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi mutaxassisligi magistranti.
E-pochta: *iqbolabintualisher@gmail.com*
ORCID: 0009-0006-8844-6569
[2]http://text-processing.com/demo/tokenize/

birliklarga bo‘lish uchun ishlatiladi. NLP jarayonining birinchi bosqichi ma’lumotlarni yig‘ish (gap) va uni tushunarli qismlarga (so‘zlarga) ajratishdir. Gapni mashina tushunib olishi uchun satrda uni alohida qismlarga ajratish uchun tokenizatsiya amalga oshiriladi. Tokenizatsiya jarayoni matnni morfologik tahlil qilish NLP vazifasining muhim bosqichi hisoblanadi [Elov, Hamroyeva, Axmedova, 2022].

Tokenizatsiya jarayoni ayniqsa katta hajmdagi matnlar uchun juda muhim bo‘lib, u mashinaga ma’lum so‘zlarning chastotalarini hamda ularning turli statistikalarini hisoblash imkonini beradi [Rai, Borah, 2021. 137].

Korpusni yaratish uchun ham tokenizatsiyani amalga oshirish zarur, chunki u tahlil qilish uchun tuzilgan va boshqariladigan ma’lumotlar to‘plamini yaratishga imkon beradi. Matn qismini alohida leksemalarga bo‘lish orqali har bir token matn ichidagi diskret ma’no birligini ifodalaydi. Bu matn ma’lumotlarini boshqarish va tahlil qilishni osonlashtiradi. Bundan tashqari, tokenizatsiya matndan keraksiz shovqin va tinish belgilarini olib tashlashga yordam beradi, ishlov berish va tahlil qilishni osonlashtiradi. Bu tilni modellashtirish va statistik tahlil kabi vazifalar uchun muhim bo‘lgan so‘z chastotalarini hisoblash imkonini beradi. Umuman olganda, tokenizatsiya tuzilmagan matnni samarali qayta ishlanishi va tahlil qilinishi mumkin bo‘lgan tuzilgan formatga aylantirish orqali korpus yaratishda muhim rol o‘ynaydi.

Korpus tuzish uchun tokenizatsiya zarurligining sabablarini quyidagi jadvalda ko‘rish mumkin:

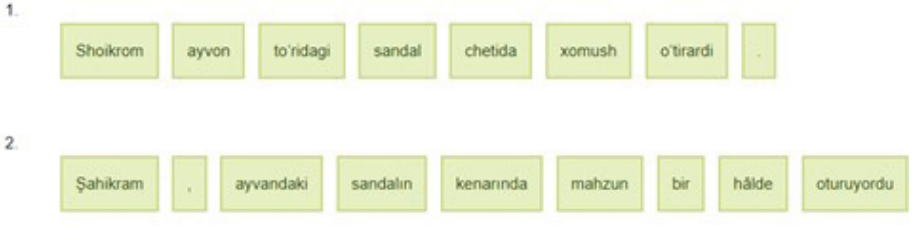
1-jadval. *Korpus tuzish uchun tokenizatsiyaning ahamiyati*

1	Lug‘at tahlili	Tokenizatsiya matn lug‘atini alohida leksikalarga bo‘lish orqali tahlil qilishda yordam beradi. Bu so‘zlarni, ularning chastotasini va korpus ichida tarqalishini aniqlash imkonini beradi.
2	Matnni oldindan qayta ishlash	Tokenizatsiya ko‘pincha matnni oldindan qayta ishlashning birinchi bosqichi bo‘lib, u xom matn ma’lumotlarini tozalash va tahlil qilish uchun mos formatga aylantirishni o‘z ichiga oladi. Matnni to-kenlarga bo‘lish orqali tinish belgilarini olib tash-lash, kichik harflarga o‘tkazish, to‘xtash so‘zlarini olib tashlash va so‘zlarni o‘zgartirish yoki lemma-tizatsiya qilish kabi turli xil qayta ishlash usul-larini qo‘llash osonroq bo‘ladi.

3	Xususiyatlarni ajratib olish	Tokenizatsiya matn ma’lumotlaridan xususiyatlarni olish uchun zarurdir. Har bir token alohida ma’no birligini ifodalaydi va bu tokenlar hissiyotlarni tahlil qilish, nomli obyektlarni tanib olish, nutq qismlarini belgilash va matnni tas-niflash kabi turli NLP vazifalari uchun xususiyat sifatida ishlatilishi mumkin. Matnni tokenlash or-qali biz matnning mohiyatini aks ettiruvchi mazmunli xususiyatlarni ajratib olishimiz mumkin.
4	Tilni tushunish	Tokenizatsiya matnning tuzilishi va mazmunini tushunishga yordam beradi. Matnni belgilarga bo‘lish orqali biz korpus ichidagi iboralar, jumlalar va paragraflarni aniqlashimiz mumkin. Bu tushunish tahlil qilish, sintaktik tahlil va semantik tahlil kabi vazifalar uchun juda muhimdir.

Umuman olganda, tokenizatsiya tuzilmagan matnli ma’lumotlarni NLP algoritmlari tomonidan osongina qayta ishlanishi, tahlil qilinishi va tushunilishi mumkin bo‘lgan tuzilgan formatga aylantirish uchun kerak. U korpusni qurish uchun asos bo‘lib xizmat qiladi va turli tillarni qayta ishlash vazifalarini bajarishga imkon beradi.

Turli tillardagi parallel matnlarni tokenlarga ajratishda bir qancha muammolarga duch kelish mumkin. Masalan, parallel matnlarni tokenlarga ajratishdagi muammolardan biri bu moslashtirish masalasidir. Parallel matnlar tarjima bo‘lgan yoki o‘xshash mazmunga ega bo‘lgan turli tillardagi juft yoki matnlar to‘plamidan iborat. Bu matnlarni leksemalarga ajratishda har bir tildagi leksemalarning to‘g‘ri kelishini ta’minlashi zarur, ya’ni bir tildagi tegishli leksemalar boshqa tildagi leksemalarga mos kelishini ta’minlash kerak. Tillar turlicha grammatik tuzilmalarga, so‘z tartibiga va tinish belgilariga ega bo‘lgani uchun moslashtirish qiyin bo‘lishi kuzatiladi. Bu tokenlashtirishda nomuvofiqliklarga olib kelishi mumkin, bunda bir tildagi token boshqa tilda bir nechta tokenlarga bo‘linishi yoki aksincha bo‘lishi mumkin. Ushbu hodisani o‘zbek va turk tilidagi quyidagi misolda ko‘rish mumkin:



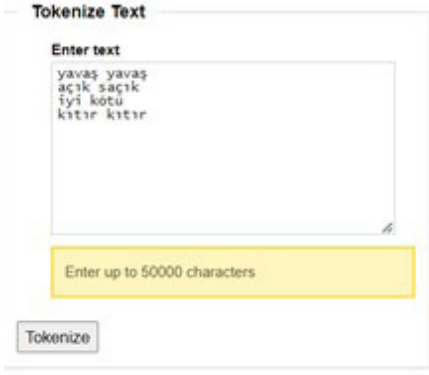
1-rasm. Tokenlarga ajratilgan parallel matn

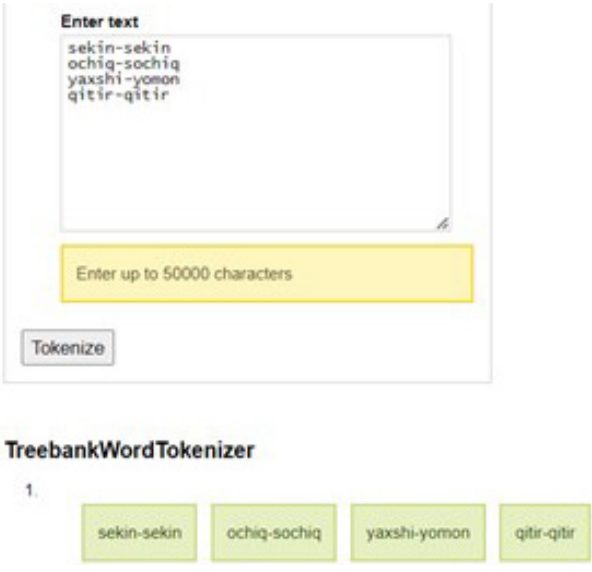
Bir tilda so‘zga ko‘makchi qo‘shish orqali ma’no hosil qilingan bo‘lsa, boshqasida so‘zga qo‘shimcha qo‘shish orqali ma’no hosil qilingan. Bu bir tilda bitta token ikkinchi tilda ikkita tokenga teng kelishiga sabab bo‘lgan. Ushbu noto‘g‘ri joylashtirish noto‘g‘ri tahlilga olib kelishi va mashina tarjimasi yoki tillararo ma’lumot olish kabi NLP vazifalarining samaradorligiga to‘sqinlik qilishi mumkin.

Parallel matnlarni tokenlarga ajratishdagi yana bir muammo - bu tilga xos muammolar mavjudligidir. Turli tillar tokenizatsiyani murakkablashtiradigan o‘ziga xos xususiyatlarga ega bo‘lishi mumkin. Masalan, ba’zi tillar qo‘shma so‘zlardan foydalanadi yoki tokenizatsiya paytida maxsus ishlov berishni talab qiladigan murakkab morfologik tuzilishga ega. O‘zbek va turk tilidagi takroriy va juft so‘zlar misolida oladigan bo‘lsak, o‘zbek tilida juft va takroriy so‘zlar chiziqcha bilan yoziladi va bir token sifatida qaraladi, ammo turk tilida takrorlar ajratib yoziladi va ikkita token sifatida qaraladi:

2-jadval. *O‘zbek va turk tilidagi juft va takroriy so‘zlarga misollar*

Turk	O‘zbek
yavaş yavaş	sekin-sekin
açık saçık	ochiq-sochiq
iyi kötü	yaxshi-yomon
kıtır kıtır	qitir-qitir





2-rasm. O‘zbek va turk tilidagi juft va takroriy so‘zlarning tokenizatsiyasi natijasi

O‘zbek va turk tilidagi so‘zlarni parallel ravishda moslashtirish uchun tokenizatsiya jarayonida duch kelish mumkin bo‘lgan yana bir muammo frazema va so‘zlarning parallel tokenizatsiyasidir. Ya’ni, bir tilda biron bir so‘z ibora yoki tasviriy ifoda tarzida ishlatilgan bo‘lsa, boshqasida birgina tokendan iborat so‘z holatida ishlatilgan bo‘lishi mumkin. Quyidagi misolda o‘zbek tilidagi 3 ta tokendan iborat “yuragi seskanib ketmoq” iborasi turk tilida 1 ta tokendan iborat “qo‘rqmoq” so‘ziga teng kelganini ko‘rish mumkin:

3-rasm. O‘zbek va turk tilidagi ibora va so‘zning parallel tokenizatsiyasi natijasi

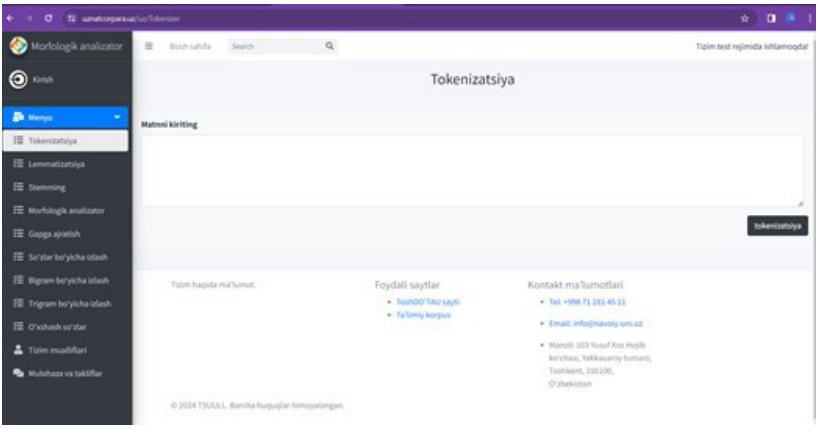
Bundan tashqari, lotin bo‘lmagan skriptlarga ega tillar matnni tokenlarga bo‘lish uchun maxsus usullarni talab qilishi mumkin. Bundan tashqari, parallel matnlar ko‘pincha shovqinli yoki nomuvofiq ma’lumotlarni o‘z ichiga oladi, masalan, formatlash farqlari, imlo o‘zgarishlari yoki tipografik xatolar. Ushbu nomuvofiqliklar matnlarni to‘g‘ri tokenizatsiya qilishni qiyinlashtirishi va tokenizatsiyadan

oldin ma’lumotlarni tozalash va normallashtirish uchun qo‘shimcha qayta ishlash bosqichlarini talab qilishi mumkin. Umuman olganda, parallel matnlarni tokenlarga ajratish moslashtirish muammolari, tilga xos murakkabliklar va shovqinli ma’lumotlar tufayli qiyin bo‘lishi mumkin. Ushbu muammolarni hal qilish tegishli tillarning lingvistik xususiyatlarini sinchkovlik bilan ko‘rib chiqishni va to‘g‘ri va izchil tokenizatsiyani ta’minlash uchun tegishli usullarni va dastlabki ishlov berish bosqichlarini amalga oshirishni talab qiladi.

Parallel matnlarni leksemalarga ajratishda bir qancha muammolarni turk va o‘zbek tillari misolida ko‘rib chiqadigan bo‘lsak, birinchidan, grammatika, so‘z tartibi va tinish belgilaridagi farqlar tufayli ikki til o‘rtasidagi moslashish qiyin bo‘lishi mumkin. Bu shuni anglatadiki, bir tildagi tegishli tokenlar boshqa tildagilarga to‘g‘ridan-to‘g‘ri mos kelmasligi mumkin. To‘g‘ri moslashishni ta’minlash uchun tokenizatsiya paytida ushbu farqlarni hal qilishga alohida e’tibor berilishi kerak. Ikkinchidan, turk tilida ham, o‘zbek tilida ham tokenizatsiyani murakkablashtiradigan o‘ziga xos lingvistik xususiyatlar mavjudligidir. Tokenizatorlar ushbu tilga xos murakkabliklarni to‘g‘ri hal qilish uchun mo‘ljallangan bo‘lishi kerak. Bu muammolarni hal qilish uchun turk va o‘zbek tillarining o‘ziga xos xususiyatlarini hisobga olgan holda tilga xos tokenizatsiya usullari va vositalaridan foydalanish muhim ahamiyatga ega.

Turk va o‘zbek tilidagi parallel matnlarni tokenizatsiya qilishning keng tarqalgan usullaridan biri tilga xos tokenizatsiya modellari yoki kutubxonalardan foydalanishdir. Turk va o‘zbek tillari uchun tokenizatsiya kutubxonalari mavjud bo‘lib, ulardan matnni tokenlarga aylantirish uchun foydalanish mumkin. Masalan, turk tili uchun maxsus ishlab chiqilgan tokenizatsiya va boshqa tabiiy tillarni qayta ishlash vositalarini taqdim etadigan Zemberek-NLP kutubxonasidan foydalanish mumkin. Xuddi shunday, o‘zbek tili uchun O‘zbek tili morfologik analizatoridagi UzTokenizatordan foydalanish mumkin. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi jamoasining tashabbuskorligi natijasida yaratilgan *uznatcorpara.uz*[1] sayti orqali tokenizatsiya amallarini bajarish imkoni yaratilgan bo‘lib, bu o‘zbek-turk tillari parallel korpusini yaratishda muhim va kerakli unsurlardan bo‘lib xizmat qiladi.

[1] https://uznatcorpara.uz/uz/Tokenizer



4-rasm. uznatcorpara.uz veb sahifasi interfeysi

Bundan tashqari, o‘zbek va turk tillari uchun “Python NLTK yordamida so‘z tokenizatsiyasi”[1] dasturidan foydalanish mumkin. *text-processing.com*[2] saytida tabiiy tillarni qayta ishlash API va Python NLTK demolarini topish mumkin. APIlar hozirda ochiq va bepul, lekin cheklangan.

Matnni qayta ishlash API quyidagi funksiyalarni qo‘llab-quvvatlaydi:

Stemlash va lemmatizatsiya

Hissiyot tahlili

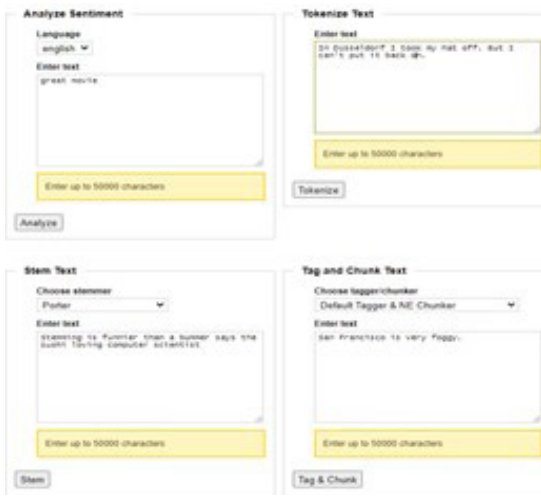
Teglash va bo‘laklarni ajratib olish

NER obyektlarini aniqlash

Tabiiy tilda matnni qayta ishlash uchun Python NLTK demolari

Hozirda 4 ta Python NLTK demolari mavjud. Yuqori chap tomonda his-tuyg‘ularni tahlil qilish mumkin, bu esa his-tuyg‘u polaritesini aniqlash uchun matn tasnifidan foydalanadi. Yuqori o‘ng tomonda turli xil so‘z tokenizatorlari qanday ishlashini ko‘rishingiz mumkin. Pastki chap tomonda 17 ta tilda matnni ajratib ko‘rish mumkin. Pastki o‘ng tomonda 22 ta nutq teggerlarining turli qismlari bilan nutq belgilarini belgilashning bir qismini, shuningdek bo‘laklarni ajratib olish va nomli obyektni tanib olish imkoniyatini bajarish mumkin.

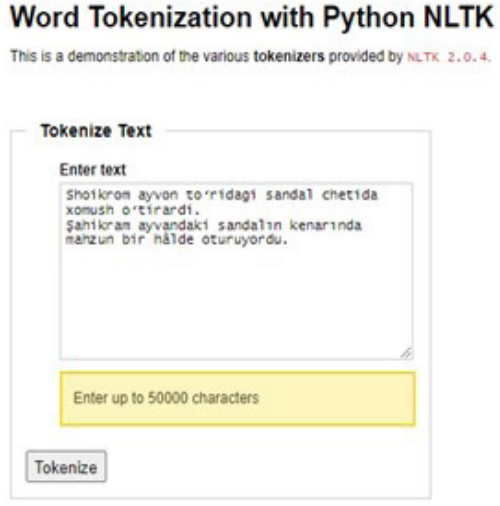
[1] http://text-processing.com/demo/tokenize/
[2] http://text-processing.com/

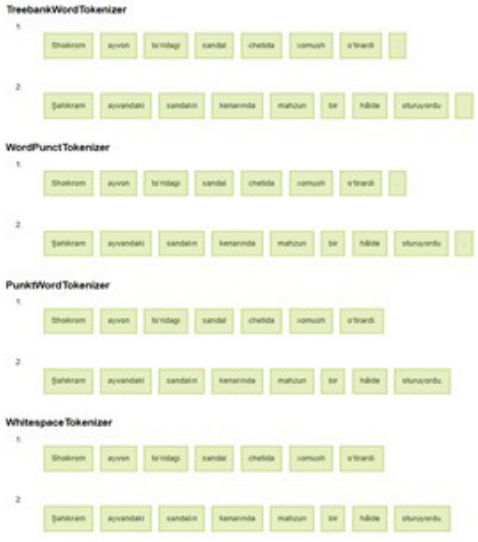


5-rasm. Tabiiy tilda matnni qayta ishlash uchun Python NLTK demolari

"Python NLTK yordamida so'z tokenizatsiyasi" dasturida matn tokenizatsiyasida matn, birinchi navbatda, PunktSentenceTokenizer yordamida jumlalarga ajratiladi . Keyin har bir jumla 4 xil so'z tokenizatorlari yordamida so'zlarga ajratiladi. Bular:

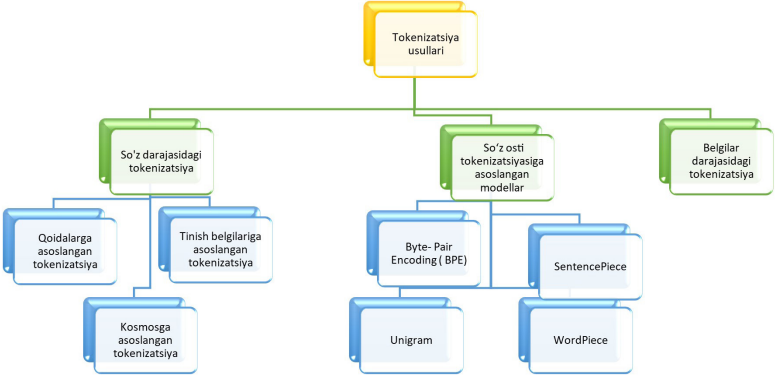
- o TreebankWordTokenizer
- o WordPunctTokenizer
- o PuncWordTokenizer
- o WhitespaceTokenizer





6-rasm. “Python NLTK yordamida so‘z tokenizatsiyasi” dasturida parallel matn tokenizatsiyasi

Bundan tashqari, neyron tarmoqlar yoki qoidaga asoslangan tizimlar asosidagi kabi oldindan o‘rgatilgan tokenizatsiya modellari turk va o‘zbek matnlarini tokenizatsiya qilish uchun ishlatilishi mumkin. Ushbu modellar katta korpuslarda o‘qitiladi va har bir tilning lingvistik xususiyatlaridan kelib chiqqan holda matnni tokenlarga samarali tarzda tokenlashtira oladi. Ba’zi ko‘p tilli tokenizatsiya modellari, masalan, so‘z osti tokenizatsiyasiga asoslangan modellar (masalan, Bayt Pair Encoding yoki SentencePiece) turk va o‘zbek matnlarini tokenizatsiya qilish uchun ishlatilishi mumkin. Ushbu modellar matnni bir nechta tillarda taqsimlanadigan so‘z osti birliklariga ajratadi, bu parallel matnlarni tokenizatsiya qilishni osonlashtiradi.



1-chizma. Tokenizatsiya jarayoni iyerarxiyasi

Turk va o‘zbek tillarida parallel matnlarni leksema qilishda har bir tilning o‘ziga xos lingvistik xususiyatlarini hisobga olish va ushbu tillarning ehtiyojlariga moslashtirilgan xos vositalar yoki modellardan foydalanish muhim ahamiyatga ega. Bundan tashqari, parallel korpusda moslashishni ta’minlash uchun tokenizatsiya jarayoni ikkala tilda ham izchil bo‘lishini ta’minlash juda muhimdir.

Xulosa qilib aytish mumkinki, turk va o‘zbek tillarida parallel matnlarni tokenizatsiya qilish moslashtirish masalalari, tilga xos murakkabliklar va shovqinli ma’lumotlarni diqqat bilan ko‘rib chiqishni talab qiladi. Tegishli texnikalar va dastlabki ishlov berish bosqichlarini qo‘llash orqali aniq va izchil tokenizatsiyaga erishish mumkin.

Foydalanilgan adabiyotlar

Aravind Rai, 2020. Webster.

Xusainova Z. Nlp: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash. O‘zbek amaliy filologiyasi istiqbollari Respublika ilmiy-amaliy konferensiyasi 2022. B-154.

Elov B., Hamroyeva Sh., Axmedova X. Methods for creating a morphological analyzer, 14th International Conference on Intellegent human Computer Interaction, IhCI 2022, 19-23 October 2022, Tashkent

Rai, A. and Borah, S. 2021. Study of various methods for tokenization. In Lecture Notes in Networks and Systems (Vol. 137). https://doi.org/10.1007/978-981- 15-6198-6 18

https://uznatcorpara.uz/uz/Tokenizer

http://text-processing.com/demo/tokenize/

http://text-processing.com/

https://www.datacamp.com/blog/what-is-tokenization

THE ISSUE OF TEXT TOKENIZATION FOR THE UZBEK-TURKISH PARALLEL CORPUS

Iqbola Xolmonova[1]

Abstract. This article provides information about tokenization, the reasons for the need for tokenization to create a corpus, programs that implement tokenization of Uzbek-Turkish parallel texts and their available options. At the same time, the problems that can be encountered in the process of text tokenization for the Uzbek-Turkish parallel corpus and their solutions are discussed. It also contains information on how to use Word Tokenization Using Python NLTK.

Key words: *token, tokenization process, parallel corpus, Python NLTK, Uzbek-Turkish parallel corpus.*

References

Aravind Rai, 2020. Webster.

Xusainova Z. Nlp: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash. O‘zbek amaliy filologiyasi istiqbollari Respublika ilmiy-amaliy konferensiyasi 2022. B-154.

Elov B., Hamroyeva Sh., Axmedova X. Methods for creating a morphological analyzer, 14th International Conference on Intellegent human Computer Interaction, IhCI 2022, 19-23 October 2022, Tashkent

Rai, A. and Borah, S. 2021. Study of various methods for tokenization. In Lecture Notes in Networks and Systems (Vol. 137). https://doi.org/10.1007/978-981- 15-6198-6 18

https://uznatcorpara.uz/uz/Tokenizer

http://text-processing.com/demo/tokenize/

http://text-processing.com/

https://www.datacamp.com/blog/what-is-tokenization

[1]*Xolmonova Iqbola Alisher qizi* – Master of degree. Alisher Navo’i Tashkent State University of Uzbek Language and Literature.
E-pochta: *iqbolabintualisher@gmail.com*
ORCID: 0009-0006-8844-6569