TABIIY TILNI QAYTA ISHLASH TIZIMLARI

Elova Dilrabo Qudratilloyevna
Alisher Navoiy nomidagi O‘zbek tili va adabiyoti universiteti
mustaqil izlanuvchisi. elovadilrabo@navoiy-uni.uz

Annotatsiya. Globalizatsiya va raqamli axborotlashtirish jarayonida o‘zbek tilining rivojida axborot texnologiyalari o‘ta muhim o‘ringa ega. Kompyuter lingvistikasining tabiiy tilni qayta ishlash (Natural Language Processing – NLP) jarayoni va tabiiy tilning kompyuter tushunadigan til (formal) shakli yaratilishi natijasida tilga oid masalalar (tahrir, tahlil, tarjima, elektron matnni ovozlashtirish, og‘zaki nutqni elektron matnga aylantirish, robot bilan muloqot qilish, yirik matnni ixcham matnga aylantirish kabi) bo‘yicha kompyuter dasturlari yaratilmoqda.

Mazkur maqolada Python dasturlash tilidagi paketlardan foydalangan holda tabiiy tilni qayta ishlash asoslari va tizimlari borasida so‘z yuritildi.

Kalit so‘zlar: tabiiy tilni qayta ishlash, NLP, NLTK, axborot texnologiyalarining lingvistik dasturlari, lingvistik tizimlar,

Abstract. In the process of globalization and digital information, information technologies play an important role in the development of the Uzbek language. In the field of computational linguistics as a result of the process of automatic processing of natural language (Natural Language Processing / NLP) and the creation of a computer-understandable language (formal) form of a natural language, many computer programs have been developed that are designed to solve linguistic problems such as editing, analysis, translation, electronic dubbing of text, transformation of oral speech to electronic text, robotic communication), such as converting large text to minitext, etc.

This article discusses the basics and systems of natural language processing using Python programming language packages.

Keywords: natural language processing, NLP, NLTK, information technology linguistic programs, linguistic systems

Jahon tilshunosligida XX asrning 50-yillaridan til va matnni avtomatik tahlil qilish muammolari bilan bog‘liq masalalar kompyuter texnologiyalari yordamida hal qilinmoqda. Natijada matnni avtomatik tushunish, mashina tarjimasi, matnni referatlash (gipermatnni asosiy mazmunni o‘zida saqlagan minimatnga keltirish, ya’ni matn hajmini qisqartirish, uni qisqa bayon holiga aylantirish), tasniflash (mavzu, uslub va janr jihatidan o‘zaro yaqin matnlarni guruhlash), matnni tahrir va tahlil qilish, matnni generatsiyalash (bir nechta tabiiy tilga oid hujjatli matnlardan formula, texnik ishlanmalar, dasturiy tizimlarni yig‘ish), matndan ma’lumotni olish, og‘zaki nutqni raqamli ma’lumotda berish va aksincha, yozma nutqni og‘zaki nutqqa aylantirish, tabiiy tilning Milliy korpusi va boshqa tur lingvistik korpuslarini yaratish, ontologik lug‘atlar bazasini shakllantirish kabi dolzarb masalalar kompyuter lingvistikasining mundarijasini egallagan va tabiiy tilni qayta

ishlash (Natural Language Processing / NLP) sohasida asosiy vazifa hamda yo‘nalishlariga aylangan [Kulkarni A., Shivananda A. 2019:5].

Kompyuter lingvistikasi, aynan, NLP har bir tilning axborot texnologiyalari makonida o‘z o‘rnini topishiga xizmat qiladi. Bugungi kunda o‘zbek tilining qo‘llanish doirasini kengaytirish, elektron lug‘atlar, tarjimon dasturlar, lingvistik dasturlar va tizimlarni, o‘zbek tili Milliy korpusi va boshqa turdagi til korpuslarini yaratish ustuvor vazifalaridan hisoblanadi. Buning uchun o‘zbek tilini qayta ishlash jarayonlari ustida ilmiy va amaliy izlanishlar olib borilmoqda va, albatta, muayyan natijalarga ham erishilmoqda. Jumladan, Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetida o‘zbek tili Milliy korpusi va ta’limiy korpus, o‘zbek nutq sintezatori yaratildi. Turkiy tillarning elektron platformasini yaratish ustida izlanishlar boshlangan. Alisher Navoiy va Zahiriddin Muhammad Bobur ijod mahsullari, shuningdek, yana bir qancha adiblar kitoblarining mobil ilovalari yaratilib, foydalanuvchilarga taqdim etildi.

Tabiiy tilni qayta ishlash (NLP) nima?

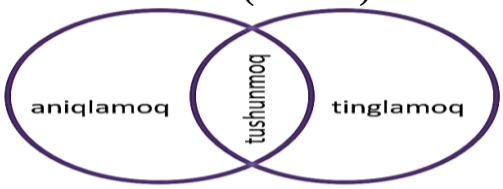
Kompyuterlar kundalik turmushning ajralmas qismi bo‘lib, jadvalli / elektron ma’lumotlar bilan ishlashda juda qulay texnik vositasi hisoblanadi. Biroq, odamlar odatda jadvallar shaklida emas, balki so‘zlar va jumlalar bilan muloqot qilishadi. O‘g‘zaki va yozma nutq matnlari xususiy bo‘lib, til umumiylik kasb etib, muayyan tuzilish (struktura)ga ega. Shu bois kompyuterlar uchun ushbu turdagi ma’lumotlarni qayta ishlash usullarini yaratish taqozo etiladi. **Tabiiy tilni qayta ishlash (NLP)dan** maqsad kompyuterlarga strukturlangan matnni tushuntirish va mazmun olishni o‘rgatishdan iborat. Tabiiy tilni qayta ishlash (NLP) sun’iy intellektning kichik sohasi bo‘lib, uning maqsadi kompyuterlar va odamlar o‘rtasidagi o‘zaro aloqalarni o‘rnanishdan iborat. Tabiiy tilni qayta ishlashga oid masalalarni yechish uchun Python dasturlash tilida yaratilgan **NLTK** kutubxonasi yuklab olinishi talab etiladi [Garousi V., Bauer S., Felderer M. 2020: 7].

NLP yo‘nalishlari

- Mashina tarjimasi (Machine Translation)
- Nutqni tanish (Speech Recognition)
- Kayfiyat / tuyg‘ularni tahlil qilish (Sentiment Analysis)
- Savollarga javob berish (Question Answering)
- Matnni referatlash (Summarization of Text)
- Chatbot (Chatbot)
- Intellektual tizimlar (Intelligent Systems)
- Matn tasniflari (Text Classifications)
- Belgilarni aniqlash (Character Recognition)
- Imlo tekshiruvi (Spell Checking)
- Spamni aniqlash (Spam Detection)
- (Matnni) Avtomatik to‘ldirish (Autocomplete)
- Nomga ega obyektni aniqlash (Named Entity Recognition)
- Taklifli terish (Predictive Typing) [Indig B., Simonyi A., Ligeti-Nagy N. 2018: 21]

Tabiiy tilni qayta ishlash (NLP) asoslari

Bugungi kunda tabiiy tilni qayta ishlash (NLP) o‘ta zarur ehtiyojga aylandi. Tabiiy tilning qayta ishlanishi natijasida raqamli texnologiyalar elektron va og‘zaki matnlarni tinglaydi, aniqlaydi va tushunadi (1-rasm).



1-rasm. NLP

Tabiiy tilning an’anaviy tilshunoslikda aniq strukturasi, til me’yorlari mukammal ravishda ishlab chiqilmaganligi natijasida formal tilni yaratishda kutilgan natijaga erishilgani yo‘q. Bu borada asosiy muammo axborotning turlicha idrok yoxud talqin qilinishi sanaladi. Masalan, quyidagi jumlani ko‘rib chiqamiz:

1-Misol:

Men teleskop bilan tepada bir odamni ko‘rdim.

Yuqorida ko‘rsatilgan jumlaning ba’zi talqinlari:

- *Tepadada bir kishi bor, men uni teleskopim bilan kuzatdim.*
- *Tepada bir kishi bor va uning teleskopi ham bor.*
- *Men tepalikdaman va teleskopimdan foydalanib, bir odamni ko‘rdim.*
- *Men tepalikdaman va teleskopga ega bo‘lgan odamni ko‘rdim.*

2-Misol:

Yaxshi ot keyin chopadi, deyishsa-da, katta maqsad yo‘lida o‘zingdagi dangasalikni o‘t

Yuqoridagi gapda “ot” shakldosh so‘zining ikki turkum (1-si ot turkumi, 2-si fe’l turkumi)ga oid shakli qo‘llangan va ularning birinchisi ***hayvon nomi***, ikkinchisi harakatni (uloqtirish; kontekstual ma’nosi ***yo‘qotish***)ni ifodalab kelmoqda. Demak, yuqoridagi misollardan ko‘rinib turibdiki, tilni qayta ishlash “deterministik” xarakterga ega emas, ya’ni tilda bir jumla yagona ma’noni ifodalab kelmaydi, bir so‘z bitta semantikada bo‘lmaydi. Chunki nutqiy vaziyat, kognitiv idrok turli interpretatsiyaning yuzaga kelishiga omil bo‘ladi. Shuningdek, tildagi aksariyat so‘zlarning ko‘p ma’no va konnotativ ma’noga egaligi, omonim va polifunksionalligi turli lingvistik muammolarni yuzaga keltiradi. Boshqacha qilib aytganda, tabiiy tilni qayta ishlash yordamida turli vaziyatlarda tilni tushunishni va gaplarni to‘g‘ri talqin qilishni biladigan **yangi aqlli tizimni yaratish** dolzarb masala hisoblanadi [Zhou M., Duan N., Liu S., Shum H.Y. 2020: 6].

Tilni qayta ishlash metodlari. Raqamli texnologiyalar dasturi va tizimlari yordamida til masalalarining aniq amaliy yechimini berish va samarali natijaga

erishishda turli usullar mavjud. Quyida shunday usullardan eng muhimlari haqida so‘z yuritildi.

Tabiiy tilni qayta ishlashda, asosan, ikki xil yondashuvga asoslaniladi:

1. Qoidalarga asoslangan NLP usuli. Mutaxassis(lar) tomonidan NLPning yuqorida berilgan yo‘nalishlari dasturiy ta’minotlari lingvistik bazalari uchun muvofiq bo‘ladigan qoidalar ishlab chiqiladi. Ammo bu jarayon ko‘p mehnat talab qilishi va qo‘lda bajarilishi bilan murakkablik hosil qiladi.

2. Tabiiy tilni stoxastik qayta ishlash usuli. Bu jarayonda mashina tomonidan katta hajmdagi ma‘lumotlardan foydalanilgan holda xulosa chiqariladi. Stoxastik NLP usulida modellarni o‘qitish uchun mashinada maxsus algoritmlardan foydalaniladi.

Qoidalarga asoslangan va statistik NLP usullarining o‘zaro taqqosi:

Qoidalarga asoslangan NLP	Stoxastik NLP
+1) moslashuvchan	+1) hisoblash oson
+2) kamchiliklarni bartaraf etish oson	+2) til o‘z-o‘zidan o‘rganiladi
+3) ko‘p tayyorgarlik talab qilinmaydi	+3) jarayon tez rivojlanadi
+4) tilni tushunish lozim	+4) keng qamrovga ega
+5) yuqori aniqlikka erishiladi	
-1) qoidalarning mukammal ishlab chiqilishiga yuqori talab qo‘yiladi	-1) katta hajmdagi ma’lumotlarni talab qiladi
-2) sekin tahlil qiladi	-2) kamchiliklarni bartaraf etish qiyin
-3) o‘rtacha qamrovga ega	-3) kontekstni tushunmaslik holatlari yuzaga keladi

Matnlar ustida ishlaydigan raqamli texnologiyalar dastur va tizimlari quyidagi lingvistik tahlillarni amalga oshirishga mo‘ljallangan bo‘lib, ular tabiiy tilni qayta ishlash komponentlarini tashkil etadi:

2-rasm. Tabiiy tilni qayta ishlash (NLP) komponentlari.

1. **Leksik tahlil.** Leksik tahlil jarayonida matnning butun qismi abzatslar, gaplar va so‘zlarga ajratiladi. Bu jarayon so‘zlarning tuzilishini aniqlash, tahlil qilish va matndagi noharfiy belgilarni aniqlab, ularni o‘chirishga erishiladi.
2. **Morfologik tahlil.** Bu jarayonda so‘zlar va so‘zshakllarni lug‘atdagi shakli (leksikon) bilan taqqoslash, so‘z asosi (lemmasi) va grammatik shakllari (formant / affiks / affiksal morfemalari) aniqlanadi hamda ularga tavsif beriladi.
3. **Sintaktik tahlil.** Sintaktik tahlil gapdagi so‘zlarni grammatik jihatdan tahlil qiladi va ular o‘rtasidagi bog‘lanish munosabatini aniqlaydi.
4. **Semantik tahlil.** Semantik tahlil so‘zlarning gapda muayyan o‘rinda qo‘llanganini e’tiborga olib, uning ma’nosini beradi va matn mazmunini tahlil qiladi. Bu jarayon lingvistik bazada so‘zlarning semantik valentliklari, aniqroq aytganda, muayyan so‘zning chap va o‘ng tomondagi semantik jihatdan maqbul birikuvchi so‘zlari berilgan baza muhim sanaladi. Shundagina “*issiq muzqaymoq*” so‘z birikmasi, “*Tog‘da mushuk o‘tlab yuribdi*” jumlasi semantik jifatdan xato ekanligi aniqlanadi.
5. **Diskurs tahlil.** Diskurs tahlilda anaforik bog‘lanishlar hisobga olinadi. Bu jarayonda muayyan gap boshida kelgan, asosan, ko‘rsatish olmoshlari o‘zidan oldingi gapdagi qaysi bo‘lakka ishora qilinayotgani aniqlanadi. Masalan, *“Salim tinim bilmaydi. U hozirda ham maktabda, ham nashriyotda ishlaydi”.* 2-gapdagi “*u*” ko‘rsatish olmoshi o‘zidan oldingi gapdagi ega (Salim)ga ishora qilmoqda.
6. **Pragmatik tahlil** gapdagi so‘zlarning umumiy aloqasi va talqinini aniqlaydi. U tilni har xil vaziyatlarda mazmunli ishlatishni keltirib chiqarish bilan shug‘ullanadi.

NLP kutubxonalari.

1. **NLTK (Natural Language Toolkit)**

NLTK Python freymvorki hisoblanib, odatda o‘quv va ilmiy jarayonlardagi masalalarni hal qilishda foydalaniladi. Uning qulay imkoniyatlari tufayli turli xildagi dasturlarni yaratish uchun NLTKdan foydalanish mumkin [Maria Razno. 2019: 9].

Vazifalari:

- Tokenizatsiya.

- So‘z turkumlarini teglash (POS tagging)
- Nomga ega obyektlarni aniqlash (NER)
- Kayfiyat / tuyg‘ularni tahlil qilish
- Chat-botlar to‘plamlarini tahlil etish

NLTK frameworkidan foydalanishning afzallik va kamchiliklari:

Afzalliklari	Kamchiliklari
1) eng ko‘p foydalaniladigan va to‘liq imkoniyatlarga ega NLP kutubxonasi	1) o‘rganish va ishlatish qiyin
2) ko‘p sonli tillar uchun foydalanish mumkin	2) so‘zning kontekstiga e’tibor berilmaydi
	3) sekin ishlaydi
	4) neyro-tarmoq modeli yo‘q

2. **spaCy** – tezkor tarzda dastur yaratishga mo‘ljallangan Pythondagi ochiq kodli tabiiy tilni qayta ishlash kutubxonasi hisoblanadi .

Funksiyalari:

- Tokenizatsiya
- So‘z turkumlarini teglashtirish (POS)
- Nomga ega obyektlarni aniqlash (NER)
- Tasniflash
- Kayfiyat / tuyg‘ularni tahlil qilish
- Tobelik tahlili
- So‘z vektorlari

Qo‘llanilishi:

- Satrlarni avtoto‘ldirish va avtotahrir
- Sharhlarni tahlil qilish
- Referatlash / umumlashtirish

spaCy frameworkidan foydalanishning afzallik va kamchiliklari:

Afzalliklari	Kamchiliklari
1) tez ishlaydi	1) moslashuvchanligi past
2) o‘rganish va foydalanish oson	
3) neyro-tarmoqlardan foydalanish mumkin	

3. **Gensim** – Python NLP frameworki bo‘lib, odatda modellashtirish va o‘xshashlikni aniqlashda foydalaniladi.

Funksiyalari:

- Yashirin semantik tahlil
- Matritsaning yashirin bo‘lmagan faktorizatsiyasi
- TF-IDF

Qo‘llanilishi:

- Hujjatlarni vektorlarga aylantirish
- Matn o‘xshashligini topish
- Matnni umumlashtirish

Gensim frameworkidan foydalanishning afzallik va kamchiliklari:

Afzalliklari	Kamchiliklari
1) qulay interfeys	1) nazorat qilinmaydigan matn modellari uchun mo`ljallangan
2) kengaytirish imkoniyati mavjudligi	2) boshqa kutubxonalarda foydalanish kerak
3) algoritmlarning joriy etilganligi	

4. **Pattern** – bu sodda sintaksisga ega bo‘lgan NLP Python frameworki hisoblanib, ilmiy va ilmiy bo‘lmagan vazifalar uchun ishonchli vositadir. Talabalar ushbu framework orqali NLPning bir qator masalalarini hal qilishlari mumkin.

Funksiyalari:

- Tokenizatsiya.
- So‘z turkumlarini teglash (POS)
- Nomga ega obyektlarni aniqlash (NER)
- Sintaktik tahlil
- Kayfiyat / tuyg‘ularni tahlil qilish

Qo‘llanilishi:

- Imlolarni tekshirish
- Qidiruv tizimini optimallashtirish
- Kayfiyatlarni tahlil qilish

Pattern frameworkidan foydalanishning afzallik va kamchiliklari:

Afzalliklari	Kamchiliklari
1) data mining	1) muayyan NLP vazifalari uchun optimallashtirilmagan

2) tarmoqli tahlil va vizualizatsiya

Data Mining – inson faoliyatining turli sohalarida zarur bo‘lgan ma’lumotlarda ilgari muhim, ahamiyatli bo‘lgan, amaliy jihatdan foydali va tushunarli talqinlarni aniqlash usullari to‘plamini belgilash uchun ishlatiladigan umumiy nom.

5. **TextBlob** – bu matnli ma’lumotlarni qayta ishlashga mo‘ljallangan Python kutubxonasi.

Funksiyalari:

- So‘z turkumlarini teglash
- Otli birikmalar ustida ishlash
- Kayfiyat / tuyg‘ularni tahlil qilish
- Matn tarjimasi
- Tasniflash
- Sintaktik tahlil
- Wordnet integratsiyasi

Qo‘llanilishi:

- Kayfiyat / tuyg‘ularni tahlil qilish
- Imloni tekshirish va xatoni bartaraf etish
- Mashina tarjimasi
- Nutqni tanish

TextBlob frameworkidan foydalanishning afzallik va kamchiliklari [Goyal P., Pandey S., Jain K. 2018: 12]

Afzalliklari	Kamchiliklari
1) foydalanish oson	1) sekin ishlaydi
2) NLTK uchun qulay interfeys	2) neyro-tarmoq modeli yo‘q
3) tarjima qilish va nutqni tushunishni ta‘minlaydi	3) integratsiyalashgan so‘z vektorlari yo‘q

Ma’lum bo‘lganidek, NLP – bu matnni qulay va foydali usulda avtomatik tahlil qilish, tushunish jarayonidir. Tabiiy tilni qayta ishlash natijasida (NLPdan foydalanib), aynan Python dasturlash tilining yuqorida imkoniyatlari ko‘rsatib o‘tilgan uning kutubxonalari yordamida, avtomatik referatlash, kompyuter tarjimasi, nomlangan obyektni aniqlash, nutq sintezatorini yaratish, tuyg‘ularni tahlil qilish, nutqni tanish va matnlar segmentatsiyasi, so‘z turkumlarini teglash, matnni tahlil qilish: tokenizatsiya, stemming, lemmatizatsiya, parsing kabi vazifalarni bajarish uchun kompyuter dasturlari va tizimlarini yaratish mumkin.

FOYDALANILGAN ADABIYOTLAR

1. Garousi V., Bauer S., Felderer M. NLP-assisted software testing: A systematic mapping of the literature. In *Information and Software Technology*. 2020. https://doi.org/10.1016/j.infsof.2020.106321

2. Goyal P., Pandey S., Jain K. SpaCy. In *Deep Learning for Natural Language Processing: Creating Neural Networks with Python*. 2018.

3. Indig B., Simonyi A., Ligeti-Nagy N. What’s wrong, python? - A visual differ and graph library for NlP in python. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*.

4. Kulkarni A., Shivananda A. Deep Learning for NLP. In *Natural Language Processing Recipes*. 2019. https://doi.org/10.1007/978-1-4842-4267-4_6

5. Lorla S. TextBlob Documentation. 2020

6. Maria Razno. (2019). Machine learning text classification model with NLP approach. *Computational Linguistics and Intelligent Systems*.

7. Morris J.X., Yoo J.Y., Qi Y. TextAttack: Lessons learned in designing Python frameworks for NLP. 2020. https://doi.org/10.18653/v1/2020.nlposs-1.18

8. Mukhopadhyay S. Advanced Data Analytics Using Python. In *Advanced Data Analytics Using Python*. 2018. https://doi.org/10.1007/978-1-4842-3450-1

9. Panchenko A., Bondarenko A., Franzek M., Hagen M., Biemann C. Categorizing comparative sentences. 2018. https://doi.org/10.18653/v1/w19-4516

10. Raj S. Natural Language Processing for Chatbots. In *Building Chatbots with Python*. 2019. https://doi.org/10.1007/978-1-4842-4096-0_2

11. Zhou M., Duan N., Liu S., Shum H.Y. Progress in Neural NLP: Modeling, Learning and Reasoning. In *Engineering*. 2020. https://doi.org/10.1016/j.eng.2019.12.014