O‘zbek tilining formal (kompyuterli) grammatikasini yaratish masalasi

O‘ZBEK TILI MATNLARIDAGI NOMUHIM SO‘ZLAR

Madatov Xabibulla Axmedovich,
Urganch davlat universiteti Axborot texnologiyalari kafedrasi mudiri, fizika-matematika fanlari nomzodi, dotsent.
habi1972@mail.ru,
Sharipov Maksud Siddikovich,
Urdu Axborot texnologiyalari kafedrasi dotsenti, texnika fanlari nomzodi.
Bekchanov Shukurla Kurbanbayevich,
UrDU Axborot texnologiyalari kafedrasi tayach doktoranti.

Annatatsiya. Ma’lumki, matnlarni tahlil qilishda uning mazmunini o‘zgartirmaydigan darajada nomuhim so‘zlarni matndan olib tashlash masalasi juda katta ahamiyatga ega. Maqolada berilgan o‘zbek tilida yozilgan matn uchun nomuhim so‘zlarni avtomatik aniqlash bilan o‘chirish, kerak bo‘lganida asl matnga qayta olish masalasi qaraladi.

Kalit so‘zlar: *nomuhim so‘zlar, deep learning, machine learning*

STOP WORDS IN UZBEK LANGUAGE TEXTS

Annotation. It is well known that in the analysis of texts it is very important to remove from the text stop words that do not change their content. The article deals with the automatic deletion of stop words for the Uzbek text and, if necessary, its return to the original text.

Key words: *stop words, deep learning, machine learning*

Matnni sinflarga ajratish yoki ma‘nosini tahlil qilish masalasi qo‘yilgan bo‘lsa biz nomuhim so‘zlarni olib tashlashimiz kerak bo‘ladi. Chunki ular biz quradigan model uchun ahamyatga ega emas. Ya’ni, ularni olib tashlash orqali model qurishni osonlashtiramiz. Modelni ishlash tezligini oshiramiz va ma’lumotlar hajmini kich-rayishiga erishamiz. Lekin til tarjimasi masalalarini yechish kerak bo‘lsa nomuhim so‘zlar ahamiyatli bo‘ladi, shuning uchun ularni o‘chirib tashlamaymiz.

Kam ma’noli ma’lumotlarga ega bo‘lgan,yoki mustaqil ma’noga ega

bo‘lmagan, yoki barcha matnlarga xos keng tarqalgan so‘zlar nomuhim so‘zlari deb ataladi.

Nomuhim so‘zlar kontseptsiyasi uzoq tarixga ega, Hans Piter Luh 1960 yilda ushbu atamani yaratgan [Luh, 1960]. Ushbu so‘zlarning ingliz tilidagi misollari: "a", "the", "of" va "not". Ushbu so‘zlar juda keng tarqalgan va odatda ba’zi masalalarni yechisda matndan so‘zlarni olib tashlash yechimga ta’sir qilmaydi[Huston, 2010].

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.*

Ko‘chamiz, degan kundan ***buyon*** *bobomda* ***qanday-dir*** *bezovtalik boshlandi. U kishining ranglari o‘zgarib,* ***biroz*** *g‘amgin* ***bo‘lib*** *qoldilar.*

Ertaga ko‘chamiz, degan kuni bobomda ***umuman*** *halovat bo‘lmadi. Buvimning aytishlaricha, tuni* ***bilan*** *bezovta* ***bo‘lib****, uxlamay chiqibdilar.*

Uydagi ***qolgan-qutgan*** *narsalarni yig‘ishtirib, tugunlarga bog‘lab, kichik qutilarga joylagach, bobom fotihaga* ***qo‘shib*** *tilovat qildilar. Bobomdagi xomushlik fotihadan so‘ng* ***bir*** *lahza hammamizni chulg‘agandek* ***bo‘ldi****.* ***Birozdan so‘ng*** *bobom aytgan Shoyim* ***tog‘a*** *mashinasida* ***yetib*** *keldi. Bobom* ***mendan boshqa*** *hammani mashinaga chiqishga taklif qildi. Yuklarni ortib bo‘lganimizdan so‘ng buvim* ***bilan*** *eng kichik ukam kabinaga joylashdi. Qolganlar mashinaning kuzoviga chiqishdi.*

Ko‘rinib turganidek yuqorida keltirilgan matnda ajratilib ko‘rsatilgan so‘zlar nomuhim so‘zlarga misol bo‘ladi. Bu so‘zlarning ko‘pchilik qismi o‘zbek tili garmatikasida olmosh, ravish, kirish so‘z va yuklmalardan iborat.

Yuqorida keltirilgan matndagi nomuhim so‘zlarni olib tashlash dasturini tuzamiz. Dasturni Python dasturlash tilida quyidagi amallar ketma-ketligida bajaramiz:

```
corpus = ['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.',
          'Ko‘chamiz, degan kundan buyon bobomda qandaydir bezovtalik
boshlandi. ',
```

```
          'U kishining ranglari o‘zgarib, biroz g‘amgin bo‘lib
qoldilar.',
          'Ertaga ko‘chamiz, degan kuni bobomda umuman halovat
bo‘lmadi.',
          'Buvimning aytishlaricha, tuni bilan bezovta bo‘lib, uxlamay
chiqibdilar.',
          'Uydagi qolgan-qutgan narsalarni yig‘ishtirib, tugunlarga
bog‘lab, kichik qutilarga joylagach, bobom fotihaga qo‘shib tilovat
qildilar.',
          'Bobomdagi xomushlik fotihadan so‘ng bir lahza hammamizni
chulg‘agandek bo‘ldi.',
          'Birozdan so‘ng bobom aytgan Shoyim tog‘a mashinasida yetib
keldi.',
          'Bobom mendan boshqa hammani mashinaga chiqishga taklif
qildi.',
          'Yuklarni ortib bo‘lganimizdan so‘ng buvim bilan eng kichik
ukam kabinaga joylashdi.',
          'Qolganlar mashinaning kuzoviga chiqishdi.']

def remove_stop_words(corpus):
    stop_words = ['O‘shanda', 'Biz', 'bilan',
'ham','buyon','qandaydir','biroz','bo‘lib','umuman','qolgan-
qutgan','qo‘shib','bir','bo‘ldi','Birozdan
so‘ng','tog‘a','yetib','mendan','boshqa','bilan']
    results = []
    for text in corpus:
        tmp = text.split(' ')
        for stop_word in stop_words:
            if stop_word in tmp:
                tmp.remove(stop_word)
        results.append(" ".join(tmp))

    return results

corpus = remove_stop_words(corpus)

print(corpus)

['uchinchi sinfda o‘qirdim.', 'oilamiz qishloqqa ko‘chadigan
bo‘ldik.', 'Ko‘chishimizdan kun oldin akam otam mollarni haydab, yangi
uyimizga ketishdi.', 'Ularga itimiz ergashdi.', 'Ko‘chamiz, degan
kundan bobomda bezovtalik boshlandi. ', 'U kishining ranglari
```

```
o‘zgarib, g‘amgin qoldilar.', 'Ertaga ko‘chamiz, degan kuni bobomda
halovat bo‘lmadi.', 'Buvimning aytishlaricha, tuni bezovta bo‘lib,
uxlamay chiqibdilar.', 'Uydagi narsalarni yig‘ishtirib, tugunlarga
bog‘lab, kichik qutilarga joylagach, bobom fotihaga tilovat
qildilar.', 'Bobomdagi xomushlik fotihadan so‘ng lahza hammamizni
chulg‘agandek bo‘ldi.', 'Birozdan so‘ng bobom aytgan Shoyim
mashinasida keldi.', 'Bobom hammani mashinaga chiqishga taklif
qildi.', 'Yuklarni ortib bo‘lganimizdan so‘ng buvim eng kichik ukam
kabinaga joylashdi.', 'Qolganlar mashinaning kuzoviga chiqishdi.']
```

Dasturda `corpus` nomli o‘zgaruvchiga matnni qiymat sifatida kiritdik. Keyin `remove_stop_words` nomli funksiya yartib, uning tarkibida nomuhim so‘zlarni olib tashlash algoritmi asosida operatorlar ketma-ketligini yozamiz. Yaratgan funksiyamizga o‘zgaruvchi orqali murojat qilib, natijani chop etamiz.

Nomuhim so‘zlar - bu har qanday tilning tarkibida mavjud, ular gapga unchalik katta ma’no qo‘shmaydi. Gapning ma’nosini yo‘qotmasdan ularni xavfsiz tarzda e’tiborsiz qoldirish mumkin. Ba’zi qidiruv tizimlari uchun bu eng keng tarqalgan, qisqa funktsiyali so‘zlar, masalan “bilan”, “ham” va boshqalar[Daowadung, 2012].

Keling, nomuhim so‘zlarini olib tashlashning ba’zi ijobiy va salbiy tomonlarini ko‘rib chiqaylik.

Ijobiy tomoni bu nomuhim so‘zlar matndan deep learning va machine learning modellarini qo‘llashdan oldin olib tashlanadi. Chunki nomuhim so‘zlar juda ko‘p uchraydi va bu so‘zlarni tasniflash yoki klasterlash uchun ishlatilishi mumkin bo‘lgan noyob ma‘lumotlarni saqlamaydi.

Nomuhim so‘zlarni olib tashlaganda ma’lumotlar to‘plamining hajmi kamayadi va modelni qo‘llashga ketgan vaqt kamayishi bilan modelning aniqligiga katta ta’sir ko‘rsatati.

Salbiy tomoni bu nomuhim so‘zlarini noto‘g‘ri tanlanganligi va olib tashlanganligi bizning matnimiz ma’nosini o‘zgartirishi mumkin. Shuning uchun biz nomuhim so‘zlarini tanlashda ehtiyot bo‘lishimiz kerak.

Masalan: "Bu film yaxshi emas."

Agar bu gapda nomuhim so‘z sifatida “emas” so‘zi olib tashlansa, gapga teskari ma’no beradi. Ya’ni, “Bu film yaxshi”. Bu esa noto‘g‘ri talqin ekanligini ko‘rsatadi[Alexandra, 2017].

O‘zbek tilining nomuhim so‘zlari [Rabbimov, 2020] maqolada ko‘rib chiqilgan va https://github.com/ilyosrabbimov/uzbek-stop-words/blob/master/uz.txt saytida ro‘yxati keltirilgan (373 ta nomuhim so‘zlar soni

ko‘rsatilgan). Shuningdek dunyo bo‘yicha ko‘pchilik tillarning homuhim so‘zlari https://www.ranks.nl/stopwords veb-sahifasida keltirilgan.

Foydalanilgan adabiyotlar:

[1] Luhn, H. P. 1960. “Key Word-in-Context Index for Technical Literature (Kwic Index).” American Documentation 11 (4): 288–95.

[2] Huston, Samuel, and W. Bruce Croft. 2010. “Evaluating Verbose Query Processing Techniques.” In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 291–98. SIGIR ’10. New York, NY, USA: ACM.

[3] P. Daowadung and Y. H. Chen. 2012. Stop word in readability assessment of thai text. In Proceedings of 2012 IEEE 12th International Conference on Advanced Learning Technologies, pages 497–499.

[4] Alexandra Schofield, Mans Magnusson, and David ˚Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 432–436.

[5] S. Matchonov, A. Shojalilov, X. G‘ulomova, Sh. Sariyev, Z. Dolimov. O‘QISH KITOBI Umumiy o‘rta ta'lim maktablarining 4- sinfi uchun darslik. TOSHKENT «YANGIYUL POLIGRAPH SERVICE» 2020 16-17 sahifa

[6] I.M. Rabbimov, S.S. Kobilov, Multi-Class Text Classification of Uzbek News Articles using Machine Learning, Journal of Physics: Conference Series 1546 (2020) 012097