

MATN TOZALASH TIZIMI AYRIM KATEGORIYALARINING TAVSIFI

Baxodirov Sanjarbek Rahmatali o'g'li

sanjarbahodirov9901@gmail.com,

ToshDO'TAU magistranti

Annotatsiya. Matnni tozalash tabiiy tilni qayta ishlash (NLP) vazifalarida muhim dastlabki ishlov berish bosqichi sanalib, bunda matn ma'lumotlarining sifati va izchilligini ta'minlash uchun maxsus belgilar, kulgichlar (emojis), raqamlar, tinish belgilari va qo'shimcha bo'shliqlar, elektron pochta manzillari, HTML teglari hamda URL (Uniform Resource Locator) manzillari olib tashlanadi. Matnni tozalash tizimlari emojilarni qayta ishlash va bo'shliqlarni olib tashlash kabi kategoriyalar bo'yicha tahlil qilish uchun matn ma'lumotlarini oldindan qayta ishlashda muhim rol o'ynaydi. Emojilar grafik belgilar bo'lib, matnni qayta ishlashni murakkablashtirishi mumkin va matnni tozalash tizimlari ko'pincha ularni o'zgartirish yoki olib tashlash vositalarini o'z ichiga oladi. Bo'shliqlar matnning o'qilishi va izchilligiga ta'sir qilishi mumkin, shuning uchun tizimlar ortiqcha bo'shliqlarni olib tashlash funksiyalarini o'z ichiga oladi. Ushbu kategoriyalar va boshqa dastlabki ishlov berish vazifalarini hal qilish orqali matnni tozalash tizimlari matn ma'lumotlarining yaxshi formatlanganligini va tahlil qilish uchun tayyor bo'lishini ta'minlaydi. Ushbu maqola matnni tozalash jarayonida ushbu elementlarni olib tashlash uchun ishlatiladigan usullar va texnikalariga qaratilgan.

Abstract. Text cleaning is a crucial preprocessing step in natural language processing (NLP) tasks. It involves removing special characters, emojis, numbers, punctuation, extra spaces, email addresses, HTML tags, and URLs (Uniform Resource Locators). Text cleaning systems are essential for preprocessing text data for categorization analysis, such as emoji processing and whitespace removal. Emojis are graphic symbols that can complicate text processing. Text sanitization systems often include tools to modify or remove them. Excess whitespace can affect the readability and consistency of text, so systems include functions to remove it. Text cleaning systems handle these categories and other preprocessing tasks to ensure that text data is well formatted and ready for analysis. This article focuses on the methods and techniques used to remove unnecessary elements during the text cleaning process.

Аннотация. Очистка текста — важнейший этап предварительной обработки в задачах обработки естественного языка (NLP). Он включает в себя удаление специальных символов, смайлик, цифр, знаков препинания, дополнительных пробелов, адресов электронной почты, HTML-тегов и URL-адресов (унифицированных указателей ресурсов). Системы очистки текста необходимы для предварительной обработки текстовых данных для анализа категоризации, например, обработки эмодзи и удаления пробелов. Эмодзи —

это графические символы, которые могут усложнить обработку текста. Системы очистки текста часто включают в себя инструменты для их изменения или удаления. Лишние пробелы могут повлиять на читаемость и согласованность текста, поэтому в системы включены функции для их удаления. Системы очистки текста обрабатывают эти категории и другие задачи предварительной обработки, чтобы гарантировать, что текстовые данные хорошо отформатированы и готовы к анализу. В этой статье основное внимание уделяется методам и приемам, используемым для удаления ненужных элементов в процессе очистки текста.

Kalit so'zlar: *matnni tozalash tizimi, matn ma'lumotlarining aniqligi, ma'lumotlarning qayta ishlash samaradorligi, regular ifoda, emojilar, bo'shliqlar.*

KIRISH

Matnni tozalash tizimi matn ma'lumotlaridan keraksiz va ahamiyatsiz ma'lumotlarni olib tashlash orqali ma'lumotlarning aniqligi va ishonchliligini ta'minlashda muhim vosita hisoblanadi. So'nggi yillarda matnlarni tozalash tizimlariga talab sezilarli darajada oshdi, chunki korxonalar va tashkilotlar ijtimoiy tarmoqlar, mijozlar sharhlari va yangiliklar kabi turli manbalardan yaratilgan katta hajmdagi ma'lumotlar bilan ishlashiga to'g'ri kelmoqda.

Matn tozalash tizimining umumiy tavsifi tozalanmagan matn ma'lumotlarini toza va tartibli formatga aylantirishga qaratilgan turli xil texnika va jarayonlarni o'z ichiga oladi. Matn tozalash jarayonidagi ba'zi bir umumiy qadamlar maxsus belgilarni, tinish belgilarini va nomuhim so'zlarni olib tashlashni, shuningdek, so'zlarni standartlashtirish va ma'lumotlarning o'lchamini kamaytirish uchun stemming va lemmatizatsiyani o'z ichiga oladi. Bundan tashqari, matn ma'lumotlarining sifatini yanada oshirish uchun ko'pincha imlo tekshiruv, tokenizatsiya va nutqning bir qismini teglash qo'llanadi.

Adabiyotlar sharhi

Mashinani o'rganish (machine learning) va data science (ma'lumotlar bilan ishlash) dagi yutuqlar matnni tozalash tizimlarining rivojlanishini yanada kuchaytirdi. Chunki sifatli matn ma'lumotlariga bo'lgan ehtiyoj NLP modellarini o'qitish va sinovdan o'tkazish uchun tobora muhim ahamiyat kasb etdi. Tadqiqotchilar matnni tozalash uchun turli xil yondashuvlarni o'rganishni boshladilar, masalan, nomuhim so'zlarni olib tashlash, HTML teglarini, URL manzillarini olib tashlash, imlo xatolarini tuzatish, formatlarni standartlashtirish, maxsus belgilar bilan ishlash va boshqalar.

Jahon kompyuter tilshunosligi NLP sohasida matnni tozalash borasida bir qator olimlar ilmiy izlanishlar olib borgan. B.Roepke, G.Spiller, J.A.Yeung, N.V.Otten, Pankaj, Vidvatek, 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. D.Jurgens va I.Augenstein matnni tozalash usullarini o'rgangan va matn sifatini yaxshilash va uni NLP vazifalari uchun ko'proq

moslashtirish uchun dasturiy vositalarni ishlab chiqqan. J.Xu va Ch.Miao hissiyotlarni tahlil qilish va ma'lumotlar olish kabi turli ilovalar uchun matnni tozalash dasturi ustida ishlar olib borishgan.

Tsyurix universiteti qoshidagi Hisoblash tilshunosligi universiteti va Qatar Hisoblash tadqiqot instituti kabi tadqiqot muassasalari ham matn tozalash va dastlabki ishlov berish texnikasi bo'yicha tadqiqotlar olib bordilar. Bundan tashqari, Allen sun'iy intellekt (AI) instituti AllenNLP kutubxonasi kabi matnni tozalash va qayta ishlash uchun vositalar va resurslarni ishlab chiqdi.

Bular matnni tozalash dasturlari sohasida o'z hissalarini qo'shgan tadqiqotchilarning bir nechta misollari. Ko'plab boshqa tadqiqotchilar va tashkilotlar ham matn sifatini va turli ilovalar uchun qulaylikni oshirish uchun ushbu sohada faol ishlashmoqda.

Matnni tozalash tizimlari nomuhim so'zlar, maxsus belgilar, kulgichlar(emojis), raqamlar, tinish belgilari va qo'shimcha bo'shliqlar, elektron pochta manzillari, HTML teglari hamda URL(Uniform Resource Locator) kabi elementlarni olib tashlash orqali matn ma'lumotlarining sifatini oshiradi. Ushbu tizimlar tabiiy tilni qayta ishlash, ma'lumotlarni qidirish, **data mining** dasturlarida keng qo'llanadi. Quyida matnni tozalash turlarining har bir turiga to'xtalib o'tiladi.

1.1 Kulgichlar (emojis)ni olib tashlash: Emojilar – yozma muloqotda his-tuyg'ularni yoki g'oyalarni ifodalash uchun ishlatiladigan kichik raqamli tasvirlar yoki piktogrammalar. Emojilar matnga chuqur emotsiyalarni qo'shishi mumkin bo'lsa-da, ularning biz xohlamagan va noo'rin holatlari ham mavjud.

Matndan kulgichlarni (emojis) olib tashlash uchun **Regular expression**dan foydalaniladi, bu orqali barcha emoji belgilarni bo'sh qator bilan qidirish yoki almashtirish mumkin. Shu bilan bir qatorda, emojilarni matndan avtomatik ravishda olib tashlash mumkin bo'lgan maxsus vositalar va dasturlar mavjud. Matndan bunday ortiqcha kulgichlar (emojis)ni olib tashlashni Python dasturlash tili asosida ko'rib chiqamiz [Yeung, 2022]

```
import re
text = u"O'shanda uchinchi sinfda o'qirdim 😊 😊. Biz oilamiz bilan boshqa
qishloqqa ko'chadigan bo'ldik 😞 😞."
print(text) # with emoji
def remove_emojis(text):
    regex_pattern = re.compile(pattern = "["
        u"\U0001F600-\U0001F64F"
        u"\U0001F300-\U0001F5FF"
        u"\U0001F680-\U0001F6FF"
        u"\U0001F1E0-\U0001F1FF"
        u"\U00002702-\U000027B0"
```

```
u"\U000024C2-\U0001F251"  
    "+", flags = re.UNICODE)  
return regex_pattern.sub(r", text)  
print(remove_emojis(text))
```

#Output

#with emoji: O'shanda uchinchi sinfdan o'qirdim 😊 😊. Biz oilamiz bilan boshqa qishloqqa ko'chadigan bo'ldik 😊 😊.

#remove emoji: O'shanda uchinchi sinfdan o'qirdim. Biz oilamiz bilan boshqa qishloqqa ko'chadigan bo'ldik.

Matndan kulgichlar (emojis)ni olib tashlashni bir qancha sabablari mavjud:

1. *Aniqlik.* Emojilar ba'zan noto'g'ri talqin qilinishi yoki tushunarsiz bo'lishi mumkin, ayniqsa professional yoki rasmiy muloqotda. Emojilarni olib tashlash ma'lumotning aniq tushunilishini ta'minlashga yordam beradi.

2. *Foydalanish imkoniyati.* Ko'zida nuqsoni bo'lgan ba'zi odamlar emojilarni talqin qilishda qiynalishi mumkin, natijada ma'lumotlar tushunilishida qiyinchiliklar yuzaga keladi.

3. *Formatlash.* Emojilar har doim ham turli platformalar yoki qurilmalarda to'g'ri ko'rsatilmaligi mumkin, bu esa tashqi ko'rinishdagi nomuvofiqlikka olib keladi. Emojilarni olib tashlash matnning izchil va professional formatlanishini saqlashga yordam beradi.

4. *Tillar o'rtasidagi nomutanosibliklar.* Kulgichlar har doim ham universal ma'noga ega bo'lavermaydi va ular turli madaniyatlarga mansub shaxslar tomonidan noto'g'ri talqin qilinishi mumkin. Emojilarni olib tashlash til to'siqlari tufayli yuzaga keladigan tushunmovchiliklardan qochishga yordam beradi.

Umuman olganda, emojilarni matndan olib tashlash muloqotning konteksti va auditoriyasiga bog'liq. Emojilar ba'zi holatda muloqotni emotsionalligini oshirishi mumkin bo'lsa-da, boshqa hollarda ularni olib tashlash ma'qul bo'ladi.

1.2 Qo'shimcha bo'shliqlarni olib tashlash. Bo'shliq matn ichidagi so'zlar va belgilarni ajratish uchun ishlatiladigan belgini (masalan, yangi qator, yorliq (tab) va ortiqcha bo'shliqlar) anglatadi. Ba'zi hollarda, bir nechta sabablarga ko'ra matndan bo'shliqlarni olib tashlashga to'g'ri kelishi mumkin:

1. *O'qishni osonlashtirish uchun.* Ortiqcha bo'shliqlarni olib tashlash, matnni o'qishni va tushunishni osonlashtiradi.

2. *Joyni tejash.* Cheklangan joy bilan ishlanganda, masalan, dasturlash yoki veb-sayt dizaynida, keraksiz bo'shliqlarni olib tashlash fayl hajmini kamaytirish va ishlashni optimallashtirishga yordam beradi.

3. *Matnni normallashtirish.* Matnni qayta ishlash ilovalarida bo'shliqni olib tashlash matn formatini standartlashtirishga va izchillikni yaxshilashga yordam beradi.

Matndan ortiqcha bo'shliqlarning olib tashlashning turli xil usullari mavjud:

1. *Trim funksiyasi yordamida*. Ko'pgina dasturlash tillari va matn muharrirlarida “Trim” funksiyasi mavjud bo'lib, u satrdan oldingi va keyingi bo'shliqlarni olib tashlash uchun ishlatilishi mumkin.

2. *Regular ifoda*. “Regular expressions” (re) matn ichidagi bo'shliqlarni qidirish va almashtirish uchun ishlatilishi mumkin. Masalan, “\s+” bo'sh joy belgilarining istalgan ketma-ketligini moslashtirish uchun ishlatilishi mumkin.

3. *Qo'lda tahrirlash (manual editing)*. Ba'zi hollarda keraksiz bo'shliqlarni olib tashlash uchun matnni qo'lda tahrirlash, ayniqsa, kichik hajmdagi matnlar uchun eng samarali yechim bo'lishi mumkin.

Bundan tashqari, matndagi bo'shliqlarni olib tashlash uchun foydalanilayotgan dasturlash tiliga qarab turli usullardan foydalanish tavsiya etiladi. Ba'zi keng tarqalgan usullar qatoriga Pythonda **strip**, **rstrip** yoki **lstrip** usullari, Javada esa, **trim** usuli yoki JavaScriptdagi **replace** usuli kiradi.

Python dasturlash tili yordami bilan turli xil usullarda ortiqcha bo'shliqlarni olib tashlashni ko'rib chiqamiz[Zia, 2024]

1-usul

```
import re
s = "    O'zbek tiliga davlat    tili maqomini berilishi, o'zbek
tiliga    bo'lgan e'tiborni yanada yuksaltirdi. "
s = re.sub(" +", " ", s)
print(s)

#Output: O'zbek tiliga davlat tili maqomini berilishi, o'zbek tiliga bo'lgan
e'tiborni yanada yuksaltirdi.
```

2-usul

```
s = "    O'zbek tiliga davlat    tili maqomini berilishi, o'zbek
tiliga    bo'lgan e'tiborni yanada yuksaltirdi"
s = s.replace(" ", "")
print(s)

#Output:
O'zbektiligadavlattilimaqominiberilishi,o'zbektiligabo'lgane'tiborniyanadayuksalti
rdi
```

2- usulda qo'llangan **replace()** usuli 1-usulga qaraganda samaradorligi past. Ya'ni bu usulda qo'shtirnoq orasiga qanday buyruq kiritgan bo'lsak faqat shu buyruq bilan ishlaydi.

3-usul

```
s = "    O'zbek tiliga davlat    tili maqomini berilishi, o'zbek  
tiliga  bo'lgan e'tiborni yanada yuksaltirdi  "  
s = s.lstrip()  
print(s)  
  
#Output: O'zbek tiliga davlat    tili maqomini berilishi, o'zbek  
tiliga  bo'lgan e'tiborni yanada yuksaltirdi
```

Bu usulda so'ralgan tomondan bo'shliqlar olib tashlanadi, ya'ni **lstrip()** – **leftstrip(chap)** va **rstrip()** – **rightstrip(o'ng)** tomondagi bo'shliqlarni tozalaydi. Bu usulda gap o'rtasidagi bo'shliqlar tozalanmaydi.

Umuman olganda, matndan bo'shliqlarni olib tashlash o'qishni yaxshilash, joyni tejash va turli ilovalar uchun matn formatini normallashtirishga yordam beradi.

XULOSA

Matnni tozalash tizimlarini bir nechta toifalarga bo'lish mumkin, jumladan emojilarni qayta ishlash va matndan bo'shliqlarni olib tashlash. Ushbu tizimlar matnli ma'lumotlarni yanada izchil, aniq va oson tahlil qilish uchun qayta ishlashga mo'ljallangan.

Matnni tozalash tizimlarida keng tarqalgan vazifalardan biri kulgichlar (emojis) bilan ishlashdir. Emojilar - bu matnli xabarlar va ijtimoiy media postlarida his-tuyg'ular, g'oyalar va tushunchalarni ifodalash uchun ishlatiladigan grafik belgilar. Ular matnga ortiqcha ma'no qo'shishlari mumkin, lekin ular matnni qayta ishlash algoritmlari uchun qiyinchiliklarni ham yaratishi mumkin. Matnni tozalash tizimlari ko'pincha emojilarni qayta ishlash vositalarini, masalan, ularni matn ekvivalentlariga aylantirish yoki ularni butunlay olib tashlashni o'z ichiga oladi.

Matnni tozalashda yana bir muhim vazifa matndan oq bo'shliqlarni olib tashlashdir. Bo'shliqlar, yorliqlar va qatorlar orasidagi bo'shliqlar matn ma'lumotlarining o'qilishi va izchilligiga ta'sir qilishi mumkin. Matnni tozalash tizimlari odatda matndan keraksiz oq bo'shliqlarni olib tashlash vositalarini o'z ichiga oladi, bu esa matnning yaxshi formatlanganligini va tahlil qilish oson bo'lishini ta'minlaydi.

Emojilarni qayta ishlash va bo'sh joylarni olib tashlashdan tashqari, matnni tozalash tizimlari matnli ma'lumotlarni oldindan qayta ishlash uchun boshqa xususiyatlarni ham o'z ichiga olishi mumkin. Masalan, bu tizimlar imlo va grammatik xatolarni tuzatish, maxsus belgilarni, HTML teglari, URL manzillarini olib tashlash va matn formatini normallashtirish vositalarini o'z ichiga olishi mumkin. Matn ma'lumotlarini tozalash va qayta ishlash orqali ushbu tizimlar matnni tahlil qilish algoritmlarining aniqligi va samaradorligini oshirishi mumkin.

Umuman olganda, matnni tozalash tizimlari matn ma'lumotlarini tahlil qilish uchun tayyorlashda hal qiluvchi rol o'ynaydi. Emojilar bilan ishlash, bo'shliqlarni olib tashlash va boshqa qayta ishlash vazifalarini bajarish orqali ushbu tizimlar matn ma'lumotlarining toza, izchil va keyingi tahlilga tayyor bo'lishini ta'minlaydi. Matnni tozalash tizimlari tadqiqotchilar va ma'lumotlar bo'yicha olimlardan tortib sotuvchilar va ijtimoiy media tahlilchilarigacha bo'lgan matnli ma'lumotlar bilan ishlaydigan har bir kishi uchun muhim vositadir.

Foydalanilgan adabiyotlar:

1. Chandra R. Text Cleaning in Python: Effective Data Cleaning Tutorial, 2023. [Text Cleaning in Python: Effective Data Cleaning Tutorial – Kanaries](#)
2. Otten N.V. Top 20 Essential Text Cleaning Techniques [Practical How to Guide in Python] / Data Science, 2023, Sep 18. <https://spotintelligence.com/2023/09/18/top-20-essential-text-cleaning-techniques-practical-how-to-guide-in-python>
3. Pankaj. How to Remove Spaces from a String in Python / DigitalOcean, 2022, December 12. How To Remove Spaces from a String In Python | DigitalOcean
4. Roepke B. How to Clean Text Like a Boss for NLP in Python / Data Knows All, 2024, Mart 3. <https://dataknowsall.com/blog/textcleaning>
5. Williams A. Remove emails from Text in Python (methods+examples) / PYtutorial, 2023, Sep 26. <https://pytutorial.com/remove-emails-from-text-in-python-methods-examples>
6. Yeung J.A. How to remove emojis from string in Python /The Web Dev, 2022, April 16. <https://thewebdev.info/2022/04/16/how-to-remove-emojis-from-a-string-in-python>
7. Zia A. How to Remove Whitespaces in Python Strings? /jQuery-AZ /2024. <https://www.jquery-az.com/how-to-remove-whitespaces-in-strings-of-python>