# MATN TOZALASH TIZIMI DASTURLARINING UMUMIY TAVSIFI. AFZALLIKLARI VA KAMCHILIKLARI

**Baxodirov Sanjarbek Rahmatali o‘g‘li**
ToshDO‘TAU, kompyuter lingvistikasi mutaxassisligi magistranti.
E-mail: sanjarbahodirov9901@gmail.com

**Annotatsiya:** Matnni tozalash dasturi matn ma’lumotlarini samarali tozalash va tartibga solish uchun mo‘ljallangan kuchli vositadir. Murakkab algoritmlari va interfeysi bilan u keraksiz belgilarni olib tashlash, ortiqcha bo‘shliqlarni yo‘qatish, matnni formatlash, takroriy jumlalarni bartaraf etishga qodir. Ushbu dasturlar katta hajmdagi matn ma’lumotlarini osonlik bilan boshqarishi mumkin. Bu marketing, tadqiqot va ma’lumotlarni tahlil qilish kabi turli sohalarda foydalanish uchun juda qulay hisoblanadi. Matnni tozalash dasturi ko‘p tarmoqli vosita bo‘lib, uni shaxsiy imtiyozlar va talablarga mos ravishda sozlash mumkin. U ommaviy ishlov berish, moslashtirilgan filtrlar va tozalangan matn ma’lumotlarini turli formatlarda saqlash va jo‘natish funksiyalarini o‘z ichiga olgan bir qator xususiyatlarni taklif etadi. Matnni tozalash dasturi tizimining qulay dizayni va mustahkam funksionalligi matnni qayta ishlash jarayonini soddalashtiradi va har bir foydalanuvchiga qiynalmasdan foydalanish imkoniyatini yaratadi. Bundan tashqari, ushbu dasturlarni avtomatlashtirish orqali foydalanuvchilarning qimmatli vaqtlari tejaladi hamda hujjatlarida izchillik va aniqlik ta’minlanadi.

**Kalit so‘zlar:** vaqtni tejash, avtomatlashtirish, html teglari, URL manzillari, agressiv tozalash, ma’lumotlarni yo‘qalishi, ommaviy ishlov berish.

**Abstract:** This is a powerful tool for effectively cleaning and organizing text data. It can remove unnecessary markers and eliminate excessive gaps, format the text, and remove repetitive sentences. These tools can easily manage large amounts of text data. Marketing is very useful in various fields such as research and data analysis. Text cleaning software is a versatile tool that can be customized to personal preferences and requirements. It offers public use, customized filters, and functions for storing and sending cleaned and formatted text data in various formats. The user-friendly design and robust functionality of the Text Refinement Program simplifies the process of revising text and provides every user with an effortless experience. In addition, automating these programs saves users valuable time and ensures precision and clarity in their documents.

**Keywords:** time saving, automation, html tags, URLs, aggressive cleaning, data loss, batch processing.

**Аннотация:** Это мощный инструмент для эффективной очистки и организации текстовых данных. Он умеет удалять ненужные маркеры и устранять лишние пробелы, форматировать текст и удалять повторяющиеся предложения. Эти инструменты могут легко управлять большими объемами текстовых данных. Маркетинг очень полезен в различных областях, таких как исследования и анализ данных. Программное обеспечение для очистки текста — это универсальный инструмент, который можно настроить в соответствии с личными предпочтениями и требованиями. Он предлагает общедоступное использование, настраиваемые фильтры и функции для хранения и отправки очищенных и отформатированных текстовых данных в различных форматах. Удобный дизайн и

надежная функциональность программы улучшения текста упрощают процесс редактирования текста и предоставляют каждому пользователю удобство работы. Кроме того, автоматизация этих программ экономит пользователям драгоценное время и обеспечивает точность и ясность их документов.

**Ключевые слова:** экономия времени, автоматизация, html-теги, URL-адреса, агрессивная очистка, потеря данных, пакетная обработка.

## KIRISH

Ma’lumotlarni tahlil qilish va tabiiy tilni qayta ishlashning jadal rivojlanayotgan davrda matn ma’lumotlarining sifati, undan olinayotgan tushunchalarning aniqligiga bo‘lgan talab oshib bormoqda. Matn tozalash dasturlari keraksiz belgilar, nomuhim so‘zlarni olib tashlash va nomuvofiqliklarni bartaraf etish orqali matn ma’lumotlarini tahlil qilish uchun tayyorlaydi. Ushbu dasturlar matnni qayta ishlash vazifalarining sezilarli darajada oshirish mumkin bo‘lgan qator afzalliklarni taqdim etadi. Biroq, har qanday texnologiya singari, ushbu dasturlar ham ma’lum cheklovlar va foydalanuvchilar bilishi kerak bo‘lgan kamchiliklarga ega.

Matnni tozalash dasturining keng tarqalgan turlaridan biri imlo tekshiruvi bo‘lib, u berilgan matndagi imlo xatolarni aniqlash va tuzatish uchun mo‘ljallangan.Yana bir turi grammatik tekshiruv bo‘lib, foydalanuvchilarga yozishdagi grammatik xatolarni aniqlash va tuzatishga yordam beradi. Boshqa matnni tozalash dasturlari takroriy jumlalar, so‘zlarni o‘chirish, matnni formatlash yoki ahamiyatsiz va ortiqcha ma’lumotlarni aniqlash va o‘chirishga qaratilgan bo‘lishi mumkin.

Matnni tozalash dasturlarining ishlatishning asosiy afzalliklaridan biri shundaki, ular matnli ma’lumotlarning umumiy sifatini yaxshilashga yordam beradi, ularni aniqroq, izchil va tahlil qilishni osonlashtiradi. Ushbu dasturlar, shuningdek, foydalanuvchilarning matndagi xatolarni aniqlash va tuzatish jarayonini avtomatlashtirish orqali ularning vaqt va qo‘l mehnatini tejash imkonini beradi. Biroq,

matn tozalash dasturlari ham ba’zi kamchiliklarga ega. Ushbu dasturlar har doim ham berilgan matndagi barcha xatolarni, ayniqsa, o‘zga tildagi so‘zlar, lahjalar qatnashganda kutilgan natijani bermasligi mumkin. Bundan tashqari, ba’zi matnni tozalash dasturlari aniqlash va tuzatish mumkin bo‘lgan xatolar turlari bo‘yicha cheklovlarga ega bo‘lishi mumkin.

## ASOSIY QISM

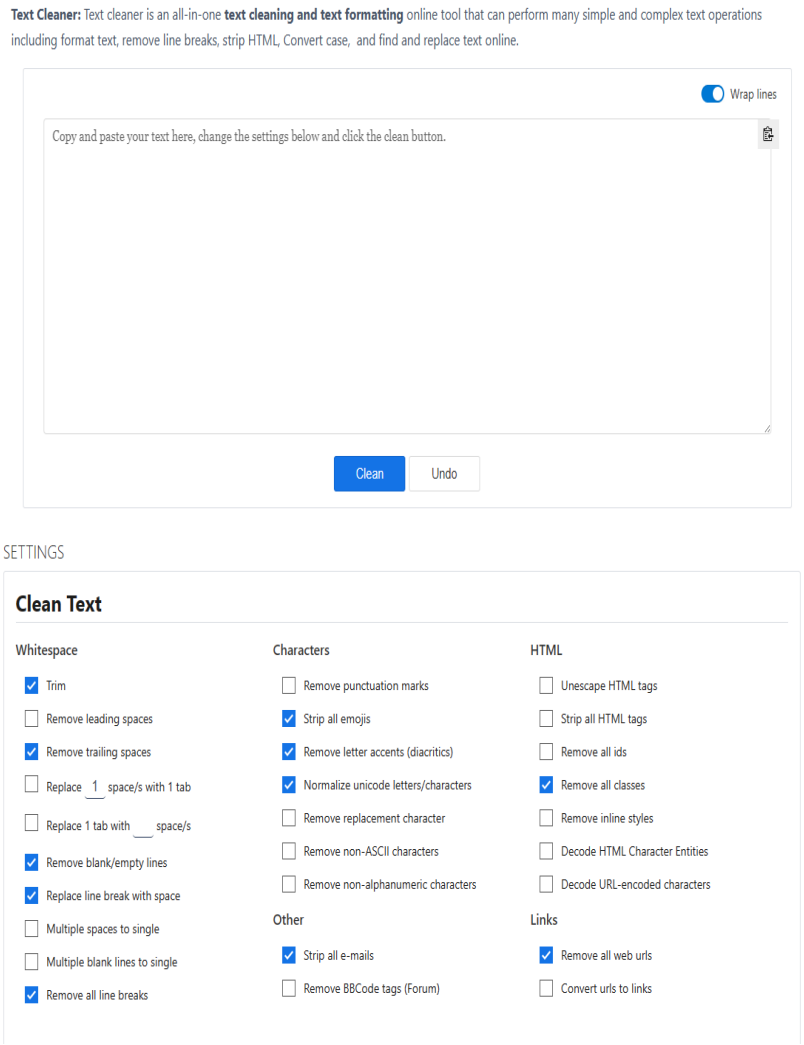
Matnni tozalash dasturlari katta hajmdagi tuzilmagan ma’lumotlar bilan ishlaydigan har bir inson uchun muhim vosita sanaladi. Ushbu dasturlar matnli ma’lumotlarni tahlil qilish, qayta ishlash va tozalash uchun mo‘ljallangan. Matnni tozalash dasturlarining bir necha turlari mavjud bo‘lib, ularning har biri o‘ziga xos xususiyat va imkoniyatlarga ega.

**1.1 Text cleaner: Text Formatter, Formatting online.** Text cleaner yoki Clean text – bu juda murakkab matnlarga ishlov berish jarayonida bajarishi mumkin bo‘lgan matnni tozalash va formatlash uchun mo‘ljallangan onlayn vosita.

Ushbu dastur ortiqcha bo‘shliqlar va keraksiz belgilarni olib tashlashi mumkin. Shuningdek, u harflarni hajmini o‘zgartirishi, tipogrfiya qo‘shtirnoqlarini o‘zgartirishi, takroriy satrlar,paragraflar va so‘zlarni o‘chirishi, qalin va kursivli Unicode harflarini oddiy harflarga aylantirishi, tinish belgilari va orasidagi masofani tuzatishi, harf urg‘ularini olib tashlashi, belgilarning kodlarini dekodlashi, HTML teglari va URL larni olib tashlashi mumkin. Bu dastur matnni onlayn formatlashda juda moslashuvchan sanalib, shaxsiy ehtiyojdan kelib chiqib

sozlamalarni o‘zgartirish mumkin. Ushbu vositani eng katta afzalliklaridan biri, shaxsiy konfiguratsiyangiz avtomatik ravishda brauzeringizda saqlanadi, shuning uchun keying safar yana tashrif buyurganingizda sozlamalarni to‘liq o‘zgartirishingiz shart emas.

Ushbu yordamchi dasturning asosiy maqsadi, formatlangan matnni formatsizlantirish va matn protsessorlari, veb sahifalar, PDF-fayllar, mijozlar qisqacha ma’lumotlari va elektron pochtalardan to‘g‘ridan-to‘g‘ri ko‘chirilgan matnlarda ko‘pincha mavjud bo‘lgan barcha ma’nosiz belgilarni olib tashlashdir. Ushbu veb-ilova tadqiqot, ishlab chiqish yoki tijorat maqsadlarida har qanday jismoniy shaxs, kompaniya, ofis yoki tashkilot uchun bepul foydalanish mumkin. Mazkur veb ilovani 1-rasmda ko‘rishingiz mumkin:

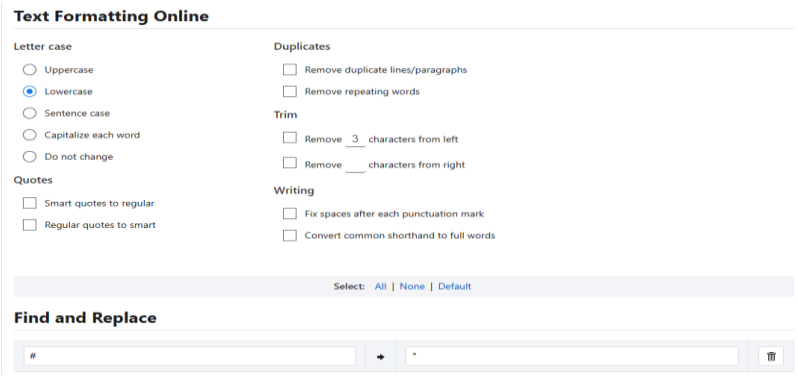



**1-rasm. Veb ilovaning ko‘rinishi**

Qanday ishlatiladi?

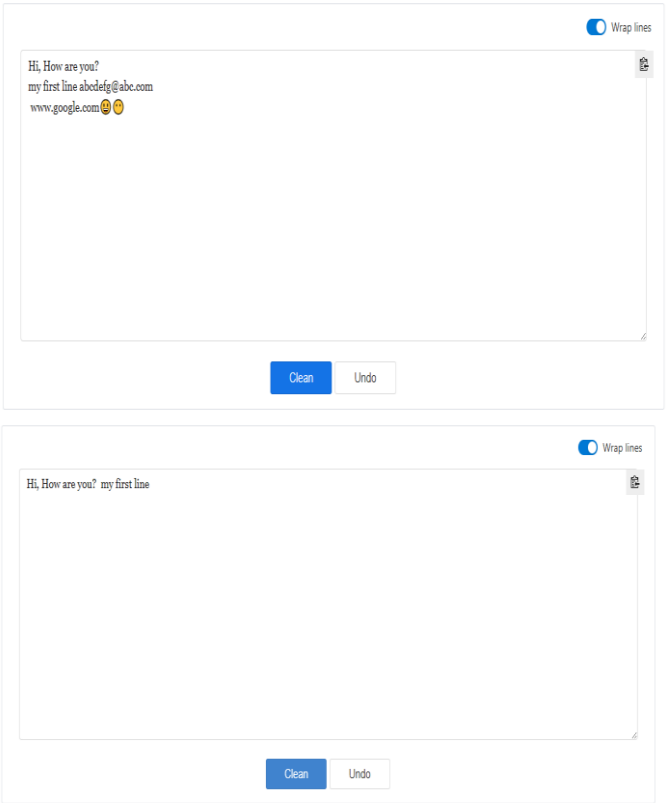
Matnni kiritish maydoniga nusxa ko‘chirasiz va joylashtirasiz, katakchalarni belgilash yoki belgini olib tashlash orqali 1-rasmdagi sozlamalarni sozlash va tozalash (clean) tugmasini bosasiz. Matningizni tozalangan versiyasi natijalar oynasida paydo bo‘lishi kerak. Agarda siz o‘ylagan natija chiqmasa, xavotir olish shart emas, siz orqaga qaytib, kiritish yorlig‘ini bosishingiz mumkin. Bekor qilish(Undo) tugmasi yordamida tozalashda xatolik yuz bersa, qayta ortga qaytishingiz va matnni oldingi ko‘rinishga keltira olasiz. Bundan tashqari, yuqorida

aytilganidek matn formatlarini o‘zgartirish, takroriy jumlalarni xoxishimizga qarab o‘zgartirsak bo‘ladi. Quyida 2-rasmda ko‘rishingiz mumkin.




**2-rasm. Matn formatlash.**

“Find and Replace” funksiyasi yordamida esa, birinchi jadvalga qidirayotgan belgi yoki so‘zni kiritishimiz va uni o‘zimizga kerakli belgiga alishtirishimiz mumkin. Bu orqali katta hajmli matnlarda yuzaga kelgan xatoliklarni bir soniyada hal etsak bo‘ladi. Quyidagi 3-rasmda **“strip all emojis”, “strip al e-mails”, “remove all web urls”, “replace line breaks with space”** katakchalarini belgilash orqali qisqa ko‘chirilgan matnni tozalab ko‘ramiz. [Zia A, 2024]




**3-rasm. Matn tozalash jarayoni**

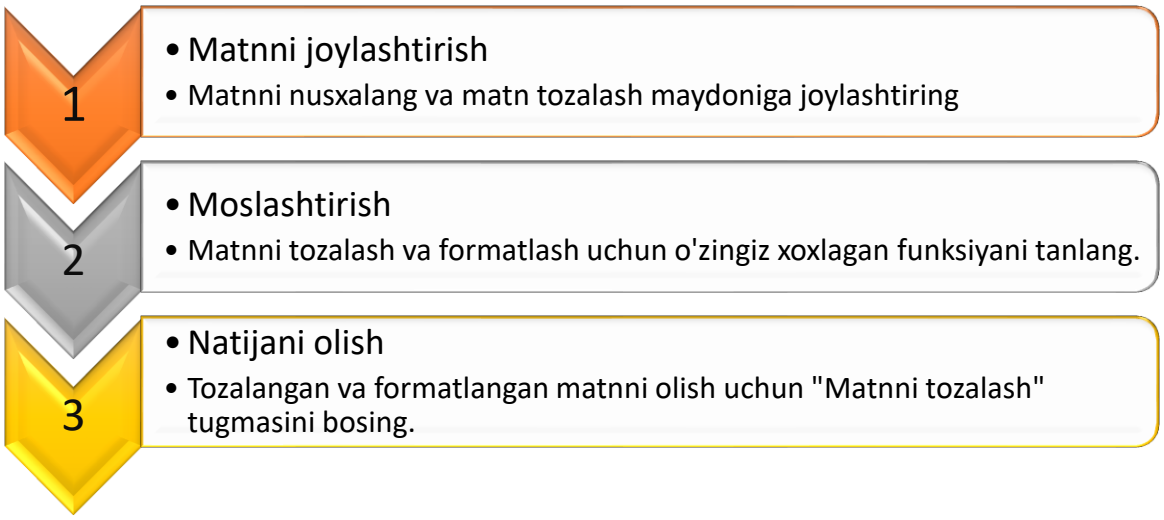
Mazkur dasturni ishlatish davomida hech qanday ishingizga halal beradigan reklamalar uchramaydi, tizimni ishlashini qiyinlashtirmaydi, veb-brauzerdan to‘g‘ridan-to‘g‘ri foydalanishingiz mumkin, hech qanday dasturni o‘rnatish shart emas. Matnni bir necha soniya ichida tozalab beradi.

Asosiy kamchiliklaridan biri, matnni faqat nusxa ko‘chirish asosida kiritish mumkin xolos, fayl(docx,pdf) shunga o‘xshash ko‘rinishlarda joylashtirib bo‘lmaydi. Bu dastur 2021-yilda tuzib chiqilgan va so‘ngi o‘zgartirish 2022-yilda kiritilgan.

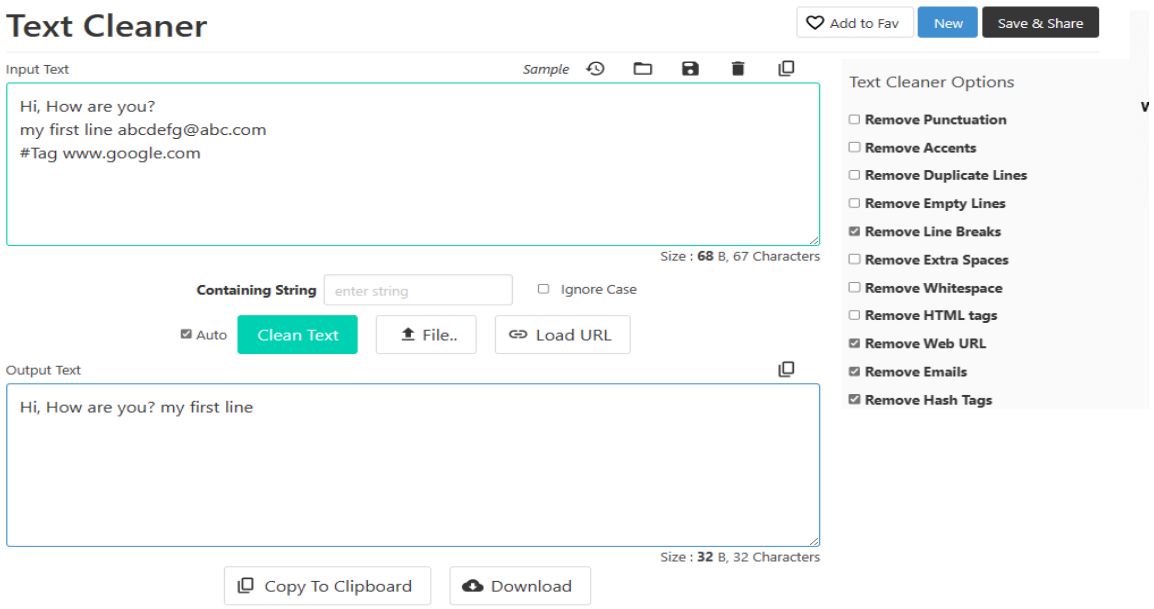
Dastur manzili: https://ravana69.github.io/textcleaner

**1.2 Code beautify. Text Cleaner.** CodeBeautify’s Text Cleaner – bu matnni tez, oson tozalash va formatlashda yordam beradigan veb ilova. Ushbu dastur matndan tinish belgilari, harf urg‘ulari, takroriy qatorlar, bo‘sh qatorlar, qo‘shimcha bo‘shliqlar, HTML teglari, veb URL manzillari, elektron pochta xabarlari va hash teglarini olib tashlashda yordam beradi. Mazkur ilova matnga ommaviy ishlov berishda yoki ma’lum bir yo‘nalishda ishlov berishda (masalan, faqat URL manzilini tozalash) qulay hisoblanadi. Shuningdek, eng katta afzalliklaridan biri matnni fayl ko‘rinishida hamda nusxa ko‘chirish asosida amalga oshirsa bo‘ladi.

Matnni tozalash vositasidan qanday foydalanish kerak:




Quyidagi 4-rasmda matn tozalash jarayonini ko‘rishingiz mumkin:




**4-rasm. Matn tozalash jarayoni**

“Text cleaner” matn formatlashni soddalashtiradi, vaqtni tejab, izchillikni taminlasa-da, ba’zi bir kamchiliklarni ham kuzatish mumkin, Misol uchun, matnga asoslangan fayllar bilan

cheklangan, murakkab formatlashda ishlamasligi mumkin va internetga ulanishni talab qiladi. Bundan tashqari, ilovadan foydalanish davomida turli xil reklamalar ish faoliyatiga ta’sir o‘tkazadi [Williams.2024]

Mazkur dastur 2024-yilda tuzib chiqilgan.

Bog‘lanish uchun: support@codebeautify.org

Manzil: https://codebeautify.org/text-cleaner

**XULOSA**

Xulosa qilib aytadigan bo‘lsak, matnni tozalash tizimlari turli tahliliy jarayonlarda qo‘llaniladigan ma’lumotlarning aniqligi va ishonchliligini ta’minlashda hal qiluvchi rol o‘ynaydi. Ushbu maqolada ko‘rib chiqilgan dasturlar, masalan, emojilar, tinish belgilari, HTML tehlari, URL manzillari, e-mail va boshqa elementlarni olib tashlash, stemming va lemmatizatsiya va imlo tekshiruvi matnli ma’lumotlar sifatini yaxshilashga yordam beradigan matnni tozalash tizimlarining muhim tarkibiy qismlaridir. Matnni tozalash tizimlaridan foydalanishning asosiy afzalliklaridan biri keraksiz va ahamiyatsiz belgilarni olib tashlash orqali ma’lumotlarni tahlil qilish samaradorligini oshirish qobiliyatidir. Bu yaxshi natijalarga va aniqroq tushunchalarga olib keladi. Bundan tashqari, matnni tozalash tizimlari matn ma’lumotlarining umumiy o‘qilishi va tushunarliligini yaxshilashga yordam beradi, bu esa foydalanuvchilarga ma’lumotni tahlil qilishni osonlashtiradi.

Ammo shuni ta’kidlash kerakki, matnni tozalash tizimlari ham kamchiliklardan xoli emas. Kamchiliklarning ba’zilari tozalash jarayonida ma’lumotlarning potentsial yo‘qolishi, shuningdek, natijalarning to‘g‘riligini ta’minlash uchun muayyan holatlarda qo‘lda aralashuv zarurligini o‘z ichiga oladi.

.

**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. 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. 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. https://dataknowsall.com/blog/textcleaning
5. Williams A. Remove emails from Text in Python (methods+examples) / PYtutorial, 2023. 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. 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