PYTHON INSTRUMENTLARI BILAN KATTA MA'LUMOTLARNI QAYTA ISHLASH

Jalolov Tursunbek Sadriddinovich
Osiyo xalqaro universiteti assistenti
E-mail: ts_jalolov@oxu.uz

ANNOTATSIYA

Ushbu maqolada katta ma'lumotlarni qayta ishlash uchun ishlatiladigan mashhur va ishlatish uchun qulay Python dasturlash tili va uning asosiy ma'lumotlarni qayta ishlash kutubxonalari tasvirlangan. Python xususiyatlari uni ma'lumotlarni tahlil qilish uchun ideal qiladi, uni o'rganish oson, ishonchli, o'qilishi mumkin, kengaytirilishi mumkin, boy kutubxonalar to'plami, boshqa tillar bilan integratsiya, faol hamjamiyat va qo'llab-quvvatlash tizimi mavjud.

Kalit so'zlar: ma'lumotlarni qayta ishlash, katta ma'lumotlar, Python, pandas, numpy.

2014-yildan boshlab jahonning yetakchi universitetlari asosiy e'tiborni Big Dataga qaratib, amaliy muhandislik va IT mutaxassisliklarini o'rgatmoqda. Keyin Microsoft, IBM, Oracle, EMC kabi IT korporatsiyalari, so'ngra Google, Apple, Facebook va Amazon kabilar kompaniyalar ma'lumotlarni yig'ish va tahlil qilishga qo'shildi. Bugungi kunda katta ma'lumotlardan sanoatning barcha sohalaridagi yirik kompaniyalar, shuningdek, davlat idoralari foydalanmoqda.

Python ma'lumotlarni samarali tahlil qilish uchun ko'plab kutubxonalar va muharrirlarni taqdim etadi. Python - bu Youtube, Google va boshqalardagi ma'lumotlar olimlari tomonidan ishlatiladigan eng tez rivojlanayotgan til. Tez tahlil qilish uchun ishlatiladi.

Xom ma'lumotlarning qiymati yo'q. Katta ma'lumotlarni o'zgartirish va tahlil qilish instrumentlari va dasturlash tillarini talab qiladi. Jarayonni avtomatlashtirish va vaqtni tejashga yordam beradigan ko'plab maxsus vositalar va dasturlash tillari mavjud.Katta hajmdagi ma'lumotlarni qayta ishlash uchun bir nechta dasturlash tillari ishlab chiqilgan va ular Paython dastur tilidagi Pandas kutubxonasi (ma'lumotlarni tahlil qilish uchun) va uning atrofida butun kutubxonalar va ramkalar ekotizimlari ishlab chiqilgan.

Katta ma'lumotlarni fanda qayta ishlash uchun oson, lekin juda murakkab matematik operatsiyalarni bajara oladigan umumiy, va moslashuvchan dastur tilni talab qiladi. Python bunday talablar uchun eng mos keladi, chunki u umumiy hisoblash

va ilmiy hisoblash uchun til sifatida o‘zini isbotladi. Bundan tashqari, u turli xil dasturlash talablariga qaratilgan ko‘plab kutubxonalarga yangi qo‘shimchalar bilan doimiy ravishda yangilanadi.

Ma’lumotlarni tahlil qilish va natijalarni vizualizatsiya qilishda Pythonni R, MATLAB, SAS, Stata va boshqalar bilan bir qatorda ko‘plab dasturlash tillari va vositalari bilan solishtirish mumkin, masalan, Python uchun ilg‘or kutubxonalar (asosan, Pandas) nisbatan yaqinda qaraganda ma’lumotlarni manipulyatsiya qilish muammolarini hal qilishda uni jiddiy raqobatchiga aylantirdi.

Ma’lumotlar bilan ishlash uchun bir nechta maxsus Python kutubxonalari yaratilgan:

NUMPY (raqamli tahlil va yaratish uchun)

MATPLOTLIB (ma’lumotlarni vizualizatsiya qilish uchun)

SCIPY (ilmiy hisoblash uchun)

SEABORN (ma’lumotlarni vizualizatsiya qilish uchun)

TENSORFLOW (chuqur o‘rganishda qo‘llaniladi)

SCIKIT-LEARN (mashinalarni o‘rganishda qo‘llaniladi)

Umuman olganda, Python ma’lumotlarni tahlil qilish uchun juda yaxshi.U ma’lumotlarni yig‘ish va qayta ishlashni avtomatlashtirish muammolarini hal qilish va ishda tahlil qilishning yangi yondashuvlarini amalga oshirish uchun ishlatilishi mumkin, masalan, neyron tarmoqlarni o‘qitishdan foydalangan holda muammolarni hal qilidi.

Pandasda siz uchta tuzilmadagi ma’lumotlar bilan ishlashingiz mumkin:

— ketma-ketliklar (Series) — bir o‘lchovli ma’lumotlar massivlari;

— ramkalar (Data Frames) — bir necha bir o‘lchovli massivlarni ikki o‘lchovli, ya’ni tanish satr va ustunlar jadvaliga birlashtirish. Ushbu format ko‘pincha tahlilchilar tomonidan qo‘llaniladi;

— panellar (Panels) — bir nechta ramkalarning uch o‘lchamli tuzilishi.

Kutubxona ma’lumotlar bilan ishlaydigan har bir kishi, ayniqsa tahlilchilar uchun foydali bo‘ladi. Pandas-dan foydalanib, siz jadvallarni guruhlashingiz, ma’lumotlarni tozalashingiz va o‘zgartirishingiz, parametrlarni hisoblashingiz va tanlov qilishingiz mumkin. Pandas kutubxonasi tuzilgan ma’lumotlar bilan ishlashni osonlashtiradigan va tezlashtiradigan funktsiyalarni taqdim etadi. Bu Pythonni ma’lumotlarni tahlil qilish uchun kuchli vosita qiladigan asosiy komponentlardan biridir.

Pandas CSV, TSV yoki SQL ma’lumotlar bazasi fayllaridan ma’lumotlarni o‘qiydi va DataFrame deb nomlangan qatorlar va ustunlar bilan Python ob’ektini yaratadi. Pandalarning asosiy ob’ekti DataFrame, satr va ustunlar bilan belgilangan ikki o‘lchovli jadvaldir. DataFrame statistik dasturiy ta’minotdagi jadvalga juda o‘xshaydi, masalan, Excel yoki SPSS.

Pandalardan foydalanish:

-Ma’lumotlar ramkalarini indekslash, tahrirlash, nomini o‘zgartirish, saralash, birlashtirish;

-Ma’lumotlar ramkasidagi ustunlarni yangilash, qo‘shish, o‘chirish;

-Yo‘qolgan fayllarni tiklash, etishmayotgan ma’lumotlarni yoki NANni tahrirlash;

Chiziqli diagramma yoki diagramma yaratish NumPy eng asosiy Python paketlaridan biri, massivlar bilan ishlash uchun umumiy maqsadli paketdir. U yuqori unumli ko‘p o‘lchovli massiv ob’ektlari va massiv vositalarini ta’minlaydi.NumPy umumiy ko‘p o‘lchovli ma’lumotlar uchun samarali konteynerdir. Asosiy NumPy ob’ekti bir xil ko‘p o‘lchovli massivdir. Bu natural sonlar to‘plami bilan indekslangan bir turdagi ma’lumotlarning elementlari yoki raqamlari jadvalidir. NumPy da o‘lchamlar o‘qlar deb ataladi va o‘qlar soni daraja deb ataladi. NumPy massiv sinfi array bo‘lib, massiv deb ham ataladi. NumPy bir xil turdagi ma’lumotlar qiymatlarini saqlaydigan massivlarni boshqarish uchun ishlatiladi. NumPy massivlarda matematikani bajarish va ularni vektorlashtirishni osonlashtiradi. Bu ish faoliyatini sezilarli darajada yaxshilaydi va shuning uchun ijro vaqtini tezlashtiradi.

Matplotlib - Ma’lumotlar vizualizatsiyasi sizga ma’lumotlaringizni vizual tarzda ko‘rsatish, ularni an’anaviy formatga qaraganda batafsilroq tahlil qilish va boshqalarga osonroq taqdim etish imkonini beradi. Matplotlib bu maqsad uchun eng yaxshi va eng mashhur Python kutubxonasidir. Foydalanish unchalik oson emas, lekin siz oddiy chiziqli diagrammalar va scatter diagrammalar uchun eng keng tarqalgan 4-5 kod blokidan foydalanib, ularni juda tez yaratishni o‘rganishingiz mumkin.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI: (REFERENCES)

1. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. — СПб.: Питер, 2017. — 336 с.: ил. — (Серия «Библиотека программиста»).
2. Ын Анналин, Су Кеннет. Теоретический минимум по Big Data. Всё, что нужно знать о больших данных. — СПб.: Питер, 2019. — 208 с.: ил. — (Серия «Библиотека программиста»).
3. Ikromova, S. (2023). FORMATION OF IDEOLOGICAL IMMUNITY TO DESTRUCTIVE INFORMATION IN TEENAGERS. Modern Science and Research, 2(5), 1009-1014.
4. Ikromova, S. (2023). CONCEPT OF IDEOLOGY AND FORMATION OF IDEOLOGICAL IMMUNITY IN YOUTH STUDENTS. Modern Science and Research, 2(6), 1223-1226.
5. Мейсман Арно. Основы Data Science и Big Data. Python и наука о данных. 2016. — 322 с. Подробнее: https://www.labirint.ru/authors/181786/