Pythonning umumiy ko'rinishi va taqqoslanishi ma'lumotlarni qazib olish va katta ma'lumotlarni tahlil qilish uchun kutubxonalar

Sa'dullayeva Sabina Rizamat qizi
sadullayeva9030@gmail.com
O'zbekiston Milliy universitetining Jizzax filiali

Annotatasiya. *Python-ning mashhurligi ayniqsa ortib bormoqda ma'lumotlar fanlari sohasida. Natijada, mavjud foydalanish mumkin bo'lgan bepul kutubxonalar soni ortib bormoqda.Ushbu sharhning maqsadi tavsiflash va taqqoslashdir turli ma'lumotlarni qazib olish va katta ma'lumotlarni tahlil qilish xususiyatlari Pythondagi kutubxonalar. Hozirda qog'oz bilan shug'ullanmaydi mavzu va barcha ushbu kutubxonalarning ijobiy va salbiy tomonlarini tavsiflash.Bu yerda biz 20 dan ortiq kutubxonalarni ko'rib chiqamiz va ularni oltita guruhga ajratamiz: asosiy kutubxonalar, ma'lumotlarni tayyorlash, ma'lumotlar vizualizatsiya, mashinani o'rganish, chuqur o'rganish va katta ma'lumotlar.*

Muayyan kutubxonaning funktsional imkoniyatlaridan tashqari, muhim omillar Taqqoslash uchun rivojlanayotgan hissa qo'shuvchilar soni va kutubxonani saqlash va jamiyatning hajmi. Kattaroq jamoalar osongina topish uchun katta imkoniyatlarni anglatadi muayyan muammoni hal qilish. Hozir tavsiya qilamiz: ma'lumotlarni tayyorlash uchun pandalar; Matplotlib, dengizda tug'ilgan yoki Plotli ma'lumotlarni vizualizatsiya qilish uchun; scikit-learn mashinaga o'tish; Chuqur o'rganish uchun TensorFlow, Keras va PyTorch; va Katta ma'lumotlar uchun Hadoop Streaming va PySpark.

Kalit so'zlar. ma'lumotlar ilmi, python, ma'lumotlarni ishlab chiqish, mashina o'quv kutubxonasi, katta ma'lumotlarni tahlil qilish.

Kirish

Ma'lumotni qazib olish (DM) ma'lumotlarni tayyorlash bilan shug'ullanadi turli

ma'lumotlar manbalaridan (masalan, ma'lumotlar bazalari, matnli fayllar, oqimlar), shuningdek, turli xildan foydalangan holda ma'lumotlarni modellashtirish texnikalar, inson xohlagan maqsadga qarab erishish (masalan, tasniflash, klasterlash, regressiya, assotsiatsiya qoidasi kon va boshqalar). DM mashinani o'rganishdan foydalanadi (ML) dan yangi bilimlarni kashf qilish usullari mavjud ma'lumotlar. Hozirgi kunda DM asosan ko'rib chiqiladi ma'lumotlar fanining kengroq doirasi doirasida, bu ham statistik ma'lumotlar, katta ma'lumotlar texnikasi va ma'lumotlarni o'z ichiga oladi vizualizatsiya. Ma'lumotlarni tayyorlash jarayonning muhim bosqichidir ma'lumotlarni tahlil qilish va u ma'lumotlarni oldindan qayta ishlash va ma'lumotlarni o'z ichiga oladi manipulyatsiya (ba'zan janjal deb ham ataladi). Oldindan ishlov berish tozalash, integratsiya, o'zgartirishga qaratilgan va dastlabki xom ma'lumotlarni kamaytirish, shunday qilib, ular bo'lishi mumkin ma'lumotlarni tahlil qilish uchun foydalanish mumkin, Wrangling esa o'zgartiradi oldindan ishlangan ma'lumotlar to'plamini ma'lumotlar formatiga osongina kiritish mumkin ma'lumotlarni modellashtirish algoritmlari bilan boshqariladi.Python-dan ma'lumotlar fanida foydalanish mavjud ayniqsa sohasida misli ko'rilmagan darajaga yetdi bepul mavjud vositalar va kutubxonalar. da chop etilgan so'rovnomada May 2018 nufuzli KDNuggets portali tomonidan [1], ostida toifasi “Eng yaxshi tahlillar, ma’lumotlar fanlari, mashinalar Learning Tools” ma’lumotlariga ko‘ra, Python’dan 65,2% foydalaniladi. taxminan 2000 ishtirokchi, 52,7% uchun RapidMiner va R uchun 48,5%, uning ikkita asosiy raqobatchisi. In amaliy nuqtai nazardan, so'nggi uch yil ichida Python bor ma'lumotlar uchun tanlangan dasturlash tiliga aylanadi ilmiy hamjamiyat, R bilan ikkinchi tanlov.Python-ning mashhurligi, ehtimol, uning nisbatan qulayligidan kelib chiqadi foydalanish (hatto kompyuter bo'lmagan olimlar uchun ham), ulkan ekotizim ma'lumotlarning har bir jihati uchun bir qator kutubxonalardan iborat fan va uning NumPy va SciPy o'ramlari orqali bog'liqligi ko'p sonli ilmiy ishlarning tez tadbiq etilishi C va Fortran tillarida yozilgan algoritmlar. 2014 yildan oldingi ishimizda biz bir umumiy DM uchun bepul mavjud vositalarni taqqoslash [2]. vaqt, Python asosidagi vositalar hali etarlicha etuk emas edi,R, RapidMiner, Weka va Knime

bo'lsa eng mashhur vositalarning oldingi o'rinlari. Buning aksincha, maqsadi bu ish umumiy ko'rinish va taqqoslashni ta'minlashdir ma'lumotlar fanlari uchun Python-ga asoslangan turli xil kutubxonalar. Xususan, biz oltita kutubxona guruhiga e'tibor qaratamiz: Python yadro, ma'lumotlarni tayyorlash, ma'lumotlarni vizuallashtirish, mashinani o'rganish, chuqur o'rganish va katta ma'lumotlar. Biz kutubxonalarni baholaymiz ularning batafsil tahlili asosidagi ahamiyati imkoniyatlar, hissa qo'shuvchilar soni va jamiyat hajmi. Chuqur o'rganish juda yaqinda rivojlanayotganligi sababli ma'lumotlar ilmi, lekin allaqachon barqaror va o'sib borayotgan vositalar bilan Python-da qo'llab-quvvatlash, biz ushbu kutubxonalarni ham o'z ichiga olamiz.

Kutubxonalarni ko’rinishi va solishtirish

A. Asosiy kutubxonalar

Python-dagi ko'plab DM va ML vazifalari tezkor va NumPy bilan samarali raqamli va vektorlashtirilgan hisoblash [3] va SciPy [4] kutubxonalari. Bulardan ko'plab funktsiyalar kutubxonalar aslida Netlib [5] atrofidagi o‘ramlar, xavfsiz va algoritmlarning mustahkam ilmiy tatbiq etilishi. Asosiy NumPy va SciPy ning afzalligi ularning qobiliyatidir samarali vektorlashtirilgan hisoblashni amalga oshirish va n o'lchovli massivlar orqali translyatsiya qilish.

Ushbu sohada Python-dan foydalanishning yana bir afzalligi shundaki uchinchi tomon kodini ulash nisbatan oson ekanligi Python tarjimoni. Ehtimol, eng ko'p ishlatiladigan Buning uchun DM kutubxonasi Cython [6]. CythonPython ustiga qurilgan til bo'lib, u ham qo'llab-quvvatlaydi C funktsiyalarini chaqirish va C tipidagi o'zgaruvchilarga ega bo'lish va sinflar. Cython-dan foydalanish ba'zi muhim qismlarni yaratishi mumkin kodni bir necha barobar tezroq.

Yuqorida aytib o'tilgan uchta kutubxonaning barchasi barqaror kodga ega va doimiy ravishda saqlash va rivojlantirishda. 1-jadval kutubxonalarning "obro'si" haqida foydali ma'lumotlarni ko'rsatadi GitHub, versiyani boshqarish uchun veb-hosting xizmati [7], yulduzlar sonini foydalanib, vilkalar, hissa va kutubxona omboridagi faoliyat. Faoliyat orqali ko'rsatiladi hissa qo'shgan mualliflar soni va soni oxirgi oyda majburiyatlarni bajaradi.

B. Ma'lumotlarni tayyorlash

Chunki ma'lumotlar fani sohasidagi hamma narsa asoslanadi ma'lumotlar, ma'lumotlarni tayyorlash kutubxonalariga ehtiyoj bor. Hozirda bu sohada eng yaxshi va eng ko'p ishlatiladigan Python kutubxonasi pandalardir [8]. pandalar kiritish/chiqarish uchun keng imkoniyatlarga ega Excel, csv, Python/NumPy, HTML, SQL kabi ma'lumotlar formatlari va boshqalar. Bundan tashqari, pandalar kuchli so'rovlarga ega imkoniyatlar, statistik hisob-kitoblar va asosiy vizualizatsiya. U boy hujjatlarga ega, ammo biroz chalkash sintaksis, Bu ko'pincha uning eng muhim kamchiligi sifatida ta'kidlanadi.

Ushbu sohadagi har bir boshqa kutubxona juda katta muammolarga ega pandalarga qaraganda. PyTables [9] va h5py [10] faqat HDF5 ni qabul qiladi ma'lumotlar turi, bu umumiy foydalanish uchun katta cheklovdir.Yana bir nechta shunga o'xshash kutubxonalar mavjud (masalan, Tabel [11]),ammo hozircha ularning hech biri pandalar bilan raqobatlasha olmaydi.

C. Ma'lumotlarni vizuallashtirish

3-jadvalda ma'lumotlarni vizualizatsiya qilishning taqqoslanishi ko'rsatilgan kutubxonalar. Plotly [12] standartning ko'p qismini qo'llab-quvvatlaydi DM va MLda qo'llaniladigan uchastkalar. dengizda tug'ilgan [13] bir necha bor imkoniyatlari Plotly dan kamroq, Matplotlib [1] esa bir nechtasiga ega dengizda tug'ilganlarga qaraganda kamroq. O'rtasida farqlar mavjud bo'lsa-da bu uchta kutubxona, ularning barchasi asosiy syujetga ega qobiliyatlar. Bokeh [7] va ggplot [16] eng kamiga ega imkoniyatlari va eng kam foydalaniladigan kutubxonalardir.

Matplotlib - bu Python ilovasi MATLAB-ga o'xshash syujetlar va past darajada yozilgan, a bilan moslashtirish uchun juda ko'p imkoniyatlar. Uning sintaksisi biroz bo'lishi mumkin dastlab chalkash, lekin uning asosiy tushunchalarini o'zlashtirgandan so'ng, deyarli har qanday grafikni chizish oson. dengiz tug'ilgan Matplotlib-ning tepasida va undan foydalanish va o'rganish osonroq Matplotlibga qaraganda yangi boshlanuvchilar. Foydalanish osonroq bo'lsa-da, ichida juda ko'p ehtiyojga ega bo'lgan ba'zi murakkab grafiklarning holatlari moslashtirish, u dengizda tug'ilgan bo'lishi mumkin amalga oshirib bo'lmaydigan variant.

Plotly ma'lumotlardagi eng kuchli kutubxonaga o'xshaydi vizualizatsiya maydoni. Uning asosiy kamchiligi nisbatan intuitivlikdir sintaksis, yangi boshlanuvchilar uchun o'rganishni qiyinlashtiradi. Biroq, kamchilik juda boy hujjatlar bilan qoplanadi ko‘plab misollar keltirish mumkin. Plotlyni birlashtirish mumkingrafiklarni Dash yordamida veb-sahifalarga kiriting [2]. Bokeh uchun mo'ljallangan interaktiv syujetlarni veb-sahifalarga integratsiyalash, bu erda foydalanuvchi ma'lumotlarni o'zi o'rganishi mumkin. ggplot bu Python-ga tegishli

R ning chizmachilik usulini amalga oshirish. Uning chegaralanganligi bor maqsadida hujjatlar va qurbonliklar xususiylashtirish oddiy va tushunarli kodga ega bo'ling.

D. Mashinani o'rganish

scikit-learn [3] eng mashhur Python kutubxonasidir mashina o'rganish. Undan tashqari mlxtend [4], a faqat bir nechta asosiy o'z ichiga olgan yangi va kichik kutubxona algoritmlar va asosan yozilgan Shogun [5] C++, lekin uning hammasi uchun Python o'rami mavjud funksionallik. Shogun mlxtendga qaraganda ko'proq algoritmlarga ega,lekin scikit-learn dan ancha kam. Faqat bir nechtasi bor Shogun amalga oshirgan va o'rgangan algoritmlar qilmaydi, buni jadvalda ko'rish mumkin. Bundan tashqari, a bor mlpy deb nomlangan kutubxona [6], qaysi jadvalda keltirilgan emas 2.uning yo'qligi sababi ham xuddi shunday kichik kutubxona

mlxtendga to'g'ri keladi, lekin u yaxshi ma'lum algoritmga ega emas boshqa kutubxonalarda yo'qligini amalga oshirdi.

scikit-learn soni bo'yicha ustunlikka ega jadvaldagi ko'pgina toifalarda amalga oshirilgan algoritmlar.

Xulosa.

Shogunning boshqa kutubxonalardan ustunligi sonida turli xil daraxtlarni amalga oshiradigan algoritmlar. mlxtend kichik kutubxona bo'lsa-da, u yagona kutubxona amalga oshirilgan assotsiatsiya qoidasi algoritmlari va stacking bilan ansambl o'rganish. Ushbu algoritmlarning etishmasligi bo'lishi mumkin.

scikit-learn va Shogun tomonidan katta e'tiborsizlik deb hisoblangan.Xuddi shu narsa induktiv qoidalarni o'rganuvchilar uchun ham amal qiladi, to'liq Bayesian

tarmoq, aylanma o'rmon va loyqa c-ma'nolarni klasterlash, ro'yxatdagi kutubxonalarning birortasida amalga oshirilmagan.

Foydalanilgan adabiyotlar

3. Nizomiddin N. et al. TA'LIMDA DASTURLASH JARAYONINI BAHOLASHGA ASOSLANGAN AVTOMATLASHTIRILGAN TIZIMNI TADBIQ ETISH //International Journal of Contemporary Scientific and Technical Research. – 2023. – С. 24-28.

4. Choryorqulov G. H., & Qosimov NS (2023) //ELEKTRON JADVAL MODELINING TAVSIFLANISHI. PEDAGOGS Jurnali. – Т. 30. – №. 3. – С. 67-73.

5. Чорркулов Г., Норматов Н., Мамараимов А. Роль анализа текстовых связей в электронных документах в информационной безопасности //Информатика и инженерные технологии. – 2023. – Т. 1. – №. 1. – С. 67-71.

6. Норматов Н., Мамараимов А. Ta'lim tizimida baholash tizimini avtomatlashtirishni joriy etish jarayonlari va foydalanish metodlari //Информатика и инженерные технологии. – 2023. – Т. 1. – №. 2. – С. 356-359.

7. Мамараимов А., Чорёркулов Г., Норматов Н. Tanib olish modullarini dasturiy amalga oshirish //Информатика и инженерные технологии. – 2023. – Т. 1. – №. 2. – С. 38-44.

8. Sanoqul o'g'li Q. N. et al. ELEKTRON HUJJAT ALMASHINUVINI AVTOMATLASHTIRISH MODELINI ANALITIK TAHLILI //Лучшие интеллектуальные исследования. – 2023. – Т. 10. – №. 5. – С. 89-100.

9. Choryorqulov G., Qobilova S. KOMPYUTER TARMOQ TEXNOLOGIYASIDA SUN'IY INTELLEKTNI QO 'LLASH //International Journal of scientific and Applied Research. – 2024. – Т. 1. – №. 3. – С. 145-151.

10. Rizamat qizi-talaba S. S. MOLIYAVIY KO'RSATKICHLARNING KORRELYATSION TAHLILI VA SUN'IY INTELLEKTGA ASOSLANGAN NARXLARNING O'ZGARISH TIZIMI.