تکلیف 0 مقدمه ای بر علم داده

TA(s):کیانوش عرشی مدرسان: دکتر بهرک، دکتر یعقوب زاده

مهلت: سه شنبه 15اسفند
ساعت 23:59

معرفی

در این تکلیف، ما قصد داریم به خراش دادن وب و تجزیه و تحلیل داده های مقدماتی بپردازیم. این تکلیف یک تمرین عملی خواهد بود که به شما کمک می کند با فرآیند استخراج داده ها از وب سایت ها و انجام تحلیل های آماری اولیه آشنا شوید.

یک دفترچه ناقص برای شما در نظر گرفته شده است که شما را در طول انجام تکلیف راهنمایی می کند.

وظایف

.1راه اندازی محیط: کتابخانه های مورد نیاز مانند سوپ زیبا، سلنیوم، پانداها، نومپی، matplotlibو seabornرا نصب کنید.

2. Web Scraping:یک اسکریپت برای خراش دادن داده های تراکنش از Etherscan.ioبنویسید. از سلنیوم برای تعامل با وب سایت و از سوپ زیبا برای تجزیه محتوای HTMLاستفاده کنید.
.3نمونه برداری و تجزیه و تحلیل داده ها: پس از جمع آوری داده ها، نمونه ای از مجموعه داده ایجاد کنید. آمار نمونه (میانگین و انحراف معیار) را با آمار جمعیت مقایسه کنید.

جمع آوری داده ها (اتراسکن)

در این بخش از وب اسکرپینگ برای جمع‌آوری داده‌های تراکنش از بلاک چین اتریوم با استفاده از کاوشگر بلوک اترسکن استفاده می‌کنیم . هدف ما جمع آوری تراکنش ها از 10بلاک آخر در اتریوم است. برای انجام این کار، از تکنیک‌های خراش دادن وب برای استخراج داده‌های تراکنش از وب‌سایت Etherscanاستفاده می‌کنیم. URLمورد نظر ما برای جمع آوری داده های خود این است: https://etherscan.io/txs.

مراحل داخل دفترچه یادداشت را دنبال کنید و همچنین ملاحظات را بررسی کنید!

تجزیه و تحلیل داده ها اکنون که داده های تراکنش را از Etherscanجمع آوری کرده ایم، گام بعدی انجام یک تحلیل اولیه است. این کار شامل مراحل زیر خواهد بود:

•بارگذاری داده ها: داده های تراکنش جمع آوری شده را به DataFrameپاندا وارد کنید. •پاکسازی داده ها: داده ها را با تبدیل انواع داده، حذف هر گونه اطلاعات نامربوط، پاک کنید
اطلاعات و مدیریت مقادیر تکراری •تحلیل آماری: میانگین و انحراف معیار جامعه را محاسبه کنید.

برای درک توزیع ارزش معاملات، این آمار را ارزیابی کنید. تجزیه و تحلیل و ترسیم بر اساس هزینه و ارزش Txnخواهد بود.

•تجسم: این مرحله شامل ایجاد نمایش های بصری برای کمک به تجزیه و تحلیل ارزش معاملات است. تجسم ها عبارتند از:

•یک هیستوگرام برای هر ستون داده، که نمایشی بصری از توزیع داده را ارائه می دهد. انتخاب اندازه سطل بسیار مهم است و باید بر اساس آن باشد

بر روی ویژگی های داده ها برای اطمینان از نمایش دقیق. در مورد انتخاب اندازه سطل توضیحی ارائه دهید! •نمودار توزیع نرمال که در کنار هیستوگرام نصب شده است تا توزیع تجربی داده ها را با توزیع نرمال نظری مقایسه کند. •طرح جعبه و طرح ویولن برای شناسایی نقاط پرت و ارائه یک طرح جامع

نمای توزیع داده ها

نمونه‌برداری و تجزیه و تحلیل داده‌ها در این بخش، به فرآیند نمونه‌گیری داده‌ها و تحلیل اولیه داده‌های تراکنش‌هایی که جمع‌آوری کرده‌ایم، خواهیم پرداخت. هدف ما درک توزیع ارزش معاملات با نمونه برداری از داده ها و مقایسه آمار نمونه با آمار جمعیت است. این کار شامل مراحل زیر خواهد بود:

•بارگذاری داده ها: داده های تراکنش جمع آوری شده را به DataFrameپاندا وارد کنید. •پاکسازی داده ها: داده ها را با مدیریت مقادیر از دست رفته، تبدیل انواع داده ها و حذف هرگونه اطلاعات نامربوط پاک کنید. •نمونه گیری تصادفی ساده :(SRS)نمونه ای از مجموعه داده با استفاده از روش نمونه گیری تصادفی ساده ایجاد کنید. این شامل انتخاب تصادفی زیرمجموعه ای از داده ها بدون توجه به ویژگی های خاص داده ها است. •نمونه گیری طبقه ای: نمونه دیگری از مجموعه داده با استفاده از روش نمونه گیری طبقه ای ایجاد کنید. این شامل تقسیم داده ها به طبقات بر اساس یک ویژگی خاص (مثلاً ارزش معاملات) و سپس انتخاب تصادفی نمونه از هر طبقه است. توضیح دهید که داده ها را بر اساس چه طبقه بندی کرده اید و چرا این ستون را انتخاب کرده اید.

•تجزیه و تحلیل آماری: محاسبه میانگین و انحراف معیار نمونه ها و جامعه. برای درک توزیع ارزش معاملات، این آمار را مقایسه کنید.

•تجسم: توزیع ارزش معاملات و کارمزدها را برای هر دو نمونه ترسیم کنید
و جمعیت برای مقایسه بصری توزیع آنها.

سوالات

.1برخی از محدودیت های بالقوه در هنگام استفاده از وب اسکرپینگ برای جمع آوری داده ها چیست؟
به طور خاص، هنگام واکشی داده ها از Etherscanبا چه مشکلاتی مواجه شدید؟ این محدودیت ها در تحلیل شما چه مشکلاتی می تواند ایجاد کند؟
.2چه چیزی می تواند تحلیل شما را غیرقابل اعتماد کند؟ راه حل های شما چیست؟
.3چگونه تجسم به شما در درک داده ها کمک کرد؟ چه چیزی را می توانید از توطئه ها تفسیر کنید؟

.4دو روش نمونه گیری در خروجی چه تفاوتی دارند؟ اینها را مقایسه کنید و توضیح دهید که کدام یک برای جمعیت مناسب تر است.

یادداشت

•کار خود را به صورت فایل فشرده در این قالب در وب سایت آپلود کنید: .DS_CA0_[Std number].zip
اگر پروژه به صورت گروهی انجام می شود، تمام شماره دانشجویی اعضای گروه را در نام ذکر کنید.

•اگر پروژه به صورت گروهی انجام شود، فقط یک عضو باید اثر را آپلود کند. •ما کد شما را در حین تحویل پروژه اجرا خواهیم کرد، بنابراین مطمئن شوید که نتایج شما درست است

قابل تکرار