تحلیل بازار رمزارزها

گزارشی از روند انجام پروژه بوت‌کمپ علوم داده کوئرا

Quera
Group 8

تحلیلگران پروژه

یوسف عبدالکریمی	منتور و راهنما
علیرضا محمدزاده	طراحی پایگاه داده
حسین زندی نژاد	تحلیل‌های آماری
حسین نادری	مصورسازی و طراحی گزارش
مهرداد عادلی	مستندسازی و اسکریپینگ
مهرشاد فلاح	استخراج داده‌ها و اطلاعات رمزارزها

Quera

بخش اول: استخراج داده

((ابتدا با استفاده از لینک‌ها و بهره‌گیری از سِلنیوم[1] و درایور فایرفاکس، اطلاعات مورد نظر مانند اسم رمزارز، سمبل، لینک صفحه اصلی، لینک تاریخچه هر رمزارز، ارزش بازار، حجم مبادله ۲۴ ساعت اخیر و حجم در دسترس مشتریان را بدست آورده و این تعداد از اطلاعات در قالب چند لیست ذخیره می‌شوند.))

Output

اطلاعات استخراج شده در DataFrame به یک فایل CSV ذخیره می‌شوند تا برای تحلیل های بعدی قابل استفاده باشند. همچنین، داده‌های تاریخ برای هر ارز از صفحات دانلود شده و به صورت فایل‌های CSV در دایرکتوری مشخص‌شده ذخیره می‌گردد.

Pandas

با اطلاعات استخراج شده از مرحله قبلی، یک DataFrame پانداس ایجاد می‌شود برای مدیریت داده‌ها که در آن اطلاعات هر رمزارز در ردیف‌های مختلف قرار دارد. این DataFrame نیازی به مدیریت داده‌های خالی ندارد زیرا هر ردیف یکتا و منحصر به فرد می‌باشد.

Scrapering

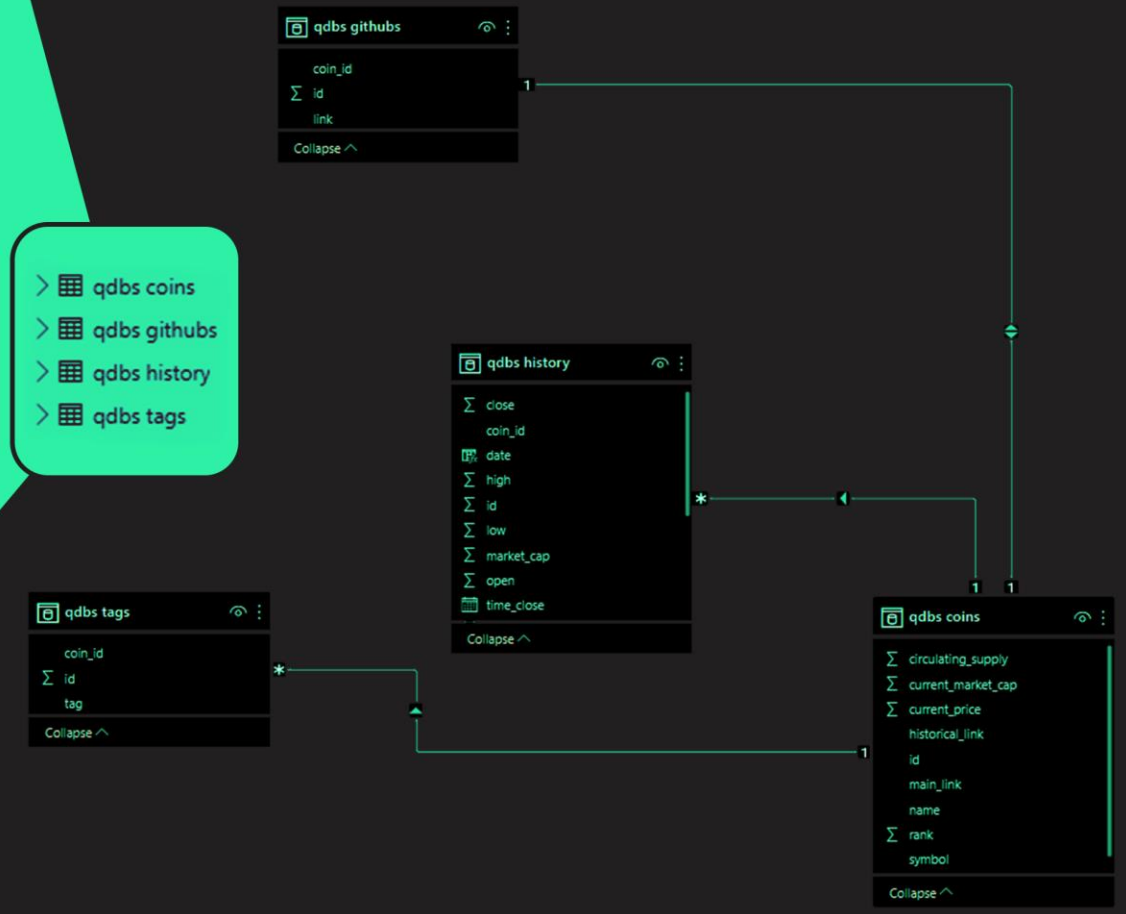
در این بخش، با استفاده از Selenium، اطلاعات مربوط به ۲۰۰ ارز دیجیتال منتخب از وب‌سایت coinmarketcap.com در تاریخ خواسته شده استخراج می‌شود. همچنین در این مرحله، لینک‌های مرتبط با صفحات GitHub و تگ‌های مربوط به هر رمزارز نیز استخراج می‌شوند.

---

۱. "Selenium" یک ابزار اتوماسیون مرورگر است که برای تعامل با وب‌سایت‌ها و استخراج داده‌ها به کار می‌رود. این ابزار می‌تواند مرورگر وب را کنترل کرده و عملیاتی مانند کلیک کردن بر روی المان‌ها و تعامل با صفحات وب را انجام دهد، ممکن می‌سازد.

Database Design

بخش دوم:
طراحی دیتابیس



ابتدا دیتاست‌ها[1] رو برای وارد کردن در دیتابیس آماده می‌کنیم. این کار شامل وارد کردن و Import فایل های ساخته شده توسط اسکریپر، ترکیب سطرهای دیتاست های تاریخچه، تغییر نام ستون‌ها، ساخت دیتاست تگ‌ها و گیتهاب می‌شود.

سپس برای موجودیت‌های موجود که شامل Coin, Tags, History, GitHub می‌شوند، schema جدول‌های مربوطه را در دیتابیس می‌سازیم. این کار شامل ساخت ستون‌ها و تعیین نوع داده، ساخت کلیدها و ایجاد رابطه میان جداول می‌شود.

---

۱. "دیتاست" مجموعه‌ای از داده‌ها است که معمولاً در جداول یا ماتریس‌ها ذخیره می‌شوند،
هر سطر در دیتاست به یک نمونه مربوط می‌شود و به تحلیل و استفاده در مطالعات و تحقیقات مختلف کمک می‌کند.

Statistical Analysis

بخش سوم: تحلیل‌های آماری

آمار توصیفی

در ابتدا با استفاده از sqlalchemy جداول به دست آمده در بخش قبلی را با کتابخانه Pandas به دیتافریم تبدیل کردیم. سپس با بررسی های اجمالی از دیتاست ها به ادامه کار پرداختیم.در این بخش، ابتدا نمودارهای پراکندگی ارزش بازار در مقابل حجم معاملاتی روزانه، نمودار توزیع حجم معاملات روزانه رمزارزهای قابل استخراج و نمودار میله ای ۱۰ رمزارز برتر از نظر تعداد افزایش قیمت در روزهای قرمز بهار سال ۲۰۲۳ با استفاده از کتابخانه Plotly رسم شده است. سپس به بررسی رمزارزهایی که در سال اخیر تغییرات قیمت همسو داشته‌اند پرداخته شد. در این بخش، رمزارزهایی با ویژگی خواسته شده به همراه تعداد روزها با تغییرات قیمت همسو در جدولی نمایش داده شده است. در آخر نیز ماتریس همبستگی و نقشه حرارتی (heatmap) تغییرات قیمت برای شانزده رمزارز برتر از نظر ارزش بازار رسم شده است. این ماتریس کمک می‌کند تا روند تغییرات قیمت این رمزارزها در یک سال اخیر را در مقابل یکدیگر رصد کنیم.

تخمین و آزمون فرض

در این بخش ابتدا با استفاده از روش‌های نمونه گیری، چهل رمزارز از میان داده های استخراج شده به صورت تصادفی انتخاب شده است. سپس میانگین حجم معاملاتی روزانه هر نمونه محاسبه شد و در نهایت بازه اطمینان ۹۸ درصدی برای حجم معاملاتی به دست آورده شد. ابتدا به بررسی روزهای متلاطم بازار برای پاسخ به سوال اول پرداخته شد. برای به دست آوردن شاخصی مناسب برای این آزمون کردن فرض خواسته شده، دو رویکرد ارائه شد. شاخص اول، تغییرات قیمت بر اساس تفاضل قیمت پایانی (close) و قیمت آغازین (open) و شاخص دوم به صورت تغییرات قیمت بر اساس تفاضل بالاترین قیمت (high) و پایین ترین قیمت (low)، در هر روز در نظر گرفته شدند. در هر دو حالت، فرض صفر که بیان کننده عدم وجود تفاوتی فاحش میان دو انتخاب ممکن برای روزهای کاری می باشد، با آزمون فرض t رد شد. سپس درستی ادعای " میانگین حجم معاملات روزانه، Ethereum، Bitcoin و USDt Tether به شدت بیشتر از میانگین حجم معاملات روزانه باقی رمزارزهاست" بررسی شد. این ادعا با آزمون فرض t رد شد. در آخر، اعتبار ادعای "نرمال بودن توزیع تغییرات قیمت رمزارزها در روزهای تعطیل (شنبه و یکشنبه)" با آزمون های فرض Anderson و Sminrov-Kolmogorov این ادعا رد شد.

Data Visualisation & Presentation

بخش چهارم: Power BI

مصـــورسازی

با استفاده از دیتابیس ساخته شده در مراحل قبل و استفاده از ابزار مصورسازی Power Bi ابتدا دیتاها را به منظور تحلیل بهتر به ستون‌های جزئی‌تر تقسیم کرده و در نهایت با استفاده از ویژوال‌های مختلف، چارت و نمودارهای موردنیاز برای ارائه این پروژه به مخاطب هدف را طراحی و اجرا می‌کنیم.

گزارش و ارائه

با بهره‌گیری از ابزارهای طراحی مثل Adobe Photoshop و PowerPoint یک گزارش حرفه‌ای از روند پیشرفت و انجام تسک‌های داده شده به همه اعضا گروه تحلیلگران به مخاطبین داده می‌شود. این گزارش باید دارای خلاصه‌ای از مراحل انجام هر بخـش از پروژه باشد و نمای کلی از تاریخچه تکمیل وظایف بدهد.

REPORT SERVER

یک سرور برای نمایش داشبورد و نمودارهای طراحی شده در قدم قبلی تهیه و تنظیم می‌گردد و قابلیت نمایش اطلاعات در انواع دستگاه‌ها امکان پذیر می‌کند.

More Information...

بخش پنجم: اطلاعات بیشتر

GitHub

روند کار اسکریپینگ گیتهاب

با استفاده از لینک های گیت‌هاب که به وسیله ی اسکریپر اصلی استخراج شده اند به گیت‌هاب هر کوین ریکوئست زده شده و اگر آدرس لینک متعلق به صفحه اول کوین نباشد، به وسیله توابع سلنیوم به صفحه ی اول منتقل شده و از بخش people تعداد کل مشارکت کنندگان و لینک پروفایل مشارکت کنندگانی که در صفحه ی اول این بخش قرار دارند استخراج میشود. سپس همین روند تکرار شده و از صفحه ی اصلی هر کوین، زبان های برنامه نویسی ای که بیشترین استفاده را در ساختن آن کوین داشته اند استخراج می‌شود. سپس با استفاده از توابع سلنیوم و قابلیت مرتب سازی و طبقه بندی ریپازیتوری ها، آنها را بر اساس تعداد استار و اینکه ریپازیتوری سورس باشند مرتب کرده و کل المان های ریپازیتوری را استخراج می‌کنیم. سپس از این المان های استخراج شده، نام ریپازیتوری، لینک آن، جزئیات آن اگر موجود باشد و همینطور لایسنس، تعداد فورک ها و استار های ریپازیتوری را استخراج میکنیم. برای هر کوین که صفحه ی گیت‌هاب ندارد رشته ی «No Github Link» ذخیره می‌شود. همچنین داده ها بر اساس رتبه ی کوین شماره دهی می‌شوند. در آخر داده ها در سه فایل csv شامل مشارکت کنندگان، زبان های برنامه نویسی و ریپازیتوری ها ذخیره میشوند.

پایان!

Quera
Group 8