IIC2413 – Bases de Datos

Control III
Data Science - Mongo DB

Entrega

El plazo para la entrega es Viernes, 2 de Julio, hasta las 20:00, mediante un cuestionario en SIDING. La entrega debe hacerse en un archivo jupyter notebook (`.ipynb`) que se pueda ejecutar teniendo el archivo de la base de datos en la misma carpeta. Si necesitas agregar algo más puedes subir un `.zip` con todos los archivos, pero debes incluir un `README` en markdown explicando a qué corresponde cada archivo.

Introducción

A principios de este año un fenómeno nunca antes visto afectó los precios de algunas acciones en USA, donde miles de personas se pusieron de acuerdo por Reddit para hacer subir los precios de acciones artificialmente. En este control usarás tus recientemente adquiridos conocimientos de `MongoDB` y de la librería `Pandas` de Python para hacer un análisis de esta extraña situación. Para eso tendrás acceso a las siguientes fuentes de datos:

- Una base de datos de `sqlite3`: `stonks.db`, con información de varias acciones del NASDAQ (y algunas otras) e historial de sus precios entre fines del año 2020 y principios del 2021. Tiene dos tablas `stonks` y `prices`.
- $\sim 700K$ Posts hechos en el sumamente popular subreddit r/wallstreetbets durante un periodo similar. Estos están disponibles para ser consultados remotamente conectándote a una instancia de `MongoDB` que levantamos (el texto esta indexado).

Explorar a más detalle qué hay (y qué no) en los datos es parte de lo que debes hacer para poder responder el control.

1. Instrucciones para manejo de datos y consultas

Queremos que uses la librería sqlite3 de Python para consultar la base de datos de acciones (`stonks.db`) para luego procesar los resultados con pandas, de forma similar a como lo hicimos en la guía de data science. Por otro lado, para consultar los posts debes conectarte a nuestro servidor de mongo usando la librería pymongo y el siguiente snippet:

```
from pymongo import MongoClient

uri = "mongodb://alumnoXX:XX@gray.ing.puc.cl/control3?authSource=perm"
client = MongoClient(uri)
db = client.get_database('control3')
collection = posts_db.wallstreet_bets_posts
```

Donde XX es tu número de alumno. Intenta ser cuidadoso al consultar a Mongo y hacer el control con anticipación para que no tengamos problemas de alta carga en el servidor. Decidir qué cosas calculas con SQL, cuáles con Mongo y cuales con pandas es decisión tuya.

2. Tareas a realizar

En esta sección se describen las tareas a realizar para el control. Hacer un análisis preliminar de los datos (ej: revisar si hay nulos o ruido) antes de completar lo que se pide es responsabilidad tuya. Puedes usar cualquier librería de visualización de python para generar los gráficos.

2.1. Análisis stonks.db

2.1.1. Volumen transado

1. Para cada acción en el dataset, calcula cuánto fue el total, media, desviación estándar y el máximo de volumen diario transado en el periodo. ¿Cuáles fueron las 5 acciones con mayor volumen transado en el periodo?

2. Calcula el volumen total transado cada día y haz un linechart de la serie de tiempo resultante. ¿Hay algún patrón interesante?

2.1.2. Correlación entre distintos activos.

1. Queremos visualizar algunas series de tiempo de las acciones, para eso crea un linechart con los precios de cierre cada día de las acciones: GME, AMC, TSLA y GOOGL.

2. Calcula la correlación de los precios de cierre entre todas las acciones. ¿Que acciones tienen mayor correlación positiva? ¿Y negativa? Haz un linechart para cada par.

2.1.3. Cambios en el precio de las acciones

1. Calcula el cambio porcentual (rentabilidad) para cada acción cada día. Esto lo puedes calcular como (precio de cierre − precio de apertura)/precio de apertura.

2. Calcula la media y la desviación estándar de la rentabilidad diaria para cada acción. ¿Cuáles tuvieron más rentabilidad media en el periodo? ¿y más variabilidad?.

2.2. Efecto de r/wallstreetbets

2.2.1. Búsqueda por texto

Para cada una de las acciones mencionadas en el punto 2.1.2.1, debes completar los siguientes análisis:

1. Cuenta la cantidad de posts y comentarios que mencionan el nombre de la empresa o su símbolo. Ojo con usar los nombres completos de las acciones, que es poco probable que aparezcan.

2. Crea dos series de tiempo agregando la cantidad de posts y comentarios por día (tip: La fecha de creación viene en el campo `created_utc` en Unix Time).

3. Finalmente, genera linecharts comparativos y calcula la correlación entre las series del punto anterior y el volumen transado, precio de cierre y rentabilidad. ¿Qué puedes concluir respecto a la relación entre la actividad en el subreddit y el comportamiento de las acciones?