APLICACIÓN DEL *WEB SCRAPING* Y EL ANÁLISIS AUTOMATIZADO A LOS MERCADOS DE DIVISAS Y DE ACCIONES

Application of web scraping and automated analysis to stock and currency exchange markets

Aplicação da *web scraping* e o análise automatizado aos mercados de divisas e de ações

Andrey Felipe Rincón Torres
anrincon35@poligran.edu.co
Politécnico Grancolombiano
Colombia

Carlos Felipe Cortés Cataño
cacortes15@poligran.edu.co
Politécnico Grancolombiano
Colombia

Oscar Leonardo Acevedo Pabón
olacevedo@poligran.edu.co
Politécnico Grancolombiano
No. ORCID 0000-0002-1270-4166
No. Autor Redalyc (solo número)
Colombia

Resumen

Este trabajo es parte de un proyecto en el que estudiantes del semillero de investigación en Ciencia de Datos del Politécnico Grancolombiano fueron entrenados en técnicas de *web scraping* y modelado de datos de mercados. Los mercados elegidos fueron el de intercambio de divisas (FOREX) y el de acciones de la Bolsa de Valores de Colombia (BVC). Usando programación en lenguaje *Python*, y una serie de librerías especializadas, los semilleristas lograron la extracción, limpieza y consolidación de datos automática con completa exactitud, sin presentar datos faltantes. Además, los semilleristas lograron desarrollar aplicaciones que permiten el análisis automático con el propósito de recomendar estrategias de *trading* algorítmico. Esto último está particularmente avanzado en el caso del mercado FOREX, pues ya se ha desarrollado un algoritmo predictivo que promete generar ganancias.

Palabras clave

Web scraping, análisis automatizado, predicción de mercados

Abstract

This work is part of a project in which students from the Data Science research hotbed of Politécnico Grancolombiano were trained in web scraping techniques and market data modeling. The chosen markets were the currency exchange market (FOREX), and the Colombian stock exchange market (BVC). Using *Python* programming, and a series of specialized libraries, the researchers achieved automatic extraction, cleanse, and consolidation of data with complete accuracy, with no missing data. In addition, the young researchers managed to develop apps that allow automatic analysis aiming to recommend algorithmic trading strategies. The latter goal was particularly advanced in the case of the FOREX market, since a predictive algorithm has already been developed and it promises to generate profits.

Keywords

Web scraping, automated analysis, market prediction

Abstrato

Este trabalho faz parte de um projeto no qual alunos do sementeiro de pesquisa em Ciência de Dados do Politécnico Grancolombiano foram treinados em técnicas de *web scraping* e modelagem de dados de mercados. Os mercados escolhidos foram o mercado de câmbio de divisas (FOREX) e o mercado da bolsa de ações colombiano (BVC). Usando programação *Python* e uma série de bibliotecas especializadas, os pesquisadores conseguiram a extração, limpeza e consolidação automáticas de dados com total exatidão, sem dados ausentes. Além disso, os jovens pesquisadores conseguiram desenvolver aplicativos que permitem análises automáticas procurando recomendar estratégias de negociação algorítmica. Este último objetivo foi particularmente avançado no caso do mercado FOREX, uma vez que um algoritmo preditivo já foi desenvolvido e promete gerar lucros.

Palavras-chave

Web scraping, análise automatizado, predição de mercados

INTRODUCCIÓN

Uno de los obstáculos a los que se enfrentan muchos proyectos de investigación en ciencia de datos es justamente la posibilidad de disponer de una base de datos lo suficientemente grande, completa e interesante para comenzar a aplicar los modelos de aprendizaje automático o de minería de datos que se tienen en mente (Broucke y Baesens, 2018). Esto es así incluso a nivel pedagógico, donde la base de datos es deseada para el entrenamiento de los futuros científicos y científicas de datos. Si bien es cierto que existen varias fuentes de bases de datos en repositorios académicos o en sitios web de aprendizaje de ciencia de datos; muchas veces estas fuentes no poseen las dimensiones adecuadas, o la tipología de los datos no es pertinente, o simplemente no concierne a los intereses investigativos de los estudiantes. El *web scraping*, es decir, cualquier técnica de extracción automática de datos de la red informática mundial accesible mediante la internet (Patel, 2020), suple esa necesidad de datos usando un arsenal técnico y conceptual muy similar al de otras áreas de la ciencia de datos (Mitchell, 2018).

Un proyecto de *web scraping* se quedaría corto si no se utilizan los datos; puesto que la aplicación y el análisis posterior deben guiar siempre el tipo de *web scraping* que se está formulando (Patel, 2020). Debido a su carácter automático, el *web scraping* podría ser el primer paso para aplicaciones en situaciones en tiempo real. En particular, el *web scraping* se podría usar para determinar y poner en marcha estrategias de inversión en línea, generando un servicio de recomendación de inversiones automático (Frankenfield, 2022).

El presente trabajo es resultado de un proyecto de iniciación científica en el que integrantes del Semillero de Ciencia de Datos del Politécnico Grancolombiano se entrenan en técnicas de *web scraping* y análisis automatizado y las aplican al mercado de divisas extranjeras (FOREX) y al mercado accionario de la bolsa de valores de Colombia (BVC). Estos casos fueron de interés para los semilleristas porque ya habían tenido contacto con estas plataformas de inversión y pueden llegar a representar fuentes de ingreso extra para sus usuarios. Recuérdese que las transacciones en estos mercados son motivadas por el ánimo de lucro. Por ejemplo, aunque los procesos directos del FOREX consisten justamente en comprar y vender divisas, este mercado no está diseñado para la compra y venta de divisas para transacciones inmediatas. En cambio, el principal objetivo de sus actores es especular sobre los precios futuros de las diferentes divisas para así poder obtener ganancias o prevenir pérdidas (Jackson y Schmidt, 2019). Por otro lado, las técnicas y conceptos desarrollados en este proyecto (*web scraping* y análisis automatizado de mercados) son de alta demanda en el mercado laboral de la Analítica de Datos, la Ciencia de Datos, entre otros contextos profesionales (Mitchell, 2018).

MÉTODO

Toda la programación se hizo en el lenguaje de programación *Python,* pero para cada uno de los dos casos (FOREX y BVC) se usaron diferentes librerías que están más especializadas para el tipo de datos y análisis que se quería realizar. La consistencia de la información obtenida automáticamente se evaluó respecto a fuentes cuya exactitud está más probada o mediante una inspección más manual de una muestra de datos. La comparación fue entrada por entrada de bases de datos comparables (es decir, al nivel de filas de una tabla). Si existía al menos una discrepancia en el valor de una de las variables, entonces se consideró como una fila mal predicha de tal forma que se pudo obtener una proporción simple de exactitud de entradas.

En el caso FOREX se usó la librería *MetaTrader 5* (MetaTrader 5, s.f.) para acceder a datos de precio, volumen y *spread* de diversos pares de divisas del mercado ofrecidos por la plataforma de transacciones también llamada *Meta Trader 5*. Este *script* de obtención de datos permitió una interacción más fluida con el análisis automatizado subsecuente. En el análisis automatizado se usó la librería *mpl-finance* (Goldfarb, s.f.) para la visualización de gráficas de velas. La estrategia de inversión recomendada automáticamente será bajo la modalidad de *scalping*: una técnica de inversión de muy corto plazo (del orden de horas o, cuando mucho, días) que se basa en la obtención de pequeñas ganancias debido a las fluctuaciones del valor (Cheng, 2007).

Para que una estrategia de inversión tenga éxito, el inversor debe reconocer momentos propicios de entrada en el mercado de compra/venta de dos divisas (zona de confluencia) y definir una estrategia que defina hasta qué momento espera llegar con ganancias y así retirarse antes de que la tendencia favorable termine. Para detectar las zonas de confluencia se usaron diversos indicadores como la alerta fractal, la varianza, el *alligator* y un modelo de tasa de rentabilidad logarítmica (Tsay, 2010). Esta estrategia está inspirada en las ideas del famoso inversor y educador en trading Billy Williams (Williams y Williams, 2004). Los detalles financieros del modelo predictivo que permite generar una estrategia automáticamente están por fuera del alcance de este artículo y son parte de la tesis de pregrado del autor A. F. Rincón. Un último avance en este caso de FOREX fue la integración del modelo de predicción de Python directamente en la plataforma de *Meta Trader 5* mediante un *socket* (Dmitrievsky, 2019). Esto es un paso más en la generación de un inversor automatizado puesto que con dichos datos se pueden tomar decisiones de compra y venta directamente en la plataforma de inversión sin necesidad de la intervención directa de un inversionista humano.

En el caso BVC las librerías usadas fueron *pandas* (Alberca, s.f.), *selenium* (selenium, s.f.) y *Beatifulsoup 4* (Beautiful soup, s.f.). Con estas librerías fue posible navegar automáticamente el sitio web de la BVC (bvc.com.co); en particular, se hizo posible explorar por completo los datos en su archivo HTML. En este caso, el análisis automático solo se avanzó a nivel exploratorio, generando indicadores ampliamente utilizados en el mercado accionario y permitiendo su visualización mediante una pequeña aplicación programada en la plataforma *dash* (Plotly Graphing Libraries, s.f.).

RESULTADOS

Caso de mercado FOREX

La plataforma de transacciones *Meta Trader 5* otorga por defecto unos datos en formato de valores separados por coma (CSV, por sus siglas en inglés) pero estos son de difícil manejo y no permiten aplicaciones en tiempo real. Debido a esto, se hizo necesario un *script* que importara automáticamente los datos. Este *script* usó la librería *MetaTrader 5*, la cual permitió acceder a todos los datos de diferentes divisas, índices sintéticos, materias primas y algunos cripto-activos segundo a segundo. En todas las pruebas realizadas en las que se compararon los archivos CSV de la plataforma y los obtenidos por el *script*, no se encontró ninguna discrepancia en la información; por lo que se tiene un porcentaje de coincidencia de 100%.

Dado el acceso a la información de precios, volumen y fechas, el siguiente paso fue la limpieza y estructuración de los mismos. Un aspecto importante fue la sincronización y cálculos de tiempos usando zonas horarias precisas y para esto se empleó la librería *pytz* (Bishop, s.f.). Por otro lado, el resto de la limpieza y estructuración se hizo mediante la librería *pandas* (Alberca, s.f.). Posteriormente, los datos

recolectados fueron usados en un modelo matemático orientado al cálculo de puntos de interés de compra y venta de activos (zonas de confluencia) mediante la estrategia de *scalping*. La verificación del modelo matemático todavía se encuentra en curso. Mediante simulaciones con datos en retrospectiva, la estrategia recomendada produce ganancias, lo cual es bastante prometedor. El modelo matemático da como resultados zonas de confluencia predichas, las cuales se pueden apreciar mejor mediante visualización en una gráfica de velas. En la Imagen 1 se aprecia un ejemplo de dicha gráfica generada mediante *mpl-finance.* Las líneas verticales son la zona de confluencia predicha por el modelo matemático. El hecho de que coinciden con un intervalo en el que es favorable cambiar USD por EUR muestra que este es un ejemplo de éxito de predicción del modelo.

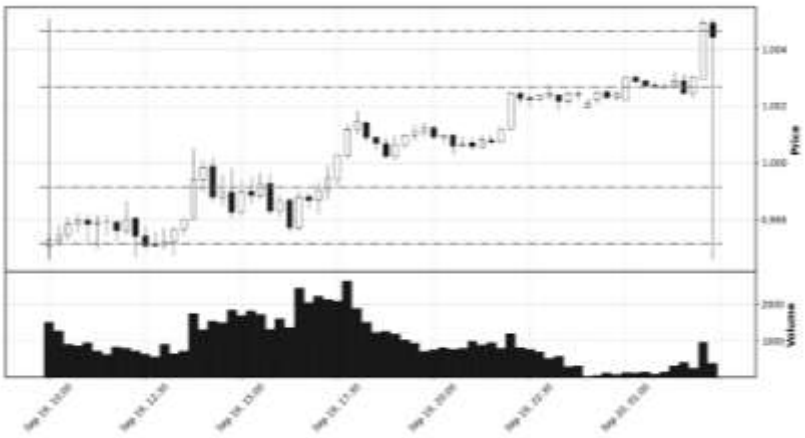

Imagen 1.
Visualización con mpl-finance en Python de gráfica de velas del precio de EUR vs. USD (divisas). Las líneas verticales marcan una zona de confluencia y son predichas por un modelo de análisis automatizado.
Fuente: realización propia.

Dado este avance, se propuso el poder visualizar y utilizar los resultados del análisis automático directamente en la plataforma *Meta Trader 5*. El objetivo era tener la información directamente en la plataforma y no acudir a fuentes externas, puesto que en ella se podría aplicar directamente la toma de decisiones de compra o venta. Dmitrievsky (2019) explica un método por el que, por medio de un *socket,* se genera una conexión para llevar datos de *Meta Trader 5* a *Python* y viceversa. La comunicación mediante este *socket* en el sentido *Meta Trader 5 - Python* está limitada a unos pocos datos que puede recibir *Python* (solo ofrece precios de cierre de cada vela). Sin embargo, este *socket* sí permite enviar los datos del modelo predictivo a *Meta Trader 5*. En la Imagen 2 se aprecia una gráfica generada directamente en *Meta Trader 5* que muestra la zona de confluencia predicha en *Python* (la misma de la Imagen 1), esto demuestra la comunicación exitosa en el sentido *Python – Meta Trader 5*.

Imagen 2
Visualización en Meta Trader 5 de la gráfica de vela y los datos predichos por el modelo matemático.
Fuente: realización propia

Caso de la bolsa de valores de Colombia

Al igual que en el caso del mercado FOREX, la BVC ofrece un servicio de descarga gratuita en archivos en formato CSV. Sin embargo, la baja eficiencia en la extracción y consolidación de los datos, sumada a limitantes como descargas de información en lotes de 6 meses y máximo un histórico de 5 años frente a la fecha actual hacen relevante crear un servicio automatizado. Para eso, se desarrolló un código de actualización que se ha puesto en ejecución desde el 1 de mayo de 2022. El código realiza un barrido dentro del contenido HTML del sitio *web* de la BVC y extrae una tabla con las variables nemotécnicas de cada acción: nombre de la acción o abreviación, último precio, precio de apertura, precio máximo, precio promedio, precio mínimo, cantidad, volumen (cantidad de acciones compradas y/o vendidas multiplicadas por el precio en cada compraventa), variación absoluta y variación porcentual. La temporalidad de estos datos es diaria, a diferencia del caso FOREX que puede hacerse segundo a segundo.

Los datos antes descritos no se presentan de forma inmediata, sino son añadidos al código HTML después de la carga de la página. Por tal razón, es necesario usar *selenium* para realizar una espera de máximo 10 segundos para poder extraer el código HTML. La transformación y consolidado de datos se realiza usando las librerías de *Beatifulsoup 4* y *pandas*. El resultado es un conjunto de datos consolidado y actualizado que permite a los usuarios del código verificar los datos extraídos con los resultados presentados en distintas páginas *web* que presentan la misma información en las mismas temporalidades. También permite visualizar e incluso aplicar técnicas de aprendizaje automático a los datos resultantes. Para el caso del presente proyecto, se realizó una verificación de los registros en su precio de cierre, los cuales no evidencian discrepancia entre los precios reportados en la descarga manual de los archivos CSV, por lo que se obtiene un nivel de coincidencia de información del 100%.

El siguiente paso consistió en diseñar una aplicación en *dash* que permita actualizar de forma automática el conjunto de datos y presentar visualizaciones usando el servicio AWS EC2 de Amazon (Amazon Web Services, s.f.). El funcionamiento es el siguiente: de lunes a viernes, a las 5:00 p.m., hora colombiana, se ejecuta automáticamente una tarea en una instancia de AWS EC2 que provee capacidad de almacenamiento y procesamiento para que el código tome el conjunto de datos inicial y realice el proceso

de extracción, transformación y carga. Luego, con esta nueva información, en la carpeta de la aplicación *web* en *dash* se actualizan los datos. Con los datos actualizados se puede hacer un llamado desde la aplicación para analizarlos y visualizarlos, el llamado se hace mediante una dirección IP asignada en un balanceado de cargas (AWS LoadBalancer).

Imagen 3
Resultados de visualización en la aplicación dash de la acción de ECOPETROL de la BVC.
Fuente: realización propia.

En la Imagen 3 se muestra el resultado de una de estas visualizaciones. Se trata de un gráfico OHLC, un estilo de gráfico financiero que describe los valores de apertura, máximo, mínimo y cierre para un determinado tiempo. La punta de las líneas representa los valores alto y bajo y los segmentos horizontales representan los valores de apertura y cierre. Los puntos de muestra donde el valor de cierre es mayor (menor) que el valor de apertura se denominan crecientes (decrecientes). Asimismo, se añaden los indicadores técnicos de media móvil exponencial de 9 y 26 días, índice de movimiento direccional promedio (ADX), media móvil de convergencia y divergencia (MACD) y la cantidad de acciones negociadas para cada día. Estos indicadores se podrían usar para proponer una estrategia de *trading* técnico que indique una orden de compra o una orden de venta cuando existen cruces de medias móviles exponenciales (Heredia García, 2016). En ocasiones, es posible que estos cruces no se den porque no hay suficientes compras y ventas que se estén realizando. En tales casos, es necesario validar que el ADX apunte a superar la cota de 0.23. Finalmente, el MACD permite identificar los cambios que podría presentar la acción del precio generando un pronóstico estimado de ganancia. En este caso todavía no se ha programado un algoritmo de *trading* automático ni evaluado su efectividad.

DISCUSIÓN Y CONCLUSIÓN

Respecto a la aplicación de *web scraping*, los dos casos (mercado FOREX y BVC) son de total éxito puesto que no se ha encontrado ninguna discrepancia en la información obtenida mediante los códigos automáticos y otras fuentes. Se ha podido corroborar la exactitud de esa información, consolidarla de forma limpia y bien estructurada, y se dispone la posibilidad de actualizarla de forma automática. Las variables

más comunes incluidas en la extracción permitieron realizar un proceso de limpieza e ingeniería de características, sin quedar con datos faltantes. En los dos casos se cuenta con una plataforma adecuada para análisis automatizado: dentro de *Meta Trader 5* en el caso FOREX y una aplicación *dash* con servicios de AWS E2 en el caso de la BVC.

En el caso de FOREX se construyó un sistema de transferencia y análisis de datos para cualquier activo ofrecido por la plataforma *Meta Trader 5*. Este sistema abre paso al estudio y análisis de datos financieros de una manera económica donde el gasto es solo a nivel computacional y no se paga por servicios que analicen estos datos. Estos servicios de análisis financiero pueden ser bastante costosos (Frankenfield, 2022). El proceso de modelado de los datos de FOREX se encuentra en un buen estado de avance como para poder predecir ciertos aspectos del comportamiento de los datos de forma acertada y ya se está evaluando su rentabilidad como estrategia de *trading* algorítmico.

En cuanto al mercado de acciones, se tiene implementado el procesado de indicadores de forma automática que harían parte de la estrategia de *trading* técnico. Para este objetivo fueron cruciales librerías como *mpl-finance*. Sin embargo, todavía no se ha diseñado un algoritmo específico de *trading*. En trabajos posteriores se recomienda a nuevos semilleristas aplicar técnicas de aprendizaje automático supervisado a las bases de datos recolectadas de la bolsa de valores con el objetivo de entrenarse en diferentes modelos y verificar la pertinencia del uso de este tipo de *trading* dentro del mercado colombiano.

REFERENCIAS BIBLIOGRÁFICAS

Alberca, A. S. (s.f.). *La librería Pandas*. https://aprendeconalf.es/ docencia/python/manual/pandas/

Amazon Web Services (s.f.). Amazon EC2. Recuperado el 29 de abril de 2022 de https://aws.amazon.com/es/ec2/.

Beautiful Soup (s.f.). Beautiful Soup Documentation. Recuperado el 1 de abril 2022 de https://beautiful-soup- 4.readthedocs.io/en/latest/.

Bishop, S. (s.f.). pytz. *Python Package Index.* Recuperado el 30 de abril de 2022 de https://pypi.org/project/pytz/

Broucke, S. y Baesens, B. (2018). *Practical Web Scraping for Data Science: Best Practices and Examples with Python*. Springer Science and Business Media Apress.

Cheng, G. (2007). *7 winning strategies for trading forex: real and actionable techniques for profiting from the currency markets*. Harriman House.

Dmitrievsky, M. (7 de junio de 2019) Integración de MetaTrader 5 y Python: recibiendo y enviando datos. *MQL5*. https://www.mql5.com/es/articles/5691

Frankenfield, J. (9 de septiembre de 2022) Robo-advisor. *Investopedia. https://www.investopedia.com /terms/r/roboadvisor-roboadviser.asp*

Goldfarb, D. (s.f.). mpl-finance. *Python package index*. Recuperado el 21 de mayo de 2020 de https://pypi.org/project/mpl-finance/

Heredia García, N. (2016). *Predicción del precio de acciones mediante técnicas de minería de datos.* Industriales ETSII UPM.

Jackson A. L. y Schmidt J. (1 de septiembre de 2019). A basic guide to forex trading. *FORBES*. https://www.forbes.com/advisor/investing/what-is-forex-trading/

MetaTrader 5 (s.f.) *Módulo MetaTrader para la integración con Python*. Recuperado el 29 de abril de 2022 de https://www.mql5.com/es/docs/integration/python_metatrader5.

Mitchell, R. (2018). *Web scraping with Python: collecting more data from the modern web*. O'Reilly Media.

Patel, J. (2020). Getting structured data from the Internet: running web crawlers/scrapers on a big data production scale. Berkeley, CA: Apress.

Plotly Graphing Libraries (s.f.). *Dash Python User Guide*. Recuperado el 15 de junio de 2022 de https://dash.plotly.com/.

Selenium (s.f.). *The Selenium Browser Automation Project.* Recuperado el 1 de mayo de 2022 de https://www.selenium.dev/ documentation/

Tsay, R. (2010). *Analysis of financial time series*. Wiley.

Williams, J. y Williams, B. (2004). *Trading chaos: maximize profits with proven technical techniques*. J. Wiley.