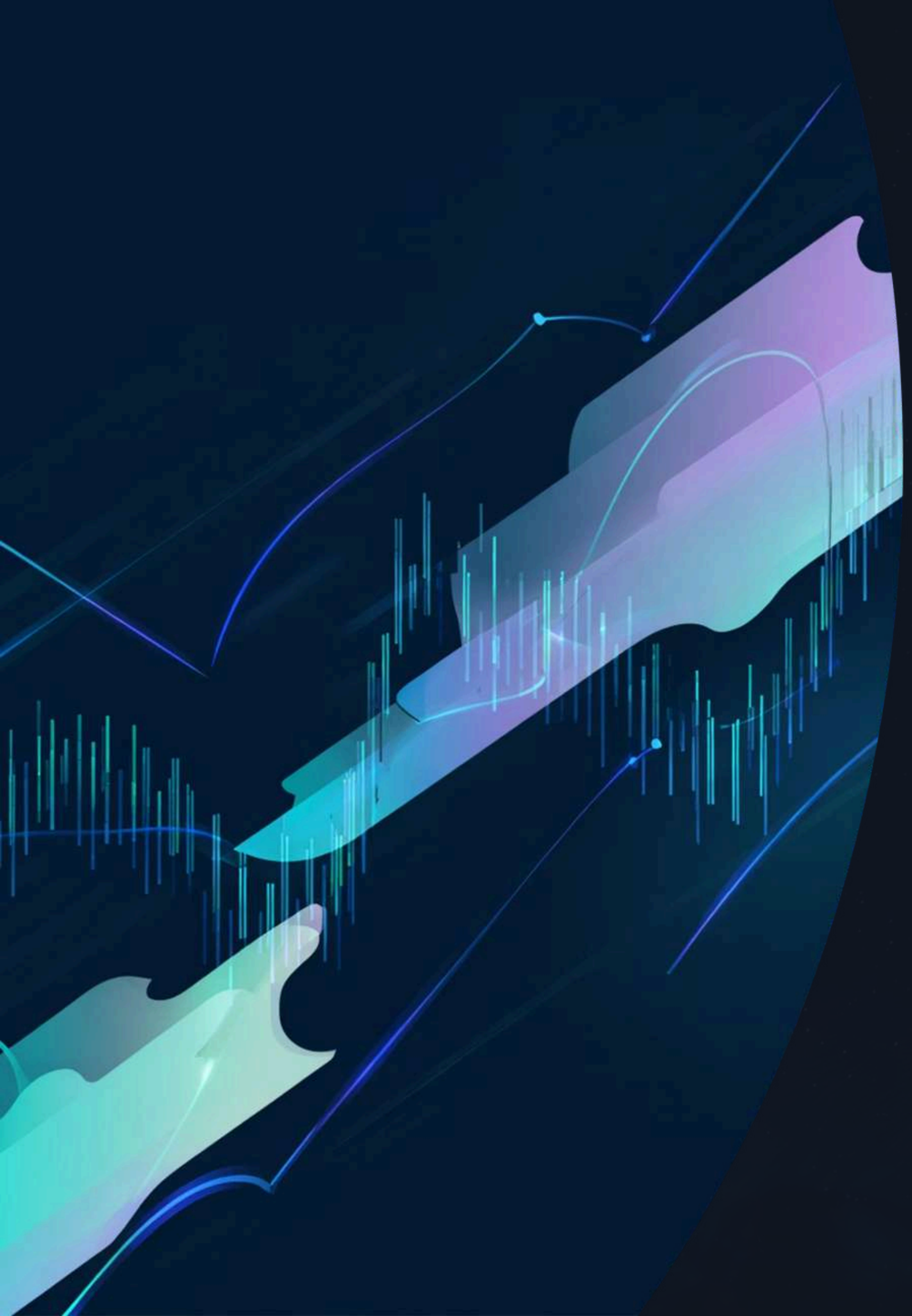

利用 LSTM 模型進行股票市場大數據分析與預測

作者：葉致均

研究動機：應對金融市場的非線性挑戰

市場資料爆發

金融市場資料量呈指數級增長，傳統統計模型難以有效處理其複雜性與高頻波動。

模型優勢互補

結合 PySpark 處理大規模歷史股價數據，並利用 LSTM 深度學習模型捕捉非線性時間模式。

建立智能系統

目標是設計一個能持續從新數據中學習、自我優化預測性能的智能金融預測系統。

> 本研究旨在克服傳統方法的局限，利用大數據與深度學習技術，從海量歷史股價中提取可預測模式。

研究目標與方法

數據管線建構

以 PySpark 為核心，設計高效能的股票大數據處理與特徵工程管線。

2

動態資料擷取

利用 yfinance API 動態獲取多家全球領先公司的歷史交易數據。

3

特徵工程強化

整合重要技術指標（如 7 日、21 日移動平均線 (MA7, MA21) 與日報酬率）作為模型輸入特徵。

4

分散式模型訓練

訓練 LSTM 深度學習模型預測下一日股價，並支持多節點分散式訓練以加快迭代速度。

效能評估與比較

精準評估單機與分散式架構在處理大規模金融時間序列數據時的效能與預測準確率（例如 RMSE）。

大數據資料來源與處理管線

資料來源與標的

來源：Yahoo Finance API (yfinance)。

標的：AAPL, MSFT, GOOGL, AMZN, TSLA, META 等科技巨頭。

時間範圍與規模

時間：跨越 2010–2024 年。

規模：累積數十萬筆（Time-Series）交易數據，形成龐大的數據集。

PySpark 並行處理

利用 PySpark 集群並行抓取、清洗數據，處理遺漏值和極端異常值。

特徵與格式生成

生成時間窗（Rolling Window）特徵與技術指標。儲存為高效能的 Parquet 格式，最終轉換為 TFRecord 供模型快速讀取。

LSTM 模型架構與訓練配置

採用 LSTM 模型，適合處理和記憶長序列數據中的時間依賴關係。

輸入與核心層

- 輸入層：50 天的歷史數據作為時間窗，共 **8 個特徵**。
- LSTM 層：**128 單元**，用於捕捉長短期記憶模式。

訓練參數

- 優化器：**Adam**，自適應學習率調整。
- 損失函數：**MSE** (Mean Squared Error)，用於量化預測誤差。
- 正則化：Dropout 層（0.2）以防止模型過度擬合。

關鍵技術

PySpark

實現大規模金融數據的 ETL（提取、轉換、加載）流程與分散式清理。

yfinance

可靠且自動化的工具，用於獲取標準化、結構化的歷史股價數據。

TensorFlow + Keras

提供靈活的深度學習框架，用於快速建構、實驗和部署 LSTM 模型。

tf.distribute

利用 MultiWorkerMirroredStrategy，實現跨多 GPU 或多機器的深度學習模型訓練，顯著縮短訓練時間。

符合大數據分析的要素

本研究涵蓋了現代大數據處理的 3V 特性，並採用了相應的工程化解決方案。

Velocity（速度）

PySpark 並行化： 大幅優化了數十萬筆數據的清理、特徵提取與轉換速度，支持高頻率模型再訓練。

Variety（多樣性）

異構數據整合： 涵蓋多家公司、多種市場指標（如價格、成交量、技術指標），數據類型豐富。

Volume（量大)

龐大的數據集： 涵蓋超過十年的多公司日頻數據，數據筆數遠超傳統單機處理能力。

- 使用 Parquet 和 TFRecord 格式：確保數據 I/O 效率，尤其是在分散式訓練環境中。
- 可擴展性：本架構設計易於部署到雲端大數據平台（如 Google Dataproc、AWS EMR），以應對更大規模數據。

研究總結與未來展望

研究結論

技術融合成功

成功整合 PySpark 大數據處理與 Keras LSTM 深度學習模型，形成高效能的預測流程。

提升工程效率

分散式架構顯著提升了資料準備速度與模型訓練的可擴展性。

預測能力驗證

LSTM 模型能夠有效地捕捉金融時間序列的複雜趨勢和波動性，預測性能穩定。

未來展望

集成**即時數據流**：例如利用 Spark Streaming 或 Kafka，將預測頻率從日級提升至分鐘級。

納入**非結構化數據**：結合新聞情感分析（NLP）和宏觀經濟指標，提高模型的資訊豐富度。

實現**全自動化雲平台**：建立 CI/CD 管線，實現模型訓練、部署、監控的全流程自動化。