TIME-SERIES FORECASTING DATASET AND SETUP EXPERIMENT ENVIRONMENT REPORT

1. Khảo sát một số time-series forecasting dataset

1.1. Nottingham dataset

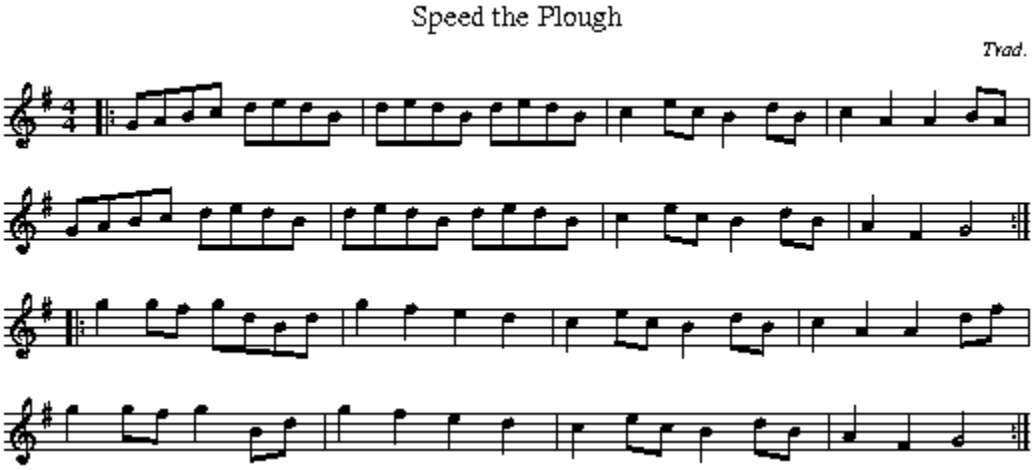
Nottingham dataset là một bộ sưu tập của hơn 1000 giai điệu dân ca Anh và Mỹ (hornpipe, jigs,…) được tạo ra bởi Eric Foxley. Sau đó, cơ sở dữ liệu này đã được chuyển đổi sang định dạng nhạc abc[1] và đã được đăng trên http://abc.sourceforge.net/NMD/.

- ***Nguồn gốc***: Được Eric Foxley – một giáo viên dạy múa, nhạc và vẽ ở Nottingham, Anh sưu tập. Dataset được đăng lần đầu vào ngày 12/10/1996. Sau đó, dataset qua nhiều lần chỉnh sửa và được đăng lần cuối vào ngày 20/06/2003.
- ***Kích thước dataset***: gồm 1037 giai điệu.
    - Jigs (340 tunes)
    - Hornpipes (65 tunes)
    - Morris (31 tunes)
    - Playford (15 tunes)
    - Reels A-C (81 tunes)
    - Reels D-G (84 tunes)
    - Reels H-L (93 tunes)
    - Reels M-Q (80 tunes)
    - Reels R-T (92 tunes)
    - Reels U-Z (34 tunes)
    - Slip Jigs (11 tunes)
    - Waltzes (52 tunes)
    - Chrismas Carols and Songs (13 tunes)
    - Ashover collection (46 tunes)
- ***Cấu trúc dataset***:

[1] http://abcnotation.com/

- Dataset được tổ chức thành từng file với phần mở rộng là .abc. Mỗi file lưu trữ những giai điệu của từng thể loại nhạc.
- Mỗi giai điệu được lưu trữ trong file đều có một cấu trúc nhất định và cấu trúc đó biểu diễn một giai điệu riêng biệt. Ví dụ:

```
X:1
T:Speed the Plough
M:4/4
C:Trad.
K:G
|:GABc dedB|dedB dedB|c2ec B2dB|c2A2 A2BA|
GABc dedB|dedB dedB|c2ec B2dB|A2F2 G4:|
|:g2gf gdBd|g2f2 e2d2|c2ec B2dB|c2A2 A2df|
g2gf g2Bd|g2f2 e2d2|c2ec B2dB|A2F2 G4:|
```



- ***Tiền xử lý:***
  - Một phân tích tiền xử lý cho nottingham dataset đã được trình bày rất chi tiết ở trang github của tác giả Jukedeck[2]. Tác giả xử lý các ký hiệu sai ở phần hòa âm, thêm ký hiệu ở phần lặp lại giai điệu,…
- ***Link download: http://abc.sourceforge.net/NMD/***
- ***Hướng áp dụng deep learning:***
  - John Gamboa [1] đã đề xuất một mô hình Deep Learning là UFCNN. Để chứng minh hiệu suất vượt trội so của mô hình này so với các mô hình đã có,

[2] https://github.com/jukedeck/nottingham-dataset

tác giả đã chạy thực nghiệm với nottingham dataset. Ở thực nghiệm này, mục tiêu là dự báo giá trị của chuỗi thời gian trong bước kế tiếp, tức là dự đoán các nốt kế tiếp trong một giai điệu.

1.2. Trading dataset

Trading dataset là dữ liệu của một cuộc thi về áp dụng machine learning trong dự đoán giá chứng khoán trong tương lai. Mục tiêu chính là đưa ra một số kỹ thuật dự báo giá và tối ưu hóa lợi nhuận trong kinh doanh.

- ***Nguồn gốc***: Cuộc thi Trading Competition[3].
- ***Kích thức dataset***:
    - Dung lượng 426 MB
    - Trading dataset bao gồm 118 file, mỗi file là một file text có định dạng như sau prod_data_yyyymmddv.txt, ví dụ prod_data_20130103v.txt.
- ***Cấu trúc dataset***:
    - Mỗi dòng trong file prod_data_yyyymmddv.txt lưu trữ một chuỗi thời gian về giá hiện tại của cổ phiếu và một số chỉ số có thể hữu ích trong việc phân tích sự thay đổi giá cổ phiếu theo thời gian.
    - Format của mỗi dòng dữ liệu có dạng sau:

milliseconds | Price-Now | BestBidPrice | BestBidSize | BestAskPrice | BestAskSize | Some columns of Indicators ...

39679059 | 4739.316982 | 4739 | 80 | 4740 | 280 | 0.225328 ... etc

Trong đó:

1) *milliseconds* là thời gian bắt đầu một ngày giao dịch theo giờ chuẩn UTC.

2) *Price-Now* là giá hiện tại của cổ phiếu.

[3] http://www.circulumvite.com/home/trading-competition

3) *BestBidPrice* là giá cao nhất của cổ phiếu mà một người sẽ trả để mua cổ phiếu trong thời gian này.

4) *BestBidSize* là tổng số cổ được mua với giá cao nhất.

5) *BestAskPrice* là giá thấp nhất mà một người muốn bán cổ phiếu tại thời điểm này.

6) *BestAskSize* là tổng cổ phiếu có thể bán với giá thấp nhất – BestAskPrice.

7) Những cột còn lại là một số chỉ số khác dùng để tham khảo khi dự báo sự thay đổi giá cổ phiếu.

- ***Link download:*** ***https://s3.amazonaws.com/dvcpublic/training_data_large.zip***
- ***Hướng áp dụng deep learning:***
    - Trading dataset được sử dụng nhiều trong các thực nghiệm của các tác giả. John Gamboa [1] và Roni Mittelman [4] sử dụng Trading dataset trong các thực nghiệm của mình. Mục đích của các tác giả là tìm được một chiến lược đầu tư phù hợp sao cho tối ưu hóa lợi nhuận mang về khi chơi cổ phiếu.

1.3. California Housing dataset

California Housing dataset là bộ dữ liệu thu thập các chỉ số về các khu vực dân cư ở California và từ đó cho biết giá trị trung bình của các căn nhà ở các khu vực dân cư thông qua các chỉ số đó.

- ***Nguồn gốc***: Tập dữ liệu này xuất hiện trong một bài báo "*Sparse Spatial Autoregressions"* vào năm 1997 của Pace, R. Kelley và Ronald Barry, xuất bản trong tạp chí *Statistics and Probability Letters*. Nhóm tác giả xây dựng dataset bằng cách dữ liệu điều tra dân số năm 1990 của California.
- ***Kích thức dataset***:
    - Dung lượng 1.3 MB dưới dạng file .csv.
    - Gồm 20,640 records, mỗi record ứng với số liệu của một khu vực khảo sát nhà ở.
- ***Cấu trúc dataset***:

- Dataset gồm 1 file duy nhất tên là housing.csv. File này gồm 20,641 dòng, dòng đầu tiên là header cho biết các thuộc tính trong tập dữ liệu, từ dòng 2 đến dòng 20,641 là dữ liệu của từng khu vực dân cư.
- Header và dữ liệu của mỗi dòng như sau:

  longitude latitude housing_median_age total_rooms total_bedrooms population households median_income median_house_value ocean_proximity

  -122.23 37.88 41 880 129 322 126 8.3252 452600 NEAR BAY

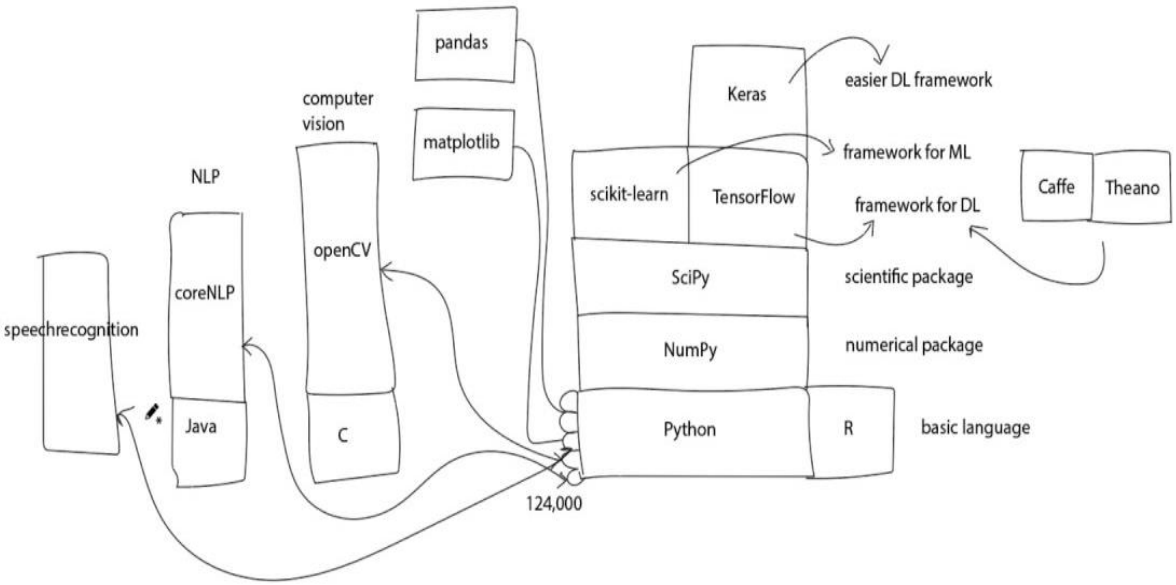
  Trong đó,

  1) *longtitude* là kinh độ khu vực khảo sát
  2) *latitude* là vĩ độ khu vực khảo sát
  3) *housing_median_age* là tuổi trung bình của các ngôi nhà trong khu vực
  4) *total_rooms* là tổng số phòng của các ngôi nhà trong khu vực
  5) *total_bedrooms* là tổng số phòng ngủ của các ngôi nhà trong khu vực
  6) *population* là dân số trong khu vực
  7) *households* là số hộ gia đình trong khu vực
  8) *median_income* là thu nhập trung bình trong khu vực
  9) *median_house_value* là giá trị trung bình ngôi nhà trong khu vực
  10) *ocean_proximity* là vị trí của khu vực ở bang California

- ***Link download:*** ***https://raw.githubusercontent.com/ageron/handson-ml/master/datasets/housing/housing.csv***
- ***Hướng áp dụng deep learning:***
  - Xueheng Qiu [3] và đồng sự sử dụng California Housing dataset để thực nghiệm cho phương pháp ensemble deep learning của mình. Trong thực nghiệm, data cuối cùng được sử dụng gồm 20,640 căn nhà với 9 inputs là các yếu tố có thể ảnh hưởng đến giá căn nhà, trong khi đó output là giá trị trung bình của một căn nhà. Với thực nghiệm này, nhóm tác giả sử dụng dữ liệu của 15,480 căn nhà làm dữ liệu huấn luyện và dữ liệu của 5,160 căn nhà còn lại làm dữ liệu test, theo tỉ lệ 75% - 25%.

2. Setup Deep Learning Environment

2.1. Tổng quan môi trường cài đặt Machine Learning và Deep Learning



Hình 1. Tổng quan môi trường cài đặt Machine Learning và Deep Learning[4]

Với sự phát triển của Machine Learning cũng như Deep Learning, nhiều framework, thư viện hỗ trợ cài đặt, thực nghiệm ra đời. Các framework, thư viện này chủ yếu sử dụng ngôn ngữ lập trình Python và R. Trong đó, Python được sử dụng nhiều hơn cả bởi tính linh hoạt, mạnh mẽ và dễ sử dụng của nó. Với ngôn ngữ Pythong, nhiều package hỗ trợ Machine Learning cũng như Deep Learning ra đời như NumPy, Scipy – thư viện tính toán khoa học. Từ đó, nhiều framework được phát triển để hỗ trợ quá trình cài đặt, thực nghiệm của các nhà nghiên cứu như scikit-learning – framework for Machine Learning, TensorFlow, Caffe, Theano, Keras – framework for Deep Learning. Ngoài ra, Python còn có khoảng hơn 124,000 thư viện hỗ trợ cho các lĩnh vực khác trong nghiên cứu khoa học.

Để bước đầu làm quen với Deep Learning, TensorFlow là framework phù hợp nhất. Với TensorFlow, người dùng không cần phải có kiến thức chuyên sâu về các mô hình toán học và các thuật toán tối ưu để cài đặt được Deep Neural Networks. Người dùng chỉ cần

[4] https://www.youtube.com/watch?v=WQt4H1Bo0jM

download một vài đoạn code mẫu, đọc qua một vài tutorials online để có thể cài đặt hoàn tất mà không hề tốn quá nhiều thời gian.

2.2. Set up Deep Learning Enviroment với TensorFlow

Để set up môi trường cài đặt Deep Learning với TensorFlow, chúng ta phải cài đặt các phần mềm sau:

- Hệ điều hành Ubuntu: hệ điều hành gọn nhẹ và nhận được hỗ trợ nhiều nhất từ các thư viện phát triển Deep Learning. Phiên bản Ubuntu mới nhất hiện nay là Ubuntu 17.10.
- Python 3: việc cài đặt Python trên Ubuntu rất đơn giản. Chúng ta mở Terminal của Ubuntu và gõ lệnh cài đặt Python 3.

```
$ sudo apt-get update
$ sudo apt-get install python3.6

python3 -V
```

- Thư viện pip: thư viện giúp dễ dàng quản lý, cài đặt các thư viện của Python

```
sudo apt-get install -y python3-pip
```

- TensorFlow: khi cài đặt TensorFlow lưu ý phiên bản TensorFlow với chỉ hỗ trợ CPU hoặc TensorFlow hỗ trợ GPU. Quá trình cài đặt TensorFlow khá đơn giản[5] nhưng phải chú ý loại mà TensorFlow hỗ trợ (virtualenv, "native" pip, Docker, Anaconda).

TÀI LIỆU THAM KHẢO

[1] John Gamboa (2017), "Deep Learning for Time-Series Analysis". CoRR, abs/1701.01887.

[2] Takashi Kuremoto, Shinsuke Kimura, Kunikazu Kobayashi, Masanao Obayashi (2014), "*Time series forecasting using a deep belief network with restricted Boltzmann machines*". Neurocomputing, Volume 137, 5 August 2014, Pages 47-56.

[5] https://www.tensorflow.org/install/install_linux

[3] Xueheng Qiu, Le Zhang, Ye Ren, P. N. Suganthan (2014), "*Ensemble Deep Learning for Regression and Time Series Forecasting*". In Proc. IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL'14), Orlando, US, Dec. 2014.

[4] Roni Mittelman (2015), "*Time-series modeling with undecimated fully convolutional neural networks*". arXiv preprint arXiv:1508.00317.

[5] http://www.circulumvite.com/home/trading-competition (Truy cập 25/12/2017).