Design on Experimental Dataset and Task for Teaching in Data Science and Machine Learning

Zhen Chen[1(✉)], Min Guo[1], Li-Rui Deng[1], Wen-Xun Zheng[2], Hao Wang[3], Jia-hui Chang[1], Yi-song Zhang[1], Xiao-Dong Ma[1], Ying Gao[1], Hao-Yu Wang[1], Xin Lu[4], Chao Li[1], Jin Zhou[1], and Shuang Shou Li[1]

[1] Tsinghua University, Beijing, China
zhenchen@tsinghua.edu.cn
[2] Ailibaba Group, Beijing, China
[3] Noyaxe Inc., Beijing, China
[4] PLS Inc., Beijing, China

Abstract. One of the key factor in teaching data science and machine learning is the well-prepared dataset. Although there are some well-known datasets, such as the MNIST, Fashion MNIST and ImageNet dataset are already used in teaching. To inspire students to learn more actively, it needs the instructors to design new experimental tasks and build dataset with the participation of the students. In this paper, we introduce our live examples to design experiments with hand-crafted dataset, crawling-base data, and live generated IoT data.

Keywords: Data Science · Machine Learning · Dataset · Performance evaluation

1 Introduction

1.1 Data Processing

Data science is the practice of using data to try to understand and solve real-world problems [1–4]. The concept of data science was put forward earlier in the last century. In the early stage, it has been in a tepid state due to the lack of data size, programming language, practical software tools, and the support of data analysis methodology.

In the 1980s, the advancement of data mining methods made the development of data science enter a new stage. Entering the 21st century, the field of data science has finally flourished due to the emergence of the ubiquity of Internet connection and mobile Internet, which has led to a sharp increase in the amount of data available. At the same time, with the fast cost reducing in the computing hardware and storage, organizations and enterprises are collecting and storing more data than ever before.

At the same time, in the aspect of the data analysis methodology, there have been breakthroughs in machine learning [5]. Machine learning techniques, especially deep learning [6, 7] have proven to be an efficient way to identify data patterns in practice.

W. Hong and G. Kanaparan (Eds.): ICCSE 2023, CCIS 2024, pp. 186–196, 2024.
https://doi.org/10.1007/978-981-97-0791-1_16

Fig. 1. Data processing steps

Machine learning technologies such as deep learning have a profound impact on the data science and technology industry. In deep learning, as more and more data are input into the Deep Neural Network, the effectiveness of deep learning has improved significantly.

As shown in Fig. 1, data science is flourish and has developed into a relatively comprehensive area. The common steps from data preparation, data management, data visualization and data analysis, a complete set of scientific and technological systems has been developed.

Since 2015, the course teaching group (CTG) has launched three courses about Big data and Machine Learning for teaching and training students in data science and machine learning. The main challenge in courses is the practice process and experiments design after theory learning. Hereafter we will introduce the principle and operation details about the dataset and experiments design in our courses.

1.2 Dataset

Dataset Used in DSML

Dataset is the key element in teaching data science and machine learning (DSML). Well-prepared dataset helps students to learn Machine learning Tasks and inspire themselves to explore the dataset space and to model the dataset with machine learning toolboxes. Although there are some well-known datasets, such as the MNIST [2], Fashion MNIST [1] and ImageNet [3] dataset already used in teaching. These datasets usually are used for toy examples or research purpose without the focus of real problems.

The need to use real data to inspire student to explore the technology frontier more actively requires the instructors to design brand-new course related experimental tasks and build their dataset with themselves, e.g., students participate in the data preparation stage.

In this paper, we introduce our three live examples to design experiments with hand-crafted dataset, crawler base data, and live generated IoT data. Our method shows with participation of student in dataset design, it helps students to learn more actively and more creatively which is helpful in improving the learning experience.

2 Experiment Dataset and Task Design

As the teaching goal in mind, we design three experimental tasks and the preparation of three different kinds of experimental dataset for teaching data science and machine learning.

The three kinds of datasets are introduced as follows:

(1) dataset for intelligent voice control task;
(2) dataset for helmet detect task;
(3) dataset for stock prices prediction task.

3 Intelligent Voice Control Experiment

Dataset for Intelligent Voice Control

This dataset is designed from real scenario with intelligent voice control devices. The intelligent voice-controlled devices are more and more popular as IoT gateway in Smart-House in recently years. The typical representative cases are Apple Home (with Siri), Amazon Echo (with Alexa), Xiaomi MI etc.

A set of typical voice directives in a well-designed intelligent voice-controlled devices are shown in Table 1. There are 24 directives includes the command to operate the device such as take phone call and play music in the Table.

Table 1. 24 Voice directives for device control.

Label	Content	
a	a Bluetooth device start	
b	b Bluetooth device take phone call	bb Bluetooth device phone call
c	c Bluetooth device answer phone call	cc Bluetooth device answer call
d	d Bluetooth device reject phone call	
e	e Bluetooth device play music	ee Bluetooth device music
f	f Bluetooth device stop play music	ff Bluetooth device stop music
g	g Bluetooth device last music	gg Bluetooth device last
h	h Bluetooth device next	hh Bluetooth device next music
i	i Bluetooth device increase volume	ii Bluetooth device increase sound
	iii Bluetooth device volume up	iiii Bluetooth device sound up
j	j Bluetooth device decrease volume	jj Bluetooth device decrease sound
k	k Bluetooth device shutdown	
l	l Bluetooth device battery notice	ll Bluetooth device battery remained
	lll Bluetooth device battery	

Handcrafted Data Description

We have collected about 400 students' voice note data. Each student recorded his or her voices with their mobile phone or laptop. Each student contributes the 24 directives' speech note. These data are voluntarily uploaded by students with privacy confirmed notice.

As speech note may leak some personal identification information, we anonymize the data by converting raw voices into spectrogram as shown in Fig. 2. This also includes the data preprocessing, as the raw data has several different formats. It takes time to handle different file formats, which is harmful for learn experience.

After the above data preprocessing, there are 2469 spectrogram pictures in dataset in total.

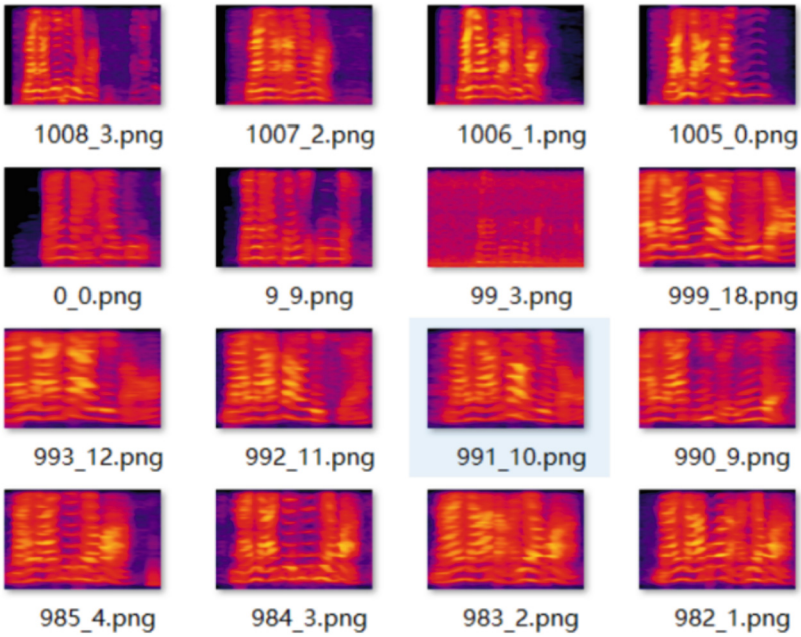

Fig. 2. Partial spectrogram dataset in intelligent voice control experiment.

Data Augmentation

Data augmentation is the common method used to expand the dataset larger in size, which is helpful to fight overfitting in the training process. The concrete approaches in data augmentation include adding random noise in images, random transformation such as image rotation, flipping, zooming etc. In the end, the size of dataset is multiplied with the factor of 6–8. The final results of data augmentation are shown in Fig. 3.

Intelligence Voice Control Task Design

Task design:

(1) Classification into 24 categories;
(2) F1-Score as metrics.

$$F_1 = \frac{precision \cdot recall}{precision + recall} \tag{1}$$

4 Helmet Detection Experiment

Helmet Dataset

The size of helmet dataset is 806 in total. Helmet Dataset includes three categories of photos. One is collected by taking photo by course group, the second is crawled in Internet, and the third is the pure photos of 7 brands. The different categories of pictures are stored in different directories as shown in Fig. 4.

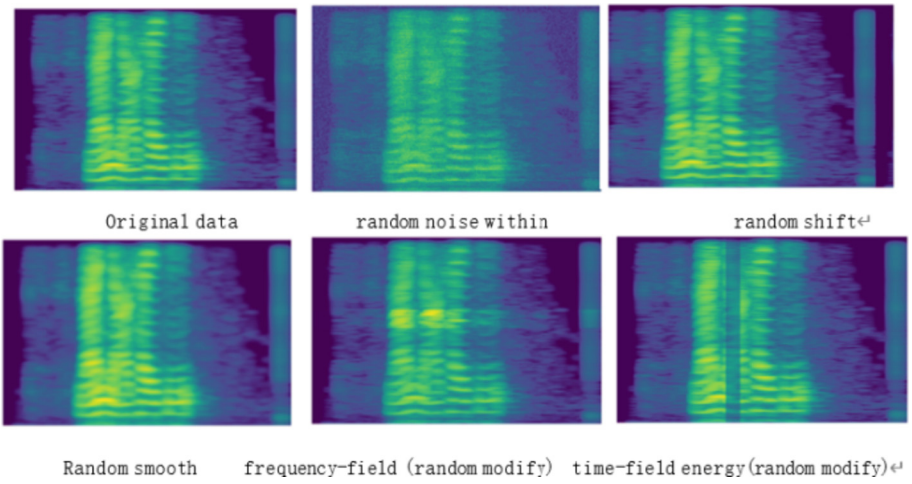

Fig. 3. Data augmentation with different methods.

Hand crafted: 216 photos, 72 students in class, each student wears a helmet and contribute the three photos: Front photo, Side-view photo and back-view photo. It is regular dataset with good definitions.

Collected Data: 200 photos with helmet wearied. With diversity of angles, colors and persons, and picture formats (*.jpg or *.png). Such dataset usually is not very in pixel level resolution.

Helmet only data: 390 with different sizes, seven brands of helmets. It is regular dataset with good definitions.

Fig. 4. Three kinds of data in helmet photo dataset.

Helmet Detection Task Design

As referred in Stanford HAI AI index report [13], TensorFlow2 remained by far the most popular open-source AI software library in 2021, which was followed by Keras, PyTorch and Scikit-learn. For this reason, TensorFlow2 is chosen as deep learning stack in our experiment tasks.

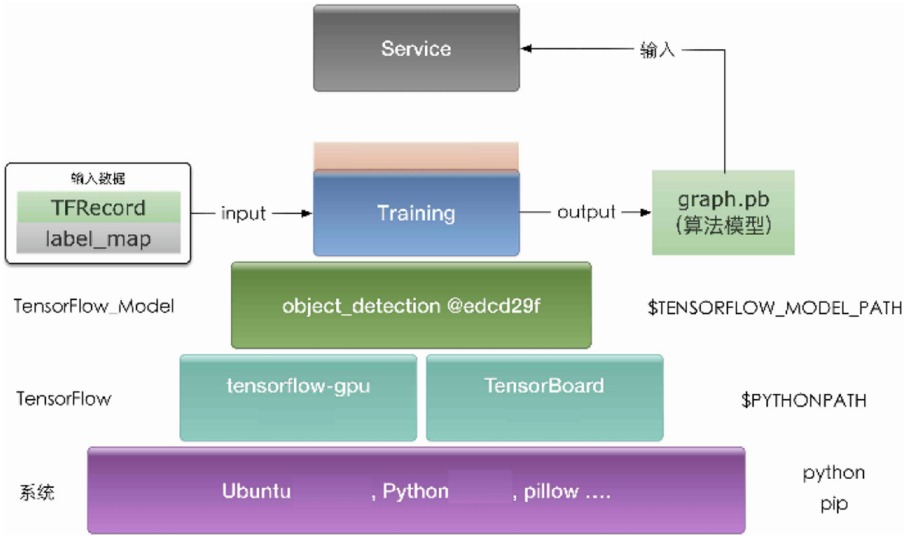

Fig. 5. Experimental environment with TensorFlow software stack.

Experimental Task Goal:

Detect the Objects in photos to find the person wear helmet or not.

(1) Step1. Preparation: Labeling the photo in train dataset, and train deep learning model. Install the requirement software and python library, the dependency in show in Fig. 5.

(2) Task steps:

There are totally 6 steps in processing the data and train the deep learning model after the preparation.

Step 2. Convert: Convert the label_map and raw photos into TFRecord format for using in TensorFlow.

Step 3. Model Training: data input into training pipeline and save the checkpoint regularly.

Step 4. Model Evaluate: Using TensorBoard to visualize the training result.

Step 5. Model Export: frozen the model file when the accuracy reach requirement.

Step 6. Model Detect: Using test data to test object detection results.

Figure 5 shows the detection results of helmets and human in heading helmet from a student's task (Fig. 6).

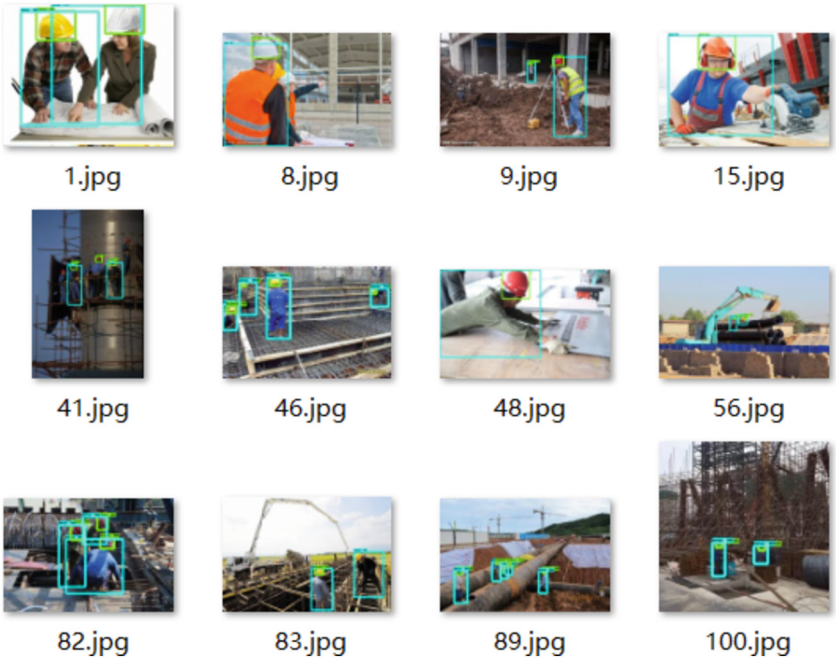

Fig. 6. Detected objects in learning results.

5 Stock Prices Prediction Experiment

Dataset for Stock Mid-Price Movement Prediction

As liquidity plays a vital role in the financial markets, the predicting of limit order book (LOB) by mining micro-structure of high frequency trading (HTF) is often considered a crucial task. LOB is a form of record listing all outstanding limit orders maintained by the exchange, which provides information on available limit order prices and their volumes. Based on these quantities, we provide a limit order book mid-price movement prediction (LOBMMP) dataset for empirical study on stock market as part of quant trading teaching courses.

Data Preparation

(1) Data gathering

To collect a more representative dataset on China stock market, we sampled 10 securities' data of 4 months' trading record (79 trading days in total) with Level-2 stock quote from both Shanghai Stock Exchange and Shenzhen Stock Exchange. The data then has been divided by continuous trading window, meaning the morning data(from 9:30am to 11:30 am) and afternoon data (from 1:00pm to 3:00pm). The daily data is stored in separated csv files and named with the form of '*snapshot_sym* $< xx >$ *_date* $< yy >$ *_am/pm.csv*', containing all 1521 csv files in total, as shown in Fig. 7.

(2) Feature enginering

In the dataset mentioned above, each row of data tables contains 31 attributes in different columns, 26 of which are features of snapshot data, with three types of information in LOB, as shown in Table 2.

For the purpose of avoiding information leaking, we masked the symbol name of security and exact trading day, keeping a sequential number as distinction. To further de-identify the symbol and standardize the price, we extract the split-adjusted factor of each security in all trading days and recalibrate their split-adjusted closing price. Next, we normalize all price data (ask/bid/mid/close) into movement ratio respect to adjusted closing price of previous trading day. For example, if the adjusted closing price of previous day was 50 and now best bid price was 50.05, it will be processed as 0.001 and 49.95 as -0.001.

(3) Data labeling

Last five columns of data provide prediction label, showing the different movement direction of mid-price in different time window. All though there are various types of weighted mid-price definition, we take the most succinct way to calculate by $(ask1 + bid1)/2$. The mid-price varied with ask and bid price is shown in Fig. 8.

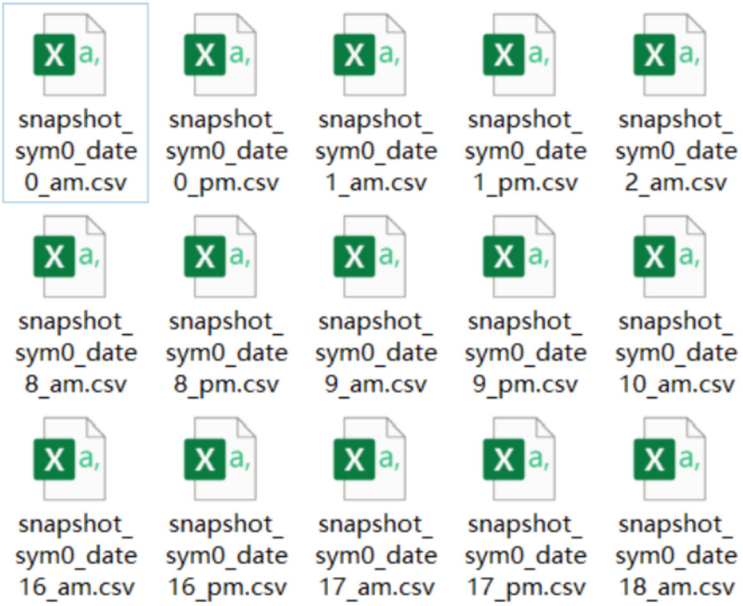

Fig. 7. Daily trade data of 10 stocks.

Table 2. List of provided features

Type	Feature	Description
basic	date	No. of trading date, from 0 to 78
	time	Time of snapshot
	sym	No. of security, from 0 to 9
price	Mid-price	Mid-price of this snapshot
	close	Latest closing price
	ask1–5	The first to fifth best ask price in LOB
	bid1–5	The first to fifth best bid price in LOB
volume	asize1–5	The volume of first to fifth best bid price in LOB
	bsize1–5	The volume of first to fifth best bid price in LOB
	amount	daily cumulative trading volume

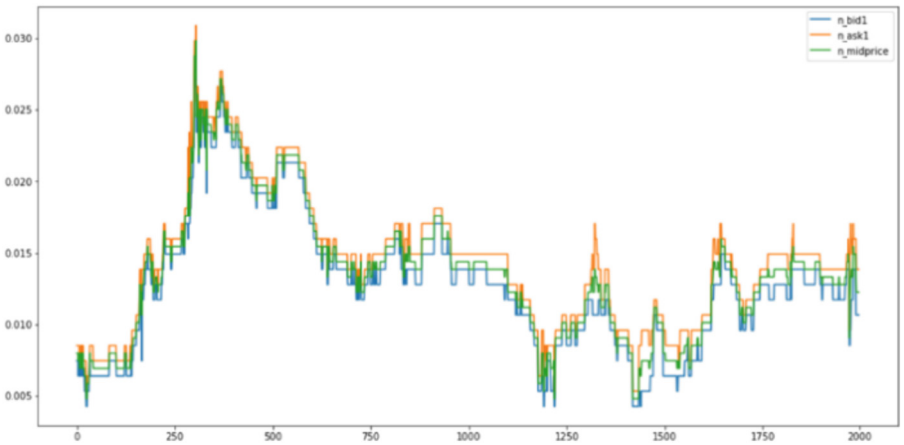

Fig. 8. The mid-price of bid and ask.

In order to better capture the movement momentum of the continuously changing stock price, we define the movement direction with $\varphi(x)$, and label with -1 for move down, 1 for move up, 0 for stationary.

$$\varphi(x) = \begin{cases} -1, & if\ x < -\alpha \\ 0, & if\ |x| \leq \alpha \\ 1, & if\ x > \alpha \end{cases} \tag{2}$$

As volatility tends to grow larger with longer time window, we take $\alpha = 0.05\%$ when predicting movement on 5-tick and 10-tick time windows, and $\alpha = 0.1\%$ with 20, 40 and 60-tick time windows.

Stock Mid-Price Movement Prediction Task Design

This experiment was designed to help students with an empirical study on high frequency stock trading with data exploration. The key-point was encouraging student with feather engineering and model selection by enhancing their domain-specific knowledge.

Task steps:

(1) Understand the task;
(2) Exploratory Data Analysis;
(3) Data check with Python pandas;
(4) Trending predication with SVM (Supported Vector machine);
(5) Trending predicting with DNN (Deep Neural network).

The task benchmark metric is F-score within the test set. As in practical HFT (high-frequency trading) algorithm, the precision usually takes more important role than recall, for this reason we choose F-0.5 score as final result.

$$F_\beta = \left(1 + \beta^2\right) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall} \tag{3}$$

6 Live Generated IoT Data

IoT Datasets

Most IoT device equipped with sensors including temperate, humidity, PM2.5 and light sensor etc. with fast and constantly connectivity, sensor data can be upload into the database in cloud and for further analytic use.

Our experiment uses raspberry Pi as the IoT devices and use DHT11 temperature sensors (Fig. 9).

Fig. 9. Temperature data acquisition.

IoT Data Visualization Task Design:

This experiment helps students to understand the principle of IoT system and implement a simple IOT system.

Experimental process:

– Step 1. Use Raspberry Pi to collect the temperature and humidity data, which are obtained by DHT11 sensor;
– Step 2. Upload it to the server of Influxdb through wireless network;
– Step 3. Configure UI to view the data in the web side.

Experimental results:

The graph of data visualizing shows the real-time temperature and humidity on the web side of the server in the experimental scene.

7 Conclusion

In this paper, we introduce spectrogram data, helmet pictures, stock prices and IoT tasks to design experiments with hand-crafted dataset, crawler base data, and live generated IoT data. All these experiments are designed and made it possible by course teaching group, and are used to teach and train students in Data Science and Machine Learning within three real courses launched by CTG since 2015. The main contribution is let students to take part in the design of dataset generating, cleaning, and to targeted real problems in DSML domain, which inspires the students to learn more actively and have better learning experiences in mastering the knowledge of Data Science and Machine Learning.

References

1. Field Cady, Data Science: The Executive Summary - A Technical Book for Non-Technical Professionals, Wiley, Hoboken (2021)
2. Godsey, B.: Think Like a Data Scientist: Tackle the data science process step-by-step. Simon and Schuster (2017)
3. Cristina Mariani, M., et al.: Data Science in Theory and Practice, Techniques for Big Data Analytics and Complex Data Sets, 1st Edn, Wiley, Hoboken (2022)
4. Neal Fishman with Cole Stryker, Smarter Data Science - Succeeding with Enterprise-Grade Data and AI Projects, Wiley, Hoboken (2020)
5. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015)
6. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
7. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
8. LeCun, Y., et al.: Handwritten digit recognition with a back-propagation network. In: NIPS (1989)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE CVPR (2009)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS (2012)
11. Graves, A., et al.: Speech recognition with deep recurrent neural networks, In: ICASSP (2013)
12. Gao, H., et al.: Densely connected convolutional networks. In: CVPR (2017)
13. Stanford HAI. https://aiindex.stanford.edu/report/
14. Big data and machine intelligence. https://www.xuetangx.com/course/thu08091010085/18960834
15. Financial big data and quantitative analysis platform. https://www.icenter.tsinghua.edu.cn/dashboard/entry.html