RD2 Engineer Training Second Review

Advertisement Content Classifier

Yen-Chun, Huang

RD2 ML Engineer

ychuang@eland.com.tw

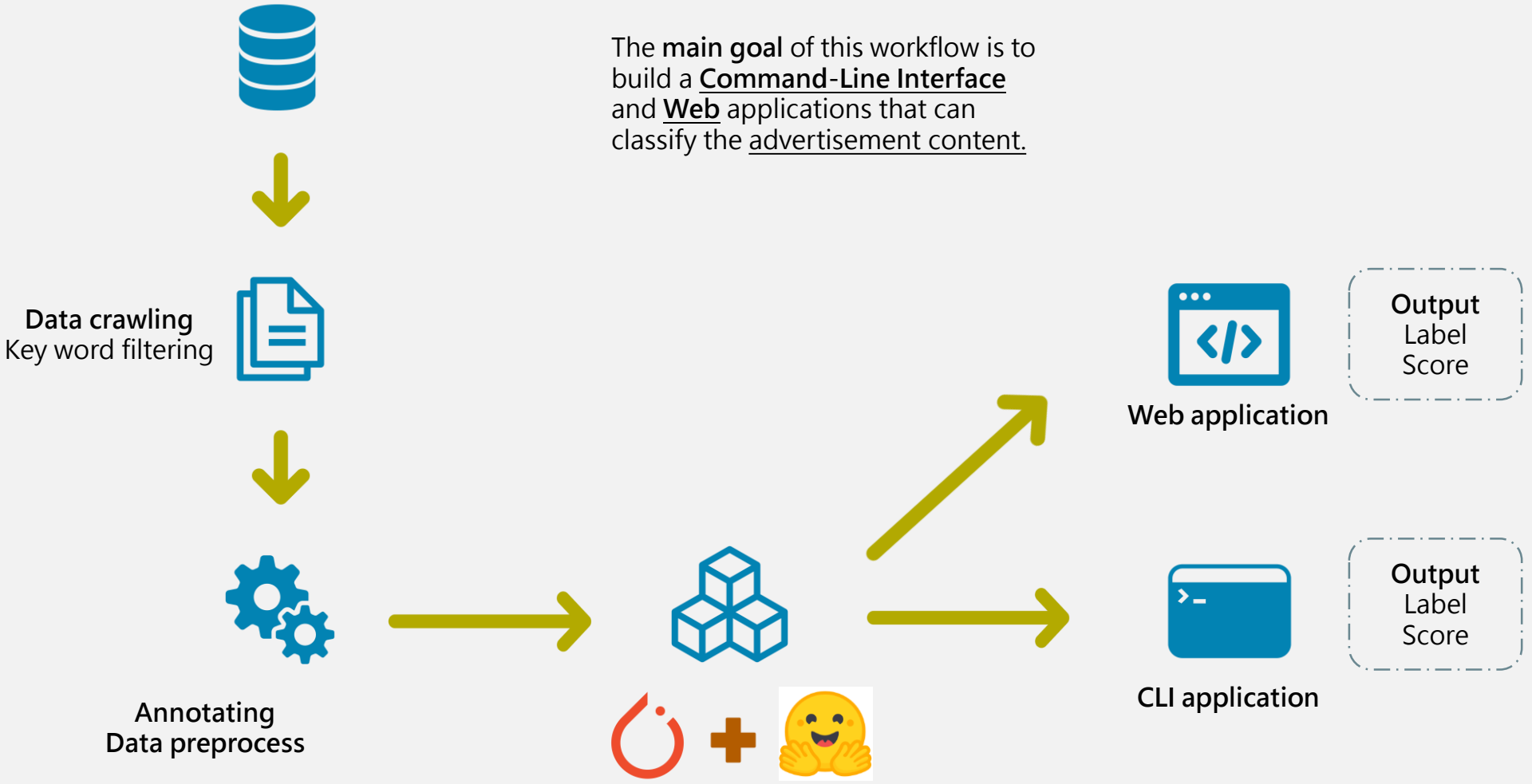
Development Goals

- **Input / Output** : String content / Label & Probability
- **CLI app**
- **WEB app**
- **Logging** : use logging in the development
- **Env var** : Set environment variable
- **Git** : Use Git track the development progress & use branch to split the subtask
- **Model evaluation** : $f_1 score > 0.8$; $precision > 0.9$

ADCC

Workflow

The **main goal** of this workflow is to build a **Command-Line Interface** and **Web** applications that can classify the advertisement content.



$precision$: 0.91
$recall$: 0.90
$f_1\ score$: 0.90

Problem statement

- **Advertisement copywriting definition:**

An copywriting which is developed from the advertisement industry is the manuscript used in newspapers, magazines, posters and other print media or electronic media, TV commercials, web banners, etc., to promote products, companies, or ideas, or people who do this.

Data source

1. Using the **pymysql** to access the database
   - Use environment variable to set host, user and password
2. Select **title** and **content** columns
3. Scrap the data and use **keyword pattern** to filter the **positive, negative** samples.
4. In this stage, I build small datasets **positive** and **negative** for each contains 250 data

Data preprocess

1. Check the dataset and annotate manually
2. Using **jupyter** and **pandas** to perform a primary data analysis
3. Merge the **title** with **content**
4. Concatenate the **positive** sample with **negative** sample
5. Generate an output file includes **500** raw data contains **text content** and **binary label**

Data description

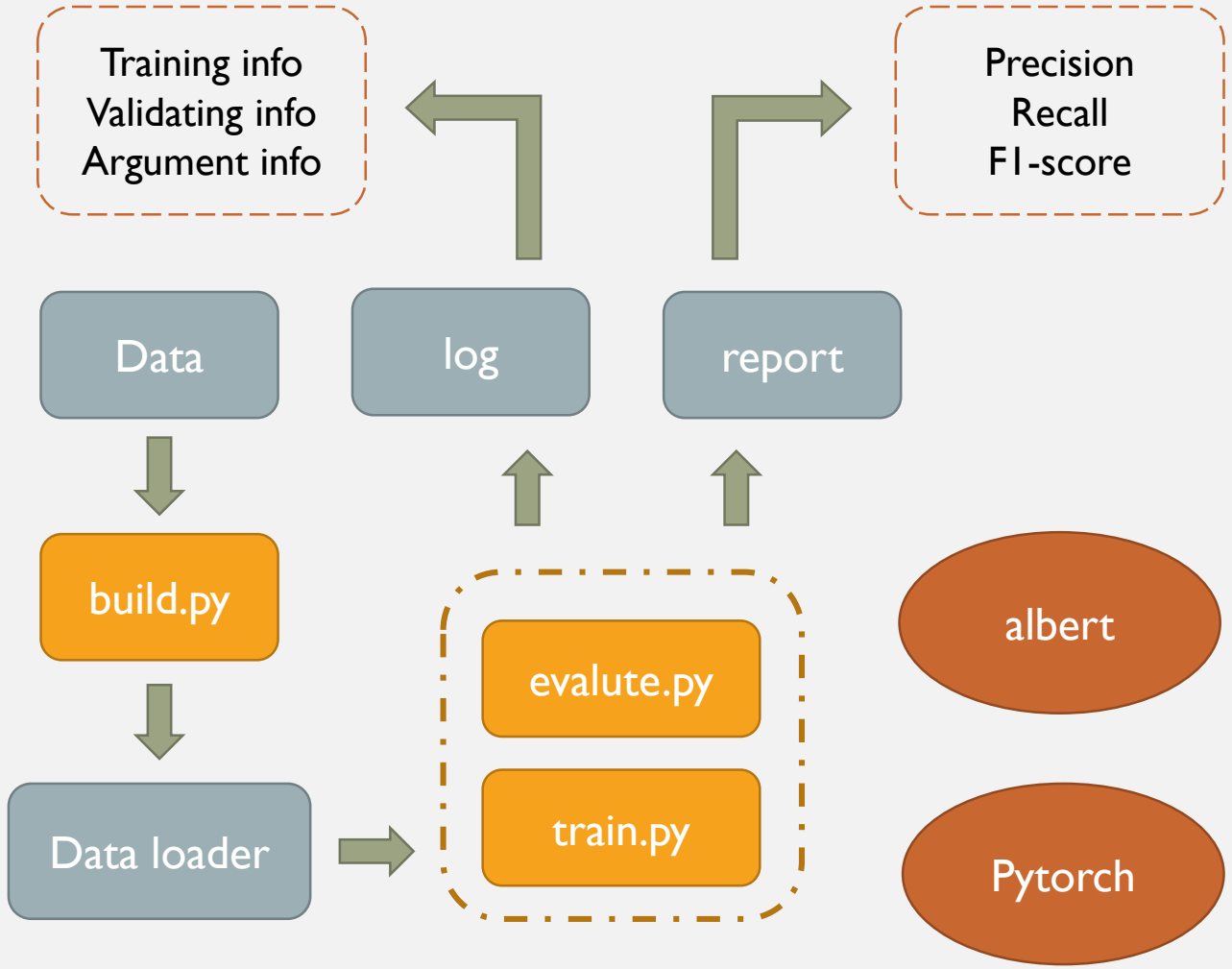
	label	title_length	content_length	text_length
count	500.00	500.00	500.00	500.00
mean	0.50	19.42	479.71	499.13
std	0.50	10.75	706.56	709.19
min	0.00	1.00	0.00	1.00
25%	0.00	11.00	69.00	88.75
50%	0.50	18.00	253.50	268.00
75%	1.00	26.00	507.75	536.50
max	1.00	68.00	5977.00	6001.00

Top words (AD):

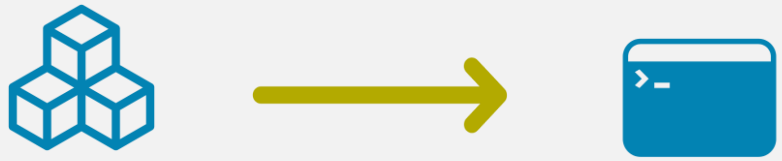
2018 館、台北、車展、健康、保濕、酒店、快樂、髮、霜、親子館、店、遊戲、公園、台灣、專業、餐廳、精、美食、親子

Training & Evaluating

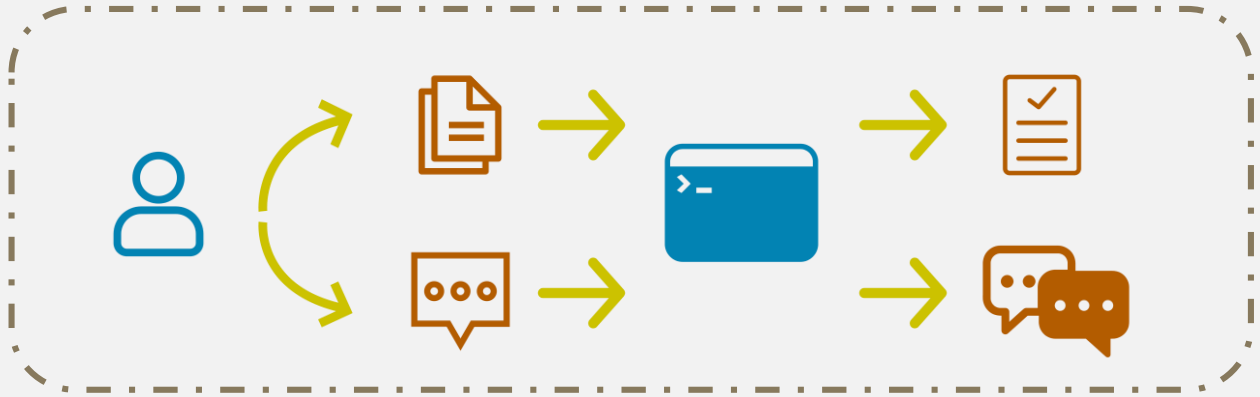
1. Introducing **albert** from **huggingface transformers**
2. Build **data loader** object to fit in the pre-trained language model.
3. Split dataset into training and validation set (8 : 2)
4. Using **Pytorch** to build classifier
5. Create **training** and **evaluation** strategies
6. Store the training **log** and classification **report**



CLI application



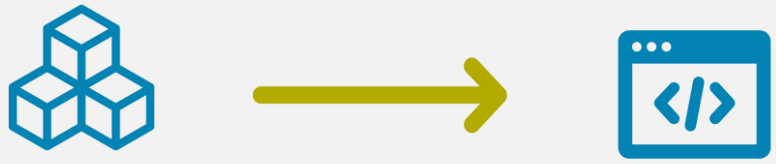
1. First, I tried to build a primary version that can allow the user to input **string content** in **Command-Line Interface (CLI)** and output the **result.**
2. How about some users who require the multiple input? Such as a .txt file input?
3. I extended the command application that can also allow user to input a file location, and generate a new file including the **{label : probability}**



台中市玩具圖書館就在...也歡迎分享出來給玩具圖書館哦。

This content is an AD article, the probability is 0.948

WEB application



1. In this application, I apply **fast API, html, css** and **jinja2** to build a WEB service of classifier.
2. Users can type the **content** inside the **text area** and add the new content by pressing ENTER.
3. The WEB service will calculate and show the result.

Advertisement Classifier

Hint: Input string and return the result (label & probability)

Type here, change line (press ENTER) if you want to add new content

Submit Reset

No result yet …

WEB application

Advertisement Classifier

Hint: Input string and return the result (label & probability)

台中市玩具圖書館就在台中市三十張犁婦女福利服務中心的二樓，這裡也是台中市新移民家庭服務中心。
來這裡玩的小朋友限6歲以下，目前只有開放週五跟週六兩天，跟其它的玩具圖書館規則差不多，玩具只能在館內使用，每次可以借3種玩具，玩完一個才能借下一個，這裡的玩具有區分為益智類、語言類、社會生活類、操作類、嬰幼兒類五種。
因為空間有限，所以每次最多只能有15位大人加小孩進場，若家裡有多的不用玩具，也歡迎分享出來給玩具圖書館哦。
我以為我高中不會有這種問題結果竟然在2017的最後一天發生事情是這樣的我們一群人一起去跨年晚會之後越來越晚有些人因為家住比較遠的關係就先回家了 最後就剩我和另外兩位很好的朋友我本來以為會剛剛的氛圍一樣大家開開心心的聊著結果 哈哈哈哈我錯了我被丟在後面因為人很多啊 我又很矮想說放慢速度他們會不會發現我沒跟上結果 也沒有哈哈哈哈哈後來是其中最常跟我膩在一起的朋友把我拉進去結果 他被另一個拉走 ...然後他們就自己講自己的自己笑自己的 啊我也不知道要怎樣就邊走邊滑手機 假裝很忙哈哈哈哈最後 那個跟我膩在一起的朋友也要回家了我跟他來個大擁抱 然後說個再見我就直接自己去搭公車回家了我就沒理另一個朋友其實這也不是第一次發生只是這次真的很尷尬有一次是在學校我們三個聚在一起我突然想到我有事情還沒做完 我就先回去我的座位結果那個同學就說哦太好了 ***（我的名字）走了 我可以跟你講事情了他要講就講 沒事講這句幹嘛好啦 我只是想抱怨可能我太玻璃心

Submit Reset

No result yet...

Success !!

Output a table of id, label and probability. Id indicate the content order

Advertisement Classifier

Hint: Input string and return the result (label & probability)

Type here, change line (press ENTER) if you want to add new content...

Submit Reset

You input 2 text contents, the results of them:

Id	Label	Probability
1	positive	0.9662
2	negative	0.9979

Result

The best model result (run 33)

	precision	recall	f1-score	support
0	0.98	0.84	0.90	55
1	0.83	0.98	0.90	45
accuracy	0.90	0.90	0.90	0.90
macro avg	0.90	0.91	0.90	100
weighted avg	0.91	0.90	0.90	100

{0: negative, 1: positive}

parameter	value
device	cpu
n_classes	2
epochs	10
batch_size	32
learning_rate	5e-6
max_len	300
loss_func	CrossEntropyLoss

Discussion

1. **False prediction**
   - Annotation Error
   - Prediction Error
2. A non-AD content has more probability to be false predicted as an AD content.
3. All of AD content which are false predicted as non-AD content are all **Prediction Error** in with sponsored post or product promotion content.

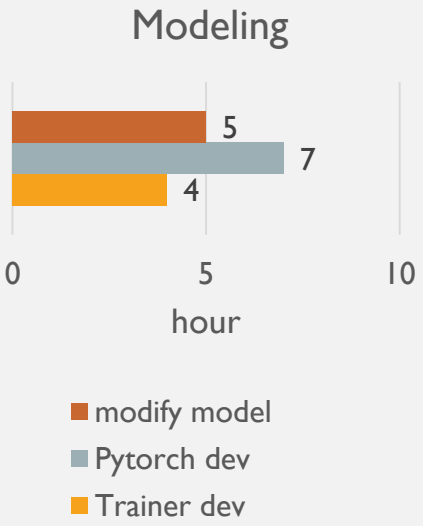
Conclusion

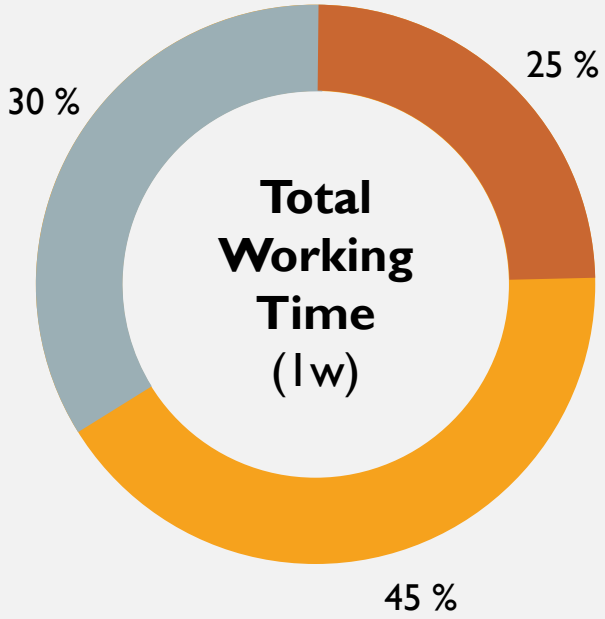
- **Data:**

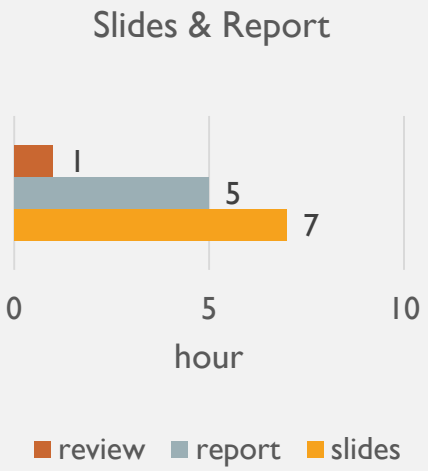
1. More thoroughly understand the data source
2. Add more data patterns to annotation step

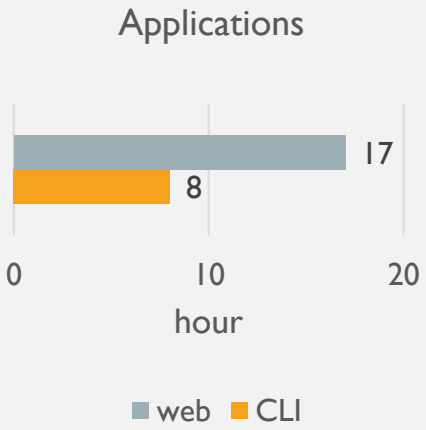
- **Model:**

1. Best model hyper parameter and dynamic tuning technique
2. Feature engineering for annotation and training
3. Tokenizing implement for optimize the tokenizer

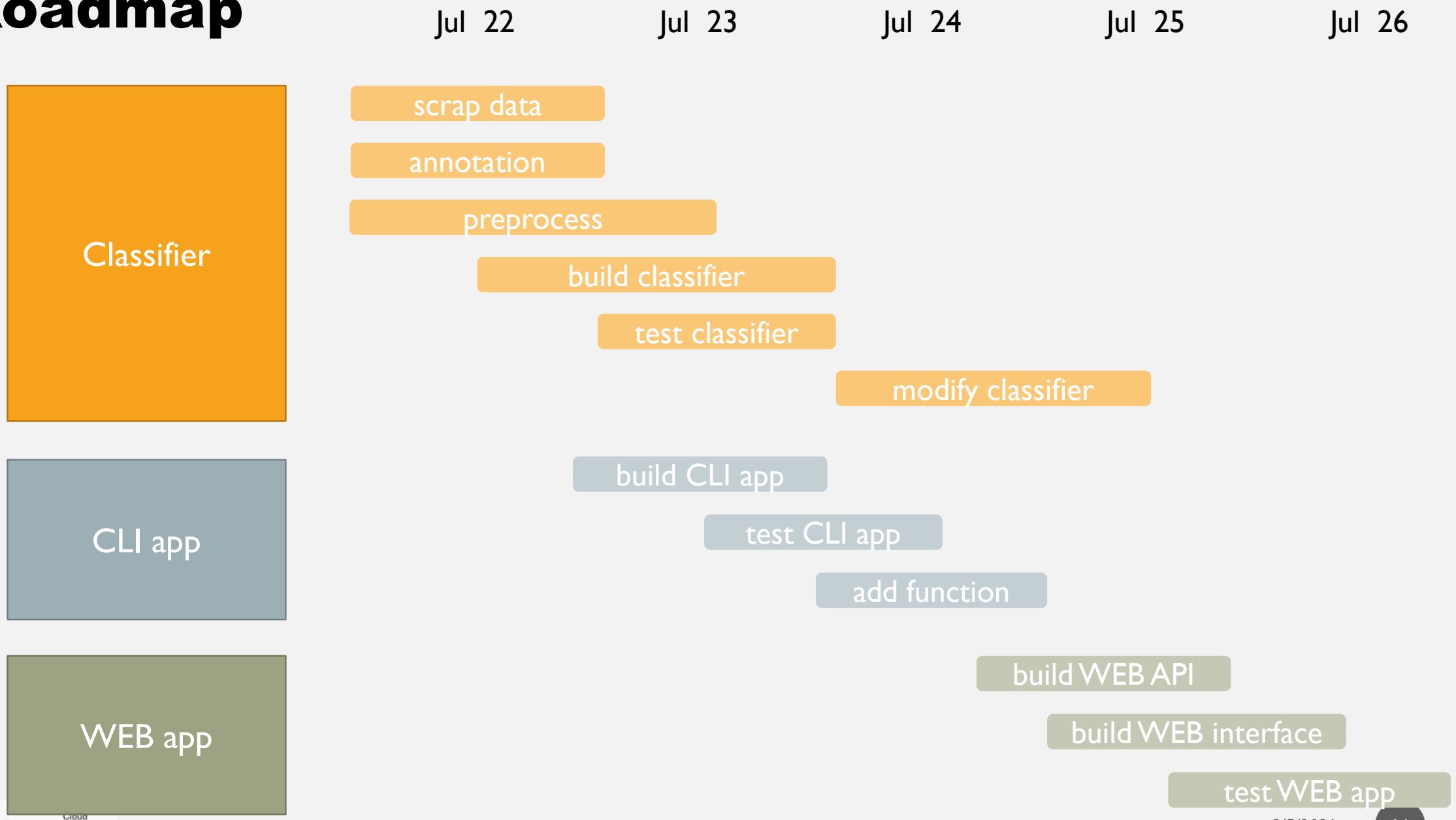

Modeling
5
7
4
0
5
10
hour
modify model
Pytorch dev
Trainer dev


30 %
25 %
45 %
Total Working Time (1w)

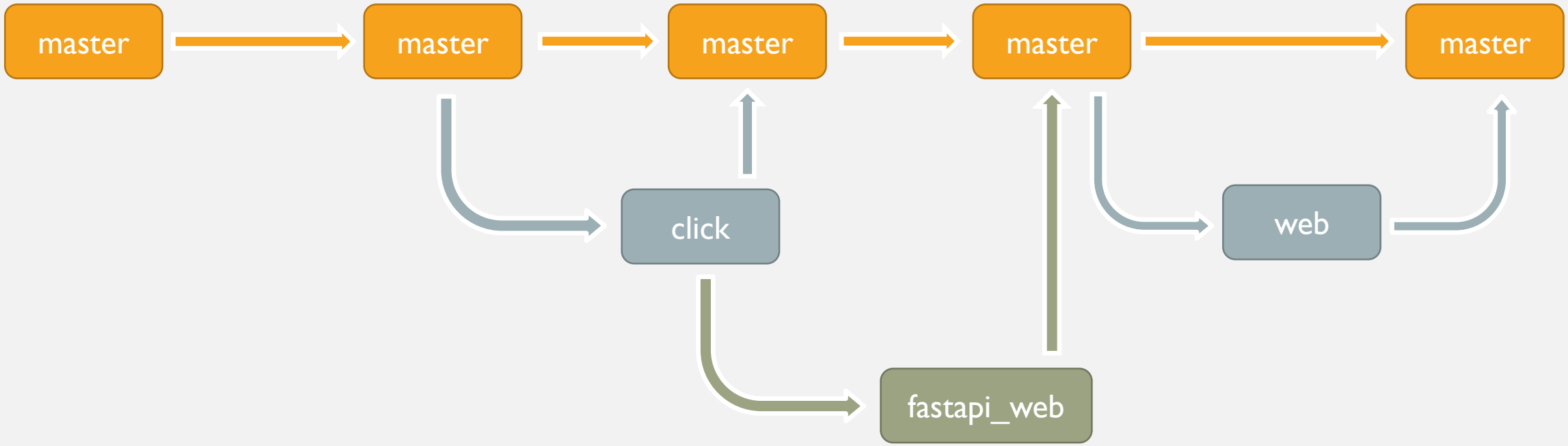

Slides & Report
1
5
7
0
5
10
hour
review
report
slides


Applications
17
8
0
10
20
hour
web
CLI

Roadmap


Jul 22
Jul 23
Jul 24
Jul 25
Jul 26
Classifier
scrap data
annotation
preprocess
build classifier
test classifier
modify classifier
CLI app
build CLI app
test CLI app
add function
WEB app
build WEB API
build WEB interface
test WEB app

Git Roadmap



Development Goals

- **Input / Output** : String content / Label & Probability
- **CLI app**
- **WEB app**
- **Logging** : use logging in the development
- **Env var** : Set environment variable
- **Git** : Use Git track the development progress & use branch to split the subtask
- **Model evaluation** : $f_1 score > 0.8$; $precision > 0.9$

Thanks for listening !!