第壱話

廣東話、自肥企画


麥當勞有幾個爸爸？

答案： 2.5個

「爸爸爸爸爸　♫
I ‘m loving it 」

柯南病咗，有邊個會探佢？

答案：明禎探柯南

Scotty Kwok

- Co-founder of Sebit
- We develop AI/Computer Vision solutions
- **We are hiring software engineers / mechanical engineer / project coordinator**

scottykwok@sebit.world

The code and materials will be shared here

https://github.com/scottykwok/cantonese-selfish-project

議程

Data Science

- **Part 1 -** 了解廣東話 (了解你嘅 data)
- **Part 2 -** 廣東話語料庫 (收集 data)
- **Part 3 -** 廣東話 **Python Library** (整理 data)
- **Part 4 -** 廣東話語音辨識 (訓練 AI 模型)

Part 1

了解廣東話

粵拼 Jyutping

- 香港語言學學會粵語拼音方案(粵拼)，是由香港語言學學會 (LSHK) 於1993年製訂的粵語羅馬化拼音方案。目的在於以一套簡單、合理、易學、易用的粵語語音轉寫方案來統一社會各界在粵語拼音使用上的混亂情況
- 優點 (參考文章：「點解要用粵拼，唔用其他粵語拼音方案？」)
  - 一個符號對應一個音位
  - ASCII 字符
  - 符合大衆習慣
  - 準確反映語言音位系統

Source: Wikipedia 香港語言學學會粵語拼音方案

聲母 19 個

b	巴
p	怕
m	媽
f	花
d	打
t	他
n	那
l	啦
g	家
k	卡
ng	牙
h	蝦
gw	瓜
kw	誇
w	蛙
z	渣
c	叉
s	沙
j	也

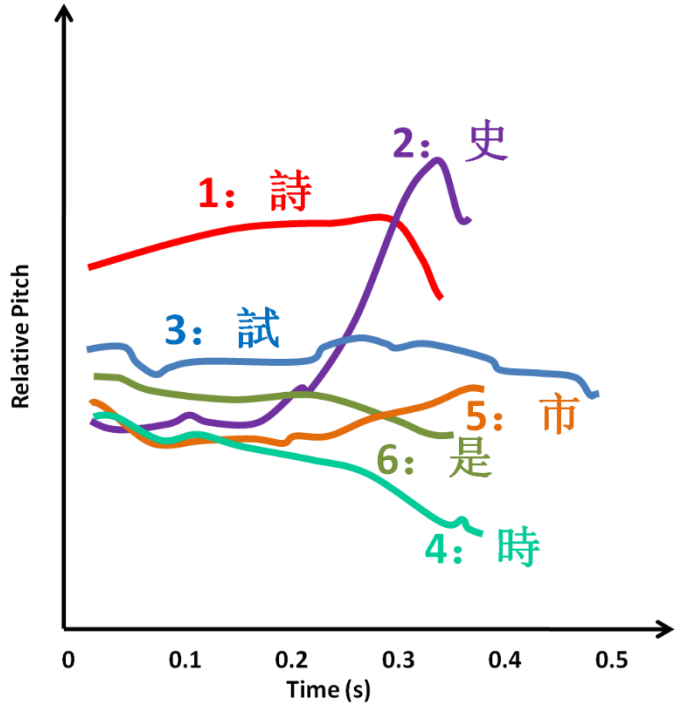
×

韻母 60 個

i 思	ip 攝	it 洩	ik 識	im 閃	in 先	ing 升		iu 消
yu 書		yut 雪			yun 孫			
u 夫	up	ut 闊	uk 福	um	un 歡	ung 風	ui 灰	
e 些	ep	et	ek 石	em	en	eng 鄭	ei 四	eu
		eot 摔			eon 詢		eoi 需	
oe 鋸		oet (*)	oek 腳			oeng 疆		
o 可		ot 喝	ok 學		on 看	ong 康	oi 開	ou 好
a (*)	ap 汁	at 侄	ak 則	am 斟	an 珍	ang 增	ai 擠	au 周
aa 渣	aap 集	aat 扎	aak 責	aam 站	aan 讚	aang 掙	aai 齋	aau 嘲

×

六調



Source: Wikipedia
https://www.lshk.org/jyutping

是

si6

詩史試時市是
色　錫　　食

si^1 si^2 si^3 si^4 si^5 si^6
sik^1 sek^3 sik^6

明	禎
ming4	*zing1*

名	偵	探	柯	南
ming4	*zing1*	*taam3*	*o1*	*naam4*

爸	爸
baa4	*baa1*

爸	吧	霸	把	霸
baa4	*baa6*	*baa3*	*baa2*	*baa6*

唱歌

一個人(刃)原來都可以盡興

jan4 jan6

潮語

唔 好 咁 hea

m4 hou2 gam3 he3

港式英文

Professor

pou6 fe1 saa4

pou6 fe1 saa2

個「Sir」字點寫？

灕 滑 梯

soe4 waat6 tai1

我個貧友去香趕痕身銀寒聽趕座，佢躉咗好耐，割得啲懶氣曾係突別凍😷

我 個 朋(貧) 友

pang4 pan4

淆演，移蛋，墨演，懶肉，麻勒山小，忍唔忍嘢？

牛 演(丸) 移(魚) 蛋

ngau4 jin2 (jyun2) ji4 (jyu4) daan2

Part 2

廣東話語料庫

何謂「 語料庫 」？(Corpus)

- 語料庫 (corpus) 在語言學上意指大量的文本，通常經過整理，具有既定格式與標記

- 語料庫語言學 (corpus linguistics ）是基於語言運用的實例（即語料庫）的語言研究。

- 語料庫語言學可以對自然語言進行語法與句法分析，還可以研究它與其他語言的關係。

References: https://zh.wikipedia.org/zh-hk/語料庫語言學

粵語語料庫

資料年份 Year	名 Name	License 版權	語音參與者 Voice Participants	連結 Link
1950 - 1970	香港二十世紀中期語料庫	(?)	21套50-60年代粵語長片	https://hkcc.eduhk.hk/v1/introduction.html
1994	Lee / Wong / Leung Corpus	For academic research	8 位兒童	https://childes.talkbank.org/access/Chinese/Cantonese/LeeWongLeung.html
1997 - 1998	香港粵語語料庫 (HKCanCor)	CC-BY	93段 2至4人的對話	http://compling.hss.ntu.edu.sg/hkcancor/
2004	CUCorpora	Industrial Research: HK$ 15,000 Academic Research: HK$ 7,500 Commercial Use: HK$ 45,000	80人	http://dsp.ee.cuhk.edu.hk/licensing/cucorpora/Documents/CUCorpora_description.pdf
2008	Paidologos Corpus: Cantonese	For academic research	80人	https://phonbank.talkbank.org/access/Chinese/Cantonese/PaidoCantonese.html
2014 - Now	粵典	Public Domain (except 粵文庫)	(文字語料)	https://words.hk/
2015 - 2018	Leo Corpus	For academic research	1 位兒童	https://childes.talkbank.org/access/Biling/Leo.html
2018 - 2020	SpiCE: Speech in Cantonese and English	CC-BY	34 位加拿大兒童	https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=doi:10.5683/SP2/MJOXP3
2020	馬來西亞粵語語料庫 (MYCanCor)	CC-BY	56個場景 2至4人的對話	https://github.com/liesenf/MYCanCor
2020	KddRES	For academic research	10間餐廳 800段對話	https://github.com/ruleGreen/KddRES
2019 - Now	Mozilla Common Voice	CC-0	>2600人	https://commonvoice.mozilla.org/zh-HK/datasets

以上只係一部份廣東話語料庫，未能盡錄，Sorry

粤典

粵典

https://words.hk/

- 《粵典》係一個大型嘅粵語辭典計劃。
- 用Crowd-sourcing 做一本大型、可持續發展嘅粵語辭典
- 用香港粵語（廣東話）做中心，編寫解釋同例句都會考慮香港嘅用法
- 務求反映真實嘅語言現狀

粵典

https://words.hk/

「騎呢 / 奇離」

解釋 #1

寫法、讀音：
- 騎呢 *ke⁴ le⁴*
- 奇離 *ke⁴ le⁴*

詞性： 形容詞

解釋： *(廣東話)* 古怪、唔正常（通常指衣着打扮、性格同行為）
(英文) odd; weird (usually used to describe outfits, personalities or behaviour)

配詞 / 用法：

ke4 le4 gwaai3
(粵) 騎呢怪
(英) weirdo

例句：

keoi5 hou2 ke4 le4 gaa3 *wui5 mou4 laa1 laa1 hai2 dou6 coeng3 go1*
(粵) 佢好騎呢㗎，會無啦啦喺度唱歌。
(英) He's really odd. He would start singing all of a sudden.

粵典

https://words.hk/

Common Voice

moz://a

Common Voice

moz://a

- Crowd-sourcing
- 建立一套**語音資料集**
- 開放原始碼
- 多國語言
- 可用於機器學習

Common Voice

moz://a

	錄音人數	已驗證鐘數	
中文（香港）	2,699	97h / 1.2k	貢獻
中文（臺灣）	1,625	61h / 1.2k	貢獻
英文	77,913	2.1k / 2.5k	CONTRIBUTE
盧旺達文	1,039	2.0k / 2.5k	FASHA, TANGA UMUSANZU
世界語	1,255	1.1k / 1.2k	KONTRIBUI
德文	16,010	1.0k / 1.2k	MITMACHEN
加泰隆文	6,304	833h / 1.2k	COL·LABOREU-HI
法文	15,724	795h / 1.2k	CONTRIBUER
卡拜爾文	1,435	549h / 1.2k	TTEKKI

https://commonvoice.mozilla.org/zh-HK/languages

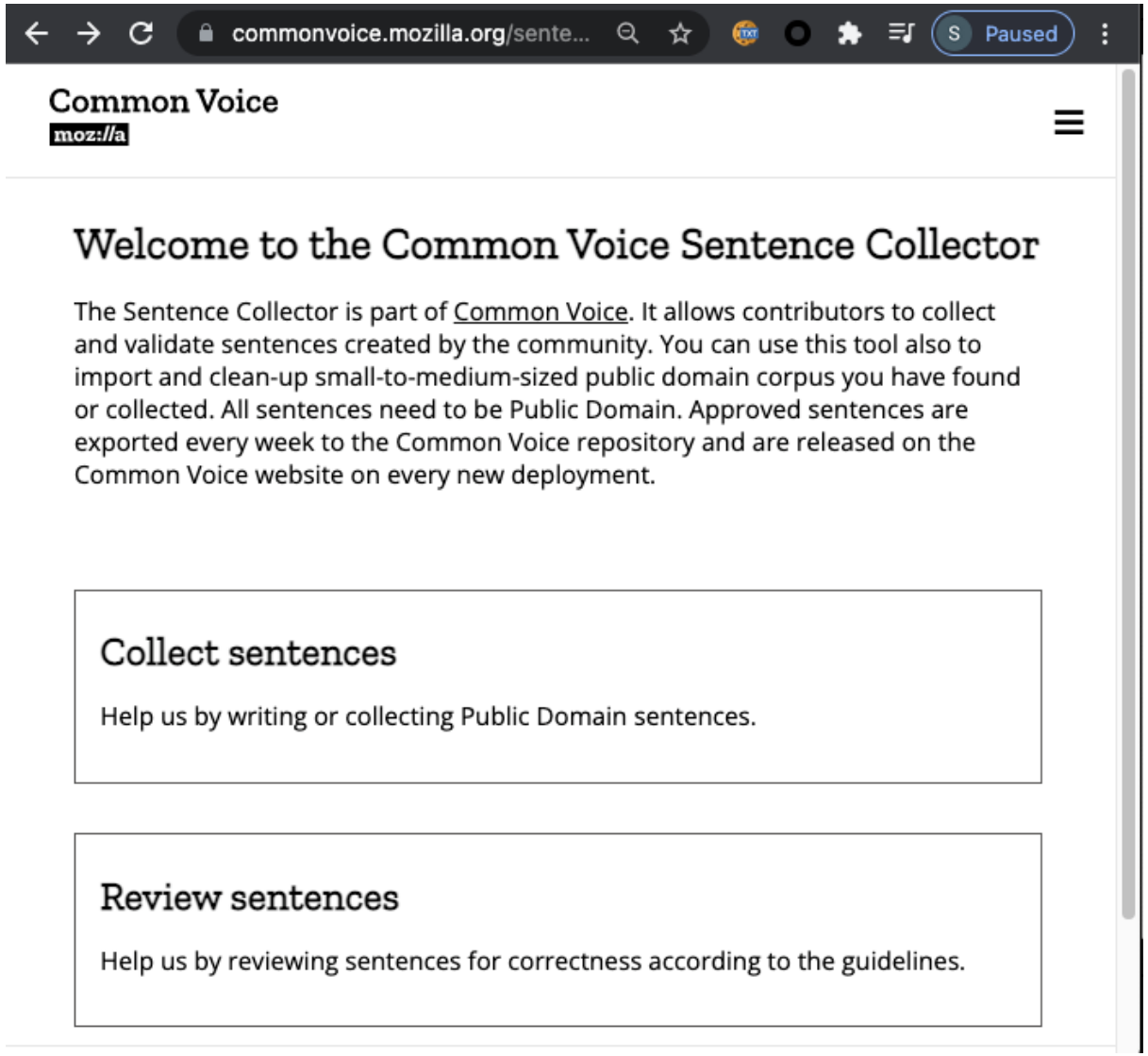
Common Voice
moz://a

講

聽

Common Voice
moz://a

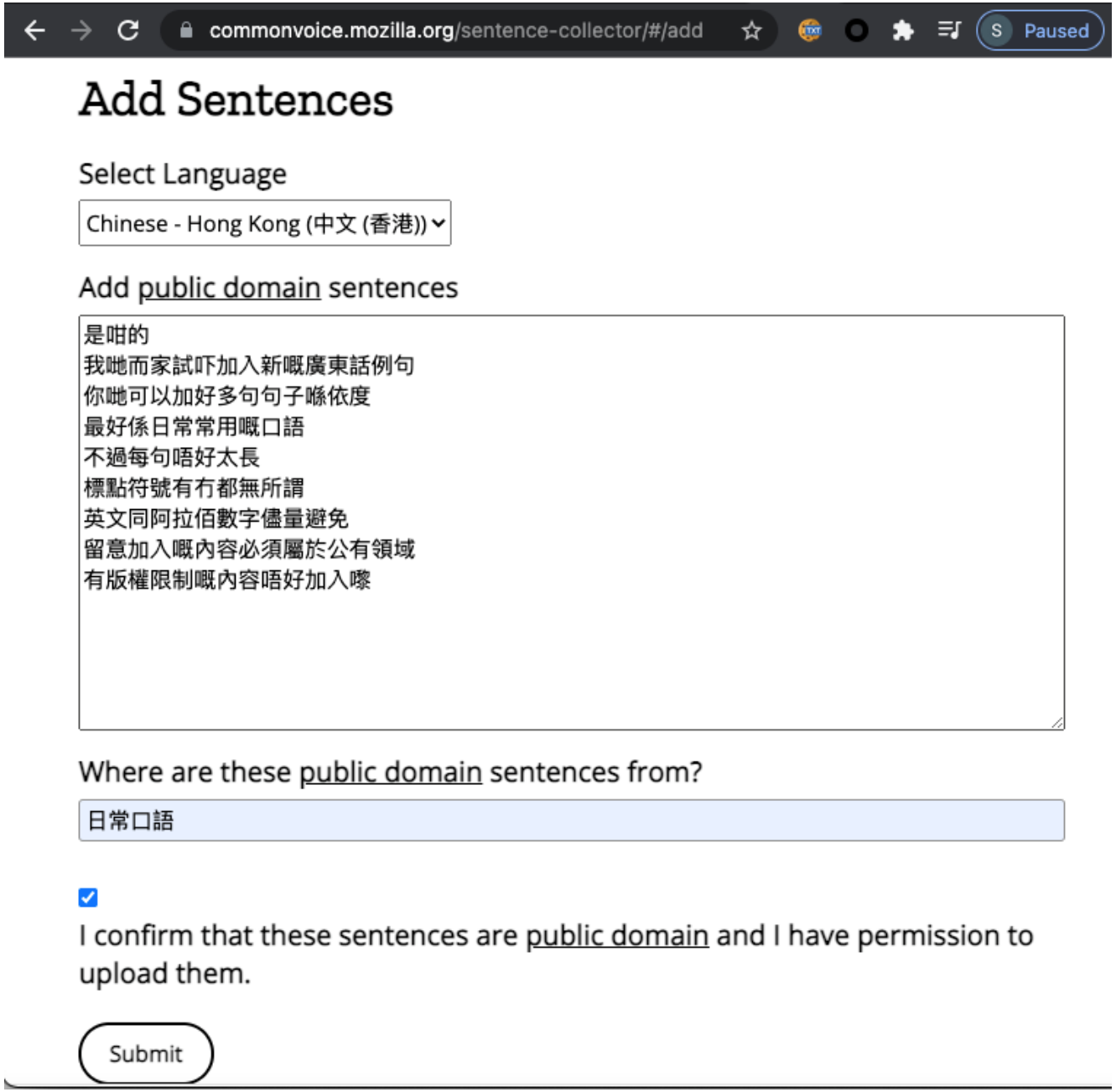
點樣用 Sentence Collector 加新嘅句子？



https://commonvoice.mozilla.org/sentence-collector

Common Voice
moz://a

點樣用 Sentence Collector 加新嘅句子？

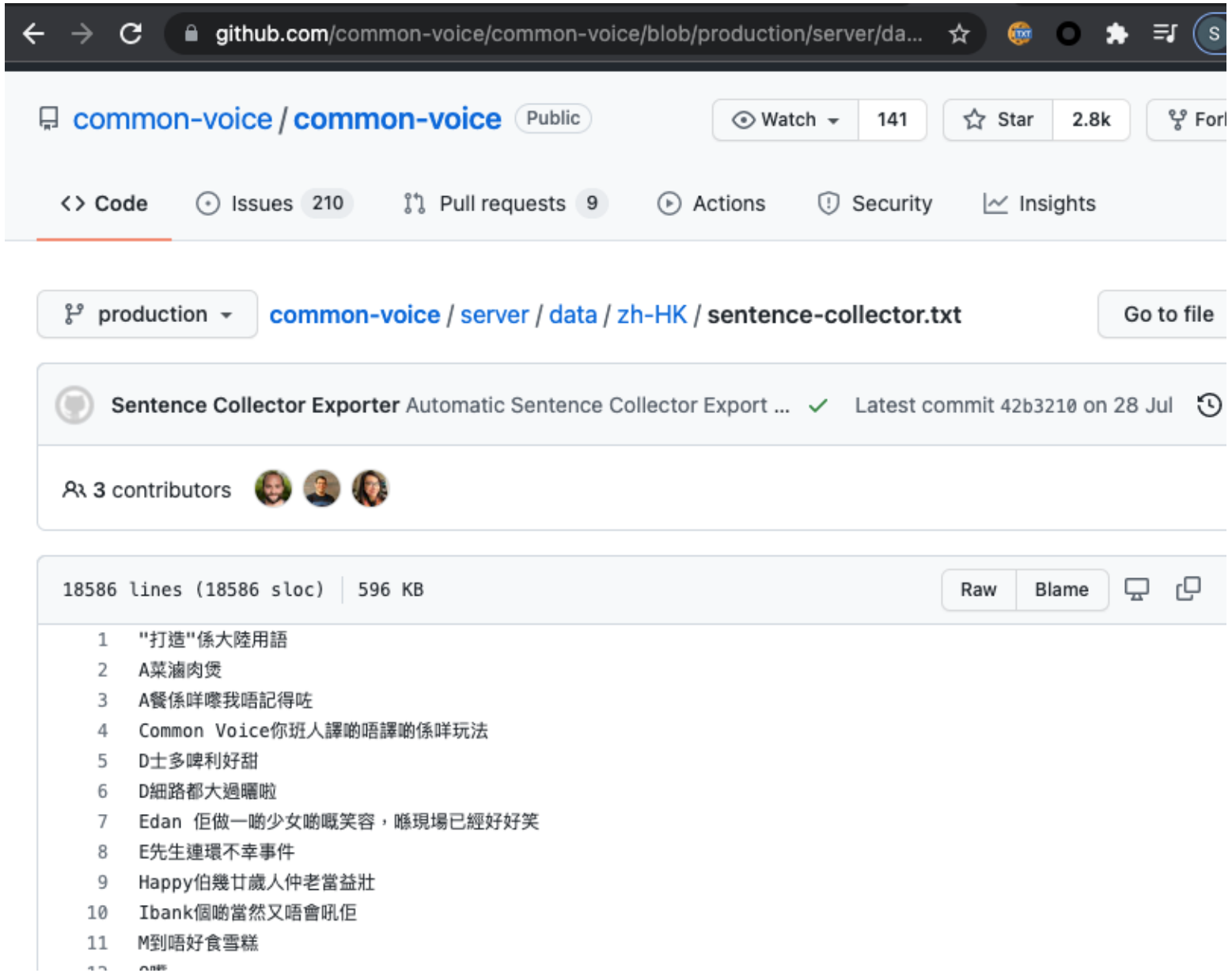


Common Voice
moz://a

有3 票 Approve
句子就會加入語料庫

在GitHub可以見到所有句子

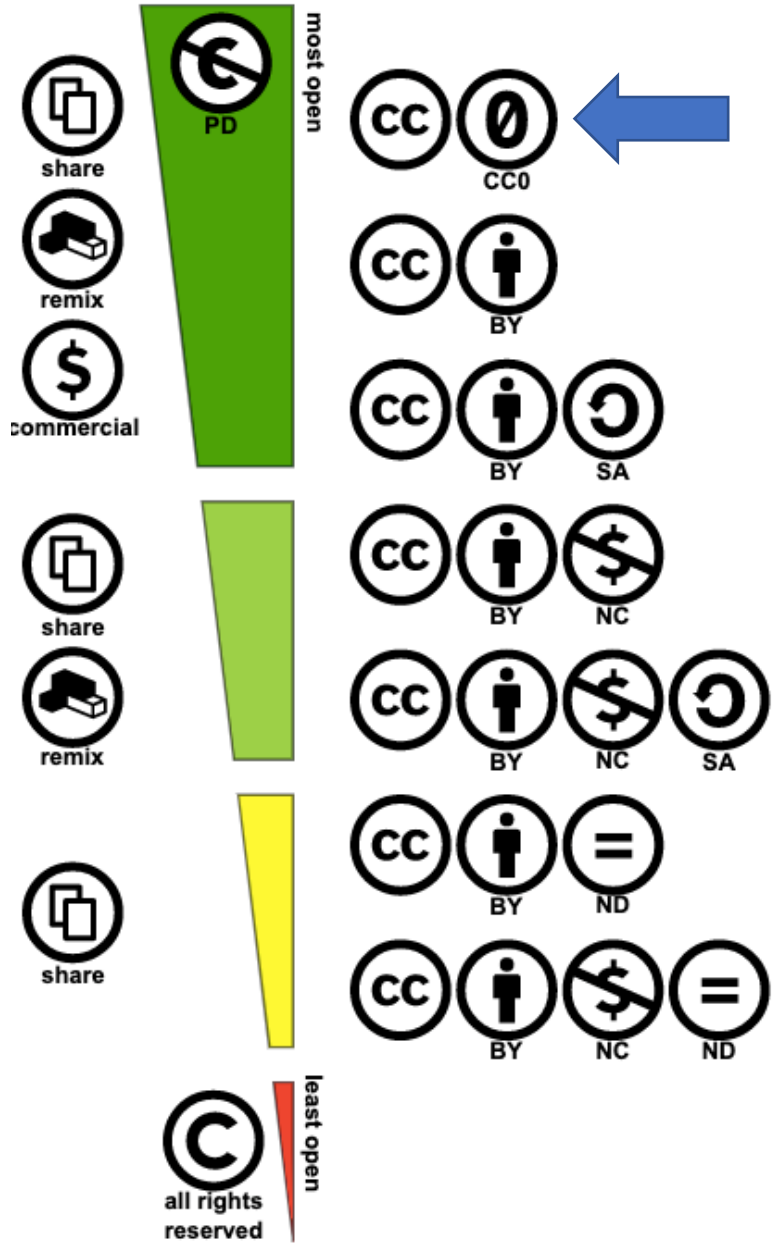
https://github.com/common-voice/common-voice/blob/production/server/data/zh-HK/sentence-collector.txt



Common Voice
moz://a

Common Voice 的版權潔癖

- All sentence must be public domain/CC0
- CC0 means relinquishing all copyrights you hold in a work and dedicating those rights to the public domain.
- 亦即係話大部份現有嘅大型粵語語料都唔可以加入去因為現有粵語語料大部份都係 CC-BY , CC-BY-SA
- 解決方法 :
  - 大家貢獻多啲原創嘅廣東話語料。例如 :
    - 自己寫嘅廣東話 blog
    - 自己 channel 嘅廣東話對白
  - 學者/作者/學術組織將語料庫貢獻到 CC0



https://en.wikipedia.org/wiki/Creative_Commons_license

Part 3

廣東話 **Python Library**

PyCantonese


```
In [4]: import pycantonese

        pycantonese.characters_to_jyutping("知我咩料啦!")
```

```
Out[4]: [('知', 'zi1'), ('我', 'ngo5'), ('咩料', 'me1liu2'), ('啦', 'la1'), ('!', None)]
```

```
In [7]:  pycantonese.parse_jyutping('zi1ngo5me1liu2la1')
```

```
Out[7]: [Jyutping(onset='z', nucleus='i', coda='', tone='1'),
         Jyutping(onset='ng', nucleus='o', coda='', tone='5'),
         Jyutping(onset='m', nucleus='e', coda='', tone='1'),
         Jyutping(onset='l', nucleus='i', coda='u', tone='2'),
         Jyutping(onset='l', nucleus='a', coda='', tone='1')]
```

```
In [16]: corpus = pycantonese.hkcancor() # get HKCanCor
         corpus.search(character='廣東話')
```

```
Out[16]: [Token(word='廣東話', pos='NZ', jyutping='gwong2dung1waa2', mor=None, gra=None),
          Token(word='廣東話', pos='NZ', jyutping='gwong2dung1waa2', mor=None, gra=None),
          Token(word='廣東話', pos='NZ', jyutping='gwong2dung1waa2', mor=None, gra=None),
          Token(word='廣東話', pos='N', jyutping='gwong2dung1waa2', mor=None, gra=None),
          Token(word='廣東話', pos='N', jyutping='gwong2dung1waa2', mor=None, gra=None),
          Token(word='廣東話', pos='NZ', jyutping='gwong2dung1waa2', mor=None, gra=None),
          Token(word='廣東話', pos='NZ', jyutping='gwong2dung1waa2', mor=None, gra=None),
          Token(word='廣東話', pos='NS', jyutping='gwong2dung1waa2', mor=None, gra=None)]
```

中文分詞

兒子生性病母倍感安慰

苦盡甘來

56歲的張婉貞，由於丈夫沉迷賭博欠債纍纍，01年時終難以忍受，與其離婚，因情緒受困，1年後更患上乳癌，面對不斷的化療及經濟問題，帶着當時14歲兒子的她承受巨大壓力，陷入谷底。幸得年幼但懂事的兒子支持，及母親的照顧，「但（兒子）始終係男仔，啲太識表達。但佢有時會同同學去超市幫手搬啲米返嚟，我哋喺陣又會倒水畀我，啲舒服會同我按摩。」而其母親亦會爲她煮飯、煲湯，令她安心養病。

感患病非失去是得着

張婉貞的兒子Jason表示，可能於單親家庭長大關係，令其年紀輕輕已十分懂事，他經常會向母親表示會照顧自己，讓她免得擔心。張婉貞認爲，患病可能不幸，但有兒子與母親的關懷，令她非常快樂，「這個病唔係失去，係得着。我而家每日都玩盡做盡，活在當下。」現時她除經營時裝店、照顧託管兒童外，更積極參與各類義工活動。

產後抑鬱母喜獲夫女諒解

另一位母親梁女士，約20年前誕下女兒，不幸患上產後抑鬱及強迫症，脾氣變得暴躁，更經常責罵女兒，丈夫不了解其病情，因而對她感到厭惡，經常外出不回家。直至去年8月，梁開始接受藥物治療，現時情況已漸趨穩定。

而丈夫報讀一些課程後，亦開始諒解她。女兒見其狀況轉好，早前更向她表示明白她一直以來的痛苦，「咁已經好開心，好過佢畀錢我用好多。」

梁女士認爲，家人與她每天共同生活，對其支持非常重要。張婉貞亦表示，現時與兒子關係如同朋友，希望天下母親亦能與子女有良好溝通，共度歡樂的母親節。

堅強媽媽照顧早產患病兒

母愛偉大

「團團提早3個幾月出世，當時佢得1磅重，得個紙巾盒咁大，仲要插喉餵住啲4個月深切治療部，會經歷過天父不如接佢走啦，可能對佢嚟講係件好事。」

兒早產得1磅患腦麻痺

同樣是早產兒的女強人媽媽陳頴娟，身體狀況如常，6年前首次誕下女兒時，因因身體未有問題，惟兩年前發現孻子漸漸不僅得轉身，平日甚少笑喊、只僅得笑，經醫生診斷下才知道患有腦麻痺及嚴重聽障。

陳頴娟(39歲)憶述，漸漸出生時醫生已說明會有問題，但起初仍未發覺有異樣，兒子飲食如常，只是覺得團團平日甚少笑喊、只僅得對我們微笑。半年後發覺團團不僅得轉身，但仍不知道出現問題。直至漸漸1歲時，醫生才證實他患上腦麻痺及嚴重聽障。她慨嘆說：「當時真係覺得好灰，嘅直係人生嘅低潮！我都係早產，但係身體一直都好健康，點諗到會發生喺團團身上？但係後來都無奈接受。」

夫支持辭工全職照顧兒子

「當知道團團出現咁嘅情況，覺得佢比份工更加緊要，希望盡量抽時間陪佢，所以毅然決定辭咗份工，全職照顧團團。好彩丈夫都好支持我，因因亦喺會因爲咁而呻嘢，仲好媽細佬，令我更加全心全意照顧團團。」可惜，經過一年辛勞照顧漸漸後，陳頴娟開始發覺身體出現異樣，腰椎經常感到無力，更曾經因左腳乏力跌倒，頓時覺得自己無力全職照顧兒子，決定與丈夫交換角色，日前重投職場出任食品銷售員。

兒子一吻是最大禮物

她又指，近期漸漸情況持續好轉，早前接受耳蝸手術後，其聽力已由嚴重聽障轉爲輕微，而且身體狀況良好，甚少出現病痛。笑言「漸漸已送咗份最好嘅母親節禮物畀我，就係嘟起嘴媽咗我一啖。由於佢呢個病好難控制肌肉部動，所以呢份禮物已經好開心、好窩。母親節願望係全家人都身體健康、快快樂樂！」

```
In [13]: import pycantonese

         pycantonese.segment("兒子生性病母倍感安慰")

Out[13]: ['兒子', '生性', '病', '母', '倍感', '安慰']
```

```
In [14]: import jieba

         [s for s in jieba.cut("兒子生性病母倍感安慰")]

Out[14]: ['兒子生', '性病', '母', '倍感', '安慰']
```

中文分詞 – Customized PyCantonese

Customize Pycantonese Segmenter

```
In [35]: from pycantonese.word_segmentation import Segmenter

         segmenter = Segmenter(allow={"病母"})
         pycantonese.segment("兒子生性病母倍感安慰", cls=segmenter)

Out[35]: ['兒子', '生性', '病母', '倍感', '安慰']
```

中文分詞 – Jieba + 粵典

In [40]:
```
import jieba
import pandas as pd
import requests
import json
```

In [41]:
```
def wget(url, encoding='utf8'):
    r = requests.get(url)
    r.raise_for_status()
    return r.content.decode(encoding)

def download_word_frequency():
    data = wget(url='https://words.hk/faiman/analysis/existingwordcount.json')#粵典詞表使用頻率
    return pd.DataFrame(json.loads(data), index=['count']).transpose() \
                .sort_values(by='count', ascending = False)

def save_word_frequency(df, filename):
    with open(filename, "w") as file:
        for word, row in df.iterrows():
            if len(word) > 1:
                file.write(f'{word.replace("*","")} {row["count"]}\n')
```

In [42]:
```
save_word_frequency(download_word_frequency(), "粵典_userdict.txt")

jieba.load_userdict("粵典_userdict.txt")

[s for s in jieba.cut("兒子生性病母倍感安慰")]
```

Out[42]:
```
['兒子', '生性', '病母', '倍感', '安慰']
```

Part 4

廣東話語音辨識

介紹其中兩個技術

DeepSpeech

wav2vec

DeepSpeech

DeepSpeech

- A project under Mozilla
- DeepSpeech is an open source embedded (offline, on-device) speech-to-text engine which
- can run in real time on devices ranging from a Raspberry Pi 4 to high power GPU servers.

DeepSpeech = Acoustic + Language model

- DeepSpeech is made up of two main parts:
  - (1) the acoustic model
  - (2) the language model
- The acoustic model takes audio as input and converts it to a probability over characters in the alphabet.
- The language model (aka. scorer) helps to turn these probabilities into words of coherent language. It assigns probabilities to words and phrases based on statistics from training data.
- E.g. “I read a book” is much more probable then “I red a book”.

Reference: https://mozilla.github.io/deepspeech-playbook/AM_vs_LM.html

How do we measure accuracy?

CER is suitable for **Mandarin** and **Cantonese**

$$CER = \frac{S + D + I}{N}$$

Character Error Rate

WER is suitable for **English**

$$WER = \frac{S_w + D_w + I_w}{N_w}$$

Word Error Rate

where:

- **S** = Number of **S**ubstitutions
- **D** = Number of **D**eletions
- **I** = Number of **I**nsertions
- **N** = **N**umber of characters in reference text (aka ground truth)

Source: https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510

Run DeepSpeech Model (English)

- docker pull nvidia/cuda:10.1-cudnn7-devel
- pip install deepspeech-gpu
- Download the acoustic model (deepspeech-0.9.3-models.pbmm) and language model (deepspeech-0.9.3-models.scorer)
- Prepare a 16k Hz mono audio input

```
$> ffmpeg -i $(youtube-dl -f 18 --get-url
https://www.youtube.com/watch?v=LTxMRQObjfs ) -ss 00:01:15 -to 00:01:59 -ar 16000 -
ac 1 audio.wav
```

- Ready to go!

```
$> deepspeech --model deepspeech-0.9.3-models.pbmm \
              --scorer deepspeech-0.9.3-models.scorer \
              --audio audio.wav
```

DeepSpeech - Let’s try it!

English subtitles generated using DeepSpeech v0.9.3 @ WER=7%

https://youtu.be/-OL-IIaRhA8

See the demo

影片來源: Marvel Studios《尚氣與十環幫傳奇》 製作花絮 https://youtu.be/LTxMRQObjfs

Subtitles generated by the model

Incorrect recognitions


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
everything was done with care with research
全都一絲不苟，經過深入研究


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
tony and michel are as legendary as you can get
梁朝偉、楊紫瓊真是傳奇中的傳奇


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
when your story telling is good it transcends the culture
有一流的敘事手法就能超越文化


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
AWKWAFINA
every one in this movie the care about the story
每個參與本片的人對故事都十分用心


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
we all bonded because it's just such an important filmore
因為這部片太重要了


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映
whether it's a language the symbolic
不管是語言或象徵符號


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映
(... failed to recognize ...)
你註定要成就大業


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
did you work with tony man today
你今日有和梁朝偉做對手戲？


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
you will protect all who came before (you)
你繼承了前人的種種


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
every models who were here as your own story
每位Marvel超級英雄都有自己的故事


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
i tell you adorable balsamides
都是關於愛，關於家庭

DeepSpeech 廣東話？

Resource about training DeepSpeech Model

- Playbook/Guidelines
  https://mozilla.github.io/deepspeech-playbook/
- Transfer Learning of new alphabet
  https://deepspeech.readthedocs.io/en/r0.9/TRAINING.html#transfer-learning-new-alphabet
- CTC Decoder, Scorer
  https://deepspeech.readthedocs.io/en/r0.9/Decoder.html#decoder-docshttps://mozilla.github.io/deepspeech-playbook/SCORER.html
- Train a Dutch Model
  https://colab.research.google.com/github/acabunoc/Tutorial-train-dutch-model/blob/master/DeepSpeech_train_a_model%2C_CV_Dutch.ipynb
- Bytes output mode for Cantonese ASR [I am still trying …]
- Paddlpaddle+DeepSpeech2 [To be investigated]

I failed to train a reasonable Deepspeech Cantonese model …
幫緊你幫緊你 …

wav2vec

Run wav2vec2 Model (English)

```python
import soundfile as sf
import torch
from datasets import load_dataset
from transformers import Wav2Vec2ForCTC, Wav2Vec2Processor

# load pretrained model
processor = Wav2Vec2Processor.from_pretrained("facebook/wav2vec2-large-960h-lv60-self")
model = Wav2Vec2ForCTC.from_pretrained("facebook/wav2vec2-large-960h-lv60-self")

# load audio
audio_input, sample_rate = sf.read('../data/selected_audio/LTxMRQObjfs.wav')

# pad input values and return pt tensor
input_values = processor(audio_input, sampling_rate=sample_rate, return_tensors="pt").input_values

# INFERENCE
# retrieve logits & take argmax
logits = model(input_values).logits
predicted_ids = torch.argmax(logits, dim=-1)
# transcribe
transcription = processor.decode(predicted_ids[0])
print("transciption:", transcription.lower())

```


Reference:

- wav2vec 2.0: A framework for self-supervised learning of speech representations (A Baevski, H Zhou, A Mohamed, M Auli - arXiv preprint arXiv:2006.11477, 2020 - arxiv.org)
- https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

wav2vec2 - Let’s try it!

English subtitles generated using facebook/wav2vec2-large-960h-lv60-self @ WER=2%

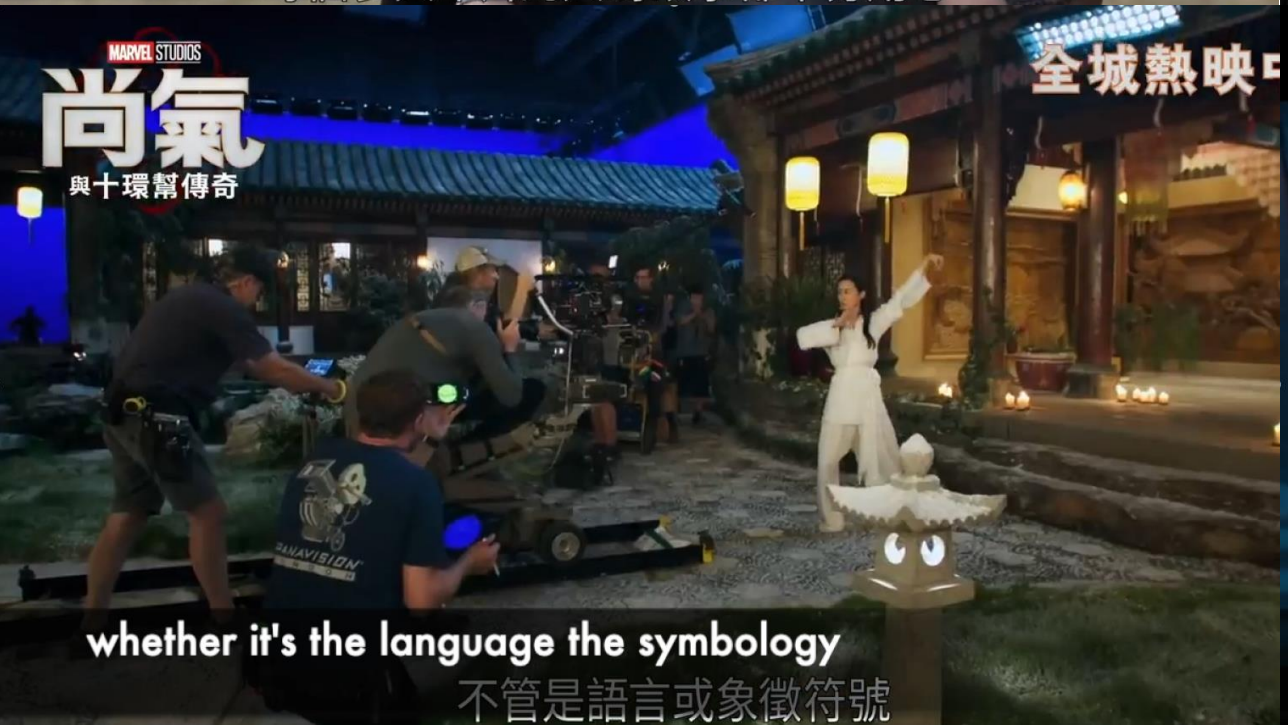
https://youtu.be/lpk4JKY2Mbo

See the demo

影片來源: Marvel Studios《尚氣與十環幫傳奇》 製作花絮 https://youtu.be/LTxMRQObjfs


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
AWKWAFINA
everyone on this movy they care about the story
每個參與本片的人對故事都十分用心


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
we all bond it because it's just such an important film
因為這部片太重要了


MARVEL STUDIOS
尚氣
與十環幫傳奇
whether it's the language the symbology
不管是語言或象徵符號


MARVEL STUDIOS
尚氣
與十環幫傳奇
you werement for greatness
你註定要成就大業

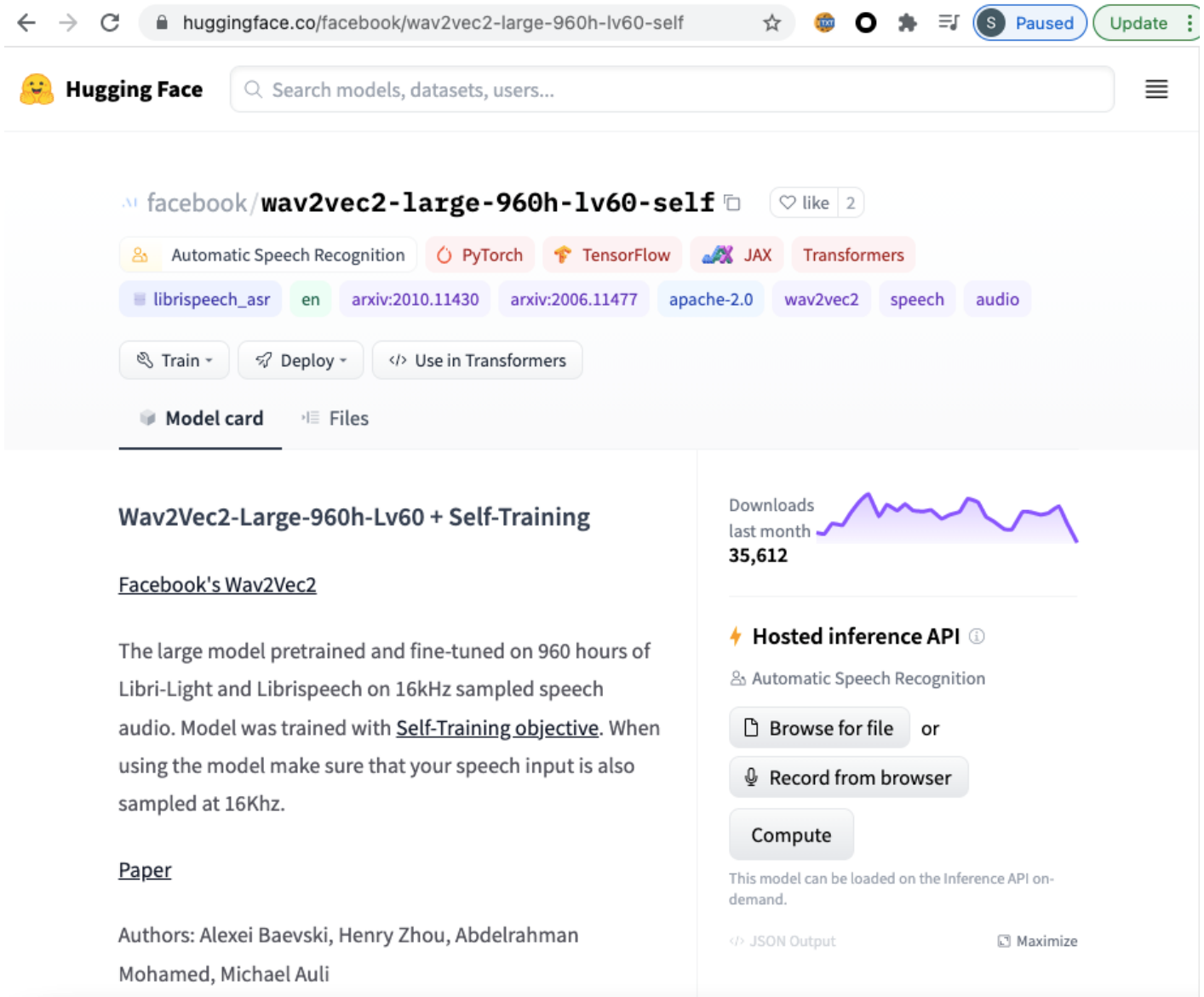

MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
did you work with tony lion to day
你今日有和梁朝偉做對手戲？


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
you are a product of all who came before (you)
你繼承了前人的種種


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
every model super hero has your own story
每位Marvel超級英雄都有自己的故事


MARVEL STUDIOS
尚氣
與十環幫傳奇
全城熱映中
we'r tell om a story about love about family
都是關於愛，關於家庭

- The large model pretrained and fine-tuned on **960 hours** of Libri-Light and Librispeech on 16kHz sampled speech audio.



wav2vec
廣東話？

wav2vec-xlsr

UNSUPERVISED CROSS-LINGUAL REPRESENTATION LEARNING FOR SPEECH RECOGNITION

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, Michael Auli
Facebook AI

ABSTRACT

This paper presents XLSR which learns cross-lingual speech representations by pretraining a single model from the raw waveform of speech in multiple languages. We build on wav2vec 2.0 which is trained by solving a contrastive task over masked latent speech representations and jointly learns a quantization of the latents shared across languages. The resulting model is fine-tuned on labeled data and experiments show that cross-lingual pretraining significantly outperforms monolingual pretraining. On the CommonVoice benchmark, XLSR shows a relative phoneme error rate reduction of 72% compared to the best known results. On BABEL, our approach improves word error rate by 16% relative compared to a comparable system. Our approach enables a single multilingual speech recognition model which is competitive to strong individual models. Analysis shows that the latent discrete speech representations are shared across languages with increased sharing for related languages. We hope to catalyze research in low-resource speech understanding by releasing XLSR-53, a large model pretrained in 53 languages.[1]

Reference: https://arxiv.org/pdf/2006.13979.pdf

wav2vec-xlsr

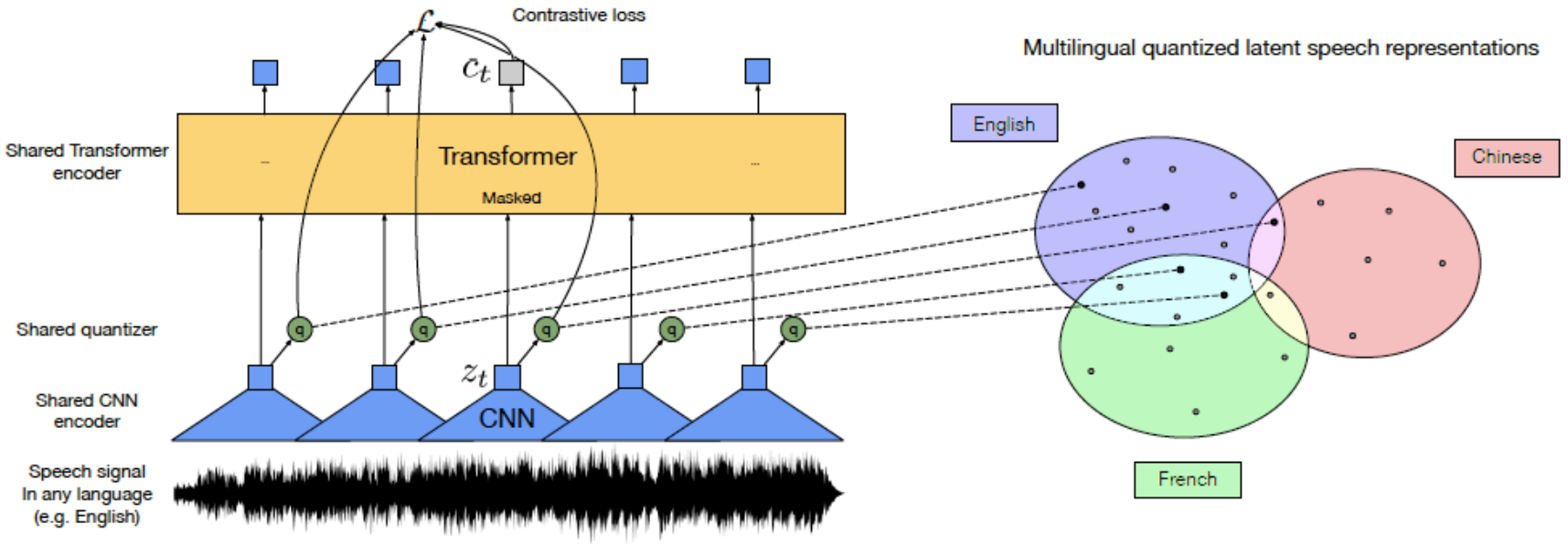


Figure 1: **The XLSR approach.** A shared quantization module over feature encoder representations produces multilingual quantized speech units whose embeddings are then used as targets for a Transformer trained by contrastive learning. The model learns to share discrete tokens across languages, creating bridges across languages. Our approach is inspired by Devlin et al. (2018); Lample & Conneau (2019) and builds on top of wav2vec 2.0 (Baevski et al., 2020c). It requires only raw unlabeled speech audio in multiple languages.

Reference: https://arxiv.org/pdf/2006.13979.pdf

wav2vec-xlsr

vectors. Finally, we cluster languages using K-Means and then perform a PCA with two dimensions.

(a) CV-12 model. (b) CV-12 model + zh-HK. (c) BABEL-17 model.

Figure 2: **Visualization of language similarities learned by the model** Figure (a) visualizes the shared discrete latent speech representations across languages for a model trained on 12 Common-Voice languages (CV-12). Figure (b) shows that adding Chinese-HongKong (zh-HK) shares relatively few latents with other languages. Figure (c) is for a model trained on 17 BABEL languages and illustrates that clusters can correspond to similar languages like Bengali and Assamese.

Figure 2a, 2b and 2c show the visualizations, where colors correspond to the clusters obtained by K-Means. Note, that we perform K-Means before PCA to avoid loss of information, and that PCA may make some points appear closer than they are in original vectors. We see that the model shares more discrete tokens for similar languages, e.g., it groups Basque, Catalan, Spanish and Italian, or English, German and French, or Arabic and Kabyle (see Figure 2a), and Mandarin (zh-CN and zh-TW), although this information is lost in the PCA visualization. Figure 2b shows that the model may also isolate a language, such as Chinese-HongKong (Cantonese), which is not close to any other language because it shares fewer discrete tokens with other languages.

Reference: https://arxiv.org/pdf/2006.13979.pdf

Resource about finetuning wav2vec-xlsr

- Finetune for English

  https://colab.research.google.com/drive/1FjTsqbYKphl9kL-eILgUc-bl4zVThL8F?usp=sharing

- Finetune for Turkish

  https://colab.research.google.com/github/patrickvonplaten/notebooks/blob/master/Fine_Tune_XLSR_Wav2Vec2_on_Turkish_ASR_with_🤗_Transformers.ipynb

- Finetune for Cantonese

  https://huggingface.co/ctl/wav2vec2-large-xlsr-cantonese [by chutaklee] 👍 👍

  https://huggingface.co/voidful/wav2vec2-large-xlsr-53-hk [by voidful] 👍 👍

  CER = 15% – 16%

wav2vec2-xlsr Cantonese finetune - Let's try it!

Cantonese subtitles generated using wav2vec2-xlsr Cantonese finetune @ CER=16%

https://youtu.be/k_9RQ-ilGEc

See the demo

影片來源: 《鏗鏘集 -廣東話》
https://youtu.be/m_pAFNtddFw
https://podcast.rthk.hk/podcast/item.php?pid=244&eid=115672&lang=zh-CN


統計署設料顯事
統計處資料顯示


大呢紀年間數至下啲咗兩個百分點
數字下跌兩個百分點


附鹽接更九成嘅香港人廣東話維冇語
雖然接近九成香港人以廣東話為母語


上反以部通話冇嘅人口有上性嘅吹世
相反，以普通話為母語的人口有上升趨勢


靜 雞
雞 雞
請請推淨啲 定靜雞雞淨
「靜靜雞」還是「靜雞雞」比較靜？


星光
STAR
咁所以啦 如窩進一部去認式 我地嘅冇語
所以若您進一步認識我們的母語


（。。。。。。。）一定該淨兩個淨程
我只聽過這個


星光
STAR
阿 我哋個個冇要嘅第層 究經仲我啲咩係
母語的底層還有甚麼未被發掘？

點解唔準？

- 要多啲口語例句 We need more Cantonese spoken sentences in the dataset
- 要多啲語音數據 Cantonese voice dataset is 10+ times smaller than English

Voice Corpus	No. of hours	File Size	No. of voices
Librispeech – English	1,000 hr	58 GB	2,484
Common Voice – English	2,637 hr	65 GB	75,879
Common Voice - Chinese (Hong Kong)	113 hr	3 GB	2,656

- 現有技術多以英語為本位， 較少針對廣東話作優化 Most of the models were designed to cope with English as priority. Other languages support came as second priority (e.g. Cantonese).

可以點做？

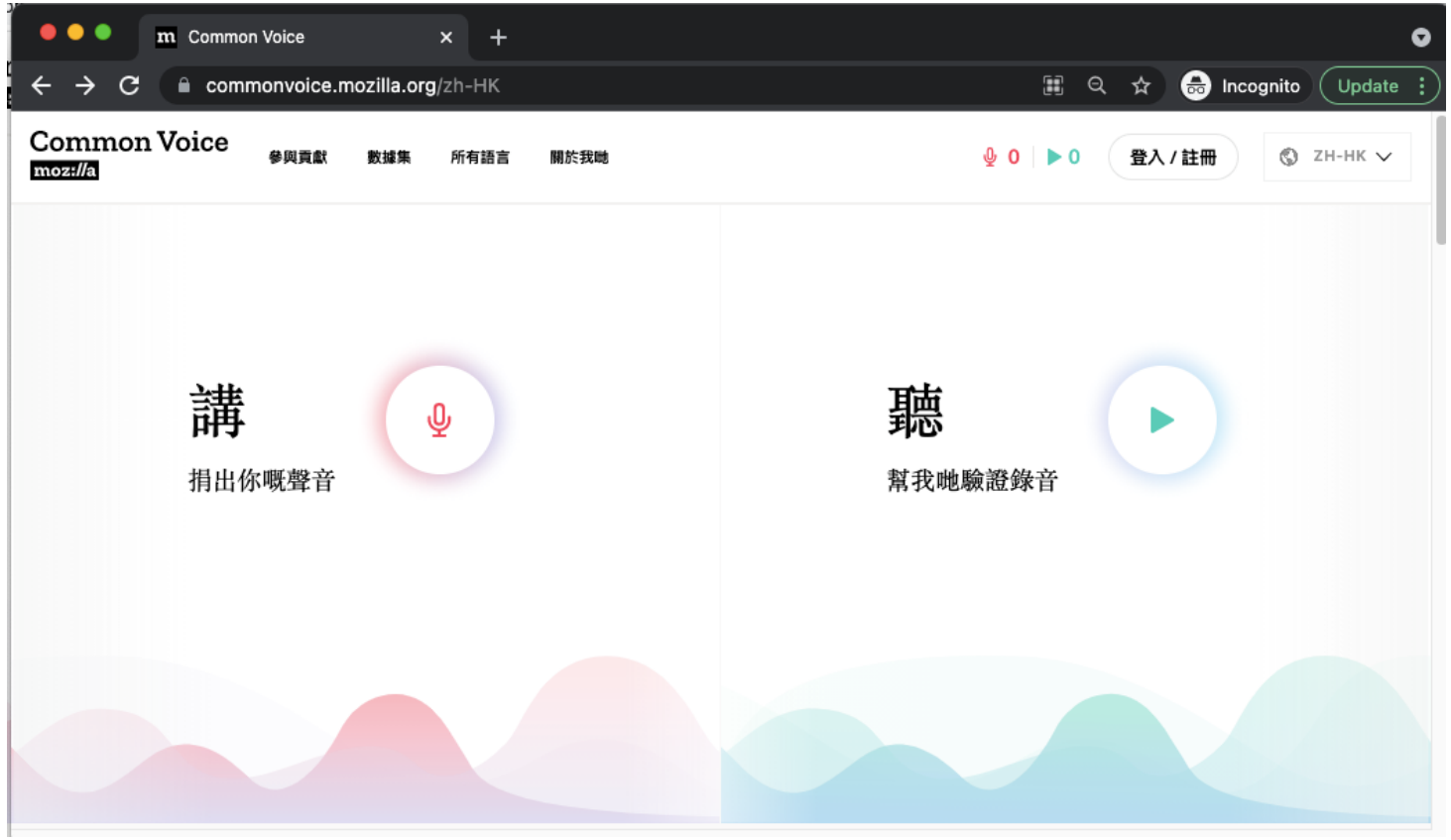
- 香港 IT 人多參與，針對廣東話對音調/粵拼嘅優化
- 香港人捐出你的廣東話句子/文本語料
- 香港人捐出你嘅廣東話語音

只要有1 % 嘅香港人，每人幫手錄 2 - 3 分鐘，加埋就係:

1%　x　700萬人　x　2分鐘 =　<u>2300 小時</u>

幫手「講 /聽」

https://commonvoice.mozilla.org/zh-HK

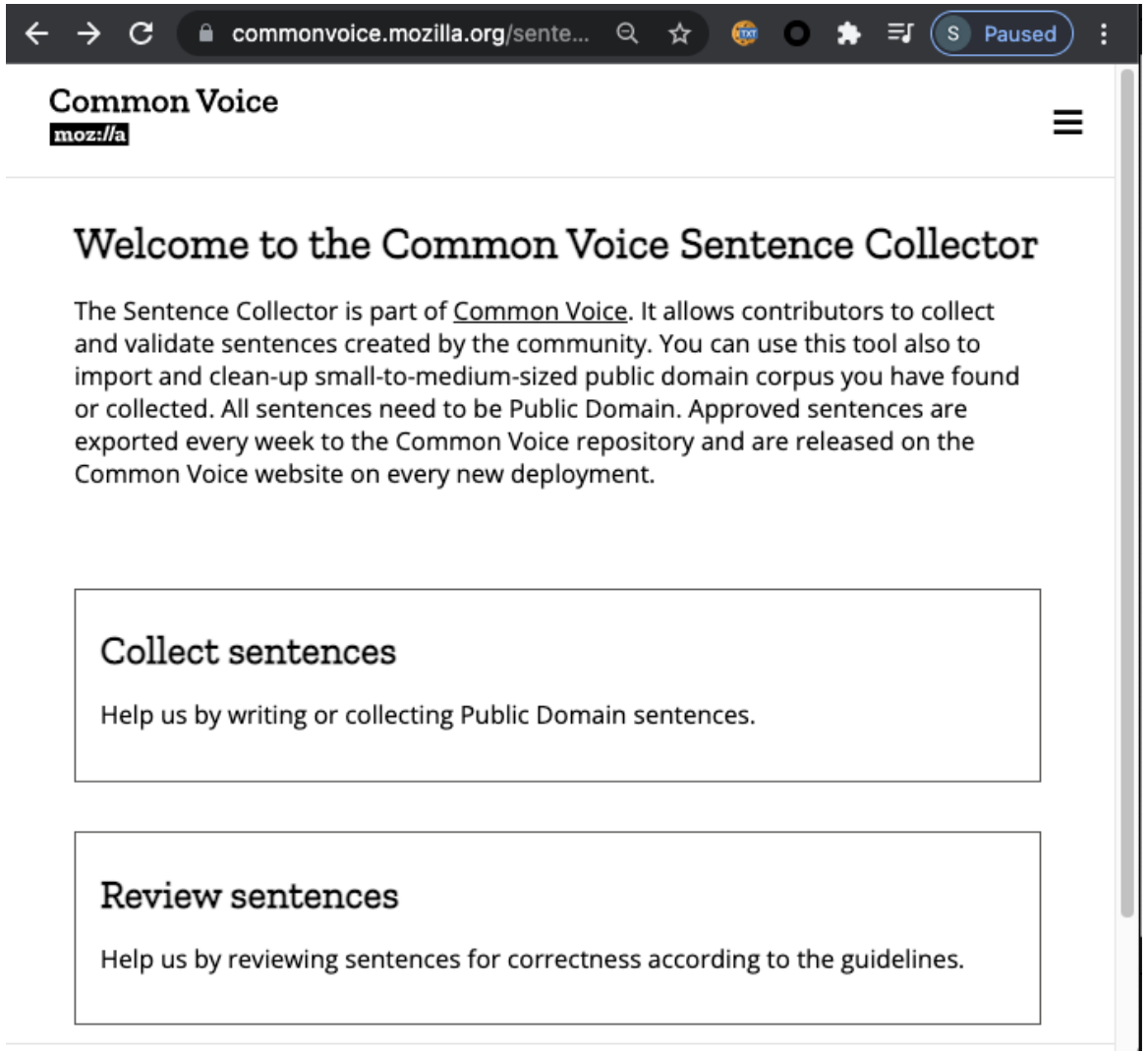


「登入/註冊」後使用更好

可以記錄自己進度，並有助系統減少句子重複出現

幫手「加句子」

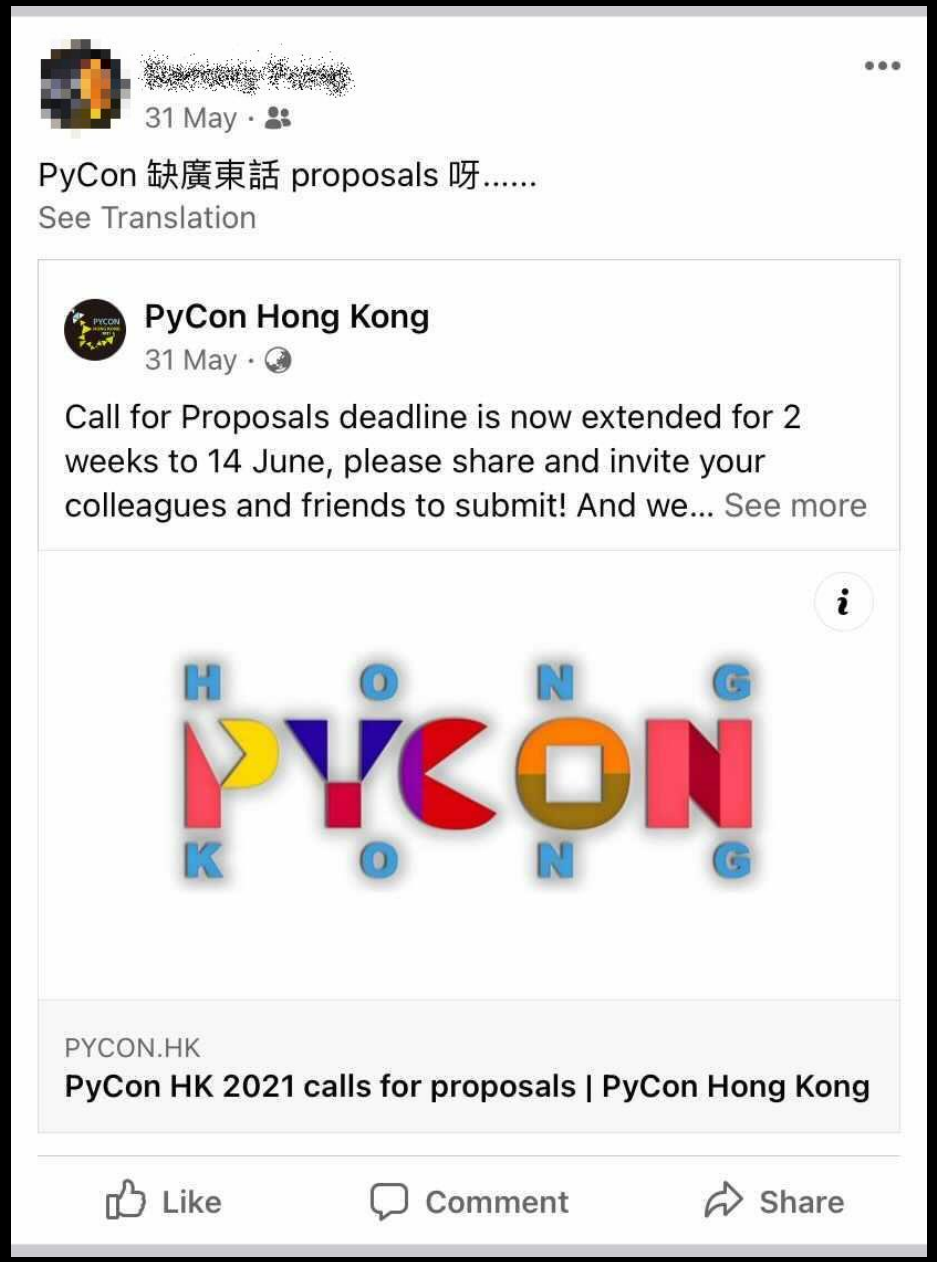
https://commonvoice.mozilla.org/sentence-collector



句子必須係
無版權限制
(public domain, cc0)

後記：

點解會有「廣東話自肥企劃」？



PyCon Hong Kong

世界上唯一一個

廣東話 PyCon

「ERROR並唔係唯一戰鬥緊嘅人，只要一日仲留喺呢片土地上，每一個人都用自己嘅方式，喺自己嘅崗位戰鬥，一個人唔會成功，但如果每一個人都盡力嘅話，每一個人都盡力嘅話 ...」

致一直還在堅持的人

廣東話作為大家嘅母語,日常使用廣東話溝通大家覺得係理所當然。但係廣東話對好多人來講並唔係理所當然!

外國人來到香港/澳門/廣州/馬來西亞想融入本地文化都花好大努力先學得識廣東話；文化創作人努力製作廣東話歌曲，廣東話電視，廣東話電影；義工努力收集廣東話語料；語言學學家花畢生努力去研究 粵 語 ，為大家剖析/整理/記錄依隻語言嘅特性；數據科學家努力去教識電腦去聽明廣東話 . . .

所以廣東話嘅保育和活化，廣東話喺科技上嘅應用，並唔係理所當然。我地每一代廣東話使用者點樣承繼前人留落黎嘅粵語文化遺產，有賴每一個人都出一分力。

我地一個人做唔會成功，但如果每一個人都努力嘅話 ， 結 果 可能唔一樣. . .

つづく

待續

補充資料

iPhone, Android, Google … 都有語音輸入啦, 點解仲要自己開發？

Common Voice

mozilla

- 語音識別技術可讓我們的裝置帶來人性，但開發者需要極為龐大數量的語音資料，才能打造出這樣的系統。
- 大部分現成嘅數據由大公司擁有，並未開放畀大眾使用。目前大部分語音資料都相當昂貴，也是專有的資料。
- 噉樣會阻礙創新
- 我們希望讓語音資料能夠公開自由使用，並確保資料反映出人們實際的多樣性

語言學參考資料

中大粵語對照資料庫
https://apps.itsc.cuhk.edu.hk/hanyu/Page/Terms.aspx

中大粵語審音配詞字庫
https://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/

粵典
https://words.hk

香港語言學學會
https://www.lshk.org/publications

PyCantonese Built-in Corpus
https://pycantonese.org/data.html

其他有趣視頻

廣東話的九聲 (一) /(二) [秒懂！]
https://www.youtube.com/watch?v=Yi-lpECdE9w
https://www.youtube.com/watch?v=EtWHAywrYJ8

港式廣東話 vs. 廣式廣東話
https://youtu.be/ZhS4zM_1nNc

廣東話保育？
https://www.youtube.com/watch?v=VG5Rg5JXdkI&t=62s

香港人竟然讀錯「香港」？語言學家：「粵語懶音問題200年前已出現」
https://youtu.be/xYsQbnFelCs

五夜講場 - 學人串社科 2020：時代擇言
https://youtu.be/PWrINbsZPT00