

Google Colab + Hugging Face: 帶你快速認識NLP

講者：Andy Chiang

自我介紹

- **Andy Chiang (江尚軒)**
- 中興大學資工系 大三升大四
- NCHU GDSC core team member
- 中興大學NLP實驗室 研究助理
- 工研院 資料服務與智慧決策部 實習生
- 主要研究領域有：網頁前後端、機器學習和自然語言處理



開始之前...

- 今天的投影片有公開，大家可以掃描QR code，待會聽演講時可以參考。



目錄

1. NLP是什麼？
2. NLP有什麼用？
3. 語言預訓練模型又是什麼？
4. Google Colab + Hugging Face 實作
5. 參考資料
6. 總結

NLP是什麼？

NLP是什麼？

自然語言處理

(**N**atural **L**anguage **P**rocessing, **NLP**)

= 電腦科學 + 語言學

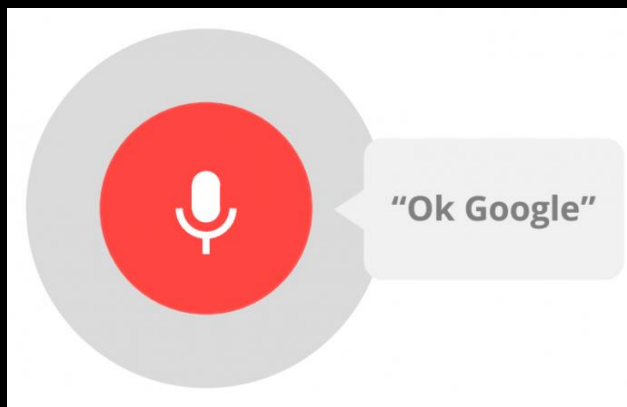
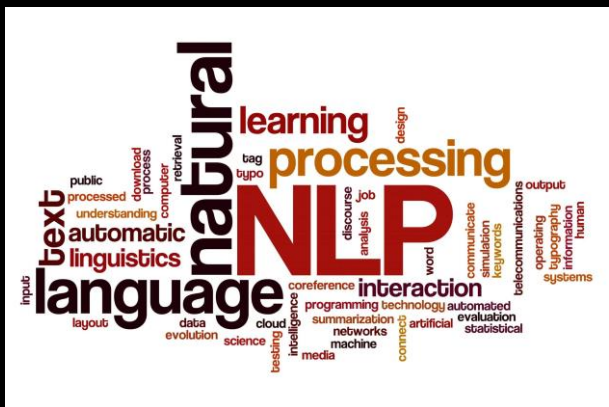
NLP是什麼？

探討如何讓電腦理解、運用**自然語言**。

NLP是什麼？

Q：什麼是自然語言？

A：人類為了溝通所創造的語言，通常有特定的文法。形式可以是**文字**、**語音**、**符號**...



NLP是什麼？

自然語言理解
(Natural Language
Understanding, NLU)

研究如何讓電腦將人
類語言轉為數值資料
(**讀懂人類語言的含意**)



自然語言生成
(Natural Language
Generation, NLG)

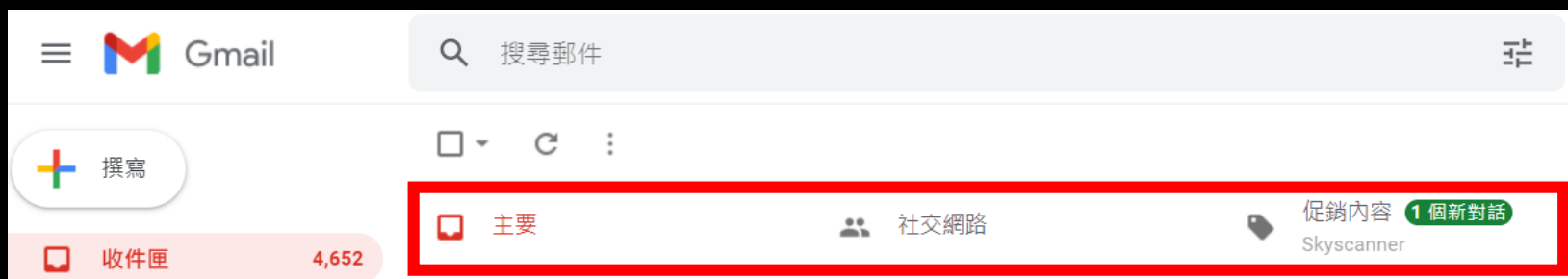
研究如何讓電腦將數值
資料轉為人類語言
(**創造有意義的人類語言**)

NLP有什麼用？

NLP有什麼用？

Email篩選器

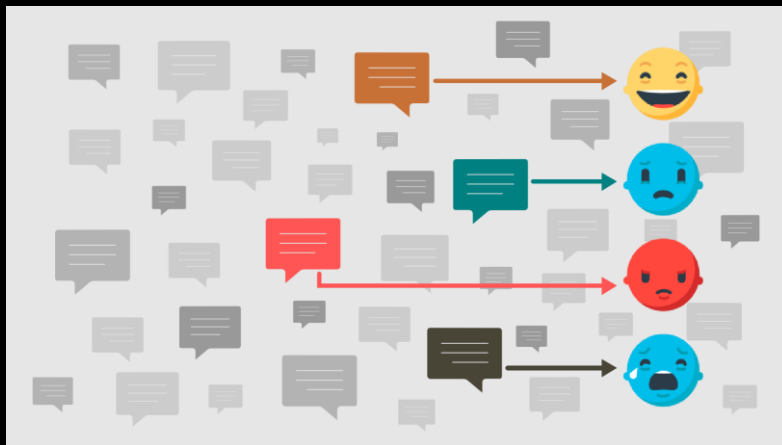
透過信件內容過濾垃圾郵件，或者像Gmail將信件分成**主要**、**社交**和**促銷**三類，讓你的收件夾不會很雜亂。



NLP有什麼用？

情感分析

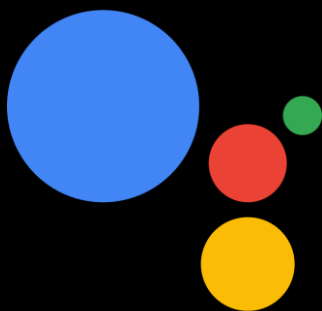
公司可以從社群媒體上蒐集客戶對該產品的相關留言或貼文，分析出**正面**、**中立**和**負面**的比例，及時改善行銷手法。



NLP有什麼用？

智能助理

像是**Google Assistant**、**Apple Siri**、**Amazon Alexa**等等，這些智能助理已經漸漸成為日常生活的一部份了，幫助我們處理各種生活瑣事，無聊也可以跟他聊聊天、講笑話~



語言預訓練模型 又是什麼？

語言預訓練模型又是什麼？

Google收集了大量的資料集 (Books Corpus + English Wikipedia 總共33億個字)，透過**非監督**的方式來**pre-train**。

之後拿這個pre-train好的模型，針對特定的下游任務作**fine-tune**，結果都比之前的模型還好！當年橫掃了很多NLP任務的排行榜。

語言預訓練模型又是什麼？

這概念其實就像我們學中文，從小到大都在接觸中文，耳濡目染之下就有**基本的語感**。如果此時再叫我們去學特定的任務 (如：接龍、照樣造句...)，對我們而言根本輕而易舉，對吧？

pre-train >> 學會基本語感

fine-tune >> 學習特定的任務

語言預訓練模型又是什麼？

但除非是大企業，不然要自己從頭開始pre-train語言預訓練模型根本是天方夜譚。

做為參考，訓練一個**1.1億**參數的**BERT-BASE**模型，**要用16個TPU跑4天!**更何況還要收集那麼大的資料集。

幸好BERT作者有開源pre-train好的模型，讓我們可以直接站在巨人的肩膀上，讓下游任務變得既有效又輕鬆。



BERT

NLP研究者

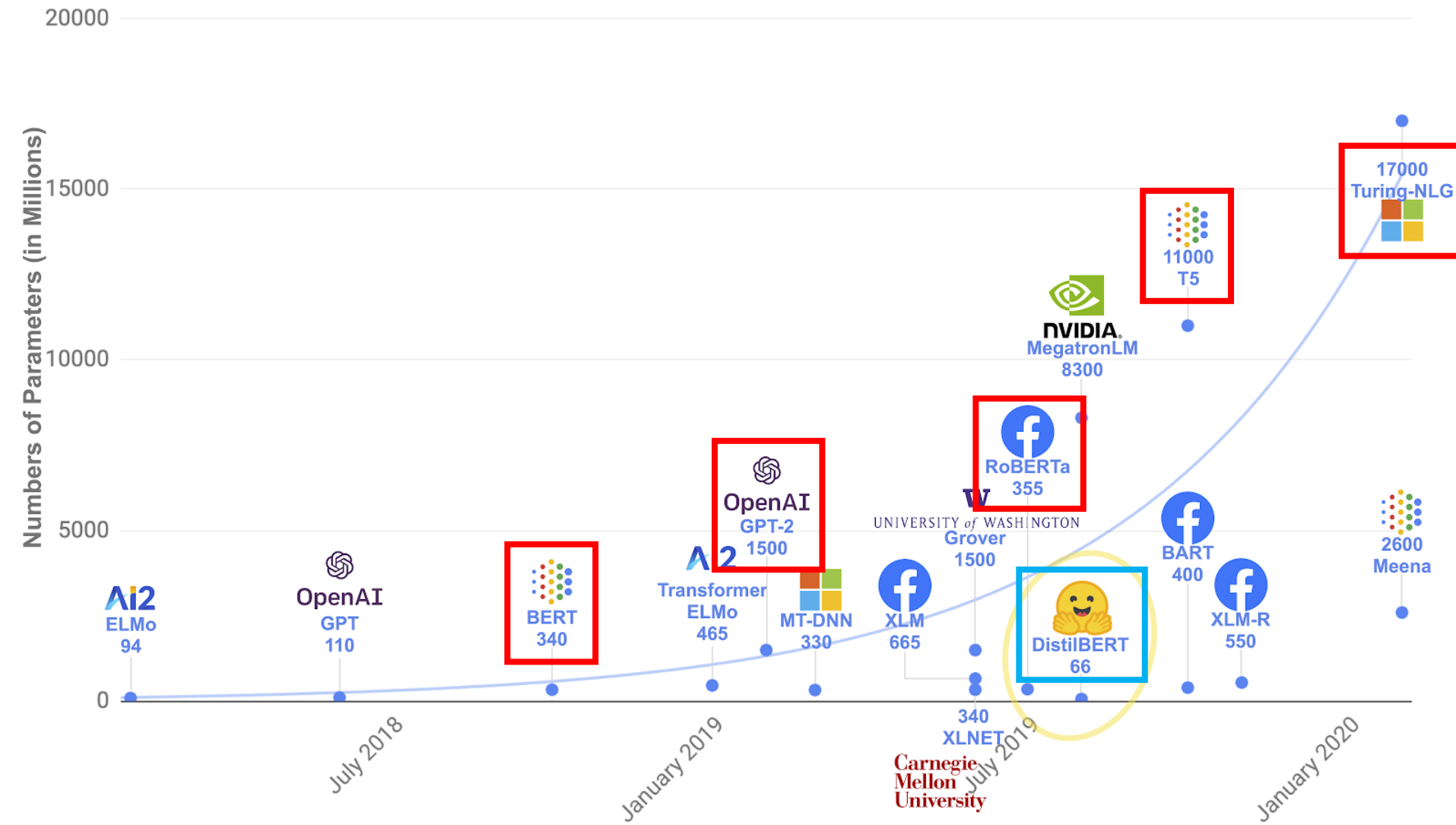
傳統模型

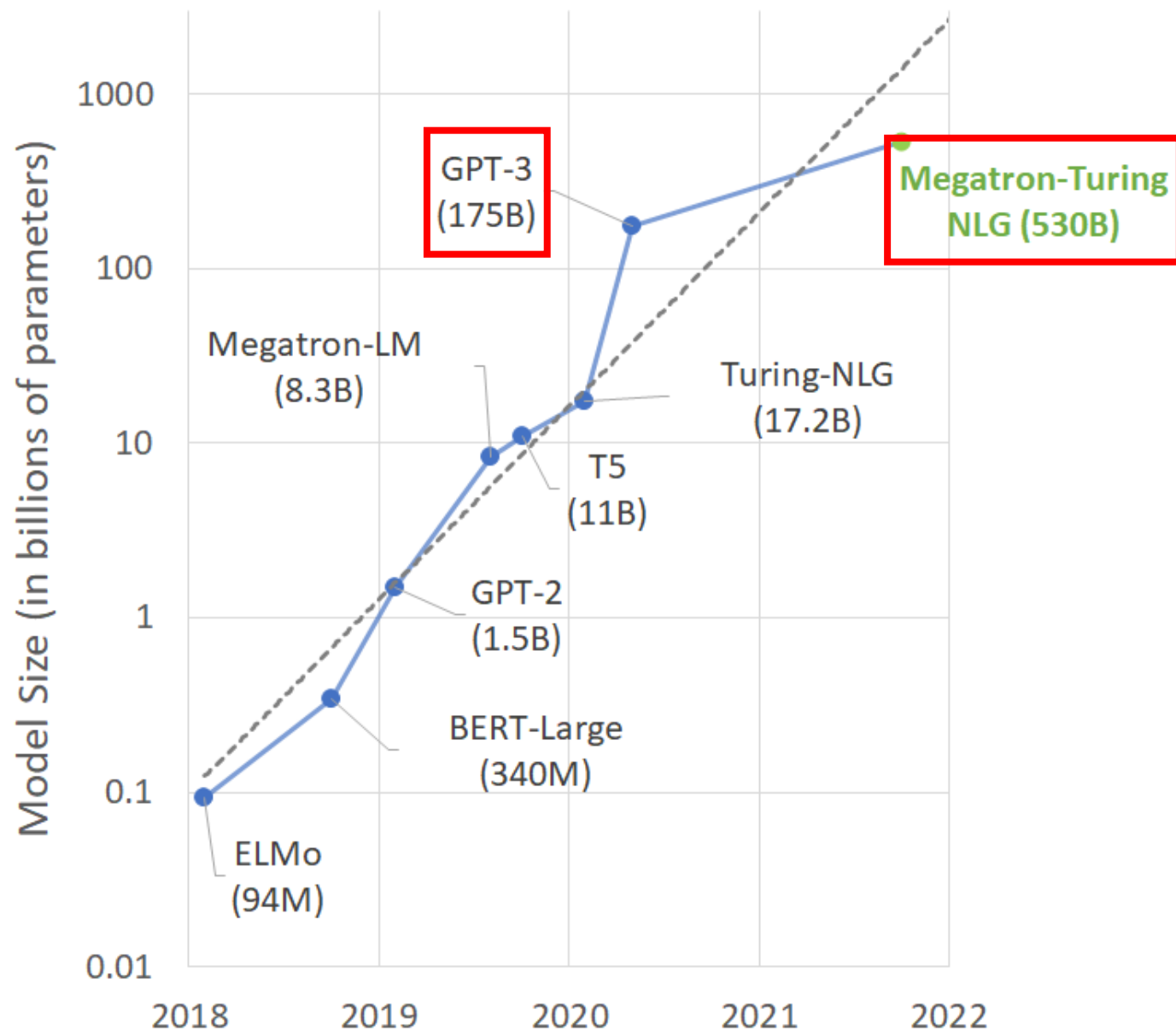
語言預訓練模型又是什麼？

Google提出了BERT後，可想而知，其他大企業或組織也競相推出了自己的語言預訓練模型。

後面就是模型**參數**一個比一個大，整個就很扯。

當然也有些模型專注在相同效果下**減少參數**。





Google Colab + Hugging Face 實作

Google Colab + Hugging Face 實作

首先介紹**Google Colab**，使用過**Jupyter notebook**的人，相信對Colab一定不陌生，下面列出一些優缺點：



Google Colab + Hugging Face 實作

優點：

- 不需要架設環境，只要有網路和瀏覽器就可以執行Python程式
- 原本就內建許多機器學習的套件
- 免費使用**GPU**、**TPU**
- 因為存放在Google Drive上，所以不管要分享還是共用都很容易
- 視覺化呈現執行結果 (圖片、表格...)

Google Colab + Hugging Face 實作

缺點：

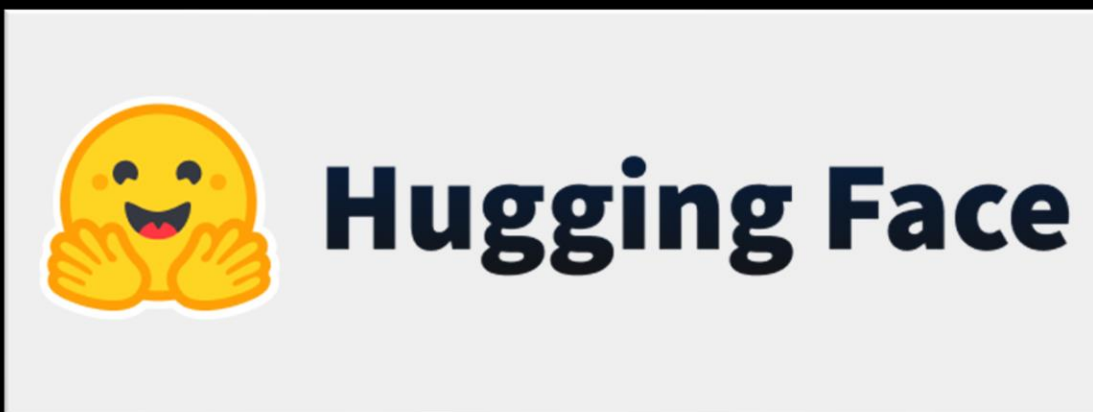
- 連續運行時間最長為**12小時**，超過就會被強制停止，而且重啟資料會被清除
- **GPU、TPU**有用量限制

雖然有缺點，但整體來說還是利大於弊。因此很推薦機器學習的初學者使用!

Google Colab + Hugging Face 實作

Hugging Face 是一間人工智慧的新創公司。

開源很多NLP領域知名的**語言預訓練模型** (如 BERT、GPT-2...)，支援**100多種語言**的文本分類、文本生成、問答等任務。



Google Colab + Hugging Face 實作


其下的**Transformers**套件，使用者可以輕易地下載、訓練、上傳語言預訓練模型。此套件目前在**GitHub**上已經有6.7萬個star，成長速度是新創公司中史上最快的。

Google Colab + Hugging Face 實作

接下來就示範怎麼使用Google Colab + Hugging Face來完成一些簡單的NLP任務吧!

Colab連結

參考資料

- 斷開中文的鎖鍊！自然語言處理 (NLP) 是什麼？
- NLP 自然語言處理 – 技術原理與其產業應用
- 進入 NLP 世界的最佳橋樑：寫給所有人的自然語言處理與深度學習入門指南
- 進擊的 BERT：NLP 界的巨人之力與遷移學習
- 台大李宏毅教授 - ELMO, BERT, GPT
-  Transformers Document

總結

今天介紹了NLP的實際應用、語言預訓練模型以及 Google Colab + Hugging Face 實作，但這不過是NLP的冰山一角而已，還有很多東西沒講。

歡迎對NLP有興趣的人自行研究，也歡迎找我一起討論，大家一起共創良好的社群環境!

總結

本議程響應開源風氣，所以投影片和範例程式碼都公開在GitHub上了，請自由使用。

這是我的GitHub，有什麼問題歡迎來聯絡我~



投影片+範例程式碼



我的GitHub

Q & A

Thanks for watching!