



Natural Language Processing

Summer 2021

#5

Chi-Jen Wu

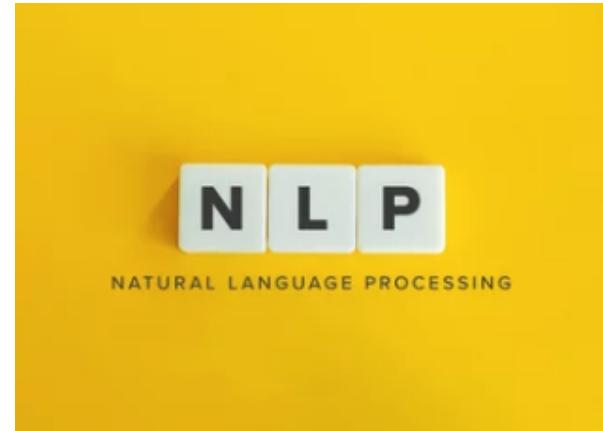


NATURAL LANGUAGE PROCESSING



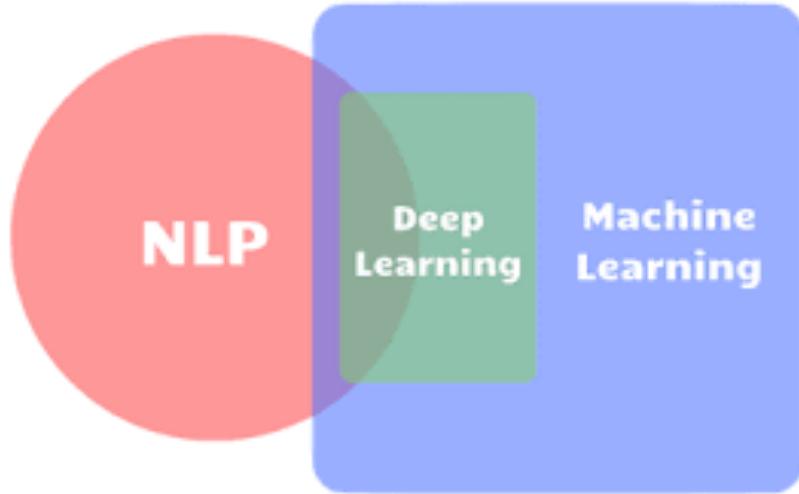
Topics

- An introduction to NLP
- Language modeling
- Representation learning
- Web crawling and indexes
- Word Embeddings
- Text classification
- Sequence modeling
- Machine learning models
- Deep learning models



Two Approaches to NLP

- Machine Learning
 - 處理語料
 - 特徵工程
 - 分類器
- Deep Learning
 - 處理語料
 - 設計模型
 - 模型訓練



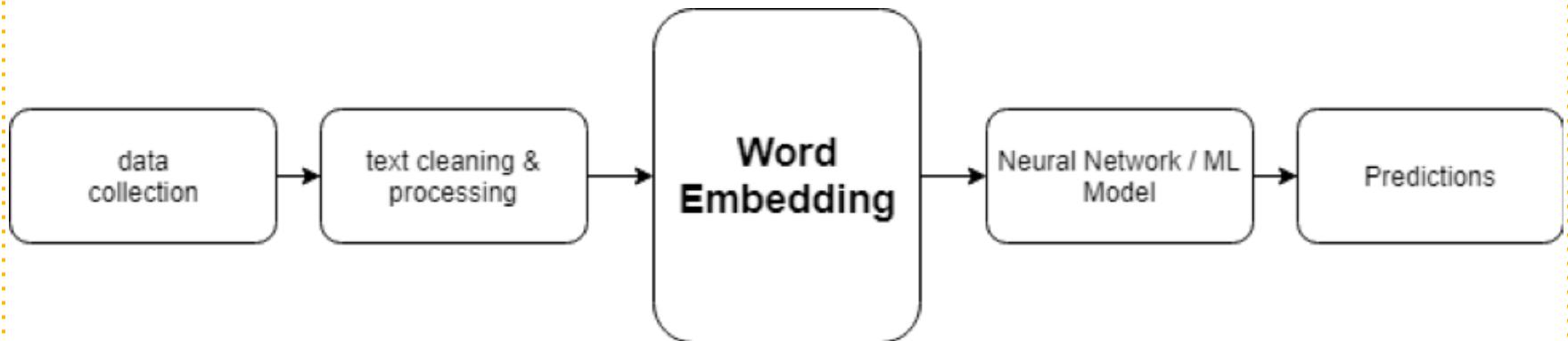


Word Embeddings

- 將字詞看作文本的最小單元
- 一個維數為所有詞的數量的高維空間映射到一個較低維數的連續向量空間中
- Embedding (映射 嵌入)
 - the representation of words
- 目前主流的方式 都以Neural Network (NN)為主
 - 之前介紹得 n-gram 就慢慢比較少人用
 - TF-IDF
- 大概分為兩大類 embedding 方法
 - Frequency based embedding
 - **Predicted Based Embedding**

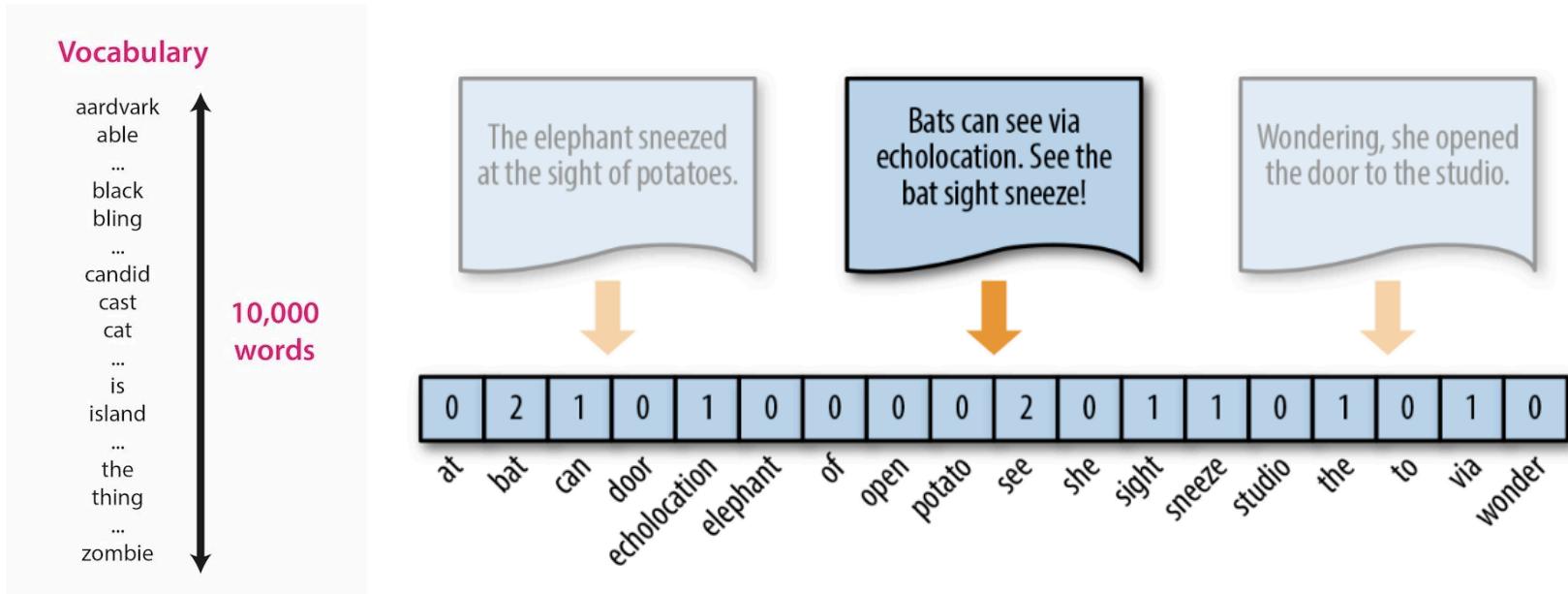


NLP flow



One Hot Encoding

- Traditional Context-Free Representations





Unsupervised learning

- K-means algorithm (算是最簡單的NLP算法之一)
 - 物以類聚
 - k是分成幾群, means就是每一群群心
 - 群心初始位置 (亂給, 所以群心初始位置會影響結果)
 - 時間複雜度 $O(NKT)$, N 是數據數量, K 是群集數量, T 是重複次數



Supervised learning

- kNN algorithm (也是最簡單的NLP算法之一)
 - k-nearest neighbors algorithm
 - 也是物以類聚,
 - 一群已經分好類別的資料 (像是yahoo電影分類)
 - 再加入未分類的資料
 - 計算資料的”距離”, 可能是座標距離或是字詞重疊程度等等 (要自行決定)

什麼是上下文

- 原出處是出埃及記
30:18-21

●《聖經預言現在進行式》

看看今天的世界發生了什麼事。
這是一些正在實現的聖經預言。

1. 廟裡的歌會消失。（阿摩司書8:3）
2. 尸體如此之多，以至於被扔掉了。（阿摩司書8:3）
3. 大地將搖動。（阿摩司書8:8）
4. 節日和慶典變得難過。（阿摩司書8:10）
5. 未來日子不好過。（阿摩司書8:10）
6. 您將無法聽到聖言。（阿摩司書8:11，12）
7. 年輕人年輕時會失去知覺。（阿摩司書8:13）
8. 婚姻將沒有慶祝活動。（耶利米書16:9）
9. 人們將死於致命的疾病。（耶利米書16:4）
10. 他們將無法為死者哀悼。
 他們將無法掩埋死者。（耶利米書16:4）
11. 他們不會去悲傷的房子。
 並且不會表現出同情。（耶利米書16:5）
12. 大大小小，老少皆宜。
 沒有人可以掩埋他們。（耶利米書16:6）
13. 禁止去參加盛宴/慶祝活動。（耶利米書16:8）
14. 來吧，我的人民，進入您的房間，然後關上門：
 隱藏自己一會兒，直到憤怒消失。（以賽亞書26:20）
15. 人的驕傲應謙卑，崇高的人應謙卑。（以賽亞書2:11）
16. 洗手，以免死亡。（出埃及記30:18-21）
17. 如果有症狀，請保持距離。
 遮住嘴並避免接觸。（利未記13:4、5、46面單）
18. 生病的人應在帳篷內停留七到十四天。（利未記13:4-5，隔離）

神準



語文學角度

- 要理解一個詞彙的語意，首先要先理解它的上下文資訊
 - 人類 = 男人 + 女人 (大概)
 - 如果要表示人類
 - 需要的維度很大
 - 好幾個向量
- 需要一個更好的表示方法 (embedding)
- Sequence modeling

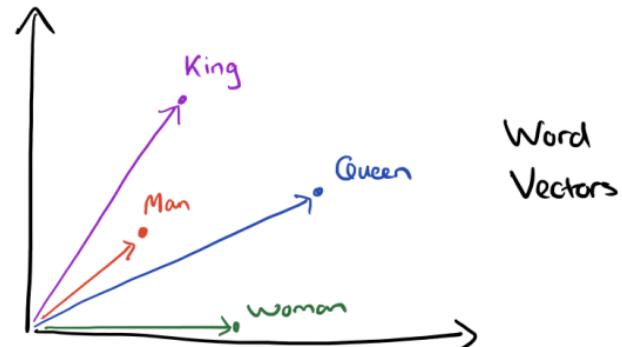


Sequence modeling

- 所以文字資料是連續的
 - 字詞的表達 (representation) 也要能連續的
 - Recurrent Neural Networks (RNN)
 - Long Short-Term Memory (LSTM)
- 很多資料都是連續的或是前後有相關的
 - 股票 (可能有關?)
 - 吃什麼晚餐 等等
 - 跟你前幾天吃的晚餐是有關係的 (基本上)

word2vec

- 根據輸入的「詞的集合」計算出詞與詞之間的距離
- 「字詞」轉換成「向量」形式
- 餘弦值 (cosine)
 - 計算距離範圍為 0–1 之間
 - 值越大兩個詞關聯度越高
- Poorly understood (目前)
 - 還在研究討論中





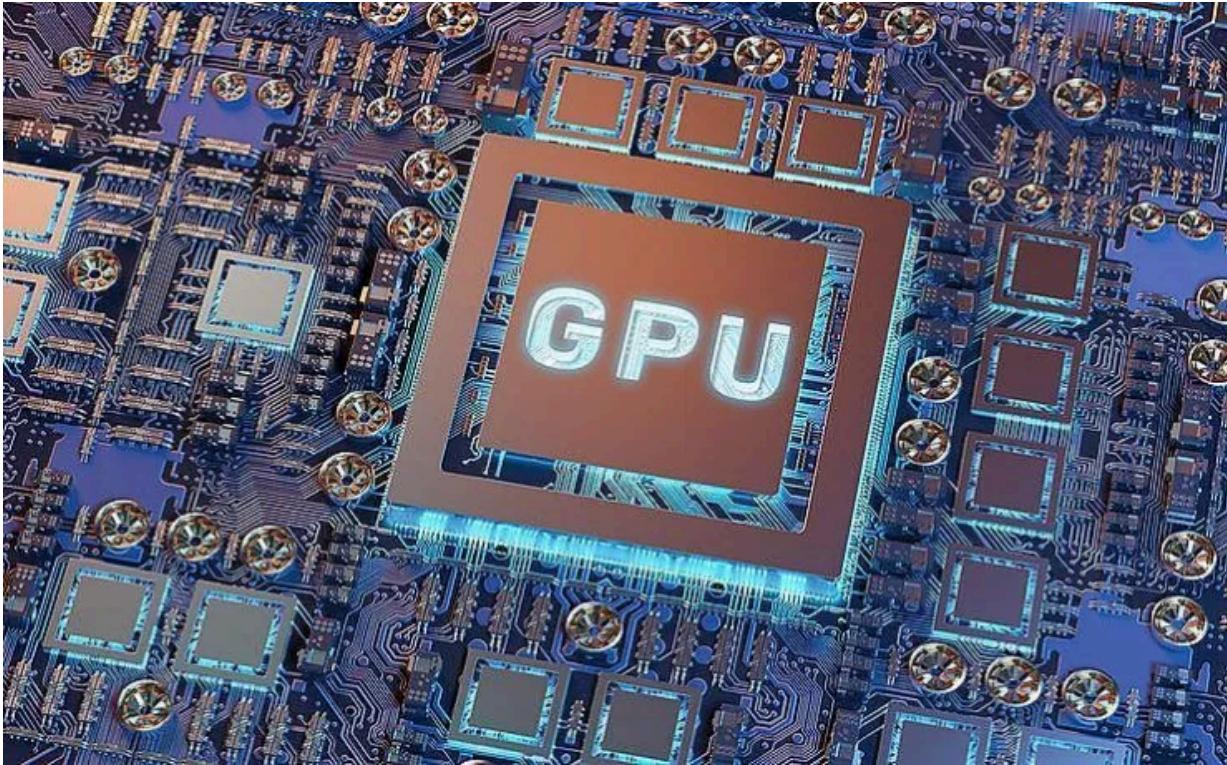
新世代的NLP工具

- Seq2Seq (2014)
 - Google 的開源工具
- Bert (2018)
 - Google 的開源工具
- GPT3 (2020)
 - OpenAI 的開源工具
- End-to-end learning





GPU 加速



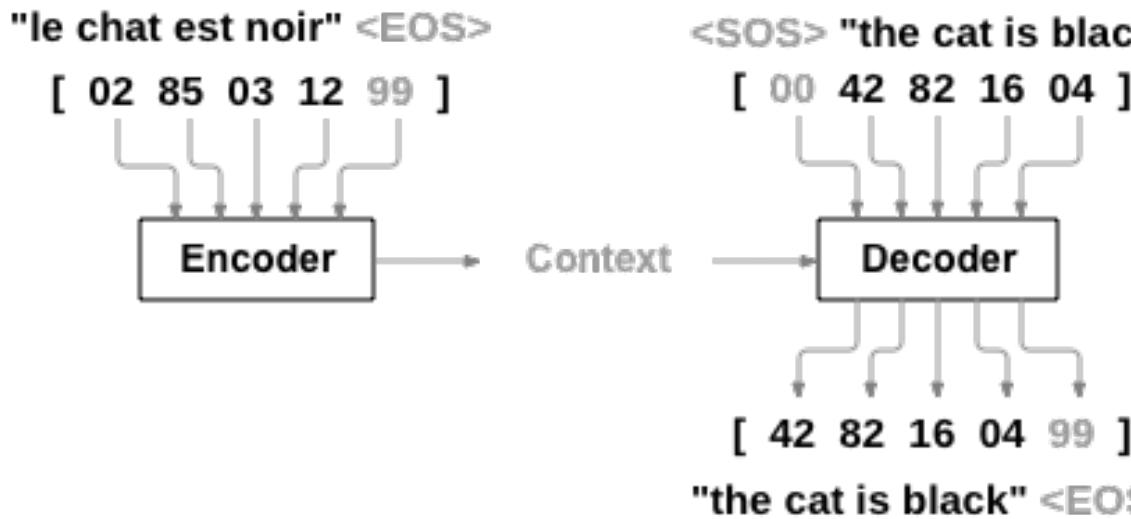


End-to-end learning

- A type of Deep learning process
- 傳統NLP過程光處理語料
 - 分詞 詞性 語意分析 等等
 - 每階段的成效會影響下個的成效
 - 將人們之前的經驗知識處理成feature 再給model訓練
- 現在
 - Raw data → model → output (Answer)

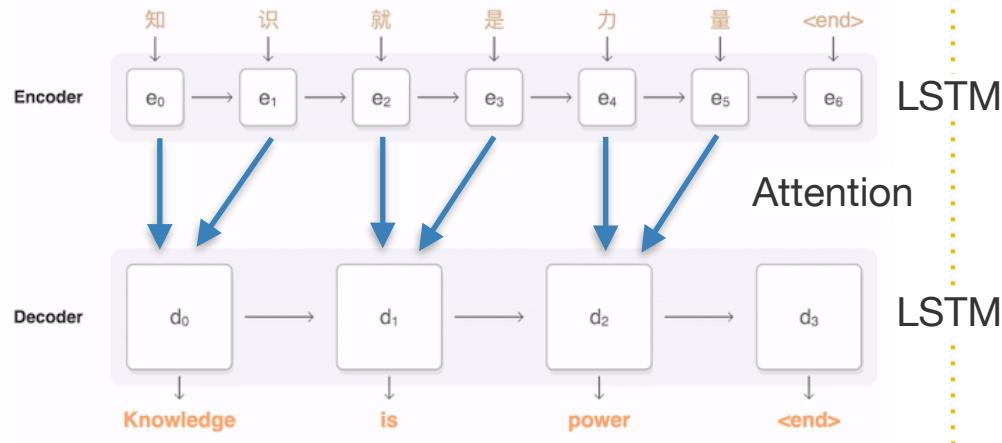
End-to-end learning

- Encoder and decoder

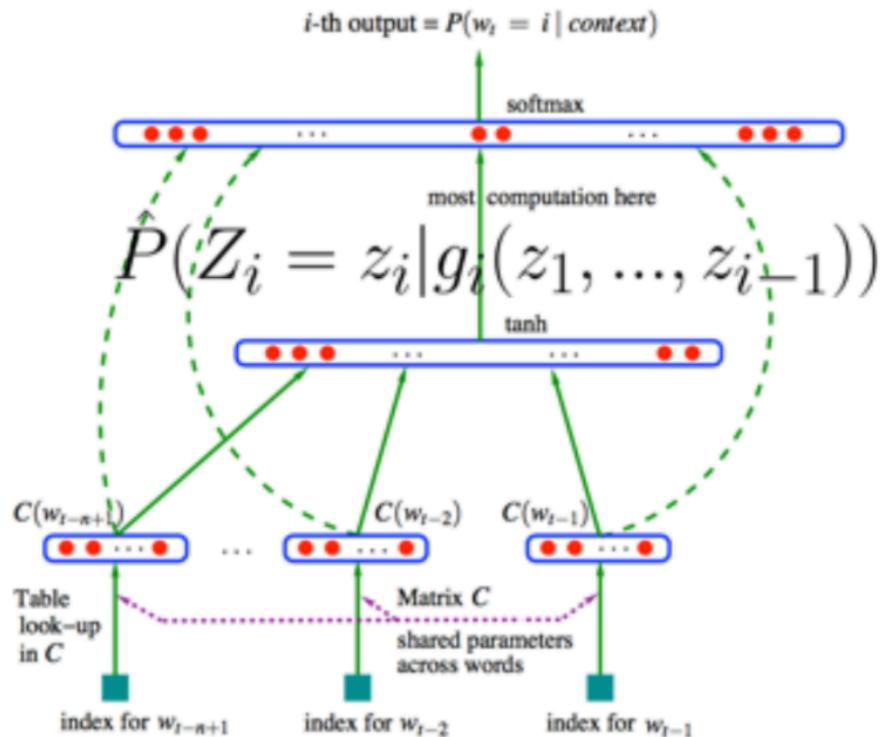


Seq2Seq

- Seq2Seq (2014)
 - Google 的開源工具
- Google 翻譯使用的模型
- 同義詞的進展
 - 拉肚子/腹瀉
 - 表示成向量
 - 向量的操作 (可微分)
 - 拉肚子 (v_1) - 0.01 = 腹瀉 (v_2)



Seq2Seq 模型



長庚大學是台灣的頂尖大學

長庚/大學/台灣/頂尖/大學

分詞轉換成向量

P(全句) =

$P(\text{長庚}|\text{大學}) * P(\text{長庚}|\text{大學}|\text{台灣}) * P(\text{長庚}|\text{大學}|\text{台灣}) * P(\text{長庚}|\text{大學}|\text{台灣}|\text{頂尖})$

從歷史的語料分析訓練神經網路 (NN)

後可以得到每一個的機率

就可以拿來預測一個詞的下一個字詞的出現機率

非常簡單的說是這樣



Bert - NLP的里程碑

- Bidirectional Encoder Representations from Transformers
- 理解上下文的語言代表模型 (Contextual Embeddings)
- 字詞轉化成多維連續向量
- NLP 任務的通用架構
- 利用別人以訓練好的模型
 - 自己訓練通常要花好幾天
 - 極需要GPU

長庚大學是 ?? 的頂尖大學





BERT-As-Service

- Tencent AI Lab 開源的服務
- 安裝之後使用bert 變的很方便
- 也有各種預先訓練好的模型可以使用
- 可以算無腦上手

```
>>> from bert_serving.client import BertClient
>>> bc = BertClient(ip='localhost')
>>> bc.encode(['hello world', 'good day!'])
array([[-0.5236851 ,  0.27427185,  0.00751604, ...,  0.09255812,
       -0.33134174, -0.166861  ],
      [-0.12277935,  0.46469706,  0.02634781, ..., -0.3485587 ,
       -0.79110664, -0.37120932]], dtype=float32)
>>>
```



GPT3

- Generative Pre-Training
- 通用語言模型
- Generative Pre-training + Fine-tuning
 - Bert 後來參考這種做法
- 資料量越大 → 越厲害
- 幾乎什麼鬼都可以做



資料量越大 → 越厲害 → 錢多多



模型	發布時間	參數量	預先訓練數據量
----	------	-----	---------

GPT	2018 年 6 月	1.17 億	約 5GB
-----	------------	--------	-------

GPT-2	2019 年 2 月	15 億	40GB
-------	------------	------	------

GPT-3	2020 年 5 月	1,750 億	45TB
-------	------------	---------	------

訓練費用預估為 1,200 萬美元

GPT-3 Demo Showcase



GPT-3 DEMO | GPT-3 showcase

[GPT-3 Market Map](#) [Youtube Channel](#) [What's GPT-3?](#)

[Get listed](#)

GPT-3 Demo Showcase, 190+ Apps, Examples, & Resources

Get inspired and discover how companies are implementing the OpenAI GPT-3 API to power new use cases

Search for apps, categories, ...

<https://gpt3demo.com/>

HW#5 seq2seq 英翻中實作



1. 訓練給定中文語料庫 (1.1G) 520萬筆資料 (JSON)

[https://drive.google.com/open?
id=1EX8eE5YWBgxCaohBO8Fh4e2j3b9C2bTVQ](https://drive.google.com/open?id=1EX8eE5YWBgxCaohBO8Fh4e2j3b9C2bTVQ)

2. 利用keras / pytorch / seq2seq-lstm / sklearn

seq2seq 英翻中實作

3. 輸入一個簡單的英文句子

3.1 “It is a nice day” → 今天天氣真好

3.2 最後100筆為testing data (需呈現testing結果)

3.3 輸出中文翻譯結果

3.4 分數以輸出中文翻譯結果評分



HW#5 seq2seq實作

4. 請 Commit 到 github
5. 請記得要commit ≥ 5 次
6. 不要改github branch name
7. 8/20 12:00 前 push github
8. 遲交或補交 扣20分
9. 一個function 不能超過50行 (超過一行扣一分)



Go to file

Code

Clone



HTTPS GitHub CLI

<https://github.com/cjwu/cjwu.github>



Use Git or checkout with SVN using the web URL.

Open with GitHub Desktop

Download ZIP

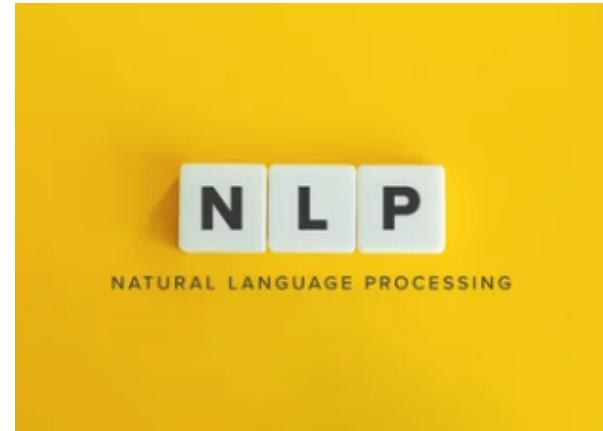
2 months ago

作業/隨堂練習
請繳交你自己
github上的這個Link



Topics

- An introduction to NLP
- Language modeling
- Representation learning
- Web crawling and indexes
- Word Embeddings
- Text classification
- Sequence modeling
- Machine learning models
- Deep learning models





下週 8/19 沒有上課

下下週 8/26
各組期末報告
五頁投影片
每組十分鐘內
第一頁 題目 組員
程式作品 demo



NLP

資料 / 算法 / 算力

窮的拼算法

有錢壓算力

資料只會越來多 越來越便宜 跟硬碟一樣

萬物皆可 embedding



Thanks!

Open for any questions

CJ Wu

cjwu@mail.cgu.edu.tw