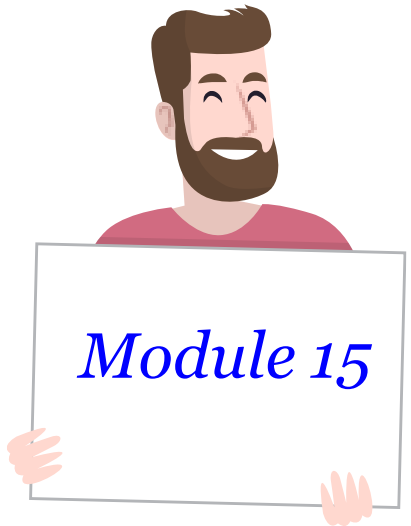
自然語言處理與Word2vec介紹



designed by freepik

Estimated time:
45 min.

學習目標

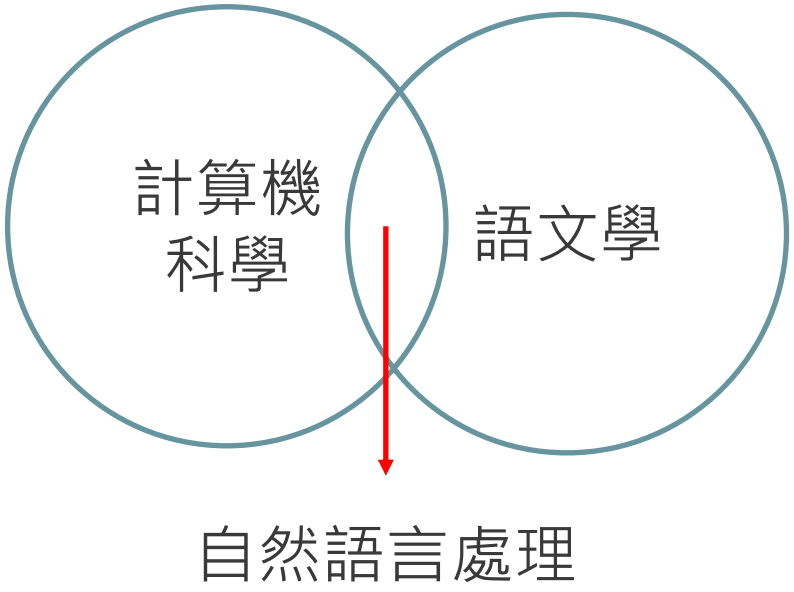
- 15-1:自然語言處理介紹
- 15-2: Word2vec介紹
- 15-3: Word2vec應用

15-1:自然語言處理介紹

- 自然語言處理
- 自然語言處理與人工智慧的關係
- 語料庫

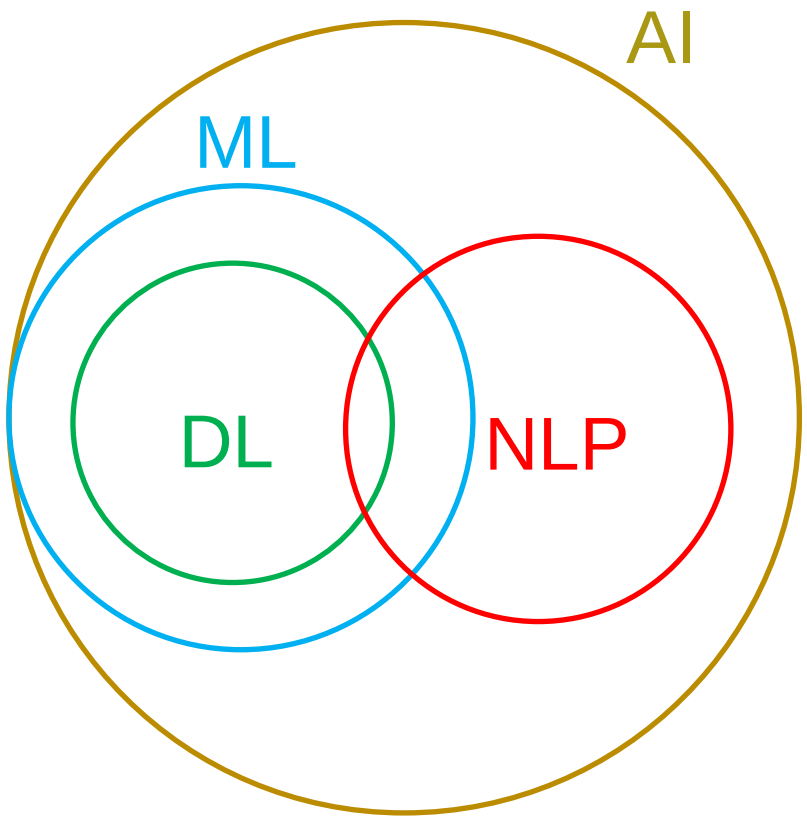
自然語言處理

- 自然語言處理(NLP)是一個幫助電腦去了解人類語言的一門技術
  - 在實際上，人類的語言滿複雜的，要讓電腦去學習理解相當不容易
  - 自然語言處理除了牽扯到計算機科學外，也與語文學有關



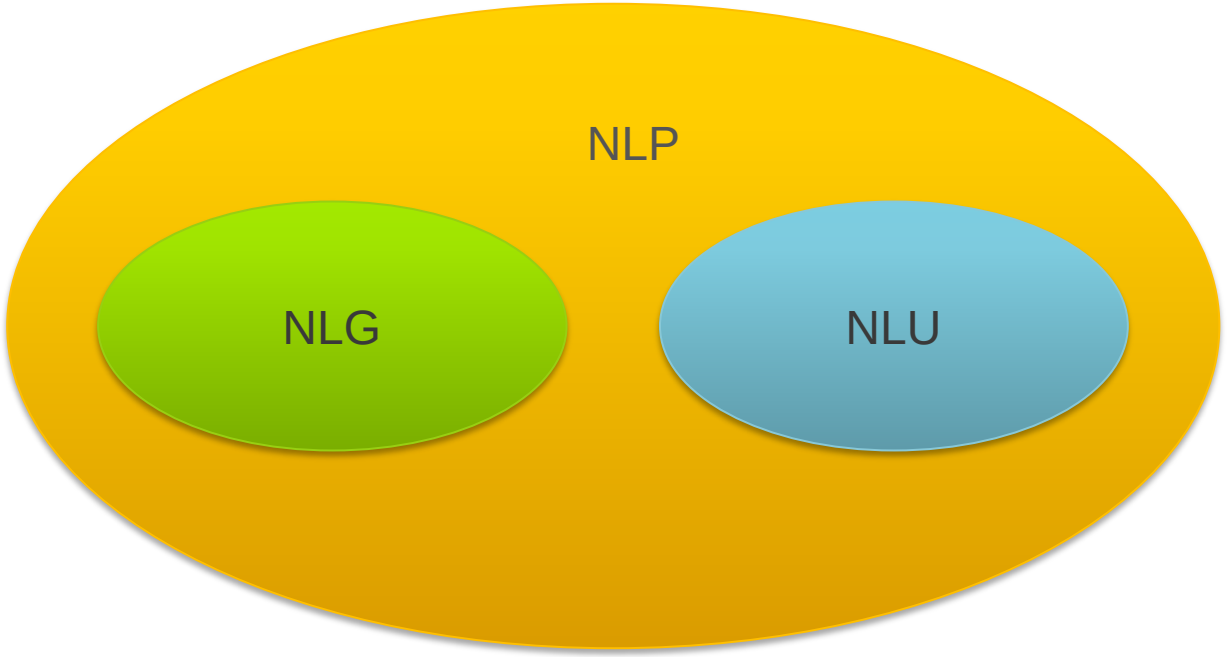
自然語言處理與人工智慧的關係

- 自然語言是屬於人工智慧領域裡的一個子領域
  - 其與機器學習、深度學習都有交集



自然語言處理

- 在自然語言處理領域裡，又包含兩種常被提及的領域
  - 自然語言理解(NLU)，即讓AI能了解語文的意思
  - 自然語言生成(NLG)，即讓AI能自動產生自然語言

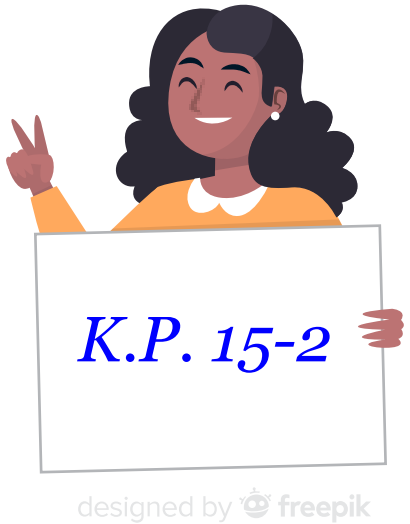


語料庫

- 在自然語言處理裡，輸入的文字資料集常被稱為語料庫
  - 語料庫可分為單一語言或多國語言
  - 像是維基百科、 Google Books Ngram、 Brown等都是非常著名的大型語料庫

15-2: Word2vec介紹

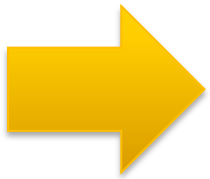
- 為什麼需要Word2vec
- Bag of words



為什麼需要Word2vec

- 因為電腦只能讀懂數字，所以當我們需要一種方法把人類的自然語言輸入給電腦，並轉換成對應的向量
  - 這就是Word2vec的由來
  - 而一個文字轉換成相對應的向量，此向量稱為word embedding

“你好”



$$\begin{bmatrix} -0.1 \\ 2 \\ 0.5 \end{bmatrix}$$

Bag of words

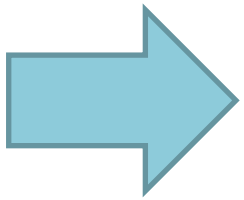
- 最基礎將文字轉成向量的方法為Bag of words
- Bags of words會先將所有語料庫閱讀一遍，並將所有出現過不同的文字給予一個字ID

文件	內容
1	[“it”, “was”, “the”, “best”, “of”, “times”]
2	[“it”, “was”, “the”, “worst”, “of”, “times”]
3	[“it”, “was”, “the”, “age”, “of”, “wisdom”]
4	[“it”, “was”, “the”, “age”, “of”, “foolishness”]

Bag of words

- 假設Bags of words閱讀完語料庫後得到以下不同字的字ID

“it”
“was”
“the”
“best”
“of”
“times”
“worst”
“age”
“wisdom”
“foolishnes
s”



word	word ID
it	1
was	2
the	3
best	4
of	5
times	6
worst	7
Age	8
wisdom	9
foolishness	10

Bag of words

- 則我們可以針對字ID給出不同的字向量，做法為產生一個與字彙量等長度之向量，並把相對應位置之英文字填寫1，其他地方寫0
  - 例如，”was”字向量為[0,1,0,0,0,0,0,0,0,0]

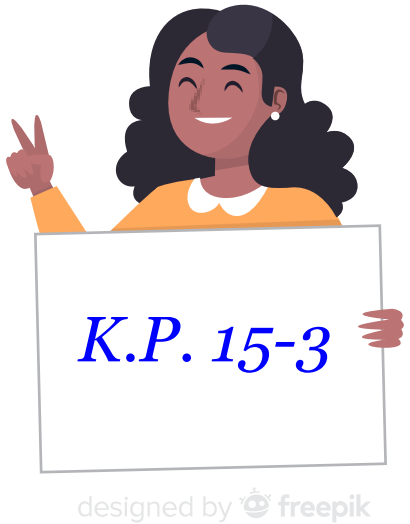
word	word ID
it	1
was	2
the	3
best	4
of	5
times	6
worst	7
Age	8
wisdom	9
foolishness	10

Bag of words

- **Bag of words雖然很快的能把文字轉成向量，但無法真的讓電腦理解每個文字的意義，因此現在通常會用更進階的方法來實作Word2vect**
  - **例如skip-gram、CBOW(之後章節會教到)**

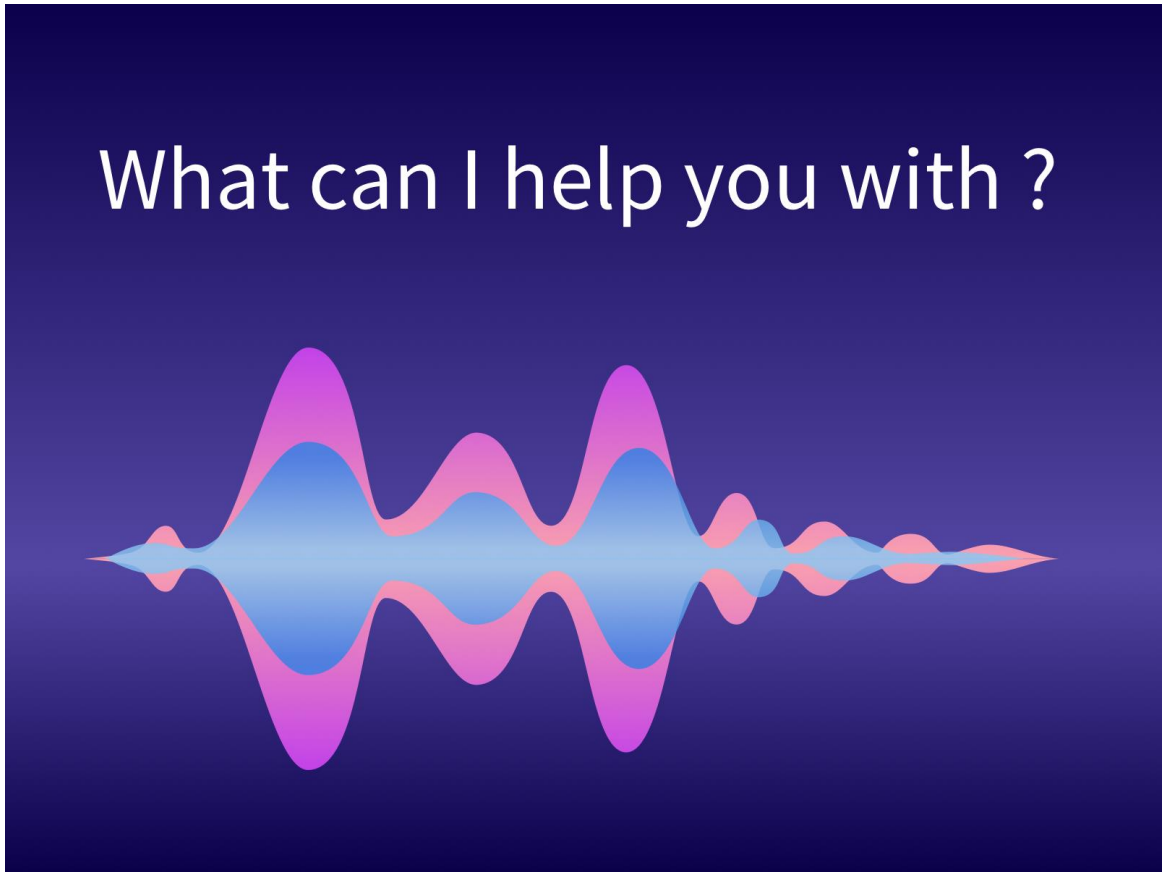
15-3: Word2vec應用

- **Word2vec應用**



Word2vec應用

- 語音助理

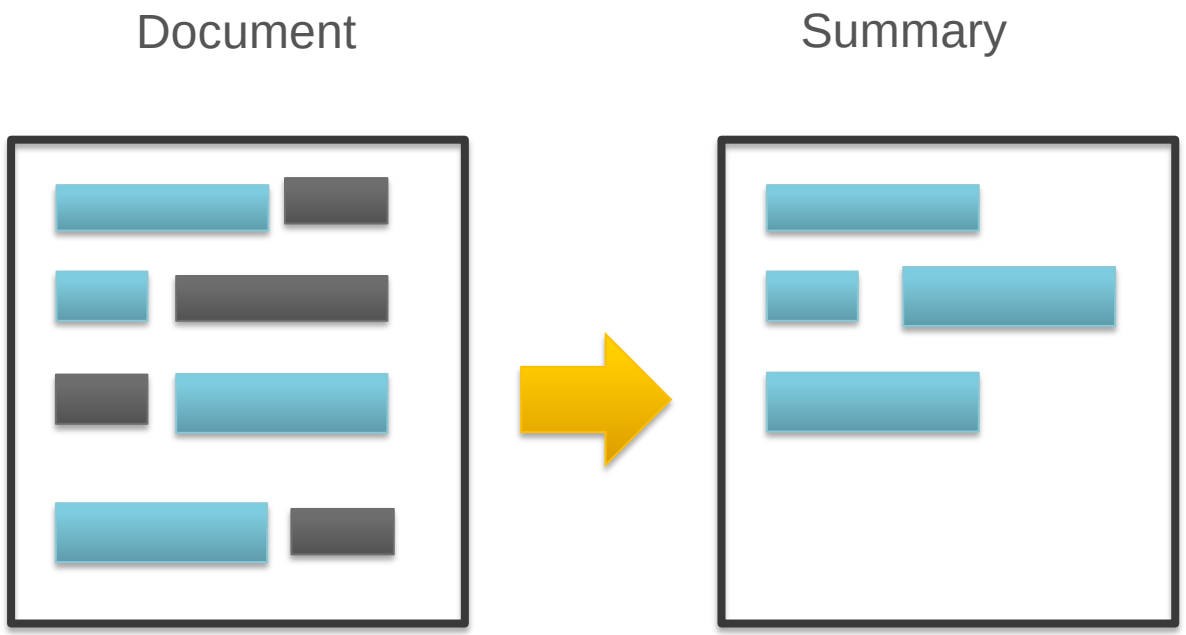


Word2vec應用

- 文章分類
  - 例如垃圾郵件分類、電影影評分析、新聞文章分類

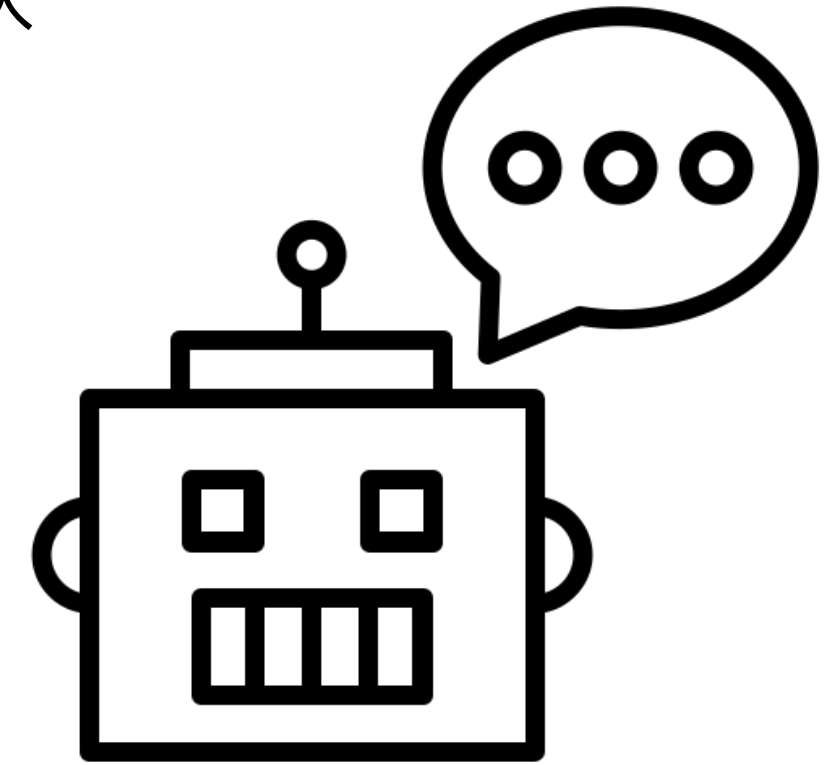
Word2vec應用

- 文章自動總結
  - 將長篇的文章總結成簡短的幾個句子



Word2vec應用

- 聊天機器人
  - 智慧客服、長照機器人



Demo 15-3

- jieba安裝
- jieba使用
- 實作bag of words

線上Corelab

- 題目1： 給予一個中文句子，使用jeiba斷詞斷出來
  - 使用精確模式
- 題目2：給予一個英文語料庫，使用scikit-learn做bag of words(基礎)
- 題目3：給予一個英文語料庫，使用scikit-learn做bag of words(進階)

本章重點精華回顧

- 自然語言處理
- NLP與AI的關係
- Bag of words
- Word2vec應用

Lab:自然語言處理套件使用

- **Lab01: jieba安裝**
- **Lab02: jieba使用**
- **Lab03: 實作bag of words**

Estimated time:
20 minutes


THANK YOU