

Fasttext

十分鐘就能讓電腦「懂人話」?

108.9.5 (四) 晚上9:00

NLP入門

實戰展示

即時QA



周凡剛 老師

點擊「提醒我」
不怕錯過直播！



自我介紹



周凡剛(Elwing)



birdfan8814@gmail.com



台大電機學士/碩士畢業



經歷：



聯發科R&D



資策會/中央大學Python講師



聯發科/日盛內訓講師

NLP



NLP(自然語言處理): Natural Language Processing



不管是什麼任務，中心思想就一個，讓電腦能夠瞭解人類語言



了解人類語言一定要搞定的兩件事！



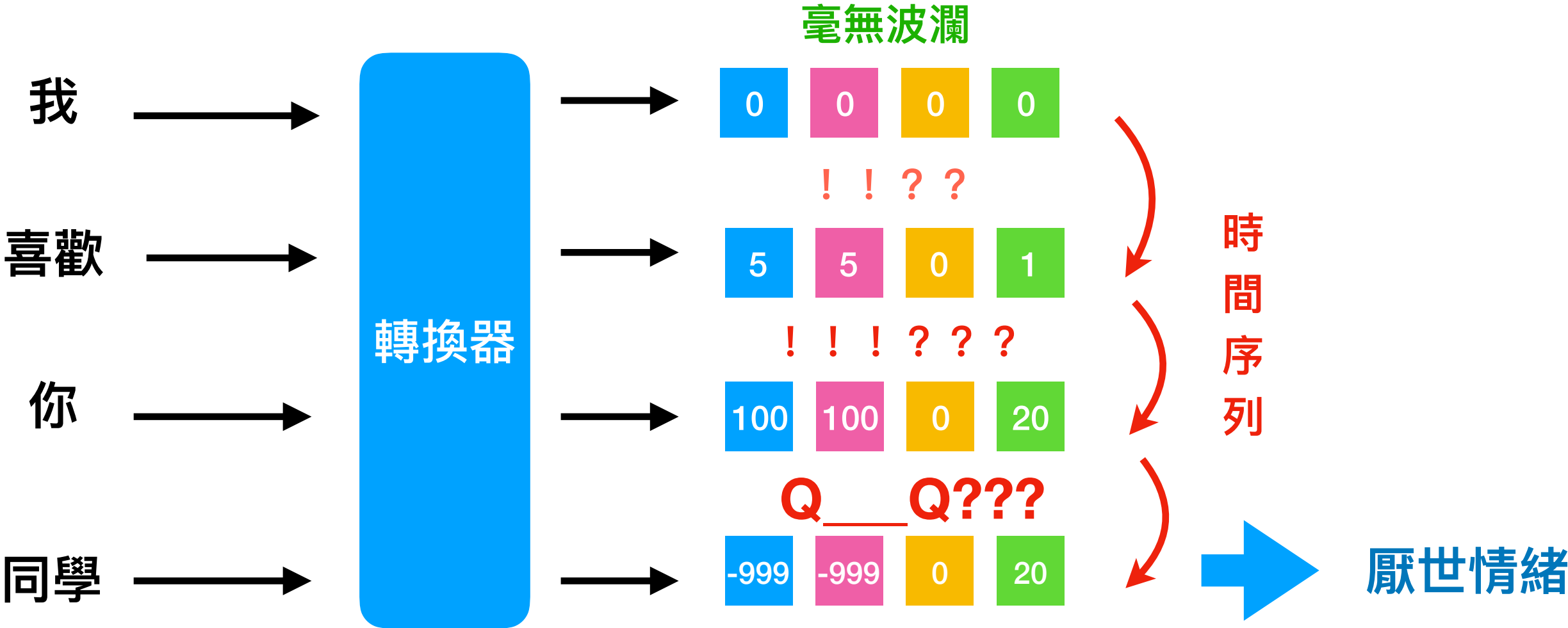
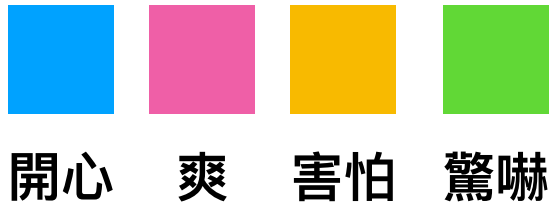
詞義(語意): 人類在聽的時候其實不是聽每個詞，而是你的大腦在聽到某個詞的時候會產生對應的情感感受



上下文(序列): 人類在聽的時候其實還會參考上下文，根據上下文的累積，我們可能會有不一樣的意思
e.g. 討厭(正常), 討厭(裝可愛), 你沒那麼讓人討厭(喜歡)

一圖流

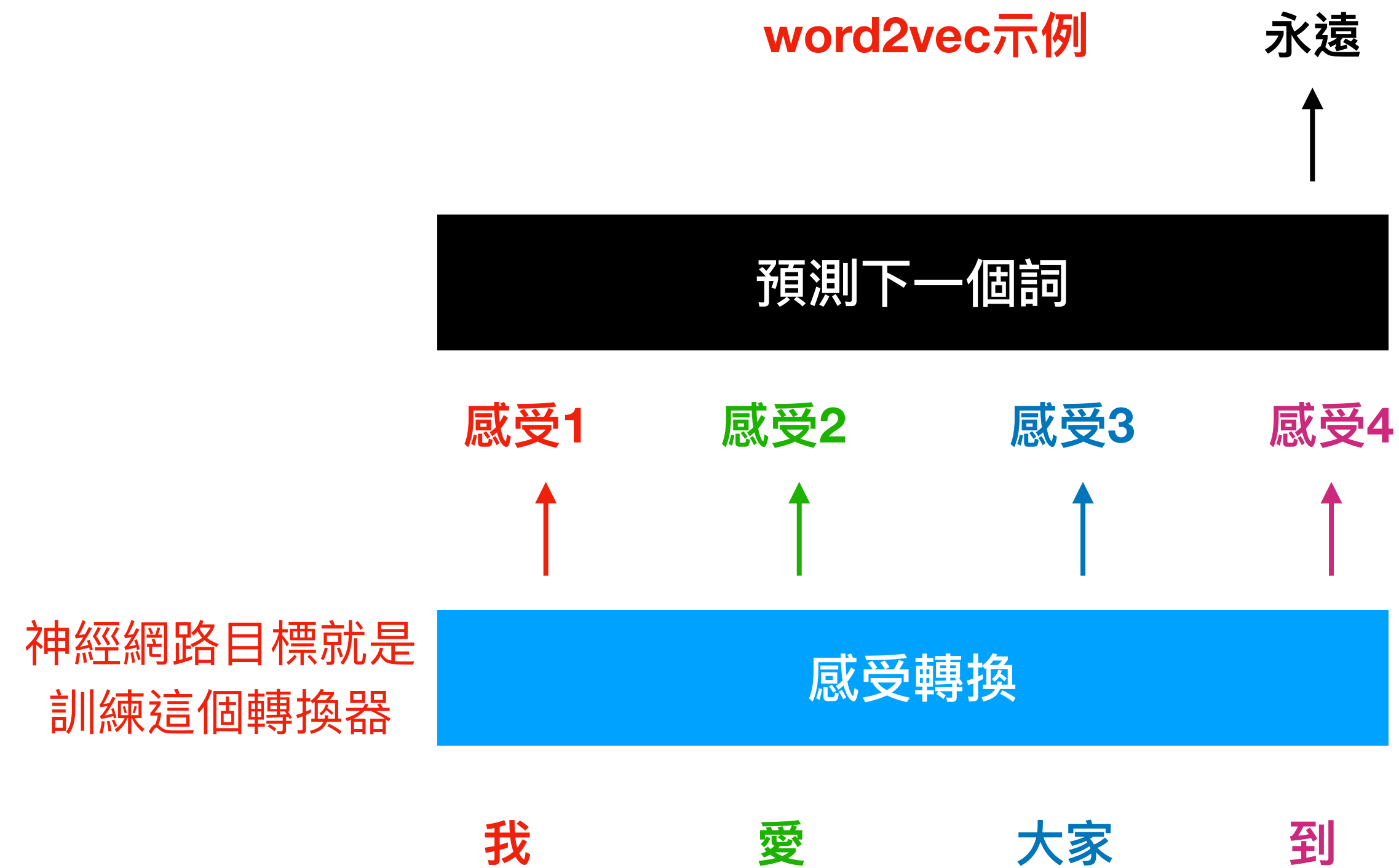
語意度量器



詞向量



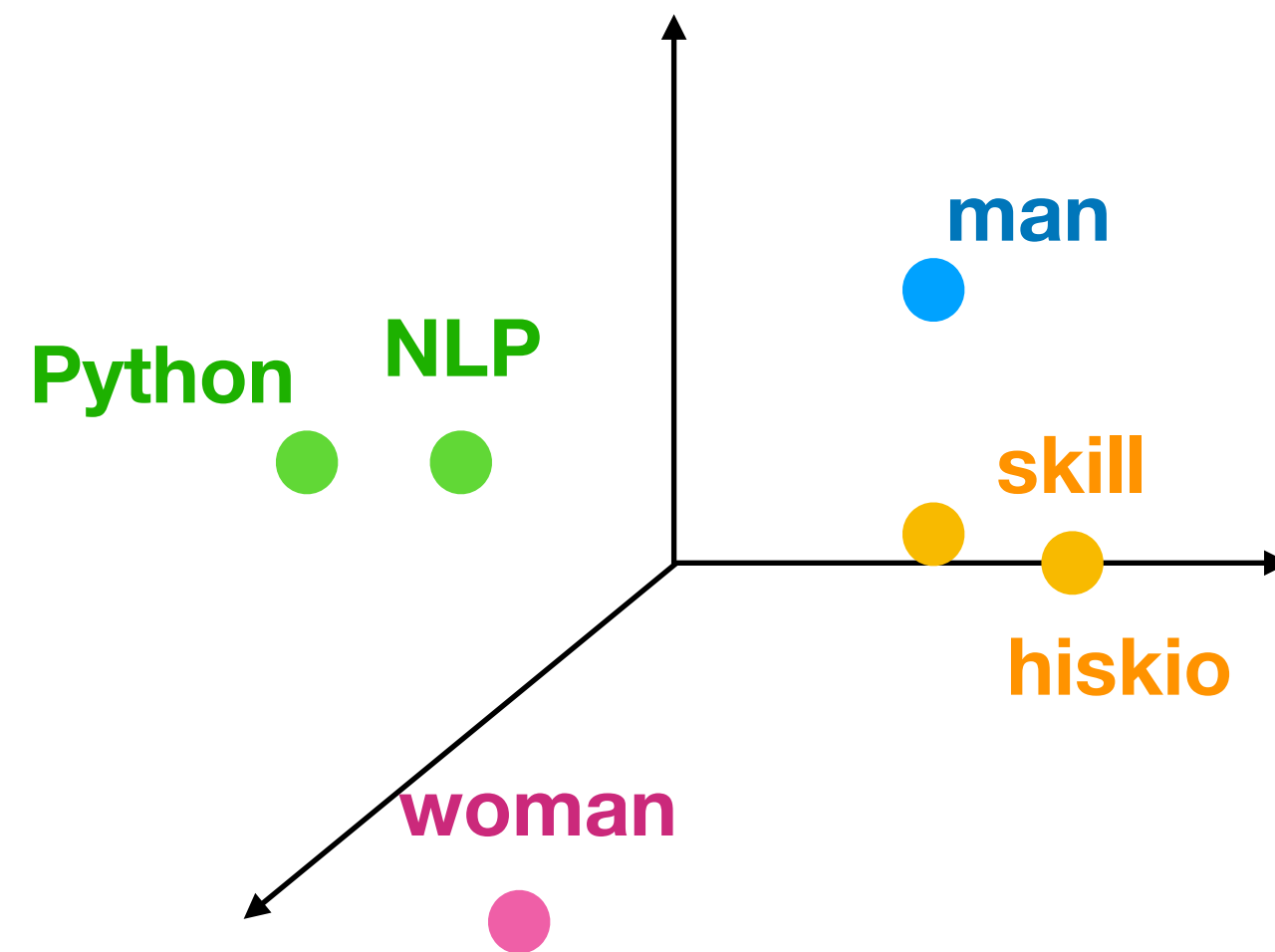
我們在NLP任務裡可能會訓練很多種不同模型，但他們在中間大部分都會先做一件事，產生語意



相似度



有了感受我們就可以把感受排列在空間中，比較兩個感受的相似度或者計算兩個感受間的差距



cos距離
只計算方向，不計算大小
-1(180度,最不相似)
1(0度,最相似)

有了詞向量
我們就可以把詞表示在空間中!

FastText



今天我們為何介紹FastText呢



NLP一個很大的問題是資料有無辦法撐得起你的模型，何不當個『**不勞而獲的人**』利用大公司訓練好的完整模型，100%比你自已訓練出的語意還準！



FastText主要以標籤來訓練詞向量，並且加入**n-gram**的機制，字詞級別的**n-gram**，尤其字級別的n-gram對我們的英文分析如虎添翼，例如tourism，模型會同時帶入tou, our, ism...等等3-gram字，那我們就會發現他跟其他以**ism**結尾的字有點類似

[Docs](#) [Resources](#) [Blog](#) [GitHub](#)

Word vectors for 157 languages

We distribute pre-trained word vectors for 157 languages, trained on *Common Crawl* and *Wikipedia* using fastText. These models were trained using CBOW with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 and 10 negatives. We also distribute three new word analogy datasets, for French, Hindi and Polish.

號稱已經訓練好157種語言的FastText

Demo Time

Colab網址: <https://reurl.cc/xDDYG1>

#@title 比較兩個特定詞的相似度

```
text1 = 'Coldplay' #@param {type:"string"}
text2 = 'OneRepublic' #@param {type:"string"}
model.wv.similarity(text1, text2)
```

0.6214945

↑ ↓ ↻ 💬 ⚙️ 🗑️ ⋮

比較兩個特定詞的相似度

text1: " Coldplay "

text2: " OneRepublic "

Python相關

