智慧運算學院
College of Intelligent Computing
人工智慧學系
Department of Artificial Intelligence

智慧運算技術導論
自然語言處理篇 - Embeddings

林英嘉 (Ying-Jia Lin)
長庚大學人工智慧學系
2026/01/26

Slido
AIMD (Link)

www.slido.com

slido

Product Solutions Pricing Resources Enterprise

Log In **Sign Up**

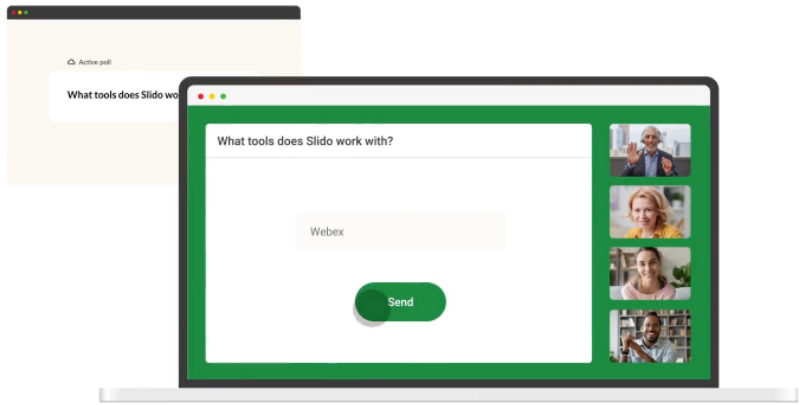
Joining as a participant? # AIMD

By using Slido you accept the Acceptable Use Policy

The easiest way to make your meetings interactive

Engage your participants with live polls, Q&A, quizzes and word clouds — whether you meet in the office, online or in-between.

Get started for free



Your privacy matters.

We use cookies to improve your experience, analyze traffic, and serve personalized ads. Cookie settings.

Ying-Jia Lin

長庚大學 人工智慧學系 助理教授 (2025/02 -)

國立清華大學 資訊工程學系 博士後研究員 (2024 - 2025)

國立成功大學 資訊工程學系 博士 (2019 - 2024)

國立陽明大學 生物醫學資訊研究所 碩士 (2017 - 2019)

長庚大學 生物醫學系 學士 (2013 - 2017)

Outline

- 深度學習與自然語言處理的概念
- Embeddings
  - Sentence embeddings
  - Word embeddings

什麼是自然語言處理（Natural Language Processing, NLP）？

文字搜尋

下一個字預測

程式語言 vs. 自然語言

程式語言

main transformers / src / transformers / loss / loss_deformable_detr.py

Code Blame 178 lines (151 loc) · 7.15 KB

```
3
4   from ..image_transforms import center_to_corners_format
5   from ..utils import is_scipy_available
6   from .loss_for_object_detection import (
7       HungarianMatcher,
8       ImageLoss,
9       _set_aux_loss,
10      generalized_box_iou,
11      sigmoid_focal_loss,
12  )
13
14
15  if is_scipy_available():
16      from scipy.optimize import linear_sum_assignment
17
18
19  class DeformableDetrHungarianMatcher(HungarianMatcher):
20      @torch.no_grad()
21      def forward(self, outputs, targets):
22          """
23          Differences:
24          - out_prob = outputs["logits"].flatten(0, 1).sigmoid() instead of softmax
```

自然語言

先生不知何許人也，亦不詳其姓字。宅邊有五柳樹，因以為號焉

閑靜少言，不慕榮利。好讀書，不求甚解，每有會意，便欣然忘食。性嗜酒，家貧，不能常得。親舊知其如此，或置酒而招之。造飲輒盡，期在必醉，既醉而退，曾不吝情去留。環堵蕭然，不蔽風日；短褐穿結，簞瓢屢空。——晏如也。常著文章自娛，頗示己志。忘懷得失，以此自終。

贊曰：黔婁之妻有言：「不戚戚於貧賤，不汲汲於富貴。」極其言，茲若人儔乎？酣觴賦詩，以樂其志。無懷氏之民歟！葛天氏之民歟！

什麼是自然語言處理？

天青色等煙雨

I like apple!

?

定義 ➡️

自然語言處理（Natural Language Processing, NLP）
是讓電腦能「理解、分析、產生」人類語言的技術。

The Revolution of ChatGPT

ChatGPT came out in November, 2022.

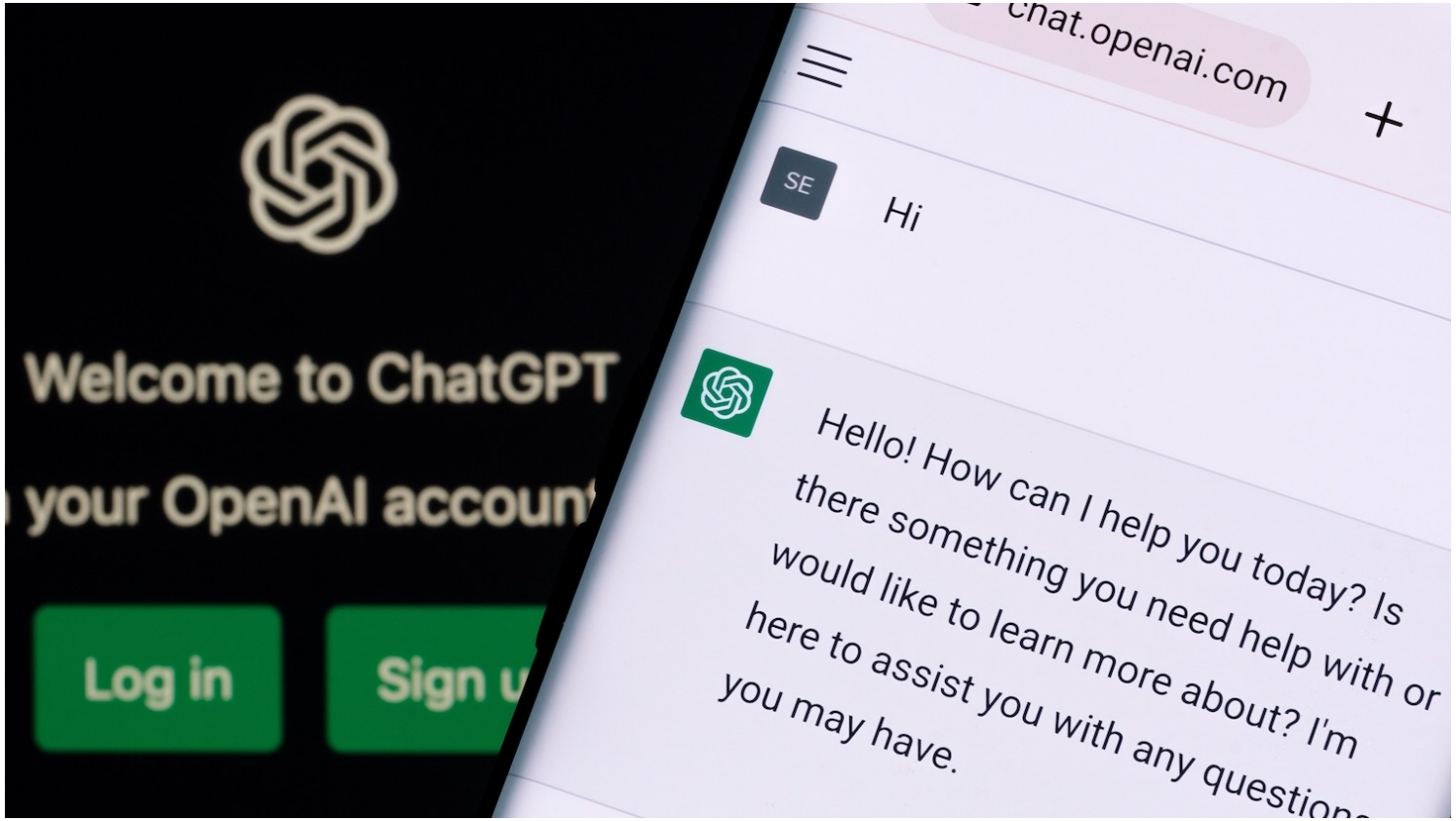
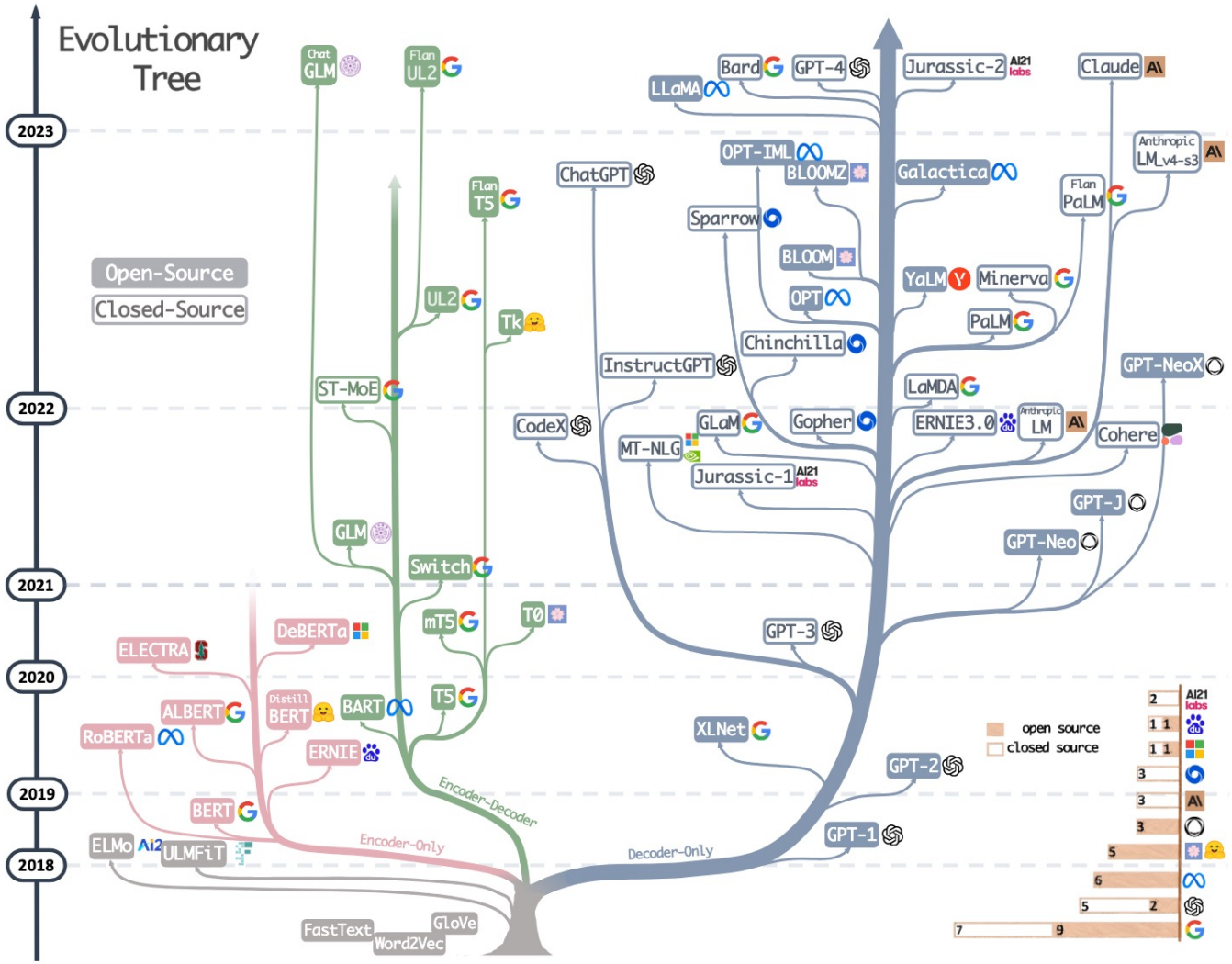


Figure source: https://poole.ncsu.edu/thought-leadership/article/lets-chat-about-chatgpt/
https://openai.com/index/chatgpt/

大型語言模型的演化樹



Yang, Jingfeng, et al. "Harnessing the power of llms in practice: A survey on chatgpt and beyond." ACM Transactions on Knowledge Discovery from Data 18.6 (2024): 1-32.

以前的 GPT 有口吃

InstructGPT: ChatGPT 的前身

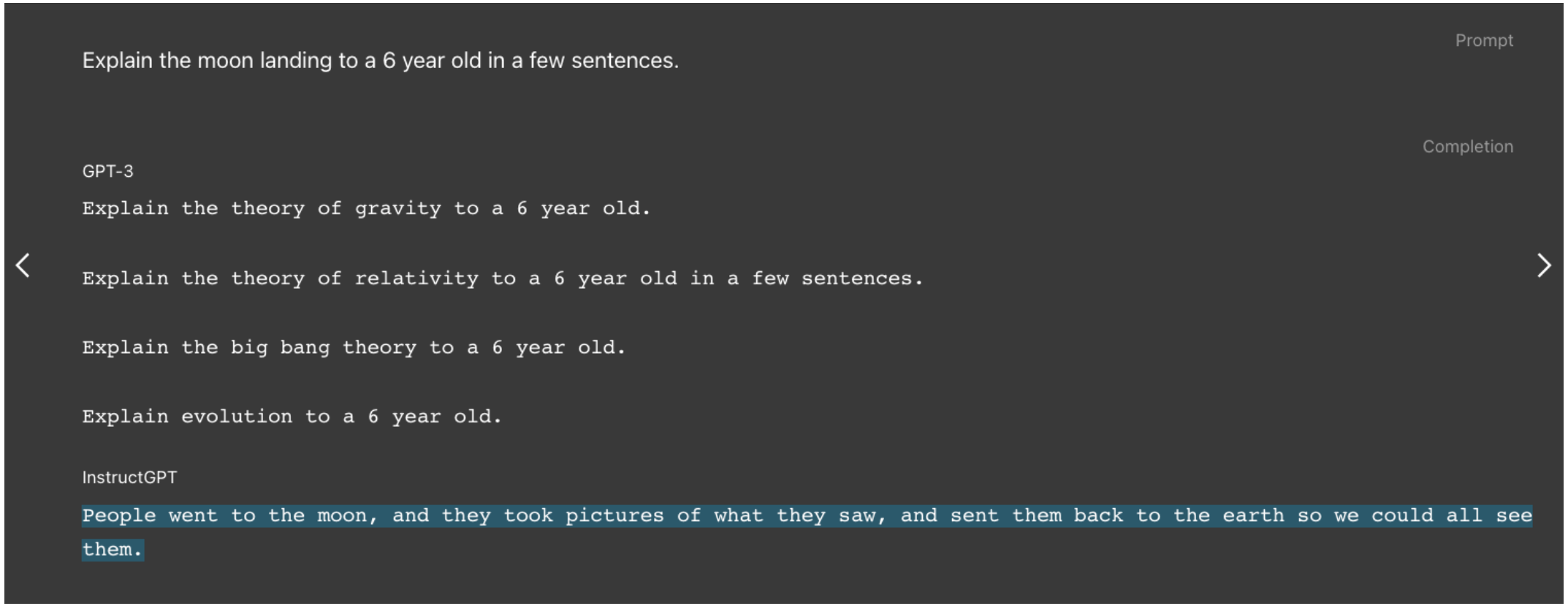


Figure source: https://openai.com/index/instruction-following/
Ouyang, Long, et al. “Training language models to follow instructions with human feedback.” NeurIPS 2022.

自然語言處理在醫學上的應用？

應用1: 醫療文本分類 (使用 ICD-10 編碼)

Input: radiology report

A 75 Y/O male; Clinical Information:LC-C, Increasing AFP, 2.9> 6.4 > 17.0, RFA on 105-8-2, Please evaluate any recurrenceCT scan of liver for F/U a patient of HCC post RFA was done by using triphasic study without and with bolus IV non-ionic contrast enhancement showed: ** Comparison CT study: 2016-06-24; 2017-3-23 ** 1. Mildly undulated surface of liver with hypertrophy of lateral segment. No ascites. Mild splenogemaly with EVs. Cirrhosis of liver with portal hypertension is considered. 2. Hypodense lesion about 4.4x3.5-cm with lobulated margin in S7, C/W post RFA change. High suspicion of newly developed enhancing viable tumor at posterior aspect with arterial wash-in (se 7, im 15-16). 3. No new liver tumor is found. A small cyst in S4.4. The portal and hepatic venous system are patent. No biliary tree dilatation. 5. No remarkable finding in the gallbladder, pancreas and both kidneys, adrenal glands. 6. No evidence of enlarged lymph node is found in the perigastric area, hepatoduondenal ligament, para-aortic area, pelvis and inguina. 7. Grossly, no abnormality is found in the GI tract. Clear mesentery and omentum. No ascites. 8. Normal contour, capacity and wall thickness of urinary bladder. Normal size and contour of seminal vesicle and prostatic gland. 9. Clear retroperitoneum. Normal contour, diameter and enhancement of abdominal aorta. 10. No active lesion is found in the visualized bilateral lower lungs. 11. Grossly, no destructive bony lesion or abnormal bone density. IMP: Cirrhosis of liver with portal hypertension is noted. A HCC in S7 post RFA with high suspicion of small viability at posterior aspect. No new HCC is found.

ICD-10 類別

- K7689 Other specified diseases of liver ✓
- C220 Liver cell carcinoma ✓
- Z4889 Encounter for other specified surgical aftercare ✓
- K7460 Cirrhosis ✓
- K760 Fatty (change of) liver, not elsewhere classified
- C770-C779 Secondary and unspecified malignant neoplasm
- D1803 Hemangioma of intra-abdominal structures

NLP

ICD-10 國際疾病分類編碼

ICD-10: 《疾病和有關健康問題的國際統計分類》第10次修訂本

The International Statistical Classification of Diseases and Related Health Problems 10th Revision

章節	編碼範圍	標題
I	A00-B99	某些傳染病和寄生蟲病
II	C00-D48	腫瘤
III	D50-D89	血液及造血器官疾病和某些涉及免疫系統的疾患
...	...	...

G43001 無預兆偏頭痛，非頑固性，伴有偏頭痛重積狀態

J00 急性鼻咽炎（感冒）

試試➡️國際疾病分類查詢系統
https://info.nhi.gov.tw/inae5000/INAE5010S01

Z開頭是跟健康狀態有關的項目

衛生福利部中央健康保險署
National Health Insurance Administration, Ministry of Health and Welfare

首頁＞健保資料站＞國際疾病分類第 10 版＞國際疾病分類查詢

國際疾病分類查詢清單

查詢條件：
全部 | z552

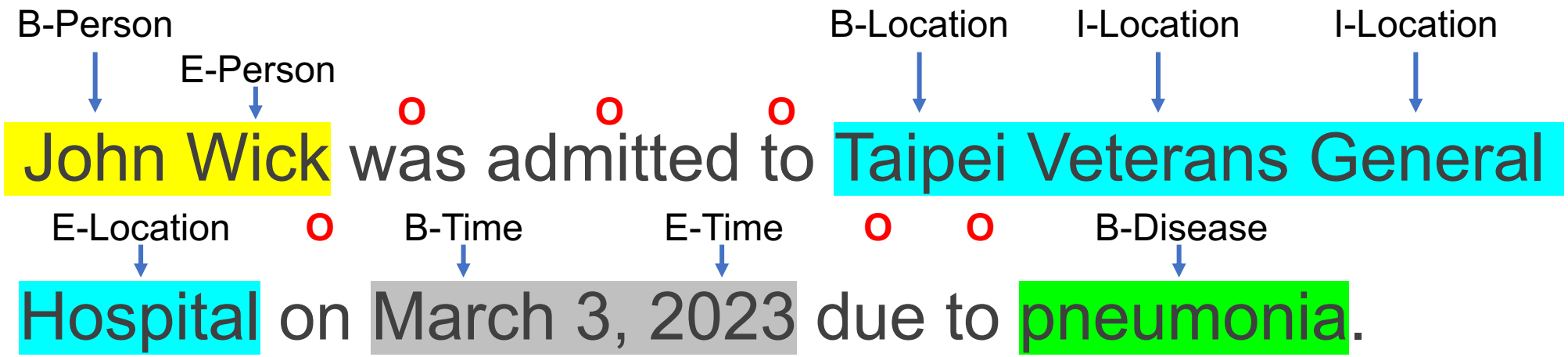
Excel下載 Ods下載 結果清單 查詢條件

每頁顯示： 10筆 20筆 50筆

診斷處置別	疾病代碼	中文名稱	英文名稱
1：診斷碼	Z552	學校考試不及格	Failed school examinations

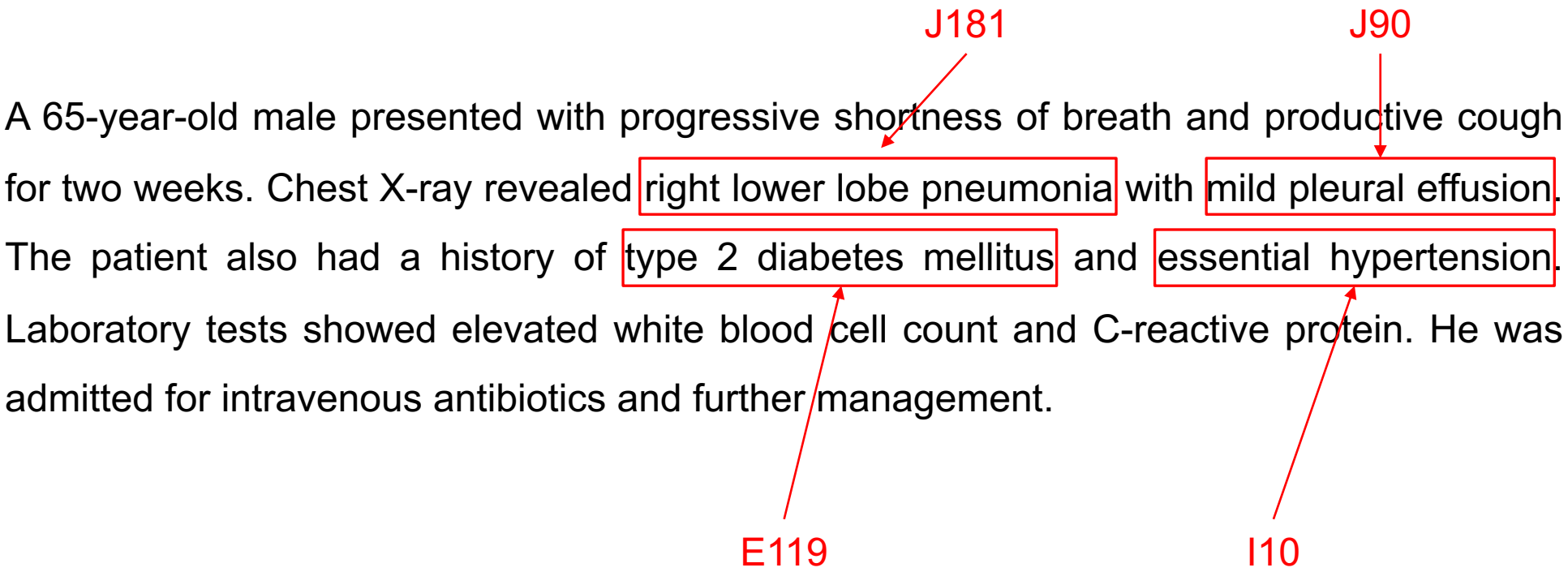
1

應用2: 命名實體辨識 (Named Entity Recognition, NER)



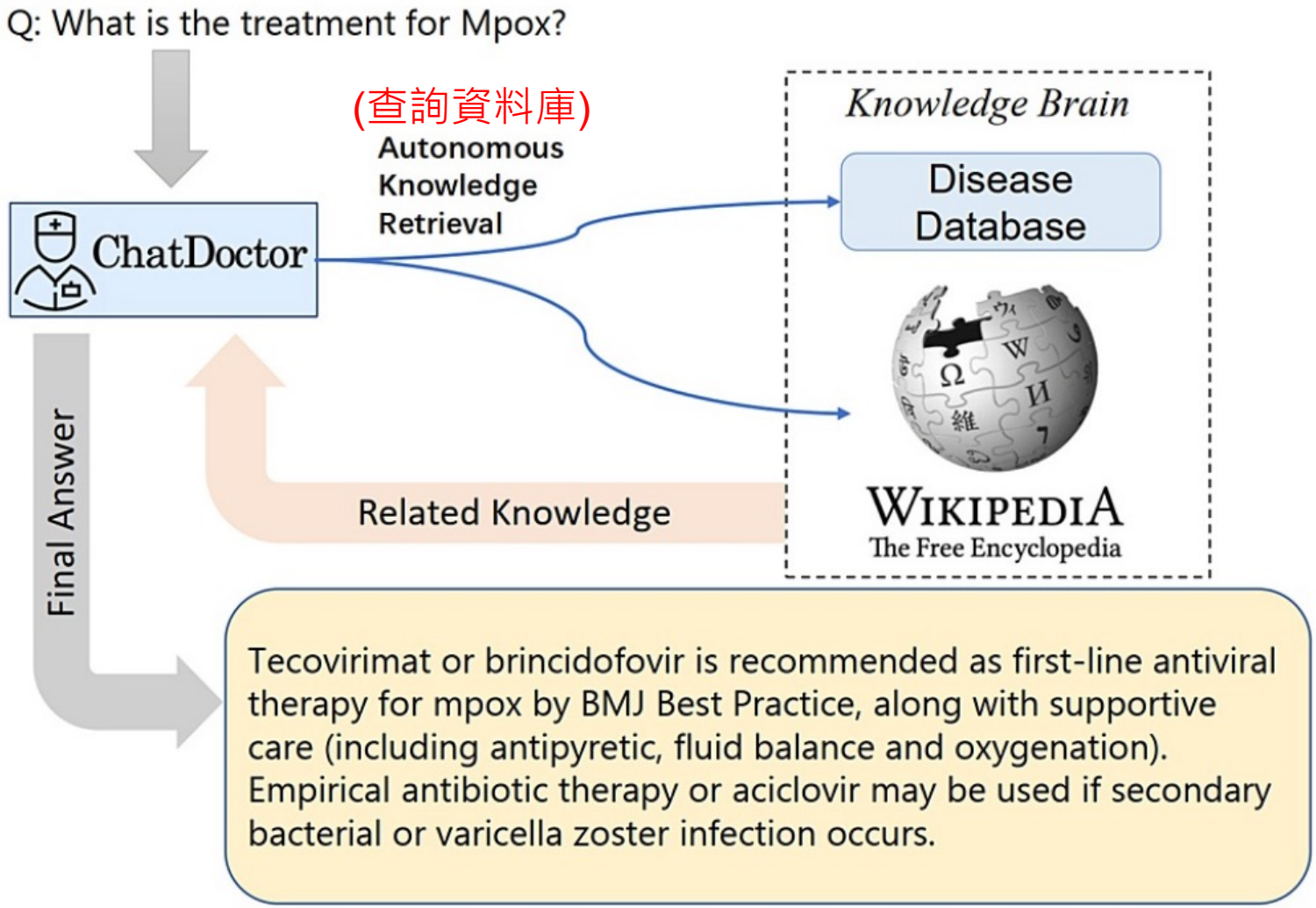
	Meaning
B	Beginning
I	Inside
O	Outside
E	End

應用2: 於文中標註出疾病分類位置



➡️ 命名實體辨識 (Named Entity Recognition, NER)

應用3: ChatDoctor (醫學問答系統)



Li, Yunxiang, et al. "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge." Cureus 15.6 (2023).

很多厲害的大型語言模型都需要上網...

Meta

Llama (7B - 65B) by Meta (2023)


市場概況 > 輝達
186.26 USD
+ 追蹤
+173.73 (1,386.51%) 過去 5 年
已收盤: 10月27日 上午4:07 [EDT] • 免責聲明
開盤前 189.84 +3.58 (1.92%)
1 天
5 天
1 個月
6 個月
本年迄今
1 年
5 年
最久
200
150
100
50
0
12.53 USD 2020年10月30日
2022年
2023年
2024年
2025年
開盤 183.84
最高 187.47
最低 183.50
市值 4.53兆
本益比 53.01
殖利率 0.021%
52 週高點 195.62
52 週低點 86.63
季度股利金額 0.010

電腦是如何學習自然語言呢？

影像 vs. 文字

⟶ 28 x 28 的矩陣

I want to study AI ⟶ ?

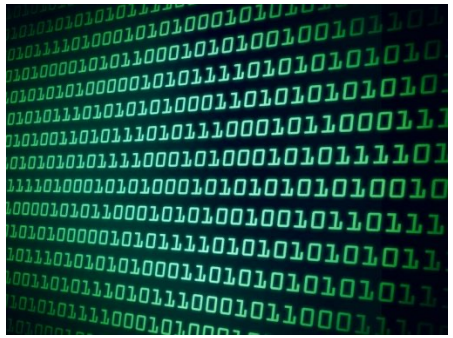


Figure source: https://kopu.chat/wp-content/uploads/2017/04/binary-507x380.jpeg

Outline

- 深度學習與自然語言處理的概念
- Embeddings
  - Sentence embeddings (Document Representations)
  - Word embeddings

Google Search (Retrieval)

- Retrieval: get relevant information from a pool (like a search engine)

Google Search vs. 向量化

NLP 的 Embeddings: 語意資訊被嵌入到（高維度）向量空間

Bag-of-words Model

Meaning: Each document (sentence) carries a bag of words.

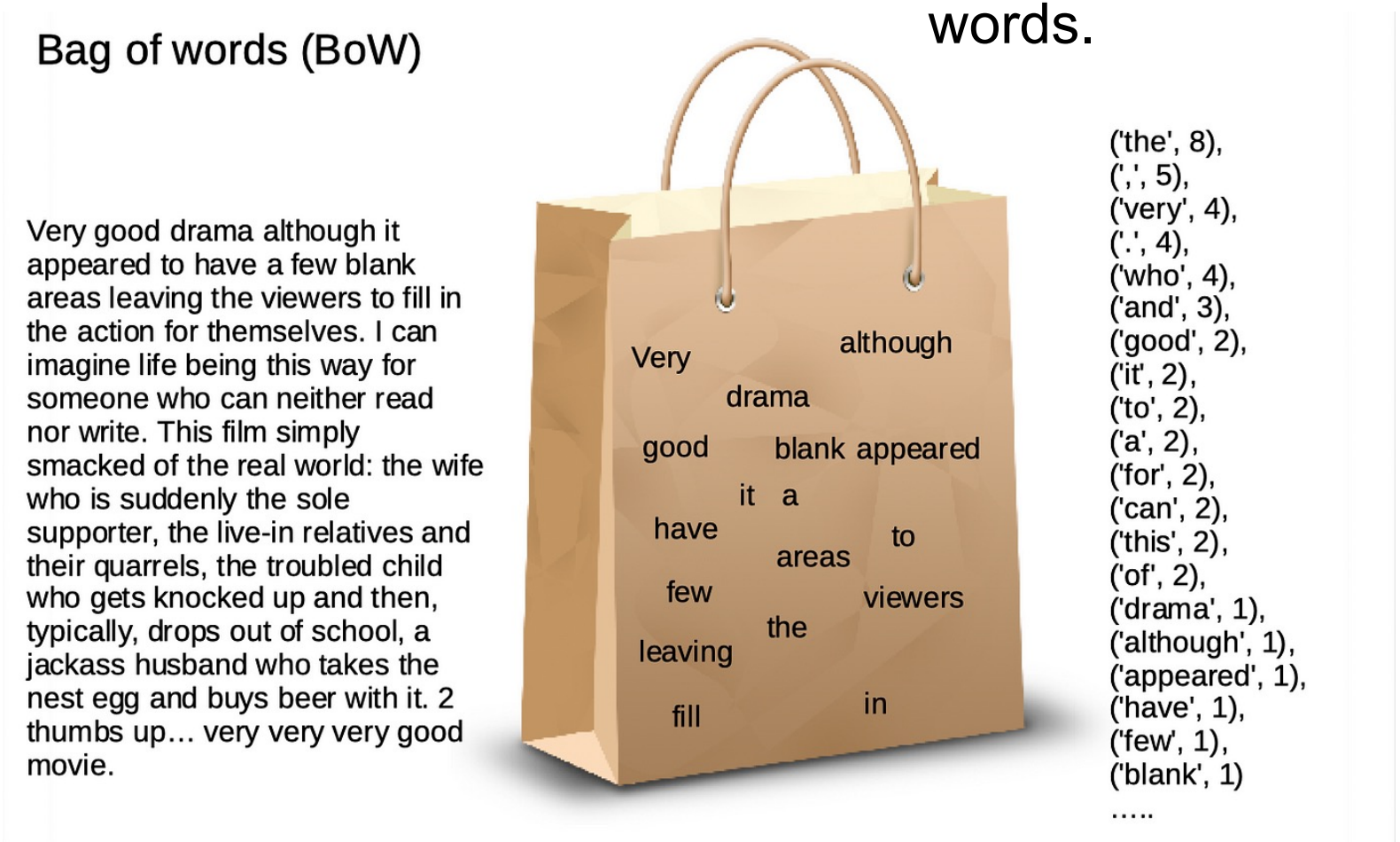
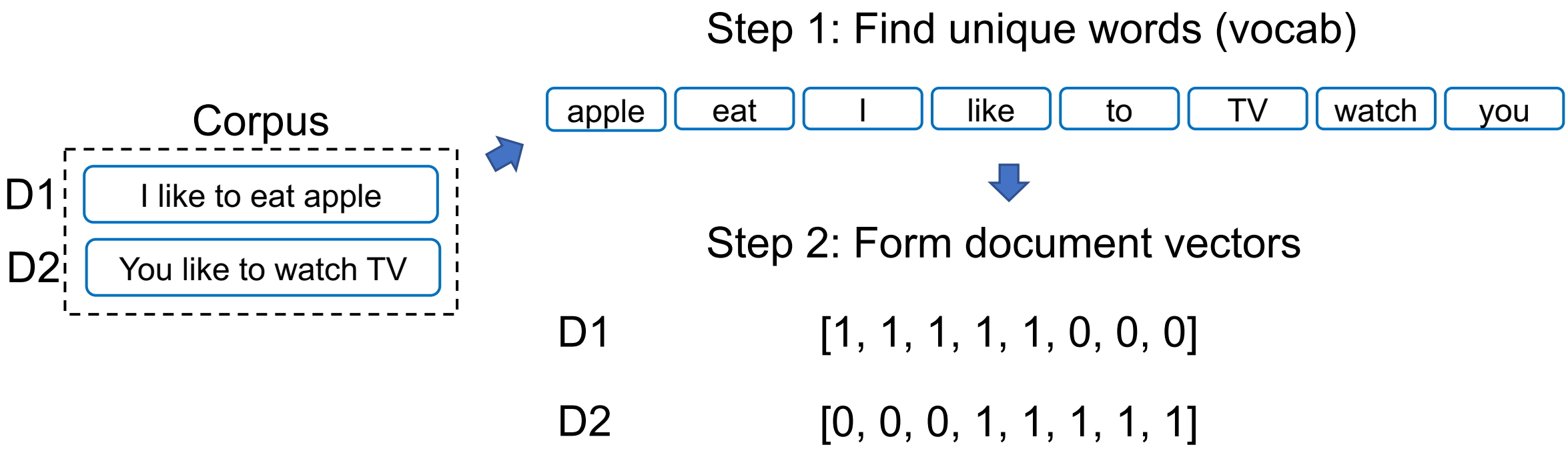


Figure source: https://sfhsu29.medium.com/nlp-入門-1-text-classification-sentiment-analysis-極簡易情感分類器-bag-of-words-naive-bayes-e40d61de9a7f

Bag-of-words Model (Example 1)

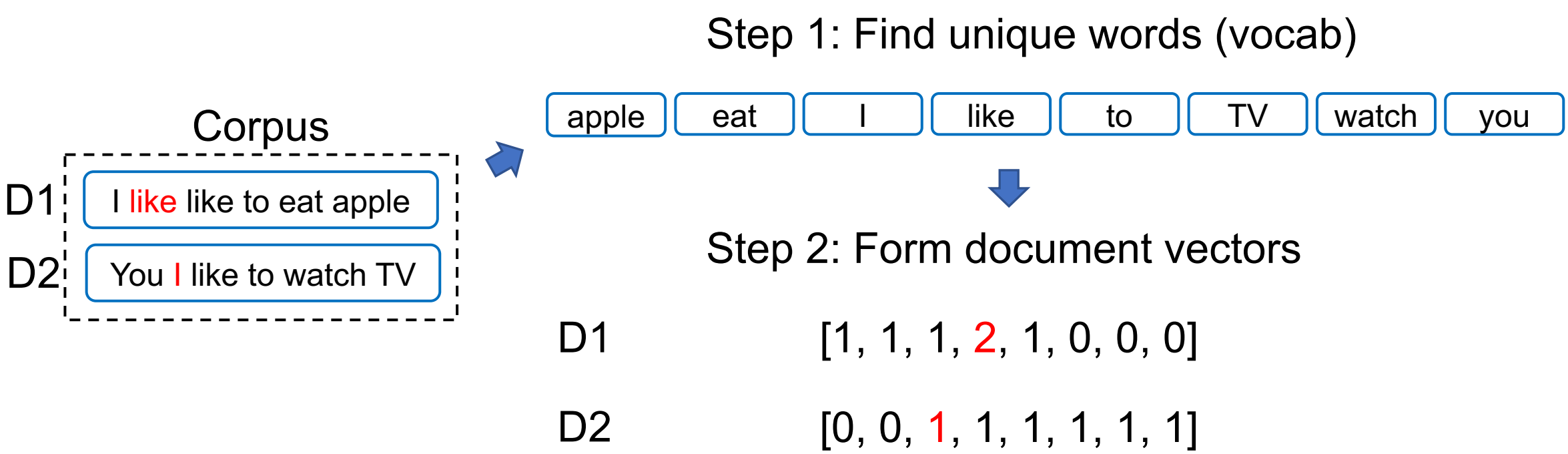
https://github.com/mcps5601/CGUNLP_2025_Spring/code/bow.py



Notes:

- The vector size is equal to the vocab size.
- Each position of a vector is corresponding to the position in the vocab.

Bag-of-words Model (Example 2)



Notes:

- The vector size is equal to the vocab size.
- Each position of a vector is corresponding to the position in the vocab.

TF-IDF

- TF (Term Frequency; 詞頻)
- IDF (Inverse Document Frequency; 逆文本頻率)

TF x IDF = 關鍵字分數

某詞出現在該句子的頻率

某詞出現在全部文本的頻率的倒數

TF-IDF 為什麼這樣設計？

想像你撿到了班上三位同學的日記：

日記 A (小明的)：「我 今天 去 吃 **漢堡**，**漢堡** 超 好吃。」

日記 B (小華的)：「我 今天 去 學校 上課，好 累。」

日記 C (小美的)：「我 今天 跟 媽媽 去 買 菜。」

日記 A 裡的內容觀察：

單字	TF (在這篇很多?)	IDF (在別篇很少?)	結果 (TF-IDF)	意義
我	是 (1次)	否 (每句都有)	**低分**	雜訊/不重要
今天	是 (1次)	否 (每句都有)	**低分**	雜訊/不重要
漢堡	**是 (2次)**	**是 (別句都沒提)**	**高分**	**關鍵字！**

Sentence Embeddings: TF-IDF

日記 A (小明的)：「我 今天 去 吃 漢堡，漢堡 超 好吃。」
日記 B (小華的)：「我 今天 去 學校 上課，好累。」
日記 C (小美的)：「我 今天 跟 媽媽 去 買 菜。」

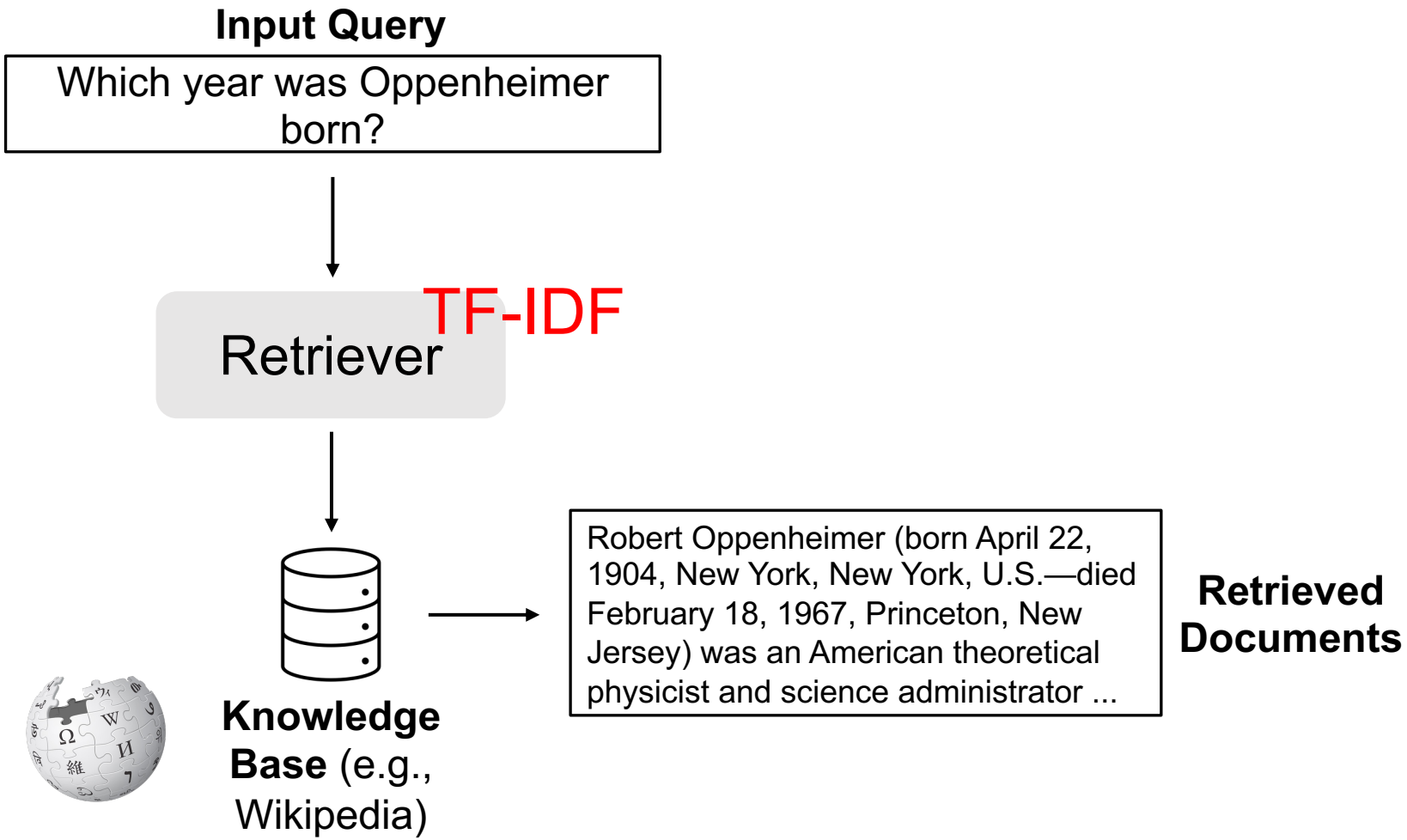
字典大小

	我	今天	去	吃	漢堡	超	好吃	學校	好累	跟	...
TF	1/8 =0.125	1/8 =0.125	1/8 =0.125	1/8 =0.125	2/8 =0.25	1/8 =0.125	1/8 =0.125	0	0	0	
IDF	log(3/3) =0	log(3/3) =0	log(3/3) =0	log(3/1) =0.48	log(3/1) =0.48	log(3/1) =0.48	log(3/1) =0.48	log(3/1) =0.48	log(3/1) =0.48	log(3/1) =0.48	
分數 (TFxIDF)	0	0	0	0.06	0.12	0.06	0.06	0	0	0	

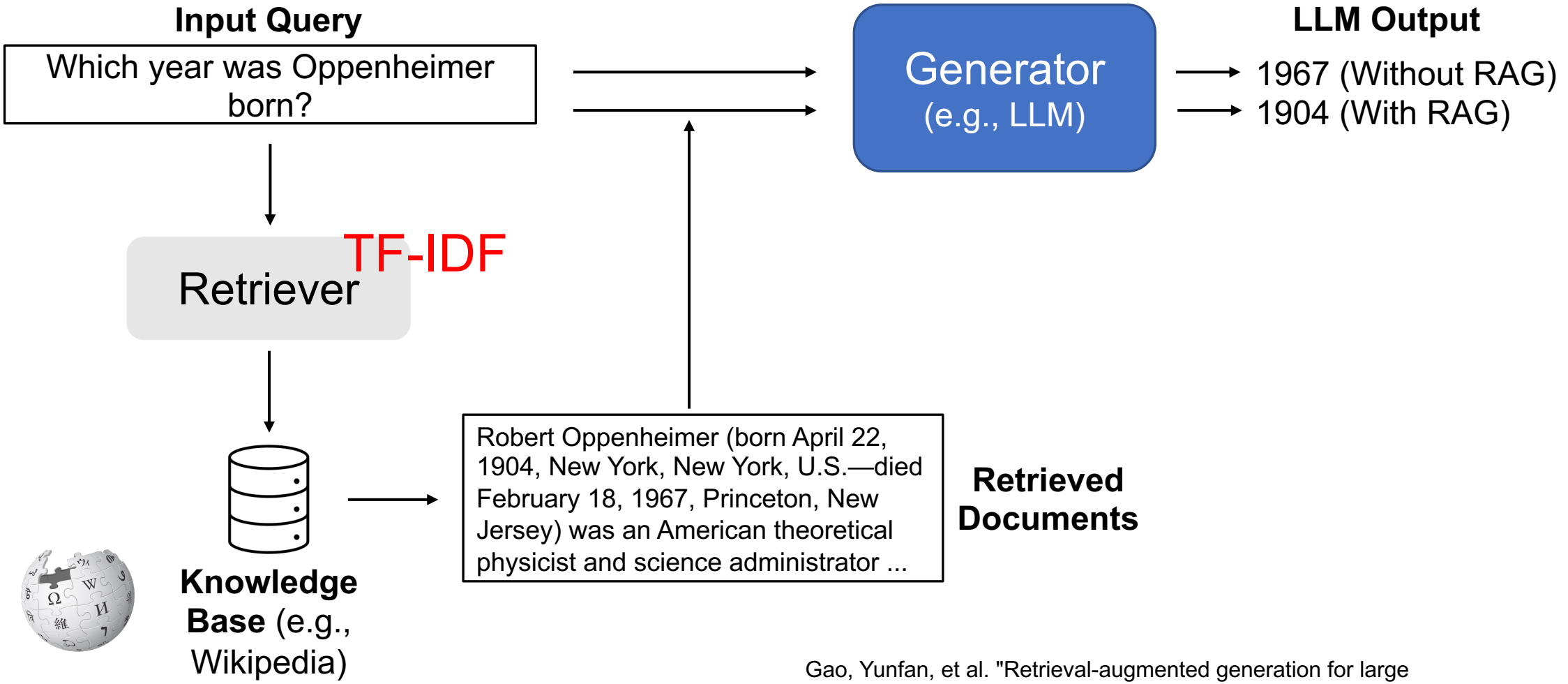
TF-IDF 的問題

1. 在 TF-IDF 的世界裡，字跟字是完全獨立的。
   - 句子 A：「我很**快樂**」
   - 句子 B：「我很**開心**」
2. 本質上仍是詞袋模型 (Bag-of-words) ，不分「順序」
   - 句子 A：「小明喜歡小華」
   - 句子 B：「小華喜歡小明」

Retrieval


Input Query
Which year was Oppenheimer born?
Retriever
TF-IDF
Knowledge Base (e.g., Wikipedia)
Robert Oppenheimer (born April 22, 1904, New York, New York, U.S.—died February 18, 1967, Princeton, New Jersey) was an American theoretical physicist and science administrator ...
Retrieved Documents

Retrieval-Augmented Generation (RAG)



Gao, Yunfan, et al. "Retrieval-augmented generation for large language models: A survey." *arXiv preprint arXiv:2312.10997* (2023).

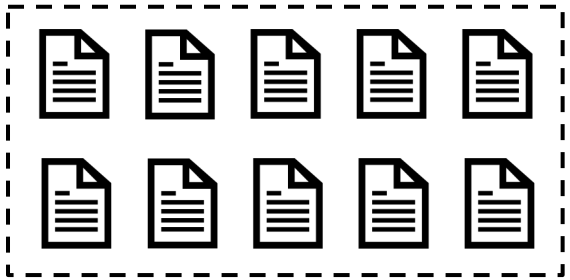
Outline

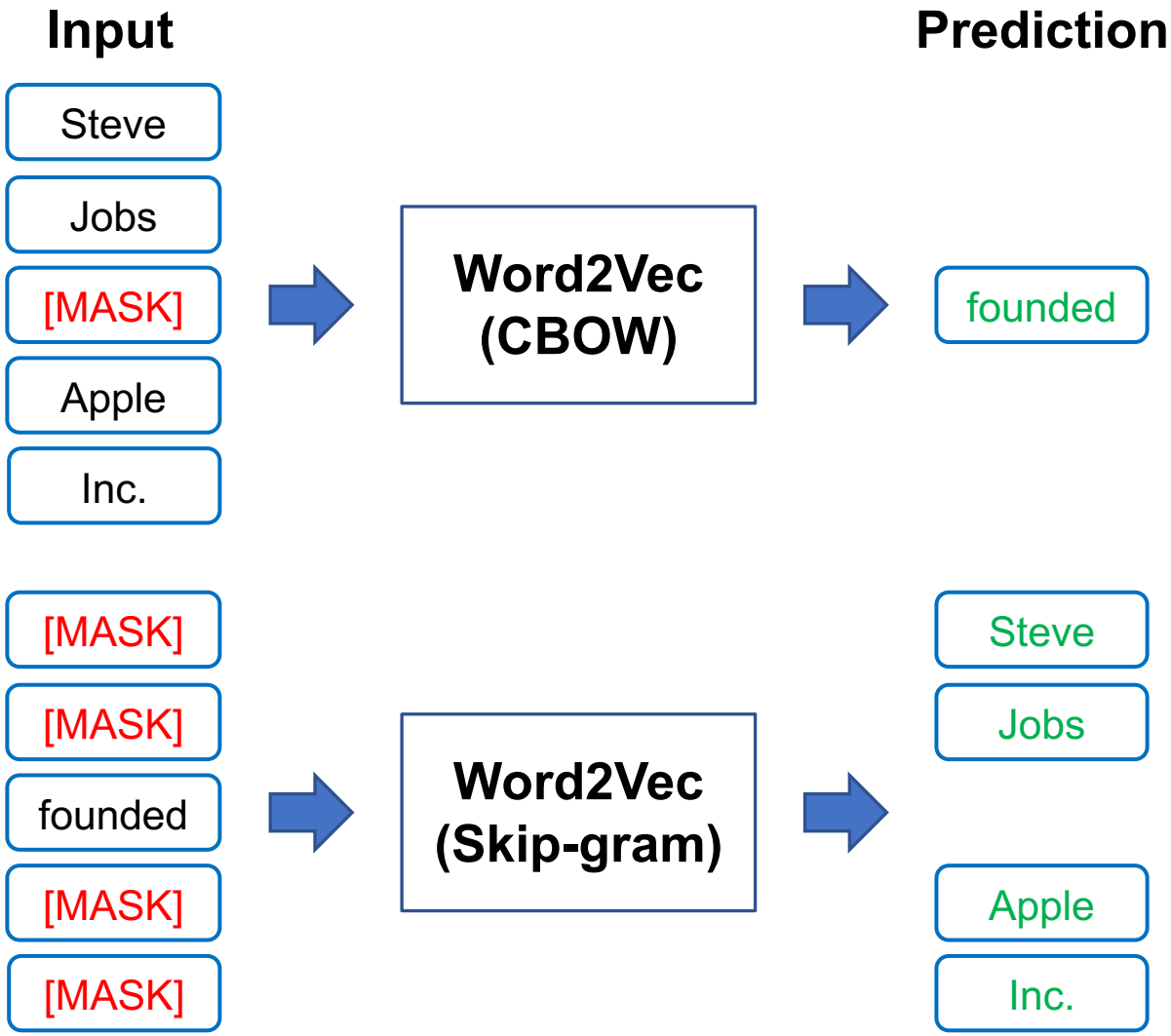
- 深度學習與自然語言處理的概念
- Embeddings
  - Sentence embeddings (Document Representations)
  - Word embeddings

詞嵌入：把詞嵌入到向量 (Word2Vec)

[1] Mikolov et al. "Distributed representations of words and phrases and their compositionality." NeurIPS 2013.

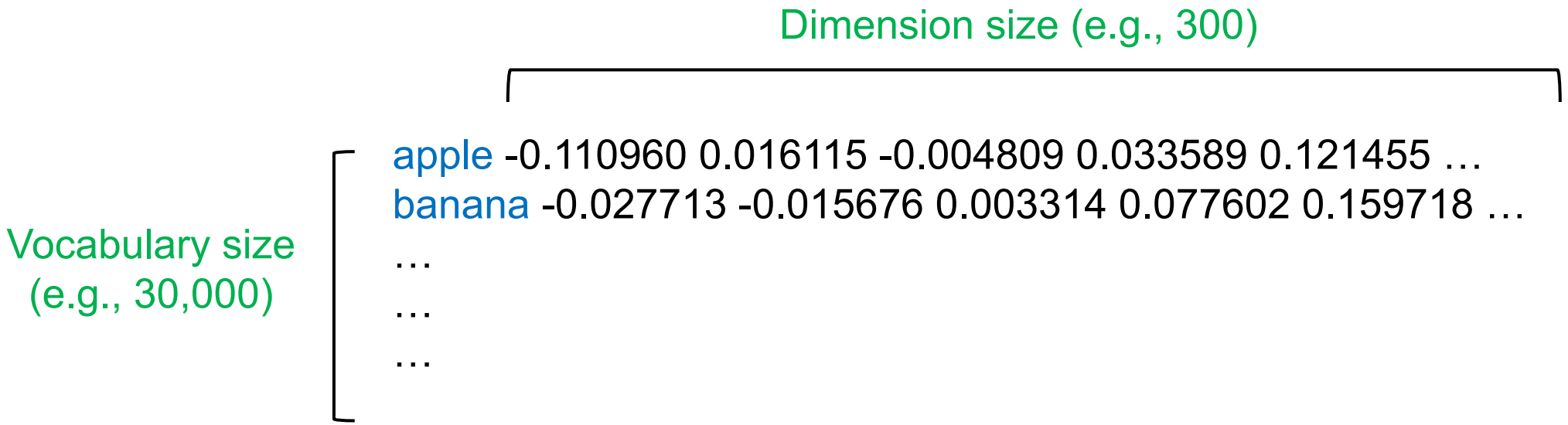
非常大量的語料庫
(例如：Wikipedia)





每個字，都是一段向量 – 詞向量

- Word2Vec 的範例長相:

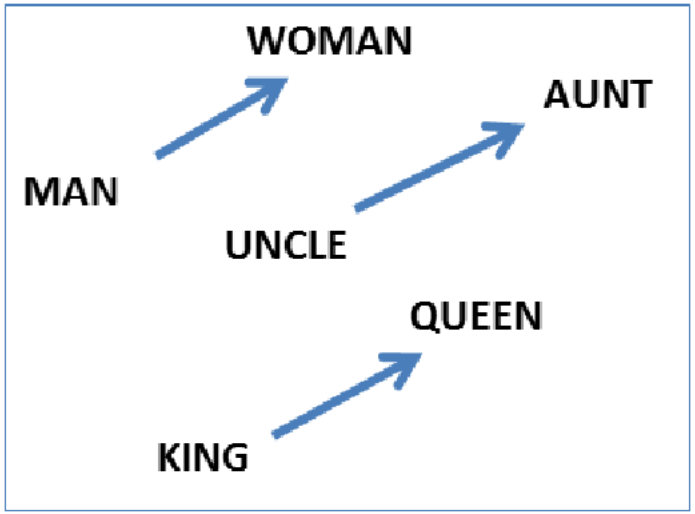


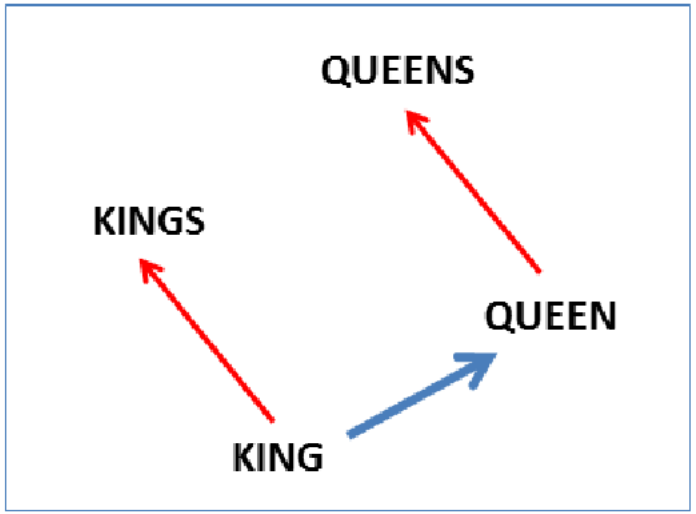
詞向量視覺化

- **Word embedding model**: glove-wiki-gigaword-100
- **Dimension reduction**: t-SNE
- **Dataset**: Mikolov et al., 2013

Try it:
https://www.cs.cmu.edu/~dst/WordEmbeddingDemo/

Word pairs illustrating the gender relation



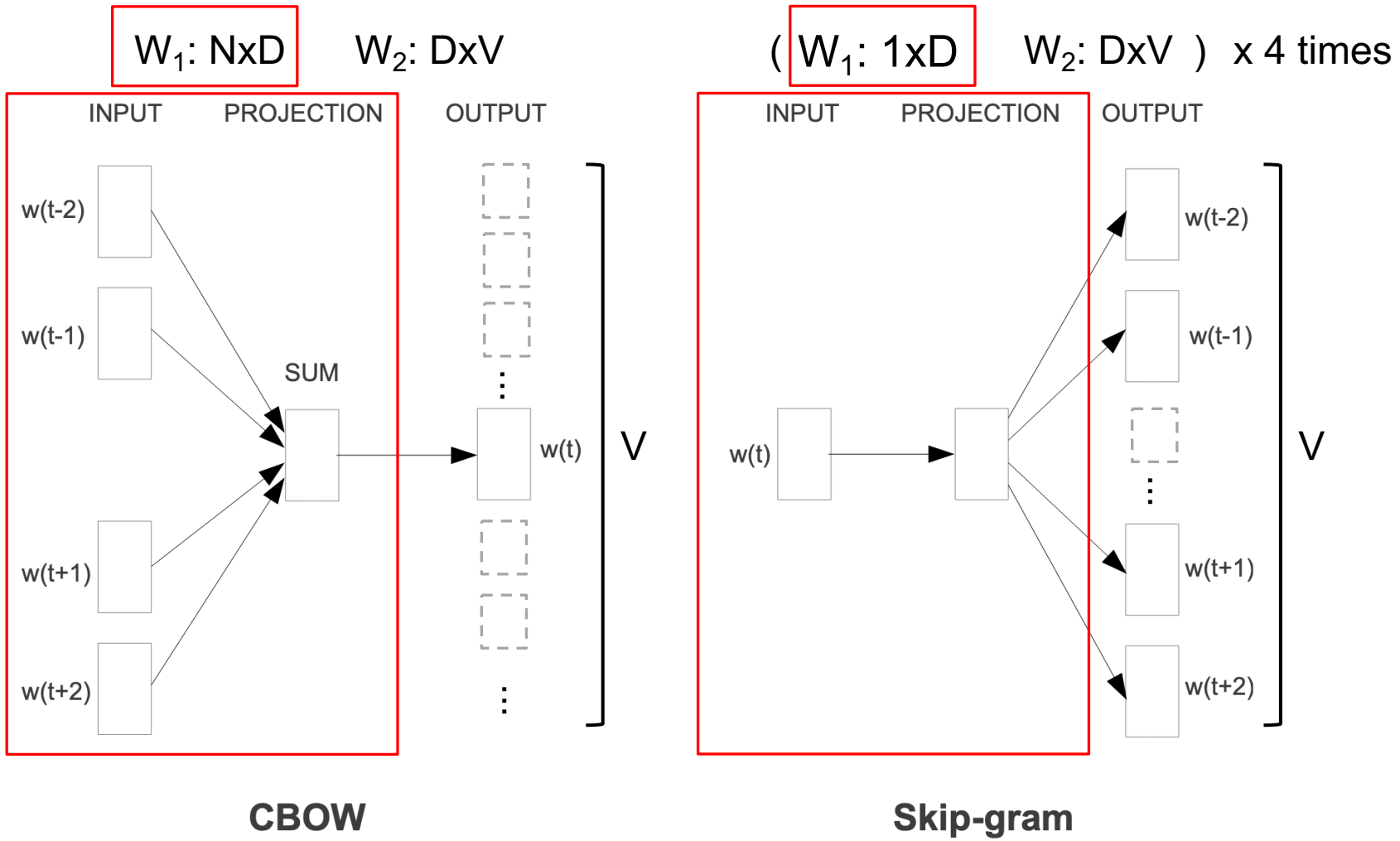


Mikolov et al. “Linguistic regularities in continuous space word representations.” NAACL 2013.

Word2Vec 是淺層神經網路

Symbol	Meaning	Example
N	輸入數量	4
D	Hidden size	100
V	字典大小	100,000

Word embeddings after training



Mikolov et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

Thank you!

長庚大學人工智慧學系 林英嘉

✉ yjlin@cgu.edu.tw