NCKU CSIE 112 Senior Project

以文會友 Find Your Best Study Partner on Facebook

分析個人在 Facebook 發佈的貼文，來媒合有共同興趣的朋友，並推薦其適合的揪課課程

Team 4-2

指導教授 李強

專題生 連思涵 F74099017

Outline

00 Introduction

研究動機的 5W1H

5W1H

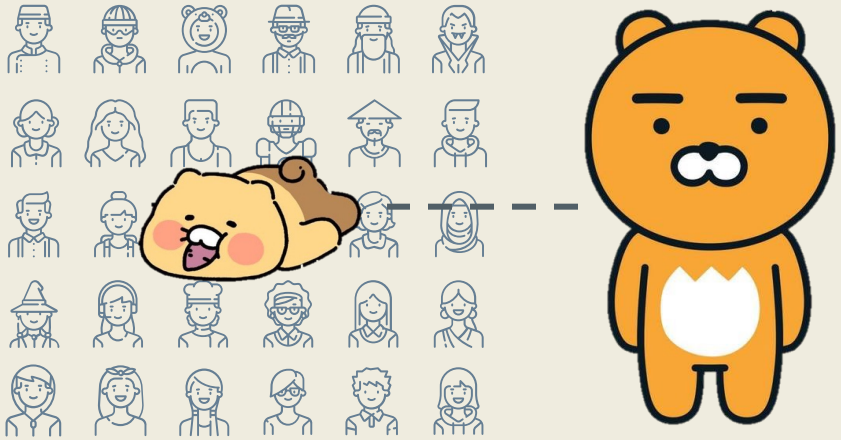
Who	• 找不到學習夥伴的同學 • 不知道要選什麼課的同學
When	• 大四需大量累積實力時 • 學期初選課時
Why	• 有很多人到大學畢業都還找不到自己偏好的領域 • 朋友圈內不一定有共同興趣的朋友
Where	• Facebook • NCKU 揪課平台
What	• 協助找到對同一領域有興趣的讀書夥伴 • 基於平時在社群發布的訊息，推薦揪課的課程
How	• 藉由平時在 FB 分享的文章找出喜好 • 與 DB 內的使用者媒合，並共同推薦課程

➡ **以文會友**

專題簡介

Database

user



以文會友

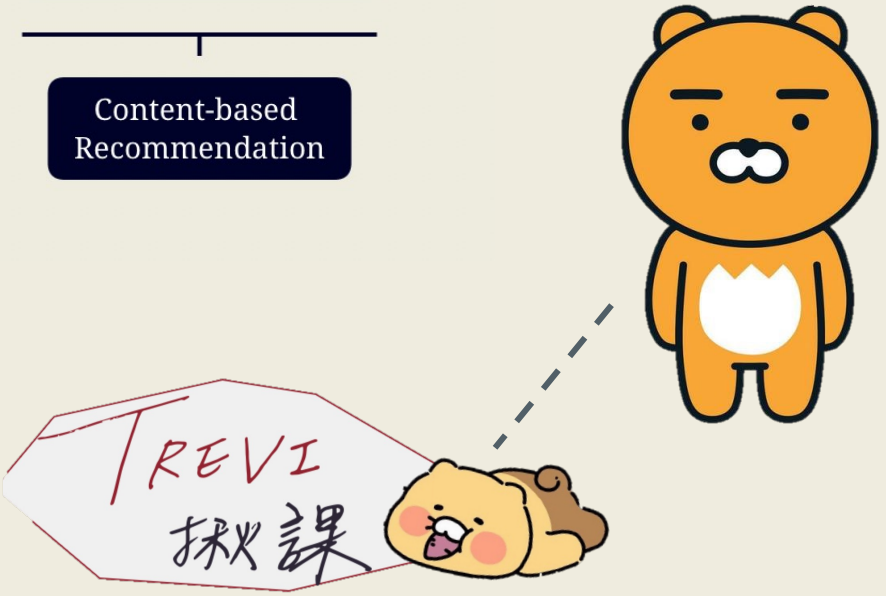
Text Compare

Post Analysis

Recommendation with Collaborative Filtering

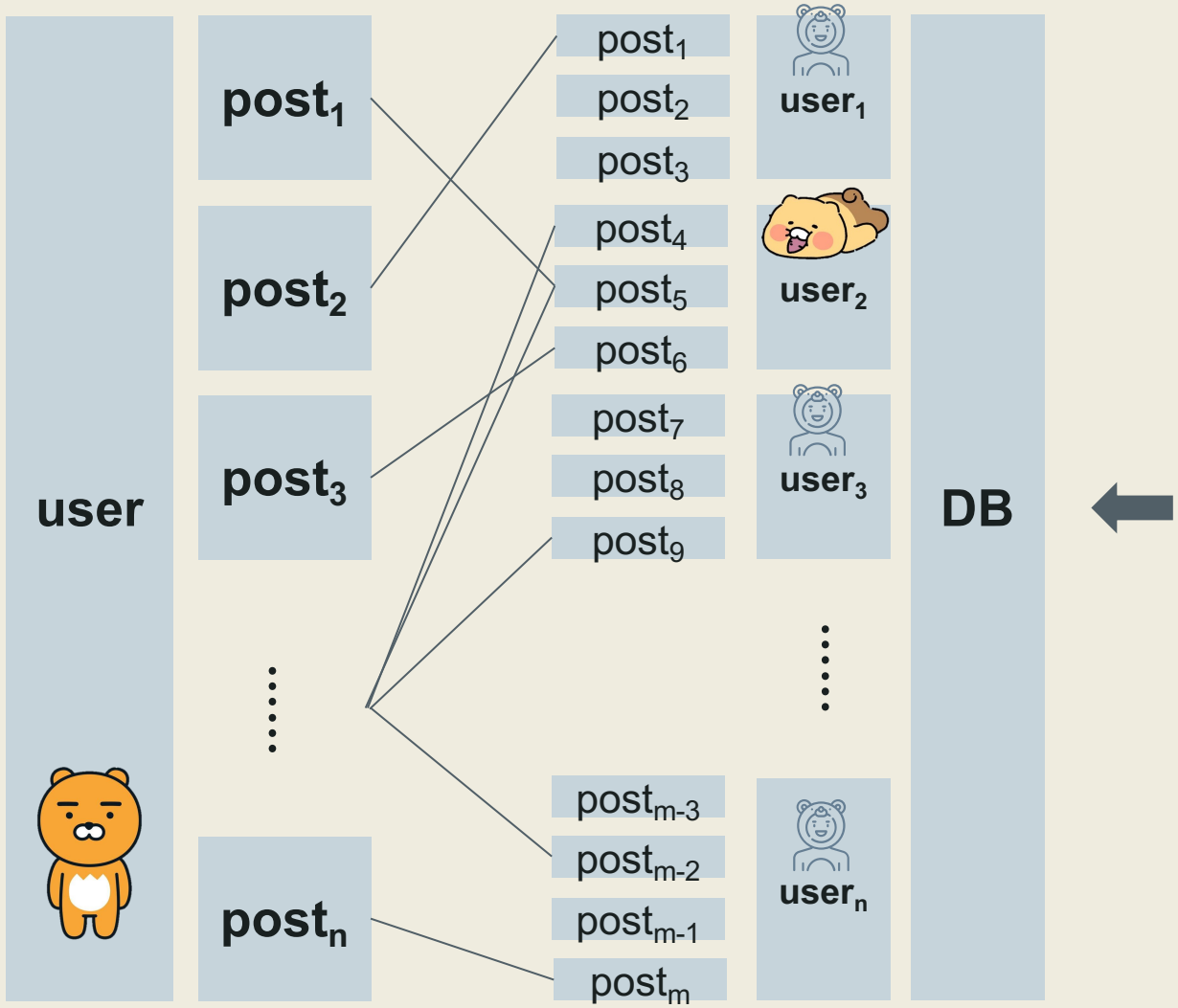
Content-based Recommendation

user



01 Text Compare and Recommendation

將使用者的 n 篇貼文與 DB 內的資料做 TF-IDF 文本比對，找出最佳讀書夥伴



	PG 財經筆記=0	PG 財經筆記=1	PG 財經筆記=2	PG 財經筆記=3	PG 財經筆記=4	PG 財經筆記=5	PG 財經筆記=6
矽谷阿雅 Anya Cheng=0	0.0	0.21061544120311700	0.0	0.015361979603767400	0.07673442363739010	0.001145594520494340	0.07391513884067540
矽谷阿雅 Anya Cheng=1	0.0	0.08701366931200030	0.04551013559103010	0.0	0.0	0.064226433634758	0.0
矽谷阿雅 Anya Cheng=2	0.014286659657955 20	0.07027796655893330	0.036756981164217000	0.0	0.04267445579171180	0.051873501390218700	0.009457391686737540
矽谷阿雅 Anya Cheng=3	0.05020962655544280	0.0	0.0	0.020146805793047 00	0.0	0.014529868029058 00	0.044604908674955 40
矽谷阿雅 Anya Cheng=4	0.0	0.0	0.0	0.043684631586074 80	0.0	0.0	0.0
矽谷阿雅 Anya Cheng=5	0.01930840127170090	0.0	0.0	0.0659237653017044	0.0	0.0	0.0
矽谷阿雅 Anya Cheng=6	0.008547567762434480	0.04473619535565380	0.001196285360492770	0.021589878946542 70	0.05155201256275180	0.0811182036995888 0	0.1028934121131900
矽谷阿雅 Anya Cheng=7	0.034765198826789900	0.0	0.0	0.006786812096834180	0.0	0.030891088768839800	0.095190815627574 90
矽谷阿雅 Anya Cheng=8	0.017523692920804000	0.13220977783203100	0.009769655764102940	0.006152628455311 06	0.03211237117648130	0.011916029267013100	0.062764607369899 80
矽谷阿雅 Anya Cheng=9	0.0	0.10345801711082500	0.03482940047979360	0.021934485062956 80	0.0	0.024576593190431600	0.008961433544754 980
矽谷阿雅 Anya Cheng=10	0.0	0.12083148956298800	0.0	0.0	0.1467435210943220	0.0	0.03252088278532030
矽谷阿雅 Anya Cheng=11	0.0	0.0	0.0	0.0	0.0	0.0	0.0
矽谷阿雅 Anya Cheng=12	0.040348924696445500	0.0	0.0	0.018096202984452200	0.060261406004428900	0.0	0.013354961760342100
矽谷阿雅 Anya Cheng=13	0.018389329314231900	0.0014181813457980800	0.000741741212550551 0	0.008587183403790900	0.03196415677666660	0.041252903640270 20	0.035349804908037200
矽谷阿雅 Anya Cheng=14	0.047663375735282 9	0.05116431042551990	0.011860466562211500	0.027229174971580500	0.05348987504839900	0.036462455987930 3	0.034092675894498800

	半路出家軟體工程師在矽谷=1	半路出家軟體工程師在矽谷=2	半路出家軟體工程師在矽谷=3	半路出家軟體工程師在矽谷=4	半路出家軟體工程師在矽谷=5	半路出家軟體工程師在矽谷=6
矽谷阿雅 Anya Cheng=0	0.03230162337422370	0.09413079917430880	0.16493353247642500	0.09936847537755970	0.07061222195625310	0.0742359310388565
矽谷阿雅 Anya Cheng=1	0.020086079835891700	0.05602323263883590	0.06831202656030660	0.03114149160683160	0.08724968880414960	0.06453309208154680
矽谷阿雅 Anya Cheng=2	0.020322639495134400	0.08869554102420810	0.06959442049264910	0.10377360880374900	0.07355383038520810	0.07031203806400300
矽谷阿雅 Anya Cheng=3	0.050991594791412400	0.00588524155318737	0.10716690123081200	0.006763797253370290	0.030549844726920100	0.02786768227815630
矽谷阿雅 Anya Cheng=4	0.007944576442241670	0.01909384876489640 0	0.11500829458236700	0.04882487654685970	0.0	0.013529804535210100
矽谷阿雅 Anya Cheng=5	0.004155758302658800	0.02840055525302890	0.07302099466323850	0.05544073134660720	0.015372841618955100	0.010579117573797700
矽谷阿雅 Anya Cheng=6	0.04847477003932000	0.09301892668008800	0.17500029504299200	0.15268391370773300	0.038790255784988 4	0.06955568492412570
矽谷阿雅 Anya Cheng=7	0.060763619840145 10	0.09727168828248980	0.15979471802711500	0.04067377373576160	0.003566874656826 2600	0.07203321903944020
矽谷阿雅 Anya Cheng=8	0.020340092480182600	0.10245874524116500	0.27022057771682 70	0.05756952613592150	0.166270464658737 00	0.060548584908247000
矽谷阿雅 Anya Cheng=9	0.020630884915590300	0.04924929141998290	0.13180261850357100	0.09415735304355620	0.14674808084964800	0.05765965580940 2500
矽谷阿雅 Anya Cheng=10	0.0	0.02388665638864040 0	0.12844525277614600	0.08582510054111480	0.0	0.016902871429920200
矽谷阿雅 Anya Cheng=11	0.0	0.0	0.02067092619836300 0	0.0	0.13742424547672300	0.0
矽谷阿雅 Anya Cheng=12	0.00883545447140932	0.04761375859379700 0	0.08018850535154340	0.09152454882860180	0.1786891222000120	0.055721905082464 20
矽谷阿雅 Anya Cheng=13	0.026880767196416900	0.08219444006681440	0.2609350383281710	0.10437951982021300	0.06579418480396270	0.052596960216760600
矽谷阿雅 Anya Cheng=14	0.03547798469662670	0.10454515367746400	0.14531861245632200	0.14667677879333500	0.06827839463949200	0.06818445771932600

	Min 的投資說書小棧=0	Min 的投資說書小棧=1	Min 的投資說書小棧=2	Min 的投資說書小棧=3	Min 的投資說書小棧=4	Min 的投資說書小棧=5	Min 的投資說書小棧=6
矽谷阿雅 Anya Cheng=0	0.0	0.031443506479263300	0.038628607988357500	0.021221471950411800	0.06796460598707200	0.012101402506232300	0.10791318863630300
矽谷阿雅 Anya Cheng=1	0.0	0.00521071208640933	0.02520398423075680	0.026119481772184400	0.00109957530945510 0	0.07181748747825620	0.020624930039048200
矽谷阿雅 Anya Cheng=2	0.0	0.03000637143850330 0	0.02385772578418260	0.025626037269830700	0.00472728395834565 0	0.092016294598579 4	0.016241729259491000
矽谷阿雅 Anya Cheng=3	0.0	0.0	0.00121073133777827 00	0.00290724355727434 00	0.02893394604325290 0	0.11216455698013300	0.02810150757431980
矽谷阿雅 Anya Cheng=4	0.0	0.18392598628997800	0.01933032274246200 0	0.00973050575703383 0	0.01030444353818890 0	0.0	0.04341162368655210
矽谷阿雅 Anya Cheng=5	0.0	0.1422995626926420	0.01495545357465740	0.012663706205785300	0.02391697652637960 0	0.027891401201486600	0.03358663618564610
矽谷阿雅 Anya Cheng=6	0.02829722873866560	0.04344220831990240	0.04476520419120790 0	0.026664482429623600	0.02546107210218910 0	0.120880804955959 00	0.04573161154985430
矽谷阿雅 Anya Cheng=7	0.0	0.06269685924053190	0.06692685931921010	0.028788655996322600	0.0391208790242672	0.135103404521942 00	0.03511188551783560
矽谷阿雅 Anya Cheng=8	0.027852758765220600	0.02865923941135410 0	0.05083517357707020 0	0.04373218119144 44	0.02817085012793540	0.058897923678159 700	0.06546416878700260
矽谷阿雅 Anya Cheng=9	0.0	0.03609131276607510 0	0.01227050647139550 0	0.03877573087811 47	0.00568592175841332 0	0.026327423751354200	0.03381908684968950 0
矽谷阿雅 Anya Cheng=10	0.0	0.00846007000654936 0	0.03932003304362300	0.036440026015043 30	0.00178526155650616 0	0.113616578280926 00	0.03408475220234
矽谷阿雅 Anya Cheng=11	0.0	0.0	0.02092422172427180 0	0.012978701852262 00	0.05704332888126370	0.0	0.013084239326417400
矽谷阿雅 Anya Cheng=12	0.0	0.08867335319519040	0.00094462488779355	0.035082407295703 90	0.02578238956630 23	0.011516711674630600	0.04393771290779110
矽谷阿雅 Anya Cheng=13	0.008939091116189960	0.00584833137691021	0.03565712645649 91	0.05703437700867650	0.05162968859076 5	0.05595276132225 99	0.06072144955396650
矽谷阿雅 Anya Cheng=14	0.025669759139418600	0.04104606434702870	0.01988599076867 1000	0.05829806625843050	0.03972093388438230	0.08640435338020330	0.05514468625187870

將使用者的 n 篇貼文與 DB 內的資料做 TF-IDF 文本比對，找出最佳讀書夥伴

user

$post_n$

1. 資料清洗

- 去除標點符號
- 去除停用詞
- 去除數字
- 以 jieba 斷字

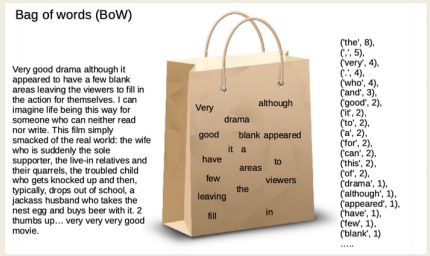
我是一個不知道要做什麼工作的人！

我/是/一個/不/知道/要/做/什麼/工作/的/人

「一個」、「知道」、「做」、「工作」

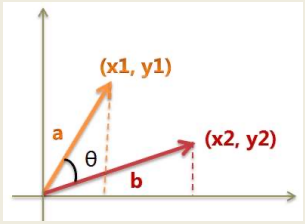
2. 建立詞袋
（ bag of words ）

- 對所有詞編號
  ex：「你好」, 4
- 將貼文都轉成二元組的向量
  ex：「你好」出現兩次
  → （4, 2）



3. 取得兩兩貼文相似度

- 得出詞袋中每個詞的 TF-IDF 值
- 分析 $post_n$ 和 DB 裡每篇貼文的餘弦相似度
  （ cosine similarity ）



TF

某字詞在某貼文中出現的頻率

ex：十萬 / 青年 / 十萬 / 肝
→ TF = 2/3 = 0.66666

X

IDF

文件數除以某特定字詞有被多少貼文所提及的數量取 log

ex：DB 總共有 m 篇貼文
其中有5個文件提及「十萬」
→ IDF = log(10/5) = 0.3

$post_1$

$post_2$

$post_3$

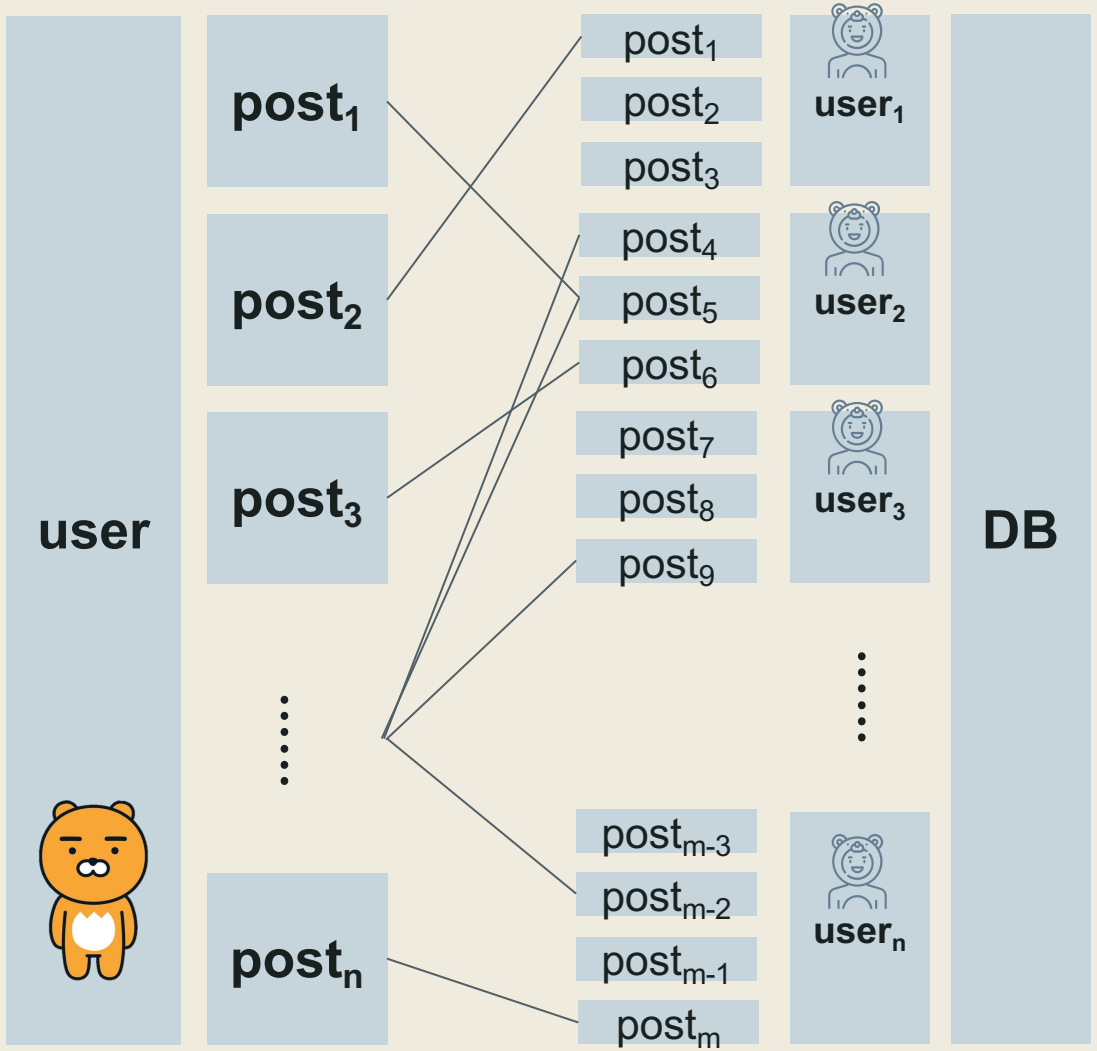
⋮

post m

Someone from DB

資料來源：NLP的基本執行步驟(II) — Bag of Words 詞袋語言模型、 TF-IDF 文字探勘 — 以NBA Ptt鄉民回文為例

將使用者的 n 篇貼文與 DB 內的資料做 TF-IDF 文本比對，找出最佳讀書夥伴



	A's post$_1$	A's post$_2$	B's post$_1$	B's post$_2$	C's post$_1$	C's post$_2$
Ryan's post$_1$	0.060	0.0	0.013	0.048	0.003	0.160
Ryan's post$_2$	0.056	0.0	0.0	0.070	0.0	0.033
Ryan's post$_3$	0.048	0.0	0.005	0.0	0.008	0.078
Ryan's post$_4$	0.048	0.0	0.005	0.0	0.058	0.008
Ryan's post$_5$	0.095	0.029	0.059	0.031	0.024	0.177
Ryan's post$_6$	0.019	0.055	0.019	0.0	0.031	0.045

將使用者與讀書夥伴 TF-IDF 值都很高的字、揪課標題透過 BERT model 轉詞向量

user Best Study Partner

揪課

$word_1$ $word_2$ $word_3$ $word_4$ …… $word_n$

words

$course\ title_1$ $course\ title_2$ $course\ title_3$ …… $course\ title_m$

encode

BERT Model

Google BERT Algorithm

- 基於 transformer 的 encoder
- 以無監督的方式，利用大量無標註文本訓練的 NLP 模型
- 可生成包含上下文語意的 context vector

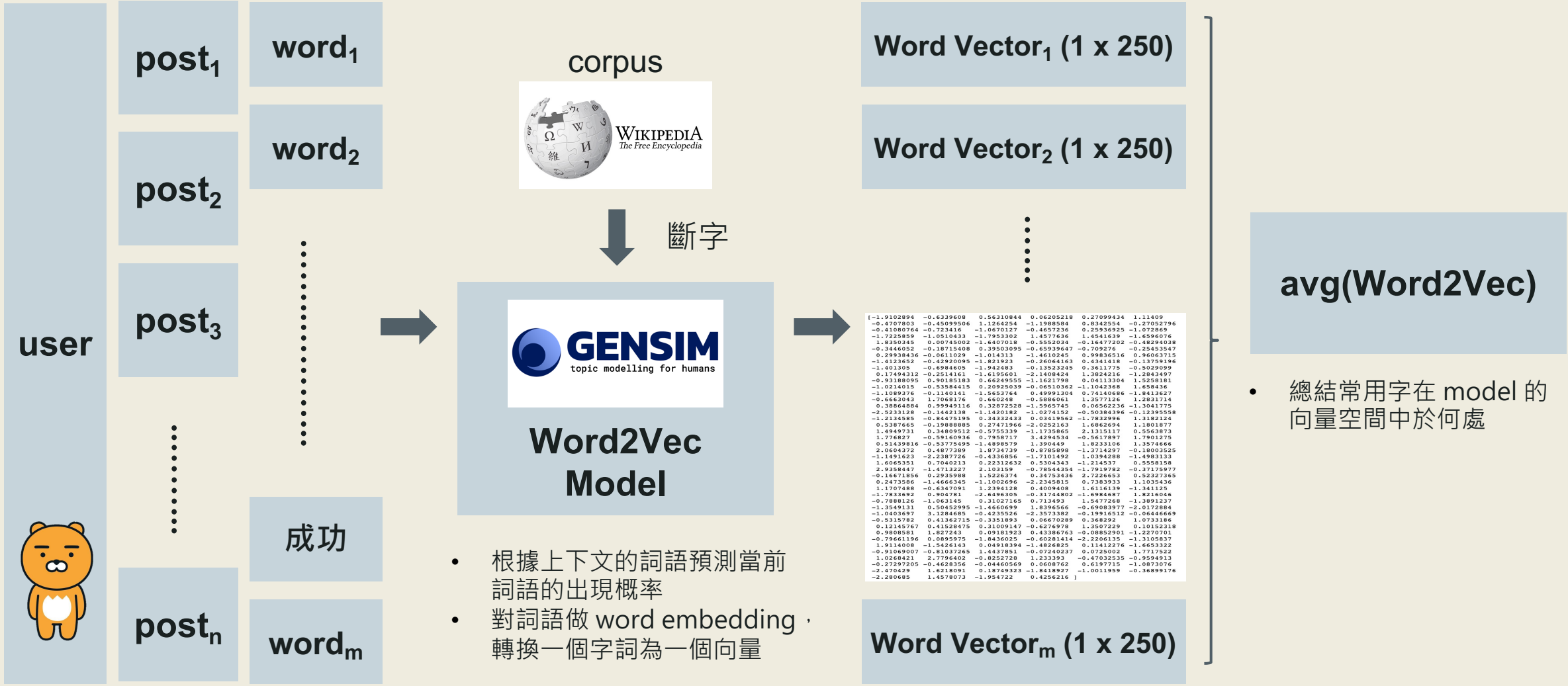
Vector (1 x 512)

$Vector_1$ (1 x 512) $Vector_2$ (1 x 512) $Vector_3$ (1 x 512) …… $Vector_m$ (1 x 512)

cosine similarity → 找出最相近的向量即為最推薦的課程

02 Post Analysis and Recommendation

找出使用者所有貼文中 TF-IDF 高的字輸入進 w2v model 後，取出所有詞向量的平均



- 根據上下文的詞語預測當前詞語的出現概率
- 對詞語做 word embedding，轉換一個字詞為一個向量

- 總結常用字在 model 的向量空間中於何處

同理到揪課上所有課程的簡介後，比對相似度以推薦課程

user

揪課

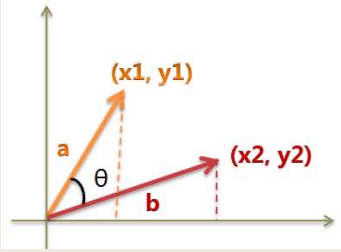
$word_1$	$word_2$	$word_3$	$word_4$	……	$word_n$	$course\ intro_1$	$course\ intro_2$	$course\ intro_3$	……	$course\ intro_m$

W2V Model

$Vector_1$ (1 x 250)	$Vector_2$ (1 x 250)	$Vector_3$ (1 x 250)	$Vector_4$ (1 x 250)	……	$Vector_n$ (1 x 250)	$Vector_1$ (k x 250)	$Vector_2$ (k x 250)	$Vector_3$ (k x 250)	……	$Vector_m$ (k x 250)

avg(Word2Vec)

						$avg(W2V)_1$	$avg(W2V)_2$	$avg(W2V)_3$		$avg(W2V)_m$

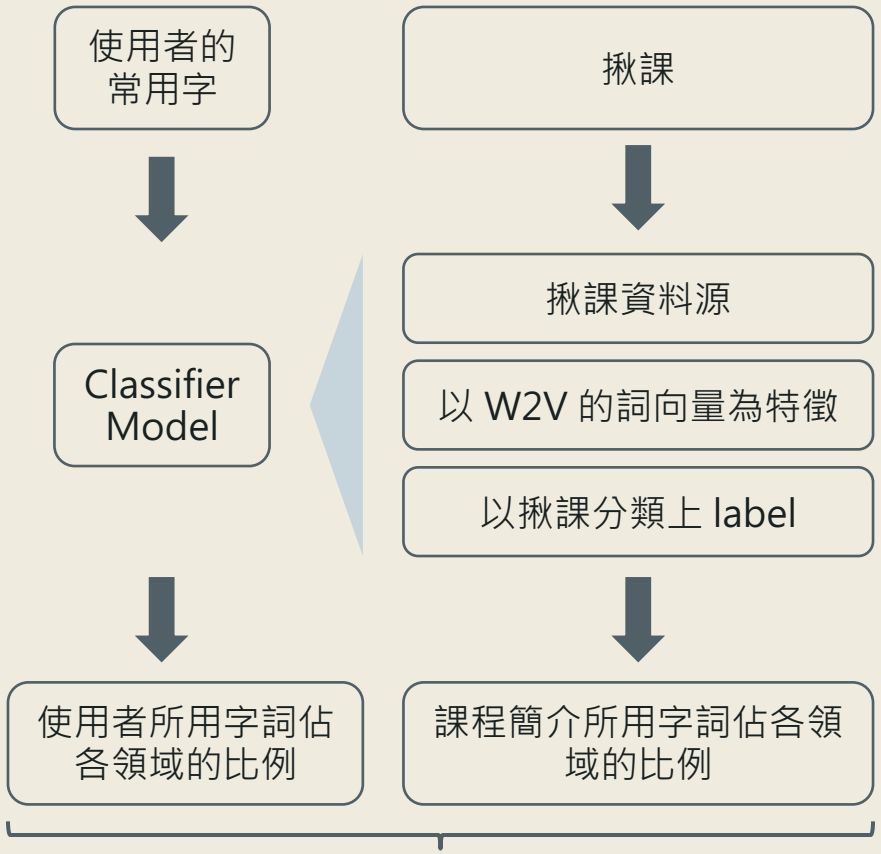


cosine similarity → 找出最相近的向量即為最推薦的課程

03 Experience and Review

實作失敗的經歷 & 本次專題需克服之處

Failure



取相似推薦 → 效果差

Flaw

W2V model

- 在處理 corpus 時，由於斷字無法處理複合字詞，會造成 model 的 key error
（ex：成功大學 →「成功」、「大學」）
- 以維基百科的資料為 corpus ，使得 model 預測的結果與想像有落差，若爬取社群媒體的文章可能可以改善
（ex：「獅子」最相關為「兩隻」）

```
print(model.wv.most_similar('成功大學', topn=10), '\n')
```

```
KeyError                                  Traceback (most recent call last)
<ipython-input-135-fde0a1741ddd> in <module>
----> 1 print(model.wv.most_similar('成功大學', topn=10), '\n')
```

```
print(model.wv.most_similar('獅子', topn=20), '\n')
```

```
[('兩隻', 0.5327986478805542), ('獨角獸', 0.519851565361023), ('公雞', 0.5154509544372559), ('烏鴉', 0.5109748840332031), ('鷲', 0.5069186687469482), ('黑毛', 0.4994606673717499), ('獅鷲', 0.4973223805427551), ('牙吠', 0.49647188186645
```

Reality

Facebook

- 大部分同學的貼文大多與生活日常有關，較少分享學術知識、想法觀點
- 分享的內容因網頁結構複雜，尚未包含在分析的貼文裡

揪課

- 資料數不夠多，分類器 precision 不到 40%
- 平台上的多分類造成字詞語意的分類不精確

Thank you!