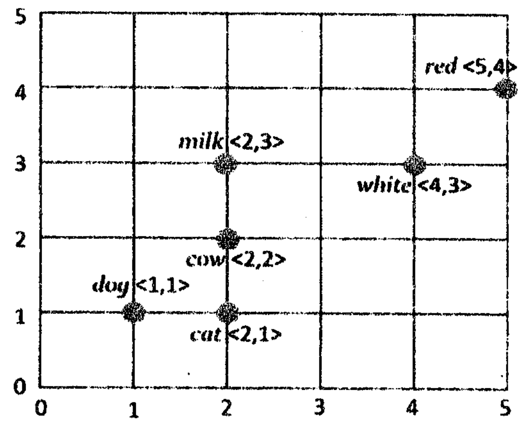
第一題以英文作答、第二題以中文作答。

1.(50%) Computational thinking

近年在計算語言學領域，常用一些技術將語詞表徵成「向量」。其中，「詞嵌入」(*word embeddings*) 是一種常用的詞義向量表徵方式。我們可以想像這是一種語詞到 n 維空間的映射。假定 $n = 2$, 那麼語詞就被映射到二維平面的點上。



舉例來說，上面這個圖呈現了 6 個英文詞彙的虛構詞嵌入。你可以看出語詞意思接近的，之間的距離上較短。(注意這個圖跟以下的問題無關)

以下是一些按字母順序排列的英文字：
boy, first, girl, grammar, language, literature, man, mathematics, mathematician, number, one, position, queen, second, time, two, woman

以下的表格從 A-Z 則是這些英文字（非按照上述順序）的山塔利語 (Santali) 翻譯，還有它們的英文詞嵌入。

#	Santali Translation	English Embedding
A	[illegible]	<3,3>
B	[illegible]	<3,2>
C	[illegible]	<5,4>
D	[illegible]	<7,2>
E	[illegible]	<7,4>
F	[illegible]	<8,2>
G	[illegible]	<8,4>
H	[illegible]	<9,3>
I	[illegible]	<10,6>

#	Santali Translation	English Embedding
J	[illegible]	<8,13>
K	[illegible]	<4,7>
L	[illegible]	<5,6>
M	[illegible]	<3,8>
N	[illegible]	<9,10>
O	[illegible]	<11,11>
P	[illegible]	<8,12>
Q	[illegible]	<10,13>

(1) 請把英文字與字母（A to Q）對應起來，並解釋理由。
(2) *king, linguist, princess* 這三個詞彙的詞嵌入為何？並解釋理由。
(3) 以下這個詞可能的英文翻譯為何？

[illegible]

Notes: Santali is a language in the Munda subfamily of Austroasiatic languages. It is spoken by around 6.2 million people in India, Bangladesh, Bhutan and Nepal. Most of its speakers live in India, in the states of Jharkhand, Bihar, Odisha, Tripura, Mizoram, Assam and West Bengal. Since Santali did not have a script, it was written using the Roman script or the Eastern Nagari (Bangla) script. However, none of the existing scripts were able to phonetically represent the Santali language. This resulted in the invention of a new script called *Ol Chiki* by Pandit Raghunath Murmu in 1925. He is popularly known as Guru Gomke among the Santals. (source: PANINI 2018)

見背面

2.(50%) Corpus-method

如果妳／你聽說在臉書的貼文中，女性在表情符號的使用上比起男性還要豐富。請問妳／你如何利用語料庫方法來確認或否證這個假說？這樣的方法可能的限制為何？

試題隨卷繳回