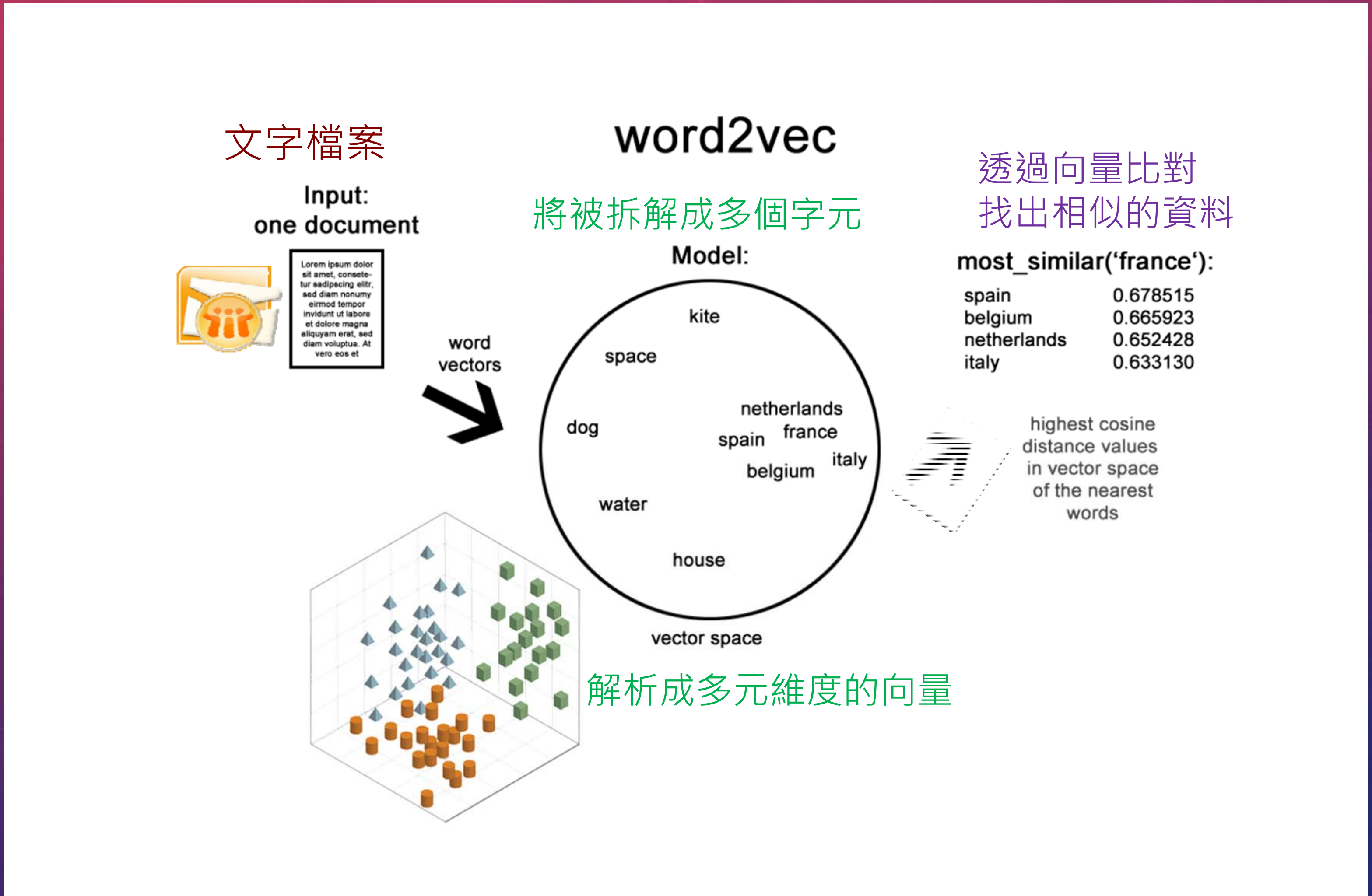
FOUNDATIONS OF NATURAL LANGUAGE UNDERSTANDING

自然語言理解的基礎

張家瑋 博士
新漢股份有限公司創新工業4.0中心顧問

SEMANTIC SIMILARITY MEASURES



文字檔案

word2vec

透過向量比對
找出相似的資料

將被拆解成多個字元

Input:
one document

Model:

most_similar('france'):

spain	0.678515
belgium	0.665923
netherlands	0.652428
italy	0.633130

highest cosine distance values in vector space of the nearest words

Lorem ipsum dolor sit amet, consete-tur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et
word vectors
kite
space
dog
netherlands
spain
france
italy
belgium
water
house
vector space

解析成多元維度的向量

VECTOR REPRESENTATION

	w_1	w_2	w_3	..	..	..	w_{n-1}	w_n	label
D_1	0.11	0.23	0	..	..	..	0.57	0	0
D_2	0	0	0	..	..	..	0.29	0.7	1
D_3	0	0.81	0.44	..	..	..	0	0	0
D_4	0	0.37	0	..	..	..	0	0.16	1
..	..	..	..	..	..	..	..	..	..
D_k	..	..	..	..	..	..	..	..	1

Machine learning

TF-IDF

TF-IDF

- TF: term frequency: $\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$
- IDF: inverse document frequency: $\mathrm{idf_i} = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$

where:

- |D|: total number of documents in the corpus
- $|\{j : t_i \in d_j\}|$: number of documents where term t_i appears

Then:

- $\mathrm{tfidf_{i,j}} = \mathrm{tf_{i,j}} \times \mathrm{idf_i}$

- The calculation of tf–idf for the term "this" is performed as follows:
  -

    ,

  - So tf–idf is zero for the word "this", which implies that the word is not very informative as it appears in all documents.