NLP and Word Embeddings

deeplearning.ai

Properties of word embeddings

Analogies

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	−1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.70	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97

e_{5391}

e_{man}

e_{woman}

Man → Woman as King → ? Queen

$e_{man} - e_{woman} \approx e_{king} - e_?$

Queen

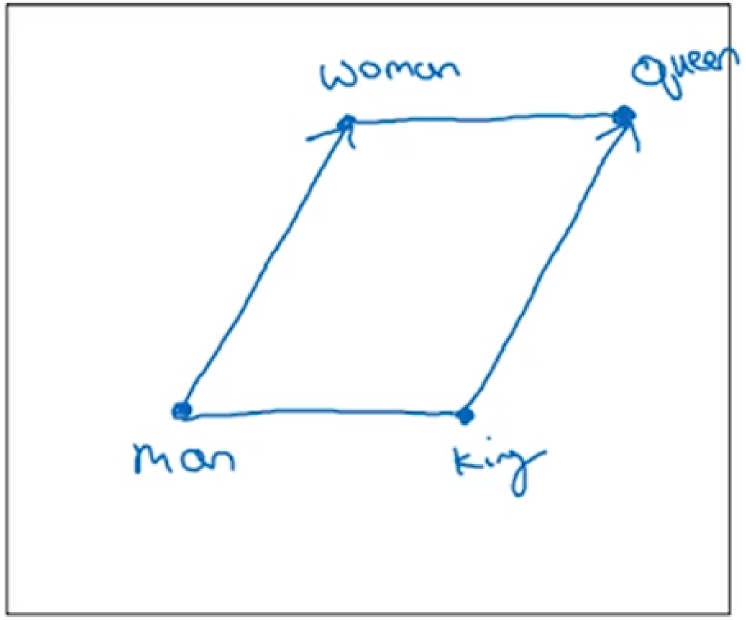
$e_{man} - e_{woman} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

$e_{king} - e_{queen} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

[Mikolov et. al., 2013, Linguistic regularities in continuous space word representations]

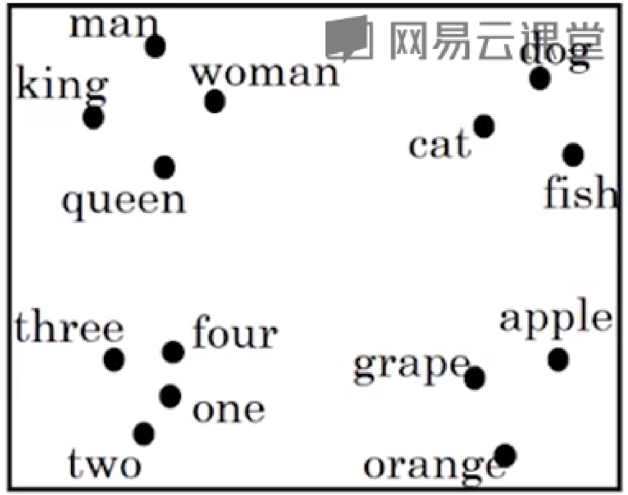
Andrew Ng

Analogies using word vectors



300 D

3000 → 2D



t-SNE

$$e_{man} - e_{woman} \approx e_{king} - \cancel{e_?}\ e_w$$

Find word w: $\arg\max_w \boxed{sim(e_w, e_{king} - e_{man} + e_{woman})}$

30 - 75%

Andrew Ng

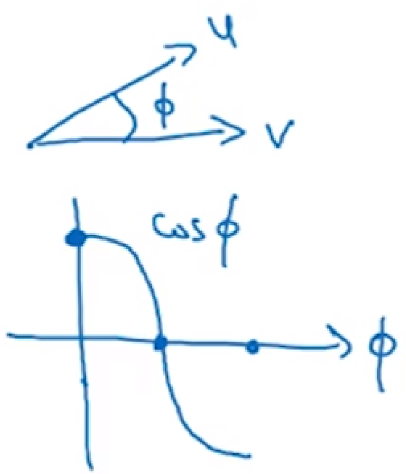
vector difference

t-SNE is non-linear mapping,and many of the parallelogram analogy relationships will be broken by t-SNE.

Cosine similarity

$$\boxed{sim(e_w, e_{king} - e_{man} + e_{woman})}$$

$$\text{sim}(u,v) = \frac{u^T v}{\|u\|_2 \|v\|_2}$$



$$\|u - v\|^2$$

Man:Woman as Boy:Girl

Ottawa:Canada as Nairobi:Kenya

Big:Bigger as Tall:Taller

Yen:Japan as Ruble:Russia

Andrew Ng

square distance or Euclidean distance ||u-v|| square
technically, this would be a measure of dissimilarity rather a measure of similarity, so we need to take the negative of this, and this will work ok as well.
Although I see cosine similarity being used a bit more often.
And the difference between this is that how it normalizes the lengths of the vector u and v.
So one of the remarkable result about word embeddings is the generality of analogy relationships they can learn.