Word - 2 - Vec

- It is technique in d.L. allows us to do mathematics with word.

eg.

give eqn to computer king - man + women
& computer will tell you - ans = Queen.

Computer cannot understand word (characters) so need to understand word by numbers.
i.e. set of numbers = vectors.

eg.

diff. ways of representing word "king".
has authority? = 1
rich ? = 1
gender (male) ? = -1
have tail? = 0 i.e [1, 1, -1, 0]

eg & Horse
authority = 0
rich = 0 ; i.e [0, 0, -1, 1]
gender = -1
have tail = 1

Suppose we have words from story...

	Battle	Horse	king	man	queen	women
autority	0	0.01	1	0.2	1	0.2
event	1	0	0	0	0	0
tail?	0	1	0	0	0	0
rich	0	0.1	1	0.3	1	0.2
gender	0	1	-1	-1	1	1

for different words we come up with diff. vectors.
now
let's do mathematics,

king - man + women ≅ Queen.

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ -1 \end{bmatrix} \begin{bmatrix} 0.2 \\ 0 \\ 0 \\ 0.3 \\ -1 \end{bmatrix} \begin{bmatrix} 0.2 \\ 0 \\ 0 \\ 0.2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0.9 \\ 1 \end{bmatrix} \cong \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

- not exactly but near to same.

But coding / Handcrafting of millions of such words is not making sense so we can use neural networks.

So

1. Take a fake problem
2. solve it using neural n/w.
3. You get word embedding as a side effect.

Let's

fake problem:- find missing word in sentence.

But this is ~~not~~ real problem but reason calling it fake is

There is a story ----

P.T.O.

There lived a king called Ashoka in India. After kalinga battle, he converted to Buddhism. This mighty king orderd his ministers to put together a peaceful treaty with their neighboring kingdoms. The emperor orderd his ministers to also build stupa, a monument with Buddha's teachings.

Fake problem;

king orderd his minister.

emperor orderd his ministers.

when we give this task to fill a missing words to computer as a side effect It will learn vectors for kings emperor

side effect

king $\begin{bmatrix} 0.7 \\ 0.4 \\ 1.2 \\ 3.8 \end{bmatrix}$ emperor $\begin{bmatrix} 0.7 \\ 0.5 \\ 0.1.2 \\ 3.8 \end{bmatrix}$ i.e. king ~ emperor

i.e. You can do synonyms, antonyms etc.

Then

① eating _____ is very healthy.

table, angry, truck, apple, pizza, ~~watch~~ walnut

② Nasa launched _____ last month.

table, angry, truck, rocket, apple, pizza.

i.e meaning of word can be inferred by surrounding words.

* Training samples from paragraph :-

window of size 3 [-, -, -]

lived, a → There

a, king → lived

:

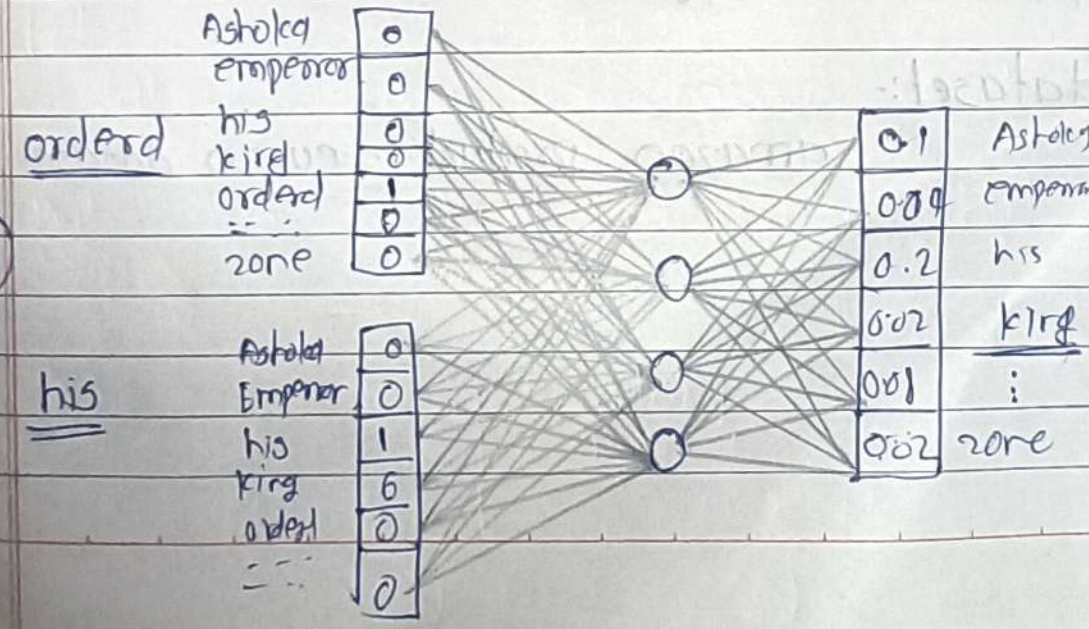
ordered, his → king

ordered, his → emperor & feed in Neural n/w.

ie

i/p

o/p.



if 5000 words in vocabolary (i.e. set of unique word) the input word will be 1 others 0.

Hidden vectors are of size of embedding vector (not fixed)

This approach is called Continuous Bag of words CBOW

CBOW - Predicting target from context.

Skip Gram - Predicting context from target.

Implementation :-

Amazon Product Review for cellphone accessories & build word-2-vec model in gensim library.

pip install gensim
pip install python-Levenshtein.

dataset:-

amazon product review dataset. 43 mb.