Deep Learning y Procesamiento del Lenguaje

Ricardo Castro
Software Engineer
uBiome

www.linkedin.com/in/ricardo-castro-a386879

Contenidos

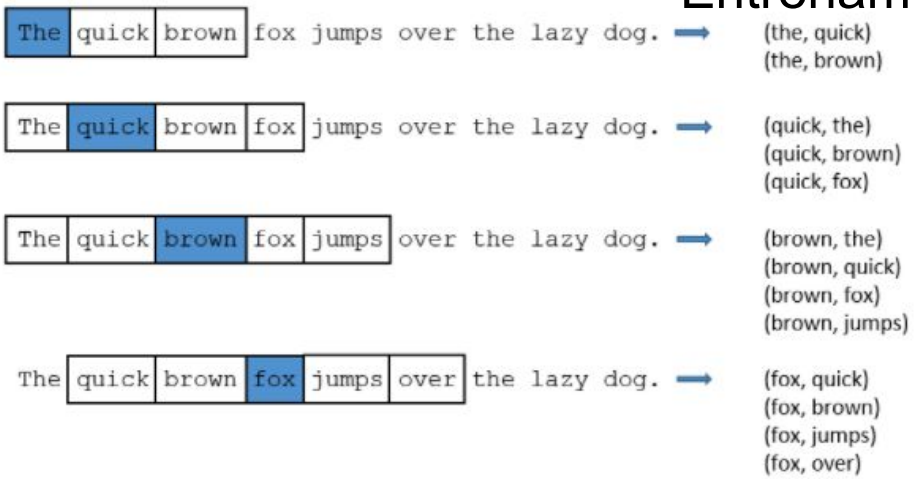
1. Como se define un Contexto?
2. One Hot Encoding no captura el significado
3. Representación de Palabras
4. Entrenamiento
5. Visualización
6. Google Projector
7. Procesar texto como una Serie
8. Redes Recurrentes
9. Preguntas

Como se define un Contexto?

Texto Original

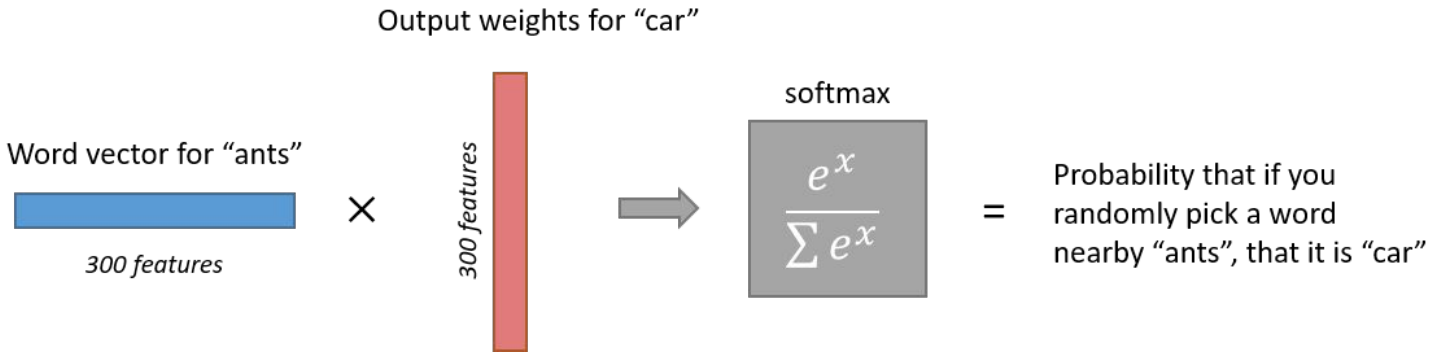
Set de Entrenamiento

Labels ??

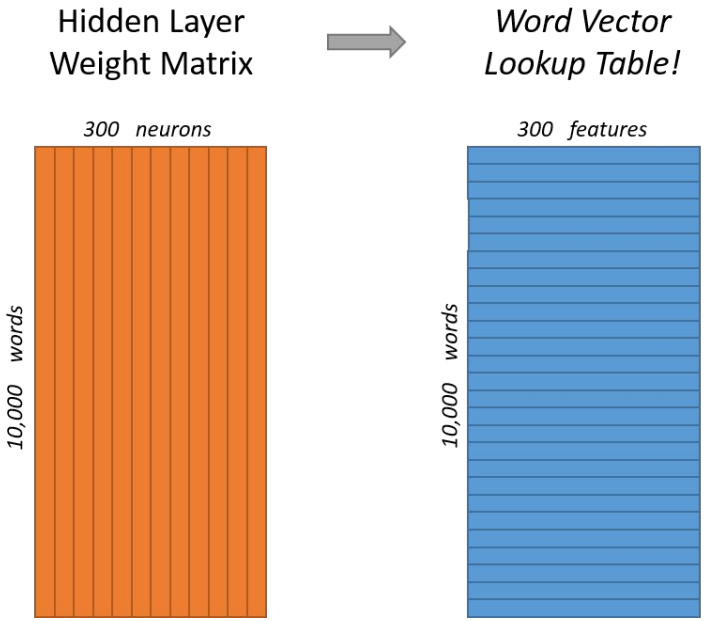


One Hot Encoding no captura el significado

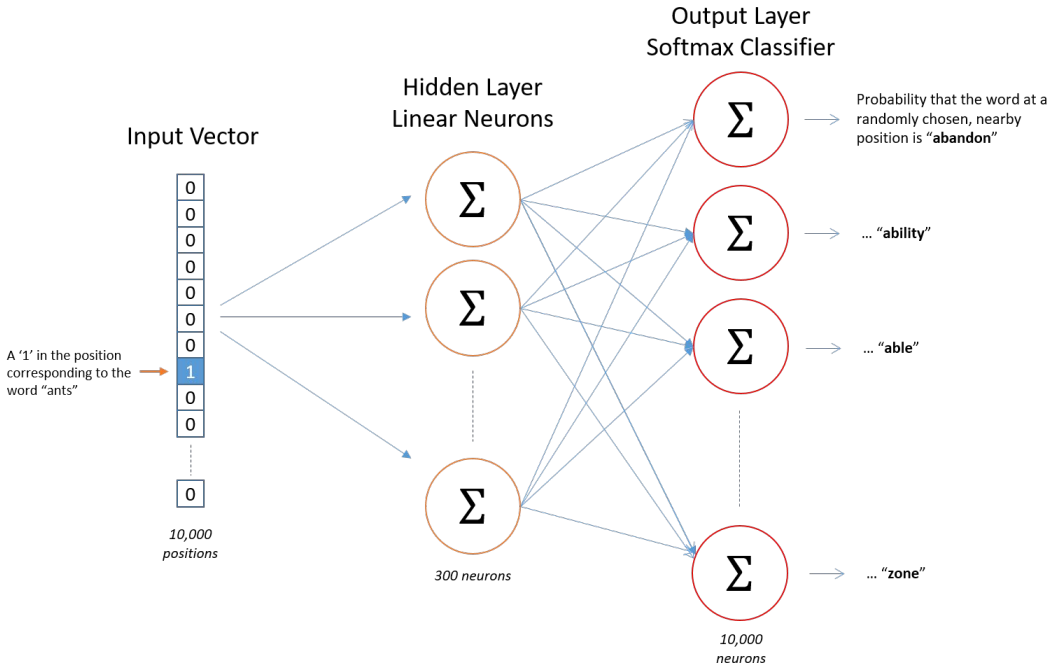
Se quiere representar una palabra como un vector, en función de sus palabras vecinas, i.e. Contexto/Significado.



Representación de Palabras



Entrenamiento



Similaridades

Nearest words to frog:

1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

litoria

leptodactylidae

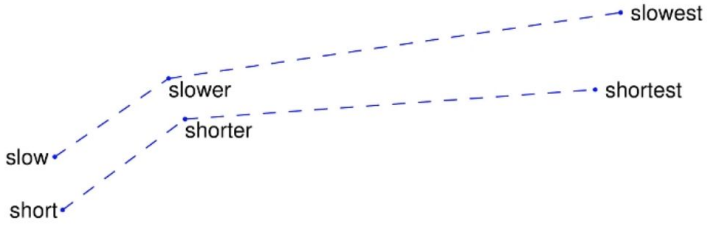
rana

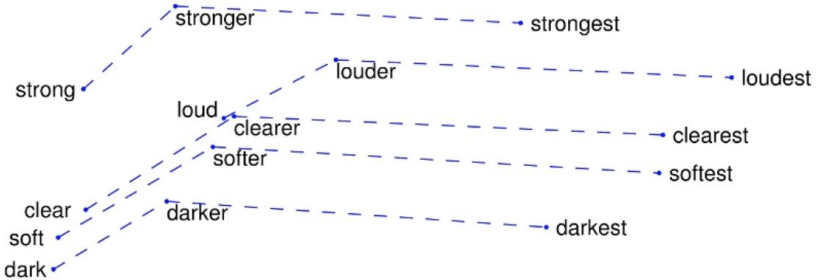
eleutherodactylus

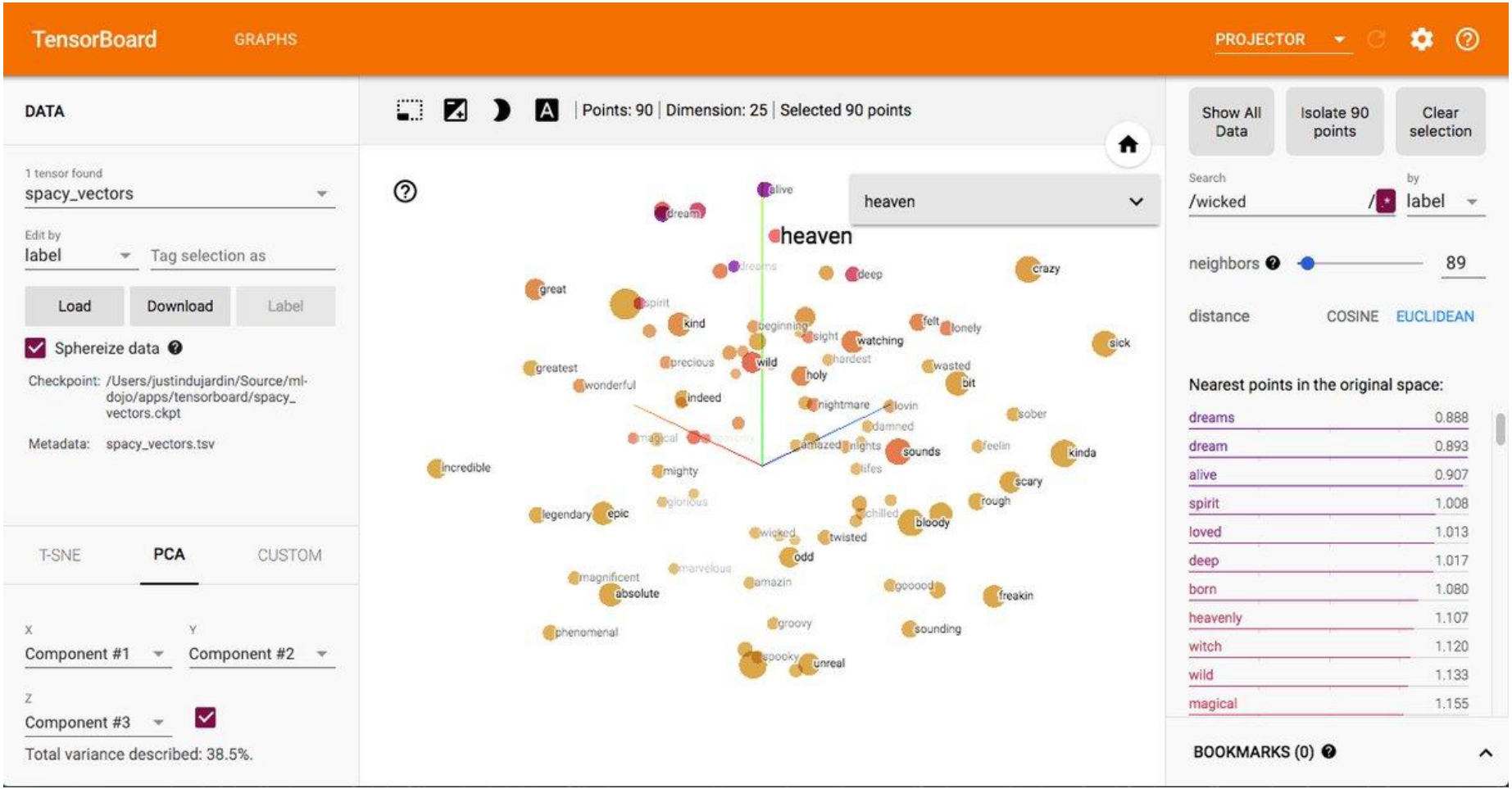
Visualización de data Multidimensional

- Palabras similares deberían estar cercanas en el hiper espacio semántico
- Aritmética vectorial de palabras
- Clusterización de palabras similares
- Queremos una métrica numérica de qué tan similares son dos conceptos

Visualización de data Multidimensional



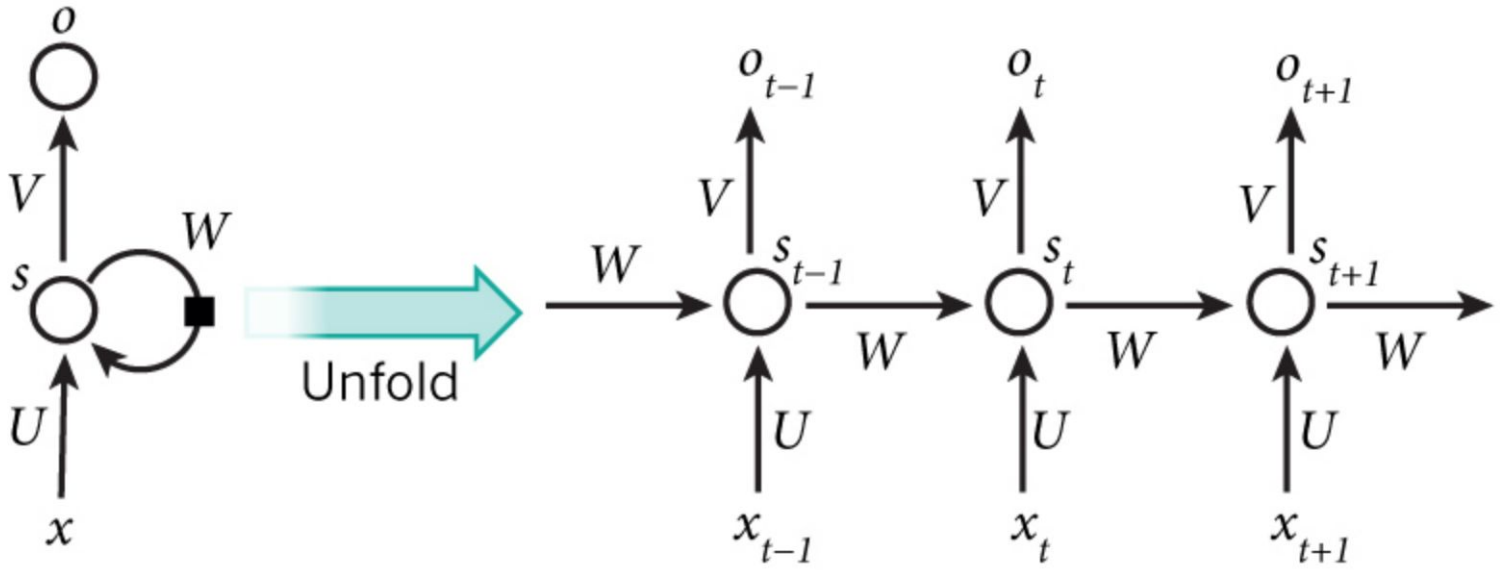



TensorBoard
GRAPHS
PROJECTOR
DATA
1 tensor found
spacy_vectors
Edit by
label
Tag selection as
Load
Download
Label
Sphereize data
Checkpoint: /Users/justindujardin/Source/ml-dojo/apps/tensorboard/spacy_vectors.ckpt
Metadata: spacy_vectors.tsv
T-SNE
PCA
CUSTOM
X
Component #1
Y
Component #2
Z
Component #3
Total variance described: 38.5%.
Points: 90 | Dimension: 25 | Selected 90 points
heaven
alive
dreams
dream
spirit
great
kind
deep
crazy
beginning
sight
watching
felt
lonely
sick
greatest
precious
wild
holy
hardest
wasted
bit
wonderful
indeed
nightmare
lovin
damned
sober
magical
amazed
nights
sounds
feelin
kinda
incredible
mighty
lifes
scary
glorious
rough
legendary
epic
chilled
bloody
wicked
twisted
odd
marvelous
magnificent
amazin
absolute
goooood
freakin
groovy
sounding
phenomenal
spooky
unreal
Show All Data
Isolate 90 points
Clear selection
Search
/wicked
by
label
neighbors
89
distance
COSINE
EUCLIDEAN
Nearest points in the original space:
dreams 0.888
dream 0.893
alive 0.907
spirit 1.008
loved 1.013
deep 1.017
born 1.080
heavenly 1.107
witch 1.120
wild 1.133
magical 1.155
BOOKMARKS (0)

Procesar texto como una Serie

- Frases no son construidas al azar
- Coherencia interna : género, número, tiempo, etc
- Solo un set de posibles alternativas para continuar una frase
- Apoyo en Big Data y conteo de co-ocurrencias

Redes Neuronales Recurrentes(RNN)



Redes Neuronales Recurrentes (RNN)

- Cada palabra es consecuencia de sus predecesores.
- Cuanto debo mirar al pasado para predecir el futuro ?
- Desambiguación de referencias

Preguntas ?