

Conclusiones

El objetivo principal de este tutorial es **representar palabras como si fueran vectores**, lo cual llaman “**Palabras embebidas**”. Cuando se habla de tareas de reconocimiento de imágenes o voz, se toman grandes conjuntos de datos (Como pixeles, coeficientes) para obtener información necesaria y poder realizar con éxito dichas tareas.

Respecto al reconocimiento de palabras o procesamiento de un lenguaje natural, usualmente se trata como **símbolos discretos** en donde, por ejemplo, la palabra ‘gato’ podría estar representada como ‘Id537’ o la palabra ‘perro’ representada por ‘Id143’. Esto provoca que el modelo en realidad **aproveche muy poco lo que ha aprendido** sobre un objeto, cuando está procesando los datos de otros. Esto provoca que los datos se **dispersen** y a la vez provoca que se requieran aún más datos para poder generar modelos estadísticos. Se justifica la implementación de este tutorial pues la **representación de vectores** puede solucionar este problema.

El punto es que los modelos de espacio vectorial, puedes representar las palabras analizadas en un espacio vectorial continuo, es decir que las **palabras semánticamente parecidas**, se pueden asignar a **puntos cercanos**. El proceso se basa en dos enfoques:

- Métodos basados en Recuento: Análisis Semántico. Calculan las estadísticas de la frecuencia con la que una palabra coincide con sus palabras vecinas y luego mapean estos conteos.
- Métodos Predictivos: Modelos de Lenguaje Probabilístico. Intentan predecir directamente una palabra de sus vecinos en términos de vectores.

En el caso de este tutorial, el modelo que se utiliza es el **Predictivo**, pues lo que hace el programa es intentar predecir una palabra en específico de sus vecinos.

Se nos introduce el modelo '**Skip-gram**', el cual es capaz de predecir palabras tomando como punto de referencia una palabra en específico. Basándome en la explicación del tutorial:

Consideramos: “el rápido zorro marrón saltó sobre el perro perezoso”

1. Se forma un conjunto de pares de datos de palabras y los contextos en los que aparecen. Algo así como:

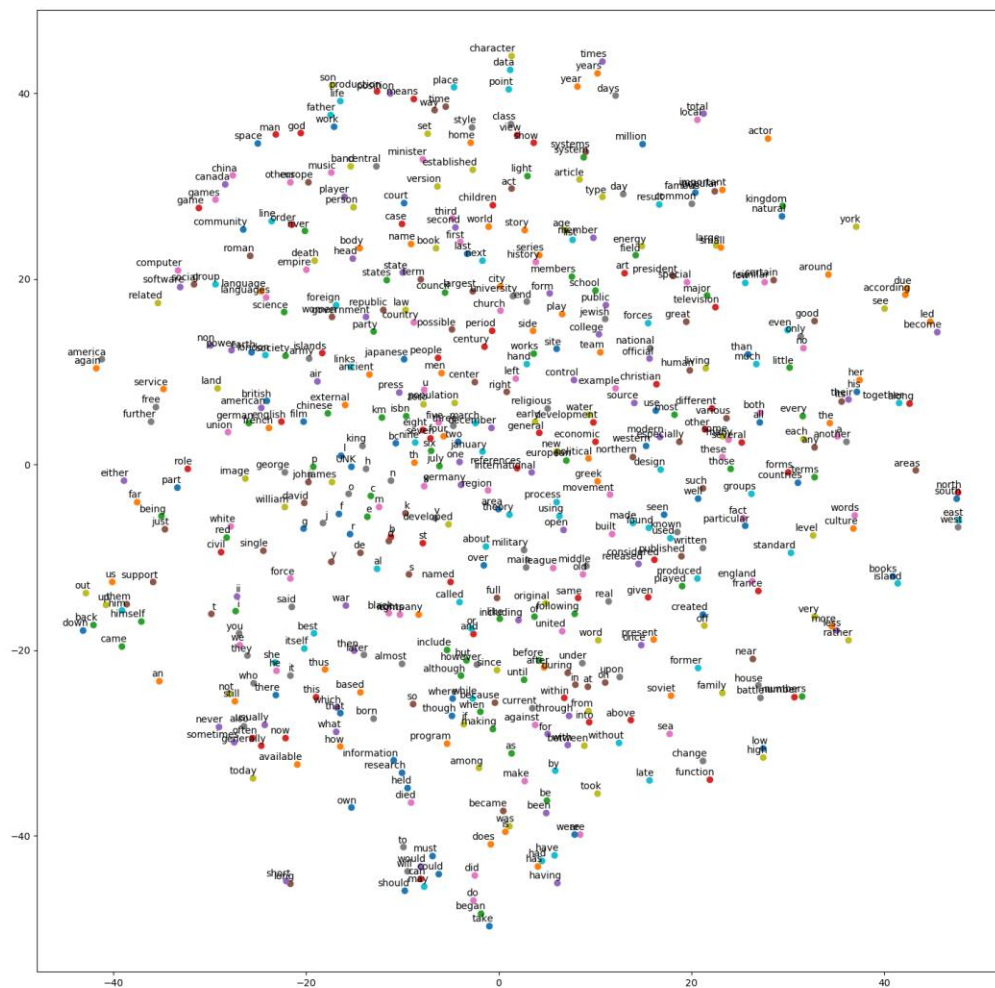
([el, marrón], rápido), ([rápido, zorro], marrón), ([marrón, saltado], zorro)

2. Lo que hace el modelo, es invertir los contextos y los objetivos, e intenta predecir cada palabra de contexto con su palabra destino. Es decir, que si una entrada a la red neuronal fuera 'rápido', la salida sea 'el' o 'marrón'.

```
C:\Users\erosh\Dropbox\TEC\8vo Semestre\Inteligencia Artificial\Inteligencia Artificial\Investigacion\vectorRepresentationOfWords>python word2vector_basic.py
Found and verified text8.zip
Data size 17005207
Most common words (+UNK) [['UNK', 418391], ('the', 1061396), ('of', 593677), ('and', 416629), ('one', 411764)]
Sample data [5234, 3081, 12, 6, 195, 2, 3134, 46, 59, 156] ['anarchism', 'originated', 'as', 'a', 'term', 'of', 'abuse',
'first', 'used', 'against']
3081 originated -> 5234 anarchism
3081 originated -> 12 as
12 as -> 3081 originated
12 as -> 6 a
6 a -> 195 term
6 a -> 12 as
195 term -> 6 a
195 term -> 2 of
```

Después de **100 000 steps**, finaliza el entrenamiento. El proceso tomó aproximadamente 15 minutos.

```
C:\WINDOWS\system32\cmd.exe
Nearest to with: between, pulau, when, in, by, displeased, through, vma,
Nearest to so: venn, collaboration, before, jungles, leontopithecus, sponsors, kumar, but,
Nearest to used: found, known, scalia, microcebus, operatorname, dasyprocta, cebus, busan,
Nearest to often: sometimes, commonly, usually, also, now, not, frequently, generally,
Nearest to known: used, explanatory, symbiotic, farsi, such, largely, dasyprocta, well,
Nearest to on: in, at, upon, yoannis, through, into, within, dasyprocta,
Nearest to he: it, she, they, who, there, but, drift, ursus,
Average loss at step 92000 : 4.67074492753
Average loss at step 94000 : 4.71248634839
Average loss at step 96000 : 4.68977924263
Average loss at step 98000 : 4.57835596848
Average loss at step 100000 : 4.68583895755
Nearest to while: but, however, and, although, when, though, bogies, is,
Nearest to by: through, pulau, be, was, with, during, microbats, operatorname,
Nearest to one: seven, two, five, four, six, three, eight, ursus,
Nearest to during: in, after, at, under, until, from, by, pulau,
Nearest to at: in, during, on, cebus, dasyprocta, ifbb, pulau, under,
Nearest to called: outage, aoc, twh, cebus, gadget, dojos, pointer, but,
Nearest to its: their, his, the, her, agouti, microcebus, kosar, iit,
Nearest to many: some, several, these, various, all, most, manure, agave,
Nearest to however: but, while, although, that, circ, tourists, two, thaler,
Nearest to with: between, pulau, in, when, by, displeased, aluma, including,
Nearest to so: venn, collaboration, before, jungles, sponsors, leontopithecus, but, however,
Nearest to used: found, known, scalia, available, referred, operatorname, considered, microcebus,
Nearest to often: sometimes, commonly, usually, also, now, generally, frequently, circ,
Nearest to known: used, explanatory, such, symbiotic, farsi, dasyprocta, largely, buried,
Nearest to on: in, upon, at, through, yoannis, within, dasyprocta, into,
Nearest to he: it, she, they, who, there, ursus, but, drift,
```



Cabe mencionar que es necesario tener las librerías de **'sklearn'**, **'matplotlib'** y **'scipy'** instaladas para poder generar dicho gráfico, pues si falta alguna de ellas, el programa no podrá generarlo.

```
Nearest to its: their, his, the, her, agouti, microcebus, kosar, iit,  
Nearest to many: some, several, these, various, all, most, manure, agave,  
Nearest to however: but, while, although, that, circ, tourists, two, thaler,  
Nearest to with: between, pulau, in, when, by, displeased, aluma, including,  
Nearest to so: venn, collaboration, before, jungles, sponsors, leontopithecus, but, however,  
Nearest to used: found, known, scalia, available, referred, operatorname, considered, microcebus,  
Nearest to often: sometimes, commonly, usually, also, now, generally, frequently, circ,  
Nearest to known: used, explanatory, such, symbiotic, farsi, dasypsecta, largely, buried,  
Nearest to on: in, upon, at, through, yoannis, within, dasypsecta, into,  
Nearest to he: it, she, they, who, there, ursus, but, drift,  
Please install sklearn, matplotlib, and scipy to show embeddings.  
No module named 'sklearn'
```

Luego de este procedimiento, se debe ejecutar el programa **'words2vec_optimized.py'** y después de un poco de investigación me di cuenta de que no puede ser ejecutado en Windows, puesto que el programa debe generar un archivo llamado **'word2vec_ops.so'** para ser utilizado más adelante, pero para generar este archivo, se deben compilar uno programas con extensión .cc, por lo que procedí a iniciar una máquina virtual con Linux, instalar Python 3, TensorFlow y sus dependencias.

Una vez listo el ambiente de Linux, se deben obtener otros archivos necesarios para el entrenamiento de la red neuronal optimizada.

- Text8: Datos de entrenamiento para la red, se obtienen del tutorial y al haber ejecutado el primer entrenamiento.
- questions-words.txt: Datos para evaluación, se encuentran en la siguiente página: <http://download.tensorflow.org/data/questions-words.txt>

Con estos archivos, se procede a hacer el **entrenamiento** de la **red neuronal optimizada**.

```
eros@ErosPC: ~/Documents/Proyectos/InvestigacionIA
k/load_library.py", line 56, in load_op_library
  lib_handle = py_tf.TF_LoadLibrary(library_filename, status)
File "/home/eros/.local/lib/python3.5/site-packages/tensorflow/python/framework
k/errors_impl.py", line 473, in __exit__
  c_api.TF_GetCode(self.status.status))
tensorflow.python.framework.errors_impl.NotFoundError: /home/eros/Documents/Proy
ectos/InvestigacionIA/word2vec_ops.so: cannot open shared object file: No such f
ile or directory
eros@ErosPC:~/Documents/Proyectos/InvestigacionIA$ python3 word2vec_optimized.py
--train_data text8 --eval_data questions-words.txt --save_path /tmp/ --epochs_t
o_train 1
2017-11-10 14:05:14.229701: I tensorflow/core/platform/cpu_feature_guard.cc:137]
Your CPU supports instructions that this TensorFlow binary was not compiled to
use: SSE4.1 SSE4.2 AVX
2017-11-10 14:05:17.930718: I word2vec_kernels.cc:200] Data file: text8 contains
100000000 bytes, 17005207 words, 253854 unique words, 71290 unique frequent wor
ds.
Data file: text8
Vocab size: 71290 + UNK
Words per epoch: 17005207
Eval analogy file: questions-words.txt
Questions: 17827
Skipped: 1718
Epoch    0 Step    943: lr = 0.025 words/sec =    8569
```

El proceso toma aproximadamente **20 minutos**.

```
eros@ErosPC: ~/Documents/Proyectos/InvestigacionIA
k/load_library.py", line 56, in load_op_library
  lib_handle = py_tf.TF_LoadLibrary(library_filename, status)
File "/home/eros/.local/lib/python3.5/site-packages/tensorflow/python/framework
k/errors_impl.py", line 473, in __exit__
  c_api.TF_GetCode(self.status.status))
tensorflow.python.framework.errors_impl.NotFoundError: /home/eros/Documents/Proy
ectos/InvestigacionIA/word2vec_ops.so: cannot open shared object file: No such f
ile or directory
eros@ErosPC:~/Documents/Proyectos/InvestigacionIA$ python3 word2vec_optimized.py
--train_data text8 --eval_data questions-words.txt --save_path /tmp/ --epochs_t
o_train 1
2017-11-10 14:05:14.229701: I tensorflow/core/platform/cpu_feature_guard.cc:137]
Your CPU supports instructions that this TensorFlow binary was not compiled to
use: SSE4.1 SSE4.2 AVX
2017-11-10 14:05:17.930718: I word2vec_kernels.cc:200] Data file: text8 contains
100000000 bytes, 17005207 words, 253854 unique words, 71290 unique frequent wor
ds.
Data file: text8
Vocab size: 71290 + UNK
Words per epoch: 17005207
Eval analogy file: questions-words.txt
Questions: 17827
Skipped: 1718
Epoch    0 Step   90484: lr = 0.014 words/sec =    6093
```


Una vez finalizado el entrenamiento con aproximadamente **150 000 steps**, se obtiene una **precisión** aproximada de **12.6%**.

```
eros@ErosPC: ~/Documents/Proyectos/InvestigacionIA
File "/home/eros/.local/lib/python3.5/site-packages/tensorflow/python/framework/errors_impl.py", line 473, in __exit__
    c_api.TF_GetCode(self.status.status))
tensorflow.python.framework.errors_impl.NotFoundError: /home/eros/Documents/Proyectos/InvestigacionIA/word2vec_ops.so: cannot open shared object file: No such file or directory
eros@ErosPC:~/Documents/Proyectos/InvestigacionIA$ python3 word2vec_optimized.py
--train_data text8 --eval_data questions-words.txt --save_path /tmp/ --epochs_to_train 1
2017-11-10 14:05:14.229701: I tensorflow/core/platform/cpu_feature_guard.cc:137] Your CPU supports instructions that this TensorFlow binary was not compiled to use: SSE4.1 SSE4.2 AVX
2017-11-10 14:05:17.930718: I word2vec_kernels.cc:200] Data file: text8 contains 100000000 bytes, 17005207 words, 253854 unique words, 71290 unique frequent words.
Data file: text8
Vocab size: 71290 + UNK
Words per epoch: 17005207
Eval analogy file: questions-words.txt
Questions: 17827
Skipped: 1718
Epoch 1 Step 150739: lr = 0.006 words/sec = 3225
Eval 2247/17827 accuracy = 12.6%
eros@ErosPC:~/Documents/Proyectos/InvestigacionIA$
```

-Fin de Documentación para Tutorial #5-