CIFAR-100 Image Classification

Benjamin Gutierrez Garcia

- Este ejercicio consiste en clasificar imágenes del dataset CIFAR-100 en sus 100 categorías (mutuamente exclusivas), 20 superclases (2 etiquetas).
- Métrica del rendimiento: *multiclass classification accuracy*
- Solución Implementada: Python3 + TFLearn http://tflearn.org
- Escrita “encima” de Tensorflow - nivel de abstracción mayor
- Menos código , mayor claridad.
- La abstracción está basada en capas (o layers)
- Visualización de algunos parametros con o Tensorboard.

Ingestión de Datos

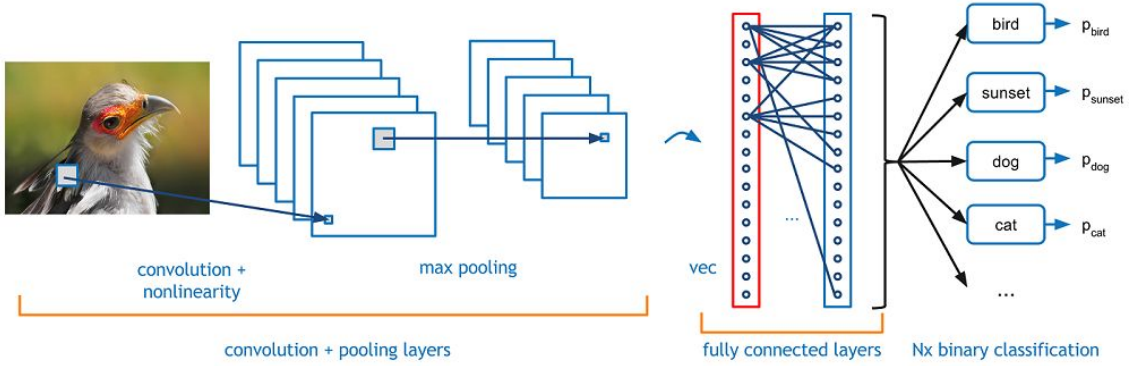
- Download automatico de https://www.cs.toronto.edu/~kriz/cifar.html a dir local
- Cada clase tiene 500 imagenes de training (50k) y 100 de prueba (10k)
- Clasificación relativamente no trivial
- Formato binario (python pickled) -> función decodifica a caracteres normales (unpickle).
- Cada imagen: 32x32x3 pixeles (3 colores RGB) ->matriz de 50,000 x 3072, una imagen por renglón.
- Cada imagen se convierte en un 4d-tensor `[records, channels, width, height]=[-1, 3, 32, 32]`
- One-Shot encoding: Datos categóricos, etiquetas->valores numéricos, ejemplo:

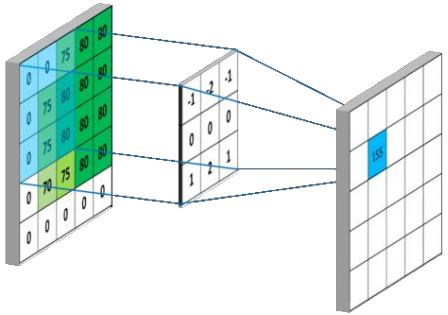
red,	green,	blue
1,	0,	0
0,	1,	0
0,	0,	1

Pre-Proceso de los Datos

- Re-escalamiento alrededor de la media y usando std de *feature values:*
  - Suprimir valores muy grandes → evitar sesgo/distorsión en el aprendizaje i.e. de los gradientes que son retro-propagados en la red
  - Más o menos a la misma escala. *Feature engineering*

- Data augmentation
  - Generar más imágenes de entrenamiento mediante reflexiones y rotaciones al azar (máx 15 grados)
  - Número de imágenes de entrenamiento relativamente pequeño.





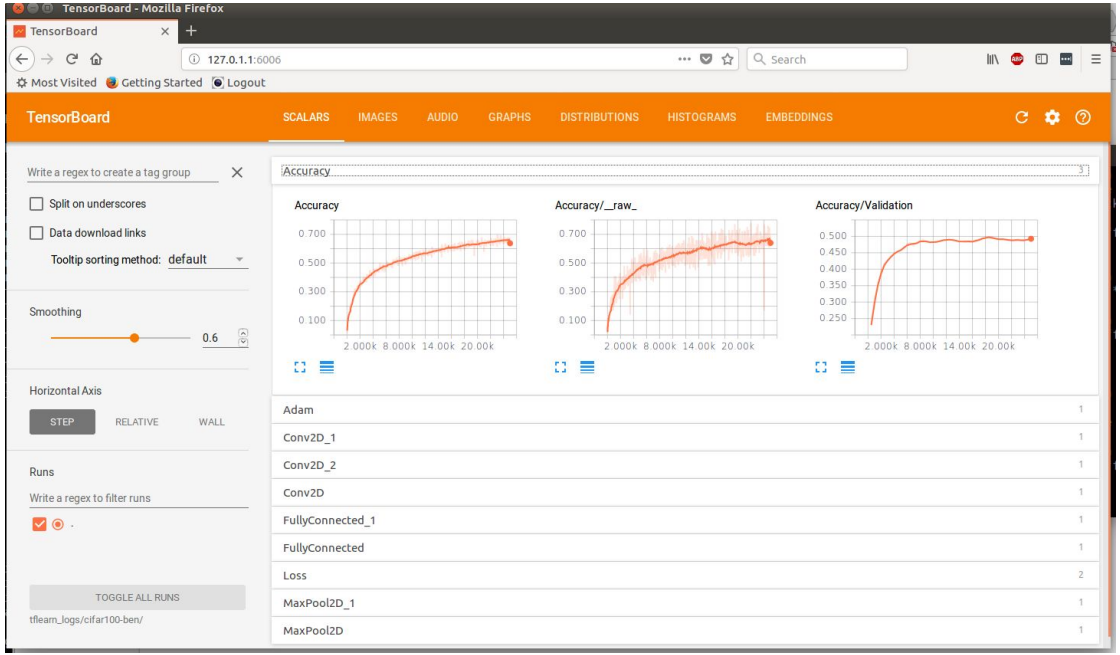
- Modelo de Aprendizaje*:* Convolutional neural network (CNN)
- De facto standard para Image/Clasificación
- 3 capas de convolución: *feature maps (coarser)* , filtros (replicados) de 3x3, stride=1, padding
- Función de activación ReLU (*Rectifier Linear Unit - layer -nonlinearity)*
- 2 capas de *max-pooling* de 2x2 (reduce comp. cost, overfitting)
- 2 capas tipo "*fully connected*", con función de activación softmax.
- El algoritmo de optimización para los pesos es ADAM
- *Dropout* al 50% (apagar nodos de las redes al azar para mitigar *overfitting)*.

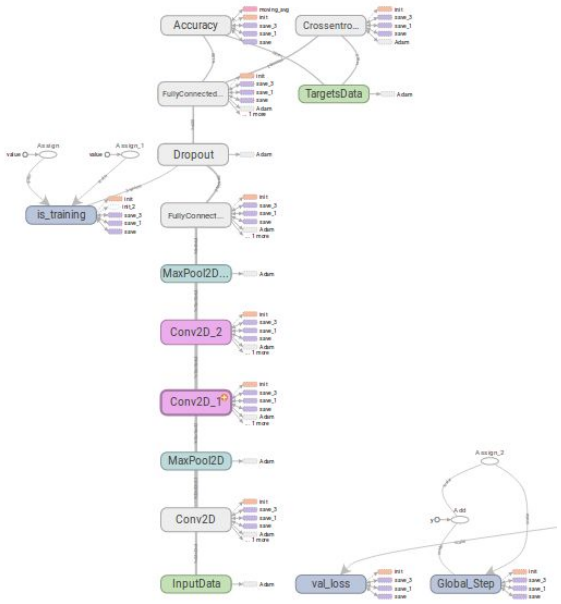
Resultados

- Ejecución:
  - ADAM optimizer (learning rate) es de 0.001.
  - El código se ejecutó por 50 épocas i.e. el dataset entero paso de adelante hacia atrás 50 veces.
  - ~2 horas de ejecución
  - Loss function*: Categorical Cross Entropy*
  - *Multiclass classification accuracy: categorical outcomes*
  - **El valor de la métrica obtenido fue de 0.4927, aproximadamente 49%.**
  - **(On testing data)**

$$Accuracy = \frac{\sum_M tp + tn}{\sum_M tp + tn + fp + fn}$$

Tensorboard





Minería de datos de Twitter para evaluar “sentimiento social”

- Objetivo: utilizar el stream de twitter para identificar tweets positivos o negativos
- El “humor” de la comunicación de tweeter es un ...
  - Reflejo de ritmos biológicos
  - Medida de los mercados bursátiles
  - Analizar “humor” a escalas nacionales o en contextos meteorológicos/desastres.
- **Los tweets representan datos muy sucios y sin estructura**: incompletos, sin texto, sin geo-información, , idioma, codificados en distintos lenguajes y formatos, etc.
- Primer Paso: capturar el stream de twitter usando su API **https://dev.twitter.com/apps**.
- Codificado en *unicode strings* --->API ---> JSON y sea humanamente legible.

Stream de datos

-
```
xpanded_url":"https:\/\/twitter.com\/QnA0304\/status\/992977255918850049\/photo\/1","type":"photo","sizes":{"thumb":{"w":150,"h":150,"resize":"
crop"},"large":{"w":575,"h":360,"resize":"fit"},"small":{"w":575,"h":360,"resize":"fit"},"medium":{"w":575,"h":360,"resize":"fit"}}}]}},"quote_
count":0,"reply_count":1,"retweet_count":47,"favorite_count":12,"entities":{"hashtags":[],"urls":[{"url":"https:\/\/t.co\/98WGM3H0gh","expanded
_url":"https:\/\/twitter.com\/i\/web\/status\/992977255918850049","display_url":"twitter.com\/i\/web\/status\/9\u2026","indices":[117,140]}],"u
ser_mentions":[],"symbols":[]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"ko"},"is_quote_statu
s":false,"quote_count":0,"reply_count":0,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[],"user_mentions":[{"screen_nam
e":"QnA0304","name":"\ube61\uccd0\uc788\ub294\ud050\uc5d4","id":890970636247482368,"id_str":"890970636247482368","indices":[3,11]}],"symbols":[
]},"favorited":false,"retweeted":false,"filter_level":"low","lang":"ko","timestamp_ms":"1525582227662"}
{"created_at":"Sun May 06 04:50:27 +0000 2018","id":992989950265253888,"id_str":"992989950265253888","text":"RT @landpsychology: We have conque
red the humans https:\/\/t.co\/nIF5HY4XvK","source":"\u003ca href=\"http:\/\/twitter.com\/download\/android\" rel=\"nofollow\"\u003eTwitter for
 Android\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_t
o_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":3306141957,"id_str":"3306141957","name":"SUE GIAIMO","screen_name":"giaimo_sue"
,"location":"Kent, OH","url":null,"description":"Love music, crazy, silly people...did I say I love music?","translator_type":"none","protected
":false,"verified":false,"followers_count":63,"friends_count":188,"listed_count":0,"favourites_count":572,"statuses_count":264,"created_at":"Tu
e Jun 02 01:03:32 +0000 2015","utc_offset":null,"time_zone":null,"geo_enabled":false,"lang":"en","contributors_enabled":false,"is_translator":f
alse,"profile_background_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_backg
round_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_link_color":"1DA1F2",
"profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true
,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/984107978306486273\/4Tl1sPcY_normal.jpg","profile_image_url_https":"https:\/\/pbs
.twimg.com\/profile_images\/984107978306486273\/4Tl1sPcY_normal.jpg","default_profile":true,"default_profile_image":false,"following":null,"fol
low_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweeted_status":{"created_at":"S
at May 05 17:24:05 +0000 2018","id":992817219984875520,"id_str":"992817219984875520","text":"We have conquered the humans https:\/\/t.co\/nIF5H
Y4XvK","display_text_range":[0,28],"source":"\u003ca href=\"https:\/\/postcron.com\" rel=\"nofollow\"\u003ePostcron App\u003c\/a\u003e","trunca
ted":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to
_screen_name":null,"user":{"id":2924593731,"id_str":"2924593731","name":"Land of cuteness","screen_name":"landpsychology","location":null,"url"
:"https:\/\/www.facebook.com\/Land-of-Cuteness-1085874071450143\/","description":"A page full of cute animals. We DON'T own any of the images.
If you want to remove one send me an email: landofpsychology@gmail.com","translator_type":"none","protected":false,"verified":false,"followers_
count":168559,"friends_count":115384,"listed_count":1493,"favourites_count":0,"statuses_count":106870,"created_at":"Tue Dec 16 09:24:05 +0000 2
014","utc_offset":7200,"time_zone":"Budapest","geo_enabled":false,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_backg
round_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_htt
ps":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_link_color":"1DA1F2","profile_sidebar_bo
rder_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url
":"http:\/\/pbs.twimg.com\/profile_images\/764035251047915520\/Jbqja7QE_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile
_images\/764035251047915520\/Jbqja7QE_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/2924593731\/1470995263","defa
ult_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null
,"place":null,"contributors":null,"is_quote_status":false,"quote_count":3,"reply_count":6,"retweet_count":60,"favorite_count":260,"entities":{"
hashtags":[],"urls":[],"user_mentions":[],"symbols":[],"media":[{"id":992817217422151680,"id_str":"992817217422151680","indices":[29,52],"media
_url":"http:\/\/pbs.twimg.com\/media\/Dccx_8NW0AASrJb.jpg","media_url_https":"https:\/\/pbs.twimg.com\/media\/Dccx_8NW0AASrJb.jpg","url":"https
:\/\/t.co\/nIF5HY4XvK","display_url":"pic.twitter.com\/nIF5HY4XvK","expanded_url":"https:\/\/twitter.com\/landpsychology\/status\/9928172199848
75520\/photo\/1","type":"photo","sizes":{"large":{"w":500,"h":667,"resize":"fit"},"thumb":{"w":150,"h":150,"resize":"crop"},"medium":{"w":500,"
h":667,"resize":"fit"},"small":{"w":500,"h":667,"resize":"fit"}}}]},"extended_entities":{"media":[{"id":992817217422151680,"id_str":"9928172174
22151680","indices":[29,52],"media_url":"http:\/\/pbs.twimg.com\/media\/Dccx_8NW0AASrJb.jpg","media_url_https":"https:\/\/pbs.twimg.com\/media\
```

Implementación de la evaluación de Sentimiento

- Asignamos un sentimiento en base a las palabras del texto del tweet.
- Cada palabra recibirá un score y el sentimiento es la suma de los scores de todas las palabras del tweet.
- Diccionario en inglés ⇨ Restricción a tweets en INGLÉS.
- El score es **cero** si:
  - La palabra no está en el diccionario
  - Tweet en otro idioma (diccionario en inglés y codificaciones e.g. Europa del Este, Asia).
  - Tweet no tiene texto (sucio)
  - El campo del lenguaje está vacío (sucio)

AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011. The file is tab-separated. Used version AFINN-111: Newest version with 2477 words and phrases.

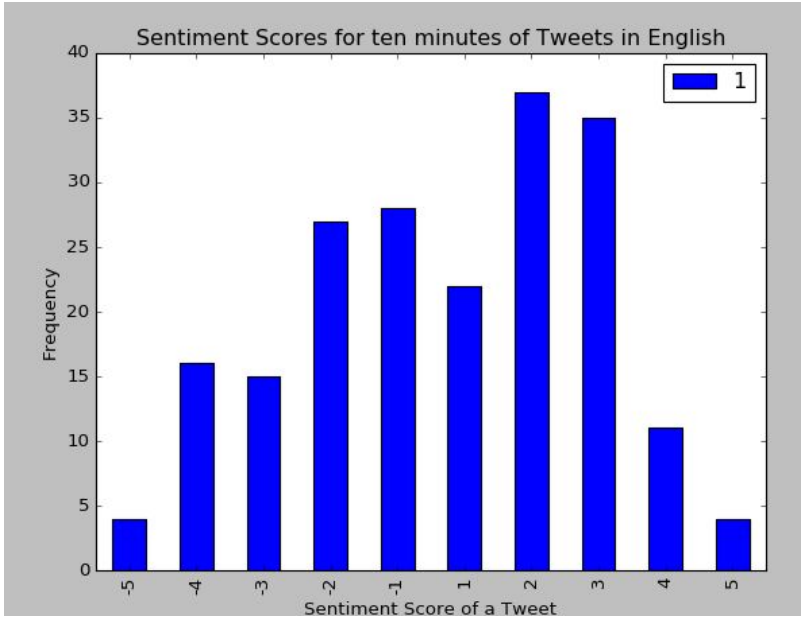
Output Stream

- Dos opciones: 0,-4,0,0,0,0,4,0,0,1,0,0,2,0,0,0,-3,0,0,5,0,0..... o bien....

('Not English', u'in', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @lilyachty: Somebody make me into a fortnite character', 'unknown', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @nganguka: Calling The Attention of all MayWard Flyers, Please Proceed to the Twitter, We Have Our Trending Party Today! \n\nLeggo, Fam!\u2026', 'unknown', 'sentiment=', 0)
('Not English', u'pt', 'sentiment=', 0)
('Not English', u'ko', 'sentiment=', 0)
('Not English', u'th', 'sentiment=', 0)
('Not English', u'pt', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @jonnysun: whenever donald glover does anything i am filled with an immense and deep-seated creative anxiety that does not go away for a\u2026', 'unknown', 'sentiment=', 0)
('Not English', u'in', 'sentiment=', 0)
('Not English', u'es', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @NoHoesGeorge: gamer girls https://t.co/XeJ3GTUJQu', 'unknown', 'sentiment=', 0)
('Not English', u'tr', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @sttepodcast: RT & Follow us to #win a @SignatureEntertainment 5-film romance movie DVD bundle to celebrate the release of #MeghanMarkle\u2026', 'unknown', 'sentiment=', 5)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @Sanemavcil: LumiWatch: On-Arm Projected Graphics and Touch Input by Robert Xiao\nOriginal Video: https://t.co/y3WlnmYzkF\n\n#Technology #A\u2026', 'unknown', 'sentiment=', 0)
('Not English', u'ko', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u"RT @_ShamGod: There is a whole generation of people who don't remember when Snoop was on trial for murder. What a world. https://t.co/Fj1Lf\u2026", 'unknown', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'@elonmusk Hello Elon... when You will bring tesla to India??', 'unknown', 'sentiment=', 0)
('Not English', u'ru', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @Penguincallme: Yes bitch I\u2019ve been trying to reach you for days wtf https://t.co/HcUXOKGTEl', 'unknown', 'sentiment=', -5)
('Not English', u'und', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'I spoke too soon. This one wins. https://t.co/6yOWXMwPid', 'unknown', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'corsets black chick dick porn teen white reggaeton sex videos jessica biel easy virtue sex scene tattoo https://t.co/rLFylD4Duk', 'unknown', 'sentiment=', -3)
HAS NO TEXT,sentiment=0
HAS NO TEXT,sentiment=0
('Not English', u'ko', 'sentiment=', 0)
('Not English', u'ja', 'sentiment=', 0)

Resultados y Posibilidades

- El análisis de sentimiento por 10 minutos (~150MB) un sábado en la noche fue



- También podemos incluir información sobre la geolocalización

('ENGLISH', u'en', 'loc:', {u'full_name': u'Queens, NY', u'url': u'https://api.twitter.com/1.1/geo/id/00c39537733fa112.json', u'country': u'United States', u'place_type': u'city', u'bounding_box': {u'type': u'Polygon', u'coordinates': [[[-73.962582, 40.541722], [-73.962582, 40.800037], [-73.699793, 40.800037], [-73.699793, 40.541722]]]}, u'country_code': u'US', u'attributes': {}, u'id': u'00c39537733fa112', u'name': u'Queens'}, 'text:', u'@ZarekValentin Hard fought win tonight!!! Thanks for leaving it all on the pitch!!! \U0001f49a\U0001f49b\U0001f49a\U0001f49b #PTFC #RCTID #lovemyteam @MLS @TimbersFC', 'unknown', 'sentiment=', 4)

Topic Modeling

- Implementación de un código que mina temas en el dataset “all the news”
- articles1.csv, ~50k líneas: Artículos publicados en diversos medios de comunicación de habla inglesa.
    - Dataset->Corpus: colección de documentos
- Latent Dirichlet Allocation (LDA): modelo probabilístico para inferir una distribución de temas
    - Es un tipo de clustering: documentos->temas, temas->palabras
    - Objetivo: Dada una bolsa de palabras (documento), determina los temas presentes.
    - Un documento es una distribución de temas
    - Un tema es una distribución de palabras que pertenecen a un vocabulario

Aprendizaje en LDA

- Escogemos un K número de temas (clusters), N documentos, M vocabulario
- Asignamos palabras a temas en <u>forma aleatoria</u>, construimos:

	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

	K1	K2	K3	K
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
Dn	1	0	1	0

	W1	W2	W3	Wm
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

- Para mejorar las distribuciones (i.e. inferir o aprender de los datos/documentos), iteramos:
- Para cada W en doc:
- Tema 1 a K:
  - ¿Cuantas palabras en el doc ya pertenecen al tema 1? P1(tema 1| doc d)
  - ¿Con qué frecuencia W aparece en el tema 1 en <u>todos</u> los docs? P2(word W| tema i)
  - P1*P2 = Probabilidad de que W vino del tema 1
  - Si (P1*P2)_tema 2 > (P1*P2)_tema 1 ----> Cambio a tema 2
  - Continua hasta iteraciones deseadas o estado estacionario

Implementación

- Este código es relativamente sencillo, K=10 temas,
  - Un solo pase, aprox 2 horas para 50k líneas de articles1.csv

- Corpus (articles1.csv) es bajado a mano

- Selección de las columnas a usar (headers)

- Limpieza del corpus
  - líneas defectuosas,
  - stop_words i.e. palabras irrelevantes e.g. conjunciones: “of”, “or”, etc.

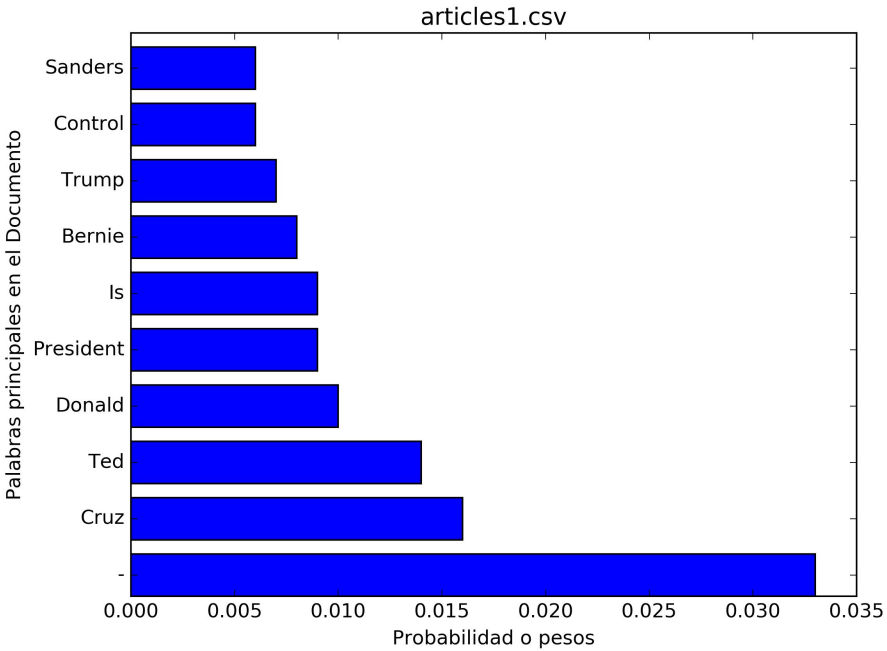
- One-Shot encoding: Datos categóricos, etiquetas->valores numéricos,

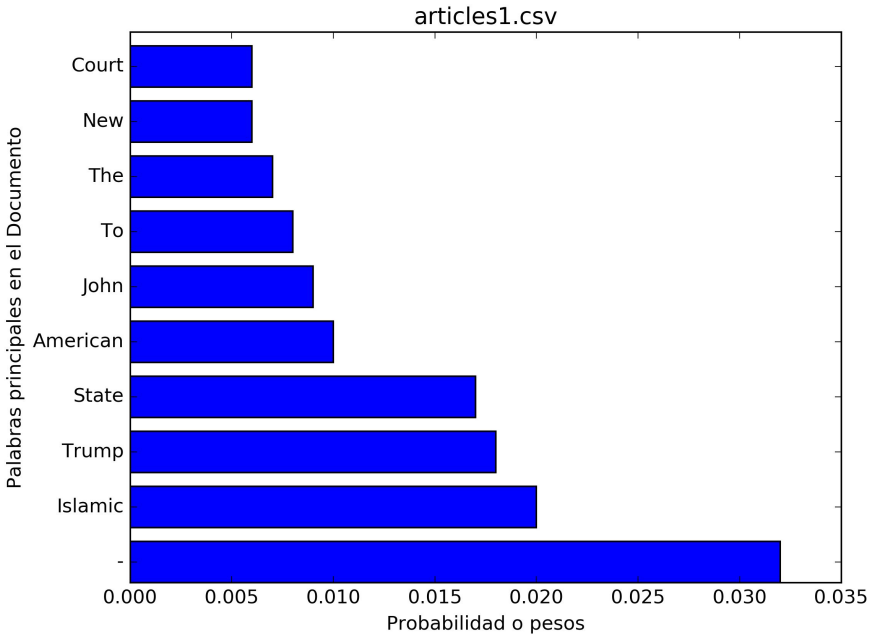
- gensim, *Topic modeling* package para construir el modelo LDA y la matriz documentos-temas.
  - Diccionario: id-strings *mappings* , doc2bow (bag of words) format, etc. Escalable.

Resultados - Temas (topic clusters) inferidos

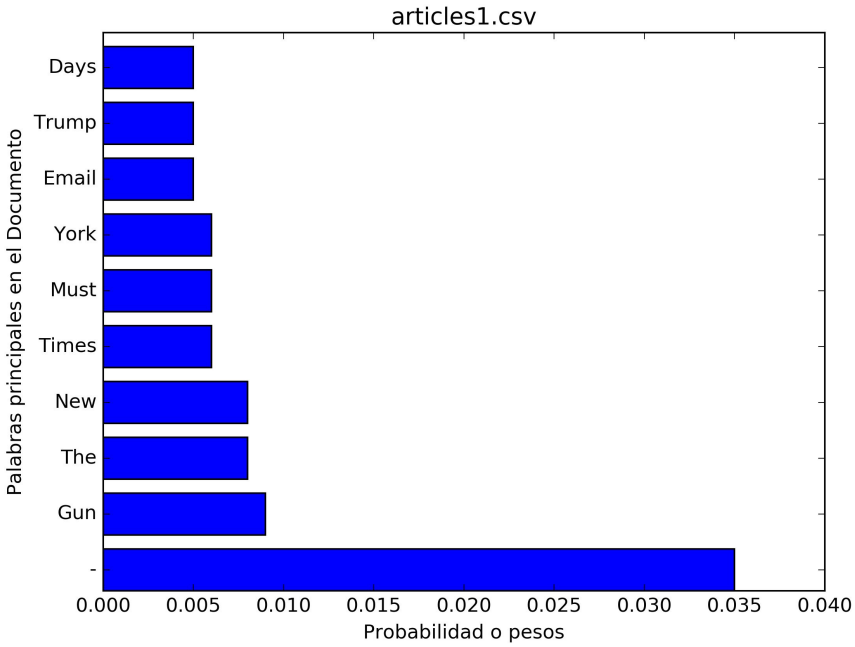
```
benjamin@higgs:~/topic$ python topic_modeling.py
c = 49999 / 49999
0.048*"-" + 0.016*"The" + 0.013*"Donald" + 0.013*"New" + 0.012*"Trump" + 0.010*"York" + 0.010*"Times" + 0.009*"Paul" +
0.006*"Because" + 0.006*"Anti-Trump"
0.034*"-" + 0.020*"Trump" + 0.013*"White" + 0.012*"Texas" + 0.011*"House" + 0.010*"Illegal" + 0.010*"Border" + 0.007*"The" +
0.007*"New" + 0.007*"After"
0.044*"-" + 0.037*"Hillary" + 0.036*"Clinton" + 0.014*"Trump" + 0.010*"The" + 0.010*"New" + 0.007*"Times" + 0.007*"York" +
0.007*"Donald" + 0.005*"Clinton's"
0.033*"-" + 0.016*"Cruz" + 0.014*"Ted" + 0.010*"Donald" + 0.009*"President" + 0.009*"Is" + 0.008*"Bernie" + 0.007*"Trump" +
0.006*"Control" + 0.006*"Sanders"
0.032*"-" + 0.020*"Islamic" + 0.018*"Trump" + 0.017*"State" + 0.010*"American" + 0.009*"John" + 0.008*"To" + 0.007*"The" +
0.006*"New" + 0.006*"Court"
0.038*"-" + 0.012*"Trump" + 0.010*"The" + 0.009*"New" + 0.007*"Rubio" + 0.007*"York" + 0.007*"Ryan" + 0.006*"Open" +
0.006*"Times" + 0.006*"Is"
0.192*"Breitbart" + 0.102*"-" + 0.021*"Trump" + 0.006*"Is" + 0.005*"Not" + 0.005*"Obama" + 0.005*"Will" + 0.005*"Trump:" +
0.005*"The" + 0.004*"GOP"
0.036*"To" + 0.026*"-" + 0.011*"Man" + 0.010*"Migrant" + 0.007*"Terror" + 0.007*"The" + 0.006*"New" + 0.006*"Following" +
0.006*"Mexican" + 0.006*"Police"
0.031*"-" + 0.014*"We" + 0.013*"Watch:" + 0.008*"After" + 0.007*"The" + 0.006*"New" + 0.006*"WATCH:" + 0.005*"Attack" +
0.005*"Two" + 0.005*"Terrorist"
0.035*"-" + 0.009*"Gun" + 0.008*"The" + 0.008*"New" + 0.006*"Times" + 0.006*"Must" + 0.006*"York" + 0.005*"Email" +
0.005*"Trump" + 0.005*"Days"
```

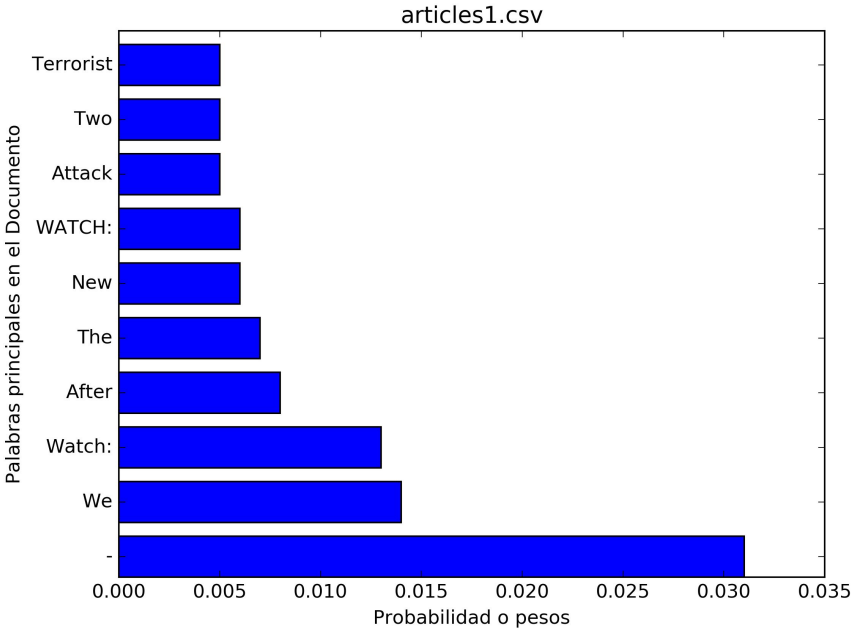
Resultados


articles1.csv
Sanders
Control
Trump
Bernie
Is
President
Donald
Ted
Cruz
-
Palabras principales en el Documento
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
Probabilidad o pesos


articles1.csv
Court
New
The
To
John
American
State
Trump
Islamic
-
Palabras principales en el Documento
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
Probabilidad o pesos

Resultados


articles1.csv
Days
Trump
Email
York
Must
Times
New
The
Gun
-
Palabras principales en el Documento
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
Probabilidad o pesos


articles1.csv
Terrorist
Two
Attack
WATCH:
New
The
After
Watch:
We
-
Palabras principales en el Documento
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
Probabilidad o pesos

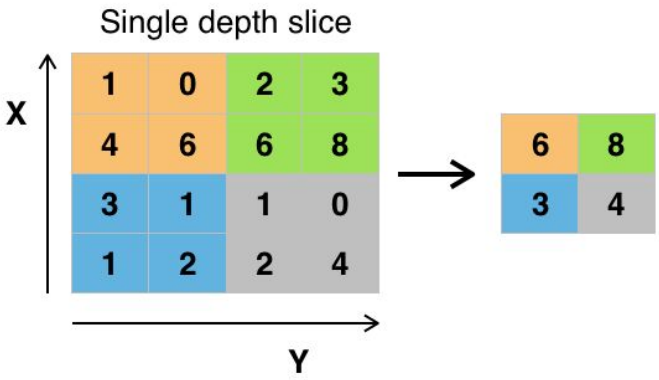
Gracias

Convolución

$$-(y\log(p) + (1-y)\log(1-p))$$

$$-\sum_{c=1}^{M} y_{o,c}\log(p_{o,c})$$

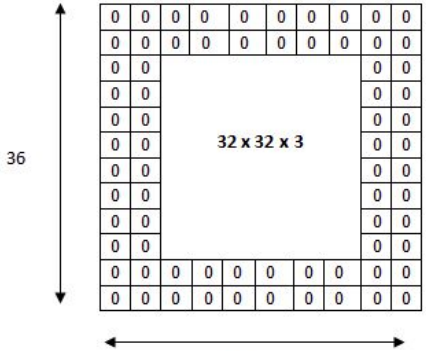
Pooling layers



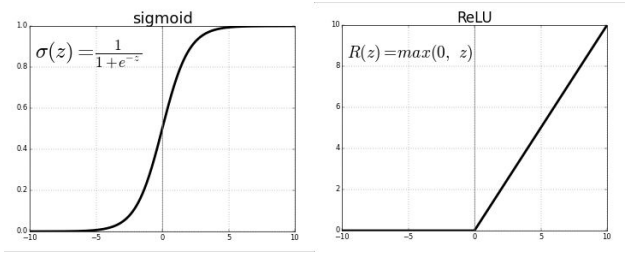
Example of Maxpool with a 2x2 filter and a stride of 2

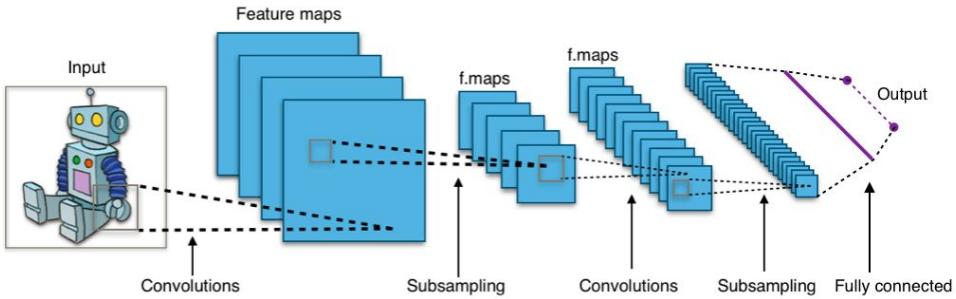
$$z = \frac{x-\mu}{\sigma}$$

Padding



ReLU





hardware

- Este código se ejecutó en Ubuntu 16.04.4 LTS, python3.5.2. En una máquina de 8 cores (Intel(R) Core(TM) i7-2600K CPU @ 3.40GHz) con 32 GB de RAM.