# Introducción al Uso del Texto en Ciencia de Datos

Steven Nichols

@spnichol @spnichol steven@prattle.co

# Propósito de la charla

Tener una mejor idea del proceso de convertir una colección de textos desconocidos en datos accionables

# Propósito de la charla

Tener una mejor idea del proceso de convertir una colección de textos desconocidos en datos accionables, a través de:

1. **Topic Models** (explorar los textos)
2. **Word2Vec/Doc2Vec** (vectoriza y amplifica los textos)
3. **Redes Neuronales** (clasificar los textos)

# Propósito de la charla

El producto final será una red neuronal que puede clasificar documentos según 90 categorías temáticas.

…para darte una idea de la escala del proyecto, y porque vale la pena, consideremos que …

# Propósito de la charla

Para poder clasificar **1.000** documentos manualmente, uno se tarda entre 3 y 4 horas. Clasificamos alrededor de **50.000** al día.

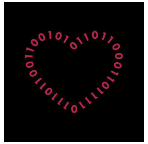
Yo me tardé alrededor de 2 semanas en preparar los datos y entrenar los modelos.

Para poder clasificar los documentos sin AI, necesitaríamos 25 personas trabajando 8 horas al día.

# Yo

Editor de contenido en Booking.com (2012-2015)

Empiezo la maestría en Gestión de Negocios de Tecnología (2015-2017)



Descubro R y Python (2015-)

Dejo de tener vida social (2015-)

Empiezo a trabajar en Prattle Analytics (2017-)

# ¿Qué hace Prattle?

En Prattle aplicamos el análisis de sentimiento a comunicaciones corporativas.

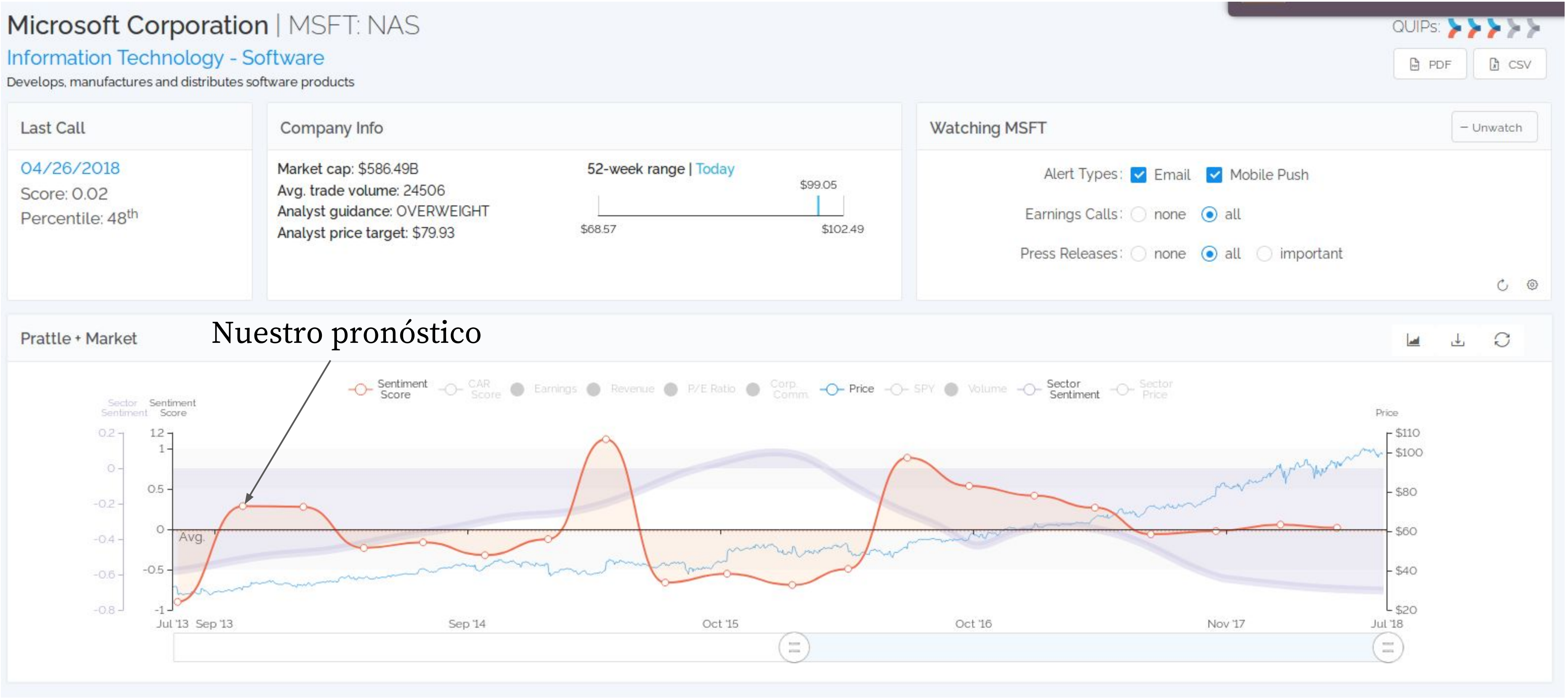
# ¿Qué hace Prattle?

1. Earnings Calls

Llamadas que hacen las empresas públicas en EEUU con los inversionistas cada cuarto.

# ¿Qué hace Prattle?

1. Earnings Calls



2. Comunicados de prensa

# ¿Qué hace Prattle?

Básicamente lo mismo, pero con una variedad más amplia de temas

# ¿Qué hace Prattle?

2. Comunicados de prensa

Básicamente lo mismo, pero con una variedad más amplia de temas

…y muchísimo más *machine learning* y *data engineering* para conseguir y preparar los datos (fun!)

# ¿Por qué más ML?

## Pues, por la calidad de los datos.

### Earnings Calls

Proveedor especializado

Curados por seres humanos

### Comunicados de Prensa

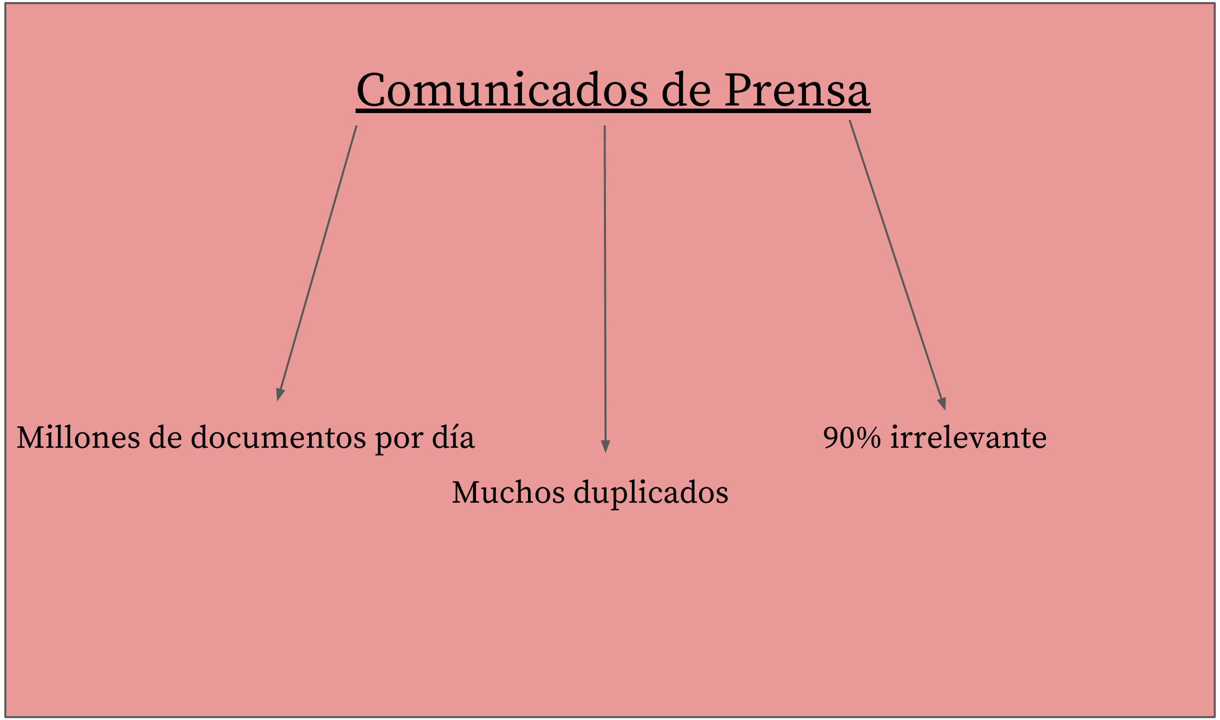
Proveedor general

Nada de datos meta

2. Comunicados de prensa

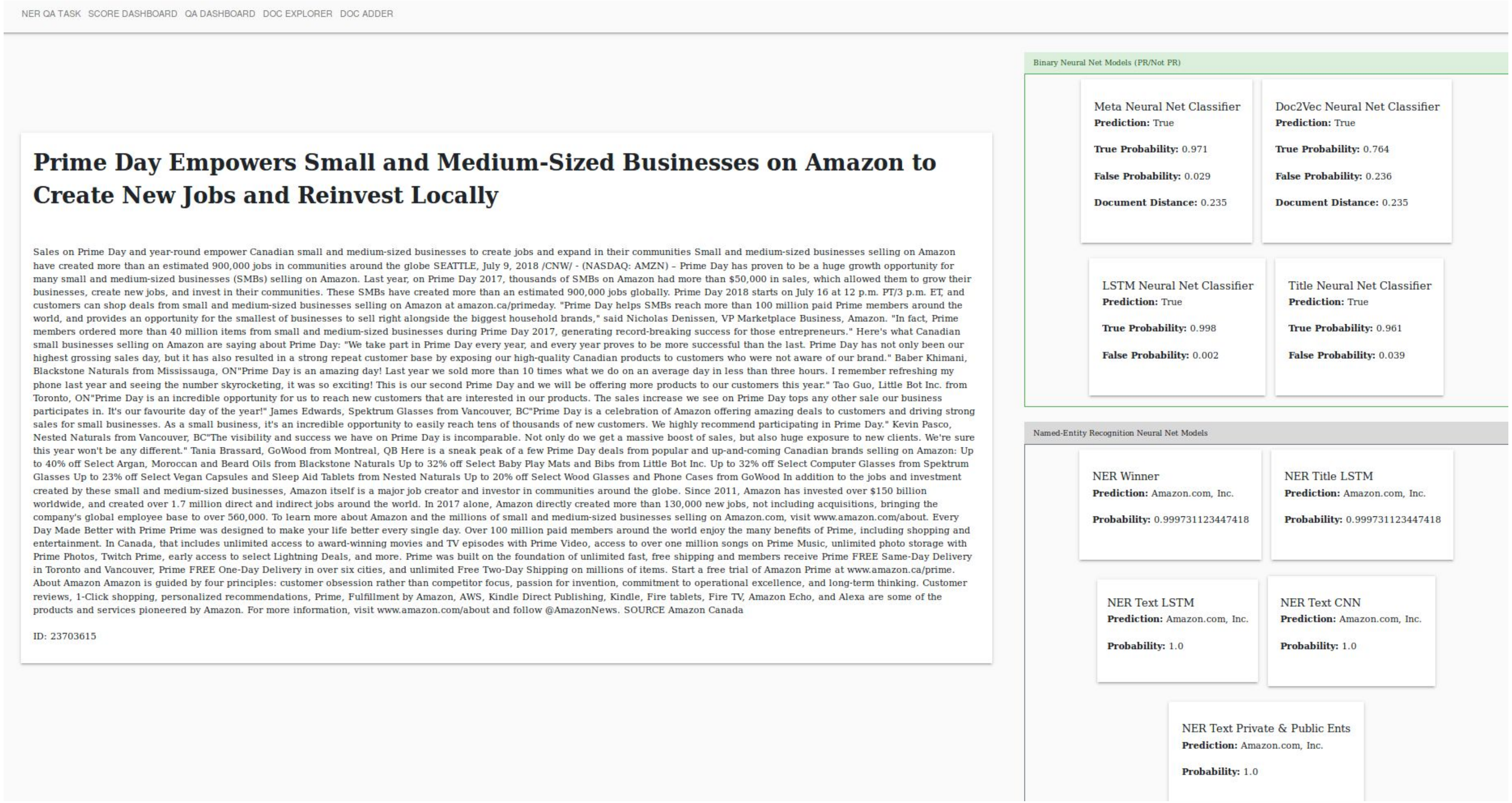
# ¿Por qué más ML?

“Tenemos lo que buscan, pero tendrán que buscarlo ustedes.” -El proveedor



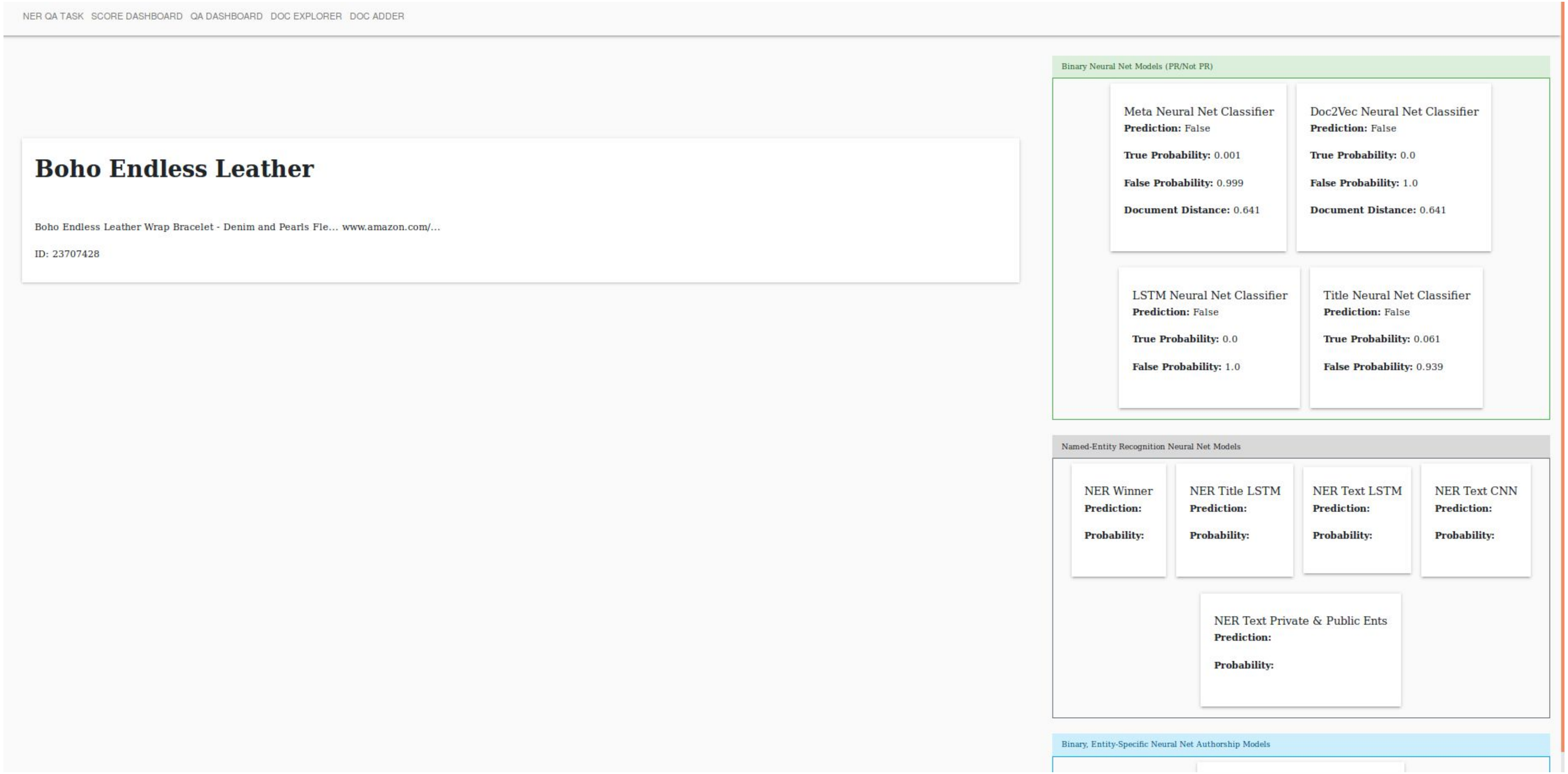
2. Comunicados de prensa

# Lo que queremos…



2. Comunicados de prensa

# Lo que no queremos…



# Nuestras metas

En resumen…

El API del proveedor consiste en un flujo constante de millones de documentos al día, sin datos meta, “escrapeados” de Internet por sus bots…

# Nuestras metas

Usando únicamente el texto de los documentos, queremos:

1. Identificar los comunicados de prensa
2. Establecer el autor/origen del documento
3. **Organizar los CP a través de categorías temáticas**

Tema de hoy

# Los datos

Exploramos los conceptos usando una muestra de CP de Prattle.

**Lo que sí sabemos:** todos los documentos son comunicados de prensa .

**Lo que no sabemos:** ¿de qué se tratan? ¿un nuevo CEO? ¿una posible fusión? ¿la copa mundial?

# Los datos

```
In [1]: import pandas as pd
        import gensim, re
        from gensim import corpora
        import numpy as np
        from nltk.corpus import stopwords
        from operator import itemgetter

In [2]: docs = pd.read_csv('talk.csv')
        docs = docs[docs.text.notnull()]

In [3]: len(docs)

Out[3]: 14964

In [4]: docs.head()
```

Out[4]:

| | id | text | title | doc_vec |
|---|---|---|---|---|
| 0 | 5576620 | T-Mobile US (TMUS) and Sprint (S) announced th... | T-Mobile, Spring enter definitive agreement to... | [-0.45392972 0.28116658 -0.17050362 0.012600... |
| 1 | 16500423 | \n JERSEY CITY and MORRIS TOWNSHIP, N.J., M... | provident financial services, inc. and first m... | [ 0.255116 0.8868336 0.5354027 -0.206942... |
| 2 | 2779308 | LONDON, 12 October 2017 /PRNewswire Policy/ --... | cma provisionally clears just eat / hungryhous... | [-0.20842455 0.59220296 0.4255448 0.085410... |
| 3 | 1086199 | \n\nHOUSTON, Dec. 16, 2015 /PRNewswire/ -- Har... | markwest energy partners to discuss recent mer... | [-4.18535650e-01 1.01070738e+00 1.20303437e-... |
| 4 | 3183287 | DALLAS & NEWTOWN SQUARE, Pa.--(BUSINESS WIRE)... | energy transfer partners, l.p. unitholders app... | [-7.5036830e-01 9.0200227e-01 2.8550895e-02 ... |

# Explorar los datos

Meta: Tener una idea de la distribución de temas dentro de nuestra muestra.

Herramientas: Topic Models (LDA y HDP) y sentido común

# Explorar los datos

## ¿Qué son los topic models?

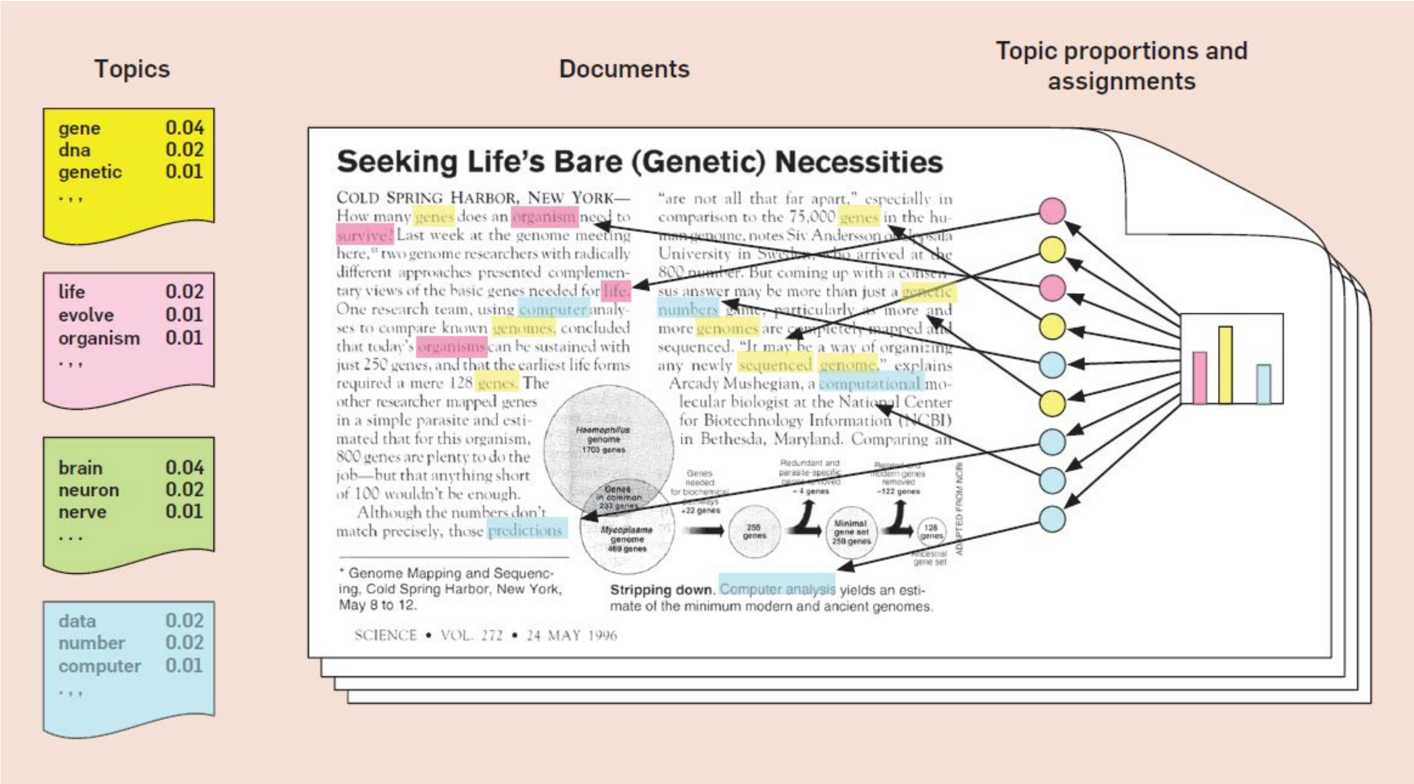
Los topic models, y LDA/HDP en particular, son algoritmos **no supervisados** que buscan:

1. Formar temas (colecciones de términos que ocurren mucho en ciertos documentos)

2. Determinar la composición de temas de cada documento

# Explorar los datos

Por ejemplo…



LDA/HDP

# Explorar los datos

Lo bueno:

- No supervisado
- Implementación en Python con multiprocesamiento (LDA)
- Proporciones de temas a nivel de documento

Lo malo:

- Debes especificar el número de temas (LDA)
- Puede ser un poco lento con corpus más grandes
- No funciona tan bien

# Explorar los datos

```
In [6]: def CleanText(text, split=True):
                text = ' '.join([line.strip() for line in text.strip().splitlines()])
                text = " ".join(re.split("\s+", text, flags=re.UNICODE))
                text = re.sub(r"[^a-zA-Z]", " ", text)
                if not split:
                    return text
                text = text.split(" ")
                stop = set(stopwords.words('english'))
                text = [x.lower() for x in text if x not in stop and len(x) > 1]
                return text

In [19]: dictionary = corpora.Dictionary([CleanText(x) for x in docs.text.tolist()])#crear diccionario. se le asigna un numero de ID unico a cada palabra
         dictionary.filter_extremes(no_above=0.20, no_below=100) #filtramos palabaras que ocurren en mas del 30% de los documentos, ya que es poco probable que esos termino
         dictionary.save('lda_dict.dict') #serializamos el diccionario por si lo queremos usar en otro momento
         print("Dictionary loaded and saved.")

         Dictionary loaded and saved.

In [23]: dictionary.token2id

Out[23]: {'always': 0,
          'america': 1,
          'april': 2,
          'behalf': 3,
          'benefit': 4,
          'both': 5,
          'bring': 6,
          'build': 7,
          'businesses': 8,
          'capabilities': 9,
          'carrier': 10,

In [26]: dictionary.doc2bow(CleanText(docs.iloc[120]['text'])) #podemos usar el diccionario para crear una representacion bag-of-words de cada documento.

Out[26]: [(2, 3),
          (5, 1),
          (11, 1),
          (13, 1),
          (15, 10),
          (19, 1),
          (26, 1),
```

LDA/HDP

# Explorar los datos

In [38]:

```
corpus = [dictionary.doc2bow(CleanText(x)) for x in docs.text.tolist()] #creamos el corpus. es una lista de listas, cada una s
this_hdp = gensim.models.HdpModel(corpus, id2word=dictionary)
this_hdp.print_topics(num_topics=100)
```

Out[38]:

```
[(0,
  '0.011*million + 0.009*net + 0.008*quarter + 0.008*income + 0.006*year + 0.006*statements + 0.005*cash + 0.005*financial + 0
005*per + 0.005*share'),
 (1,
  '0.004*services + 0.003*data + 0.003*technology + 0.003*solutions + 0.003*global + 0.003*management + 0.003*world + 0.002*or
+ 0.002*cloud + 0.002*products'),
 (2,
  '0.034*shares + 0.024*stock + 0.021*quarter + 0.021*rating + 0.015*th + 0.011*research + 0.011*price + 0.009*ratio + 0.009*b
y + 0.008*last'),
 (3,
  '0.008*million + 0.007*income + 0.007*net + 0.007*quarter + 0.006*operating + 0.005*cash + 0.005*gaap + 0.005*share + 0.005*
tatements + 0.005*per'),
 (4,
  '0.015*fitch + 0.014*ratings + 0.013*of + 0.012*fitchratings + 0.012*and + 0.010*this + 0.009*available + 0.008*are + 0.007*
eport + 0.006*on'),
 (5,
  '0.003*million + 0.003*technology + 0.003*store + 0.003*channel + 0.002*quarter + 0.002*solutions + 0.002*cloud + 0.002*mana
ement + 0.002*year + 0.002*services'),
 (6,
  '0.003*health + 0.002*schizophrenia + 0.002*statements + 0.002*forward + 0.002*looking + 0.002*products + 0.001*solutions +
.001*we + 0.001*commission + 0.001*sales'),
 (7,
  '0.005*rigrodsky + 0.004*long + 0.002*shares + 0.002*shareholders + 0.002*legal + 0.002*share + 0.002*stock + 0.002*wilmingt
n + 0.002*health + 0.002*common'),
 (8,
  '0.009*goldberg + 0.008*law + 0.007*class + 0.006*pc + 0.006*goldberglawpc + 0.003*rights + 0.003*los + 0.003*if + 0.003*inf
+ 0.003*contact'),
 (9,
  '0.005*million + 0.005*quarter + 0.004*cash + 0.003*net + 0.003*entrance + 0.003*second + 0.003*operating + 0.003*fee + 0.00
*income + 0.002*loss'),
 (10,
  '0.004*document + 0.003*analysts + 0.003*review + 0.003*notes + 0.002*analyst + 0.002*analystsreview + 0.002*free + 0.002*av
ilable + 0.002*report + 0.001*financial'),
 (11,
  '0.003*software + 0.003*cloud + 0.002*data + 0.002*market + 0.002*solutions + 0.002*management + 0.002*customers + 0.002*app
```

# Explorar los datos

In [52]:
```
test_topics = docs.sample(500)
test_corpus = [dictionary.doc2bow(CleanText(x)) for x in test_topics.text.tolist()]
tops = [this_lda[x] for x in test_corpus]
tops = [max(x,key=itemgetter(1))[0] for x in tops]
test_topics['largest_topic'] = tops
```

In [54]:
```
test_topics[['title', 'largest_topic']].sort_values('largest_topic')
```

Out[54]:

| | title | largest_topic |
|---|---|---|
| 17409 | genco shipping & trading limited to participat... | 0 |
| 2604 | achaogen to present at 14th annual needham hea... | 0 |
| 8631 | biolase reports 2015 fourth quarter and year-e... | 0 |
| 16990 | new agreement, financial results, senior level... | 0 |
| 16769 | Inter Parfums, Inc. Reports 2015 First Quarter... | 0 |
| 8593 | canadian natural resources limited reports vot... | 0 |
| 1418 | sally beauty holdings, inc. reports fourth qua... | 0 |
| 4189 | Alexander & Baldwin Reports Fourth Quarter And... | 0 |
| 10642 | Cara Therapeutics Reports First Quarter 2017 F... | 0 |
| 3218 | silver creek capital management adds maine pub... | 0 |
| 9246 | the cooper companies reports first quarter res... | 0 |
| 8879 | leap therapeutics reports third quarter 2017 f... | 0 |
| 11305 | Mohawk Industries, Inc. Announces Fourth Quart... | 0 |
| 2293 | enbridge income fund holdings inc. to hold ann... | 0 |
| 4969 | Progress Software Reports 2013 Fiscal First Qu... | 0 |
| 13084 | Markel Estimates Third Quarter Catastrophe Losses | 0 |
| 2899 | J & J Snack Foods Reports First Quarter Sales ... | 0 |
| 19055 | incontact reports fourth quarter and year end... | 0 |

LDA/HDP

# Explorar los datos

```
In [56]: test_topics[test_topics.largest_topic == 2][['title', 'largest_topic']].sort_values('largest_topic')
```

Out[56]:

| | title | largest_topic |
|---|---|---|
| 10100 | Cove Street Capital LLC Decreases Holdings in ... | 2 |
| 19477 | lci industries (lcii) holdings cut by envestne... | 2 |
| 3411 | American International Group Inc. Cuts Holding... | 2 |
| 11169 | analyst at citigroup maintained yelp inc (nyse... | 2 |
| 18200 | ubs group analysts give general electric (ge) ... | 2 |
| 17295 | Fisher Asset Management LLC Has $4.37 Million ... | 2 |
| 14100 | recent research analysts' ratings updates for ... | 2 |
| 8080 | fibrogen (fgen) stock rating reaffirmed by mizuho | 2 |
| 18275 | quadrature capital ltd raises holdings in palo... | 2 |
| 13894 | Urban Outfitters Inc (URBN) Files 10-K for the... | 2 |
| 13193 | JPMorgan Chase Declares Preferred Stock Dividend | 2 |
| 15118 | lindsay corporation (lnn) analysts see $1.39 eps | 2 |
| 14961 | baker hughes investor alert by the former atto... | 2 |
| 9503 | westpac banking has raised by $4.87 billion it... | 2 |
| 2761 | Webster Bank (WBS) Holdings Trimmed by Schrode... | 2 |
| 8570 | here's how analysts see virtu financial, inc. ... | 2 |
| 2131 | hudson technologies (hdsn) stock rating upgrad... | 2 |
| 18920 | brokerages set posco (pkx) price target at $98.00 | 2 |
| 6080 | columbus hill capital management l.p. grows ho... | 2 |

LDA/HDP

# Explorar los datos

```
In [57]: test_topics[test_topics.largest_topic == 3][['title', 'largest_topic']].sort_values('largest_topic')
```

Out[57]:

| | title | largest_topic |
|---|---|---|
| 7711 | pilgrim's pride shareholder alert by former lo... | 3 |
| 821 | shareholder alert: levi & korsinsky, llp remin... | 3 |
| 9716 | shareholder alert: levi & korsinsky, llp annou... | 3 |
| 7101 | equity alert: rosen law firm announces filing ... | 3 |
| 9301 | shareholer alert: pomerantz law reminds shareh... | 3 |
| 3308 | shareholder alert: investigation of emergent b... | 3 |
| 7178 | investor alert: levi & korsinsky, llp announce... | 3 |
| 13015 | wow internet, cable & phone to host first quar... | 3 |
| 4863 | shareholder alert: brower piven encourages inv... | 3 |
| 3494 | deadline alert: rigrodsky & long, p.a. reminds... | 3 |
| 19077 | vocera communications inc. investor alert: sco... | 3 |
| 16809 | Orexigen Therapeutics to Host Full Year and Fo... | 3 |
| 1710 | tempur sealy to present at financial conference | 3 |
| 11577 | SunOpta Inc Schedules First Quarter 2018 Finan... | 3 |
| 984 | Antero Midstream Reports Second Quarter 2016 F... | 3 |
| 9683 | pittsburgh law office of alfred g. yates jr., ... | 3 |
| 5555 | raam global energy company announces date of a... | 3 |
| 14062 | shareholder alert: the law offices of vincent ... | 3 |
| 16946 | ryan & maniskas, llp announces class action la... | 3 |
| 15853 | shareholder alert: levi & korsinsky notifies i... | 3 |
| 5448 | shareholder alert: law firm of levi & korsinsk... | 3 |

# Explorar los datos

Los *topic models* nos dan una buena idea de los diferentes temas, pero funcionan 2,3.

…vamos a hacer *un poquito* de trabajo manual para refinar los datos antes de proceder (perfect for interns).

# Explorar los datos

## Buscando cadenas

```
In [81]: docs['title'] = docs.title.str.lower()
         top_0 = docs[(docs.title.str.contains('reports') & (docs.title.str.contains('quarter')))][['title']]
         print(len(top_0))
         top_0
```

```
1407
```

Out[81]:

| | title |
|---|---|
| 20 | depomed reports fourth quarter and year-end 20... |
| 44 | global self storage reports third quarter and ... |
| 47 | netsol technologies reports fiscal 2014 first ... |
| 73 | excellon reports 2015 annual and fourth quarte... |
| 75 | celadon group reports second fiscal quarter fi... |
| 76 | hanwei energy services reports third quarter f... |
| 86 | imax corporation reports first-quarter 2018 re... |
| 149 | caleres reports second quarter 2017 results |
| 151 | schnitzer steel reports second quarter 2010 fi... |
| 154 | recro pharma reports first quarter 2017 financ... |
| 158 | johnson & johnson reports 2013 second-quarter ... |
| 166 | stepan reports first quarter earnings |
| 210 | identiv reports second quarter 2017 results |
| 212 | geron corporation reports fourth quarter and a... |
| 243 | mge energy reports first-quarter earnings |
| 247 | proassurance reports results for first quarter... |
| 270 | tegna inc. reports 2017 fourth quarter and ful... |
| 279 | smic reports 2018 first quarter results |
| 299 | strattec security corporation reports fiscal 2... |

## Buscando cadenas

# Explorar los datos

```
In [92]: top_1 = docs[(docs.title.str.contains('shareholder')) & (docs.title.str.contains('alert'))][['title']]
         print(len(top_1))
         top_1

692
```

Out[92]:

| | title |
|---|---|
| 40 | connectone shareholder alert: faruqi & faruqi,... |
| 64 | shareholder alert: levi & korsinsky, llp remin... |
| 72 | shareholder alert: the law firm of levi & kors... |
| 93 | rh shareholder alert by former louisiana attor... |
| 125 | fitbit (fit) shareholder alert: johnson & weav... |
| 126 | shareholder alert: bronstein, gewirtz & grossm... |
| 153 | important shareholder alert: khang & khang llp... |
| 209 | shareholder alert: levi & korsinsky, llp annou... |
| 219 | shareholder alert: faruqi & faruqi, llp encour... |
| 224 | shareholder alert: levi & korsinsky, llp notif... |
| 252 | shareholder alert: levi & korsinsky, llp annou... |
| 295 | lululemon athletica shareholder alert: levi & ... |
| 307 | shareholder alert: levi & korsinsky, llp notif... |
| 328 | shareholder alert: levi & korsinsky, llp annou... |
| 391 | shareholder alert: the law firm of levi & kors... |
| 400 | shareholder alert: brower piven commences an i... |
| 409 | shareholder alert: goldberg law pc announces s... |
| 427 | shareholder alert: spector, roseman & kodroff,... |
| 449 | barrick gold shareholder alert by former louis... |
| 491 | susquehanna bancshares, inc. shareholder alert... |
| 496 | shareholder alert: levi & korsinsky, llp remin... |

# Explorar los datos

Ya tenemos una forma de juntar datos para entrenar nuestra red neuronal.

…ahora veremos una forma eficiente y divertida de vectorizar nuestros textos: **Doc2Vec.**

# ¿por qué Doc2Vec?

# Doc2Vec

**¡Dimensionalidad!**

- La forma más "común" de vectorizar el texto es a través de las matrices dispersas.
- Esto se puede salir de control muy rápido, sobre todo con corpus más grandes.
- Con Doc2Vec puedes representar tus documentos con un vector de 50, 100, 200, etc.

## Document/Term Frequency Matrix 1

RAW COUNTS:
the actual number of times the term appears in each document.

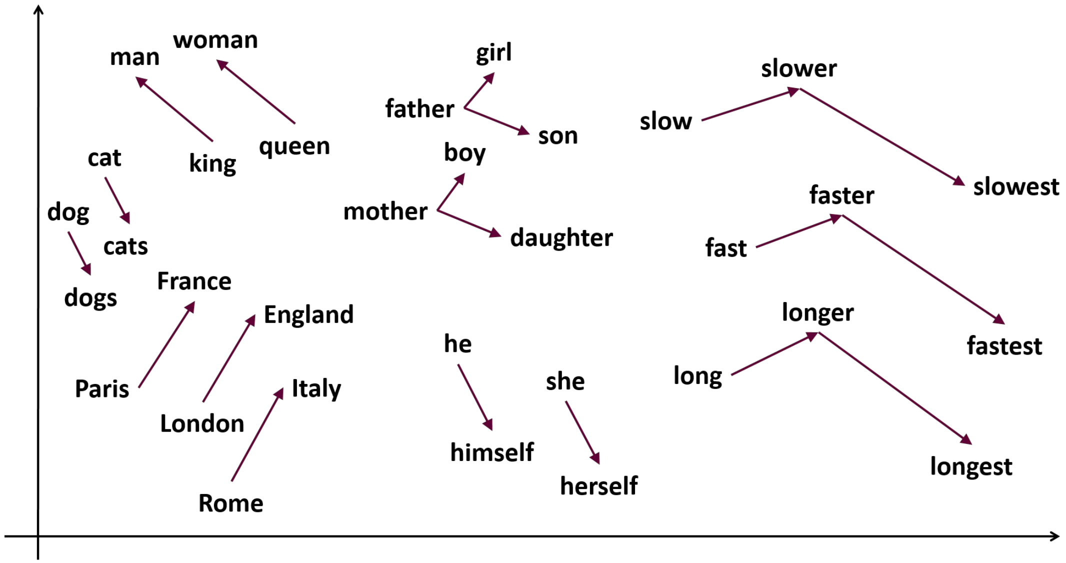
| | Term 1 | Term 2 | Term 3 | Term 4 | Term 5 | Term 6 | Term *n* |
|---|---|---|---|---|---|---|---|
| Doc 1 | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| Doc 2 | 2 | 0 | 9 | 8 | 7 | 3 | 1 |
| Doc 3 | 49 | 39 | 28 | 73 | 64 | 100 | 92 |
| Doc 4 | 0 | 0 | 1938 | 27362 | 2737 | 1162 | 283 |
| Doc 5 | | And so on... | | | | | |
| Doc 6 | | | | | | | |
| Doc *n* | | | | | | | |

# ¿qué es Doc2Vec?

# Doc2Vec

Así como Word2Vec, Doc2Vec es un modelo de ML no supervisado que utiliza coincidencias de palabras para codificar un rango muy amplio de información semántica en el espacio vectorial.

A diferencia de Word2Vec, Doc2Vec toma en cuenta el contexto del documento durante el entrenamiento y puede generar vectores para representar oraciones o documentos enteros.



¿qué es Doc2Vec?

# Doc2Vec

Puedes entrenar tu propio modelo o usar un modelo pre-entrenado.

… para nuestra tarea, usaremos un modelo que entrené con más de 5 millones de comunicados de prensa y notas periodísticas.

# Por ejemplo...

# Doc2Vec

Tomamos un documento del Topic 3 que salió del *Topic Model*

```
In [96]: top_1.iloc[0]['title']

Out[96]: 'connectone shareholder alert: faruqi & faruqi, llp announces the investigation of connectone bancorp, inc. (cnob) over the merger of the company with center bancor
p, inc. (cnbc)'
```

Usamos nuestro modelo Doc2Vec para inferir el vector del documento.

```
In [9]: inferred_vector = ngram_d2v.infer_vector(fourgram[trigram[bigram[CleanText(doc)]]])
```

Asi se ve - un vector 200D de floats (¿números reales?)

```
In [20]: inferred_vector

Out[20]: array([-0.18396847,  0.31871045,  0.18566443,  0.3900829 , -0.2984329 ,
                 0.2179191 , -0.16856088,  0.63058877, -0.185305  ,  0.12956811,
                 0.528154  ,  0.22385058,  0.75957346,  0.03451149, -0.2707205 ,
                -0.43803596, -0.15663587, -0.02322973, -0.2679045 , -0.49819857,
                -0.14879376,  0.3486168 , -0.66976   ,  0.57864946,  0.34423536,
                -0.01113601,  0.5252365 ,  0.6374084 , -0.09568592, -0.2904873 ,
                 0.14951782,  0.646801  ,  0.56326777, -0.44480854,  0.23417753,
                -0.00241793,  0.08879661, -0.3140996 , -0.06733612, -0.12067023,
                -0.05988081, -0.14663288, -0.06246691, -0.14932702, -0.16626866,
                 0.1508719 ,  0.25913426, -0.47160316,  0.2818549 ,  0.20241715,
                 0.62744844,  1.194507  , -0.18959107,  0.13607477,  0.31529155,
                -0.36656785, -0.06796684,  1.0985713 , -0.16100188, -0.6020609 ,
                 0.24390087, -0.17434259,  0.00199472, -0.82270247,  0.915137  ,
                 0.3695208 , -0.46176144, -0.3589787 ,  0.04568369,  0.44371253,
                 0.00720044,  0.4091336 ,  0.21035066, -0.64135265, -0.04186049,
                 0.03758976,  0.7526915 , -0.15085205,  0.04828598,  0.14110236,
                 0.9865262 ,  0.17184614,  0.71646893,  1.0186096 ,  1.2800798 ,
                -0.30674213, -0.7366955 ,  0.5557243 ,  0.02732648, -0.24632677,
                -0.14757127,  0.62456095,  0.33776855,  0.05174606, -0.7732837 ,
                -0.26145813, -0.53264654, -0.11274428,  1.0134214 , -0.483543  ,
                 0.83906144, -1.0713114 ,  0.02970999,  1.0606651 , -0.40898696,
                -0.03713605,  0.39203703,  0.6514021 , -0.47696456,  0.29080063,
                -0.01121997,  0.6043745 , -0.05571247,  0.28092343, -0.23942436,
```

Por cierto …

# Doc2Vec

Aparte de usarlo como input a tu RN, puedes hacer otras cosas padres con los vectores

…como encontrar documentos parecidos comparando los vectores de tu corpus calculando similitud coseno.

In [15]:
```
sims = ngram_d2v.docvecs.most_similar([inferred_vector], topn=100)

sim_list = list()
sim_amnt = list()
for sim in sims[0:100]:
    sim_list.append(sim[0].split('_')[1])
    sim_amnt.append(sim[1])

res = pd.DataFrame(Session().query(PressRelease.id, PressRelease.factset_entity_id, PressRelease.title, PressReleaseData.text)
                   .join(PressReleaseData, PressReleaseData.pr_id == PressRelease.id)
                   .filter(PressRelease.id.in_(sim_list)).all())
res['doc_sim'] = sim_amnt
```

In [18]:
```
res[['title', 'doc_sim']]
```

Out[18]:

| | title | doc_sim |
|---|---|---|
| 0 | Faruqi & Faruqi, LLP Launches An Investigation Against ACADIA Pharmaceuticals Inc. (ACAD) For Po... | 0.668067 |
| 1 | Faruqi & Faruqi, LLP is Seeking More Cash for the Shareholders of Aerosonic Corporation (AIM) | 0.665711 |
| 2 | Faruqi & Faruqi, LLP is Seeking More Cash for the Shareholders of BBX Capital Corporation (BBX) | 0.664521 |
| 3 | Faruqi & Faruqi, LLP, Partner Juan E. Monteverde Launches an Investigation of Omthera Pharmaceut... | 0.646003 |
| 4 | Faruqi & Faruqi, LLP, Partner Juan E. Monteverde Launches an Investigation of StellarOne Corpora... | 0.639375 |
| 5 | VIROPHARMA INVESTOR ALERT: Faruqi & Faruqi, LLP Is Investigating the ViroPharma Inc. Board of Di... | 0.637875 |
| 6 | KKR FINANCIAL HOLDINGS INVESTOR ALERT: Faruqi & Faruqi, LLP Announces the Investigation of KKR F... | 0.636365 |
| 7 | Faruqi & Faruqi, LLP Launches An Investigation Against Lannett Company, Inc. (LCI) For Potential... | 0.634125 |
| 8 | Faruqi & Faruqi, LLP Launches An Investigation Against Winmark Corp. (WINA) For Potential Breach... | 0.633485 |
| 9 | Faruqi & Faruqi, LLP Launches An Investigation Against Peapack-Gladstone Financial Corp. (PGC) F... | 0.632067 |
| 10 | Faruqi & Faruqi, LLP Launches An Investigation Against Exactech Inc. (EXAC) Potential Breaches O... | 0.631228 |
| 11 | ADVENT SOFTWARE, INC. INVESTOR ALERT: Faruqi & Faruqi, LLP Announces the Investigation of Advent... | 0.623869 |
| 12 | MICROFINANCIAL, INC. INVESTOR ALERT: Faruqi & Faruqi, LLP Announces the Investigation of MicroFi... | 0.622484 |
| 13 | MWI VETERINARY SUPPLY, INC. INVESTOR ALERT: Faruqi & Faruqi, LLP Announces the Investigation of ... | 0.622413 |

# Redes Neuronales

Crear una red neuronal para categorizar los documentos.

...no vamos a usar las categorías que hicimos con los *Topic Models*.

## ¿por qué?

- En mi experiencia, para que funcione bien tu red, tus categorías deben aproximar la población real. Entre más categorías, más elección, mejor.

- Si solo te interesan un par de temas, a veces es mejor una red con output binario.

Por ejemplo…



Empecé este proyecto con una sola categoría para temas de responsabilidad corporativa -> ‘resp_don’ (donativos)

…pero, la capacidad de las redes neuronales de generalizar es increíble, y lo confundía otros tipos de responsabilidad corporativa. Ya no.

| | | |
|---|---|---|
| resp_disab | 800 | Comany talks about helping people with disabilities. |
| resp_disease | 744 | Company raises awareness or money for a disease. |
| resp_don | 3299 | Company donates money to a charitable cause. |
| resp_env | 1852 | Comapny talks about green-topics or dedication to environment. |
| resp_gen | 32344 | Catch-all for corp responsibility topics. |
| resp_lgbt | 849 | Comapny talks about gay pride or LGBT topics in a favorable light (i.e., not Chick-fil-A) |
| resp_lit | 387 | Comapny talks about helping kids read good. |
| resp_refugee | 403 | Company talks about refugee crises. |
| resp_stem | 1123 | Company talks about STEM issues for kids or students. |
| resp_vet | 752 | Comapny talks about hiring or helping veterans. |
| resp_vol | 2119 | Company forces employees to volunteer for a charitable cause. |
| resp_women | 7533 | Company talks about #MeToo or how they are great place to work for women. |

# Checamos los datos

# Redes Neuronales

```
In [124]: all_tops.groupby('topic').count()
```

Out[124]:

| topic | id | title | text | doc_vec | fs_sector_code | sector | top_cat |
|---|---|---|---|---|---|---|---|
| biz_acq | 4986 | 4986 | 4986 | 4986 | 4986 | 4986 | 4986 |
| biz_acq_majstake | 484 | 484 | 484 | 484 | 484 | 484 | 484 |
| biz_advrt | 3170 | 3170 | 3170 | 3170 | 3170 | 3170 | 3170 |
| biz_bankruptcy | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| biz_benefits | 1602 | 1602 | 1602 | 1602 | 1602 | 1602 | 1602 |
| biz_boast | 1258 | 1258 | 1258 | 1258 | 1258 | 1258 | 1258 |
| biz_bybck | 3759 | 3759 | 3759 | 3759 | 3759 | 3759 | 3759 |
| biz_collab | 38314 | 38314 | 38314 | 38314 | 38314 | 38314 | 38314 |
| biz_collab_jv | 3245 | 3245 | 3245 | 3245 | 3245 | 3245 | 3245 |
| biz_collab_privpub | 151 | 151 | 151 | 151 | 151 | 151 | 151 |
| biz_ctrct | 13208 | 13208 | 13208 | 13208 | 13208 | 13208 | 13208 |
| biz_deb_loan | 8949 | 8949 | 8949 | 8949 | 8949 | 8949 | 8949 |
| biz_debt | 2057 | 2057 | 2057 | 2057 | 2057 | 2057 | 2057 |
| biz_debt_refin | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| biz_debt_repricing | 172 | 172 | 172 | 172 | 172 | 172 | 172 |
| biz_div | 18846 | 18846 | 18846 | 18846 | 18846 | 18846 | 18846 |
| biz_dlst | 604 | 604 | 604 | 604 | 604 | 604 | 604 |
| biz_dwnsize | 895 | 895 | 895 | 895 | 895 | 895 | 895 |
| biz_ern | 102726 | 102726 | 102726 | 102726 | 102726 | 102726 | 102726 |
| biz_ern_sales | 3645 | 3645 | 3645 | 3645 | 3645 | 3645 | 3645 |
| biz_ern_sales_neg | 422 | 422 | 422 | 422 | 422 | 422 | 422 |
| biz_ern_sales_pos | 1671 | 1671 | 1671 | 1671 | 1671 | 1671 | 1671 |
| biz_ern_sched | 12313 | 12313 | 12313 | 12313 | 12313 | 12313 | 12313 |
| biz_expan | 3569 | 3569 | 3569 | 3569 | 3569 | 3569 | 3569 |
| biz_expan_locat | 7896 | 7896 | 7896 | 7896 | 7896 | 7896 | 7896 |
| biz_expan_prop | 2062 | 2062 | 2062 | 2062 | 2062 | 2062 | 2062 |
| biz_guide | 2177 | 2177 | 2177 | 2177 | 2177 | 2177 | 2177 |
| biz_invest | 1737 | 1737 | 1737 | 1737 | 1737 | 1737 | 1737 |

Checamos los datos



In [40]:
```
smaller_df = pd.DataFrame()
for cat in all_tops.top_cat.unique().tolist():
    try:
        this_tmp = all_tops[all_tops.top_cat == cat].sample(1500, replace=True, random_state=i)
        smaller_df = smaller_df.append(this_tmp)
    except Exception as e:
        print(e)
        continue
```

In [125]:
```
smaller_df.groupby('topic').count()
```

Out[125]:

| topic | id | title | text | doc_vec | fs_sector_code | sector | top_cat |
|---|---|---|---|---|---|---|---|
| biz_acq | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_acq_majstake | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_advrt | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_bankruptcy | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_benefits | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_boast | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_bybck | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_collab | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_collab_jv | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_collab_privpub | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_ctrct | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_deb_loan | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_debt | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_debt_refin | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_debt_repricing | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_div | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_dlst | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_dwnsize | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_ern | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_ern_sales | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_ern_sales_neg | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_ern_sales_pos | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_ern_sched | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |
| biz_expan | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |

# Checamos los datos



Preparamos los datos, quitando unas categorías que no sirven

```
In [30]: smaller_df = smaller_df.fillna(-1)
smaller_df['sector'] = smaller_df.fs_sector_code.astype('category')
smaller_df['sector'] = smaller_df['sector'].cat.codes
sect_map = smaller_df.drop_duplicates(subset=['sector'])
sect_map.to_csv('models/{}_top_sector_cat_map.csv'.format(mod_name))
smaller_df = smaller_df[~smaller_df.topic.isin(['biz_pks', 'resp_vet', 'ind_oil_production'])]
smaller_df['top_cat'] = smaller_df.topic.astype('category')
smaller_df['top_cat'] = smaller_df.top_cat.cat.codes
cat_map = smaller_df.drop_duplicates(subset=['top_cat'])
cat_map.to_csv('models/{}_catmap.csv'.format(mod_name))
```

Quitamos documentos con poco contenido

```
In [ ]: smaller_df = smaller_df.drop_duplicates(subset=['id', 'topic'])
smaller_df = smaller_df[smaller_df.text.notnull()]
smaller_df = smaller_df[smaller_df.text.str.len() > 2]
smaller_df = smaller_df[smaller_df.title.notnull()]
smaller_df = smaller_df[smaller_df.title.str.len() > 2]
smaller_df = smaller_df[smaller_df.doc_vec.notnull()]
```

# Checamos los datos



Apartamos los datos test/train y creamos dos inputs:

1. Doc2Vec
2. BOW/Matriz Dispersa

```
In [ ]: max_words = 50000
        batch_size = 128
        max_review_length = 100
        embedding_vector_length = 100
```

```
In [42]: x_train, x_test, y_train, y_test = train_test_split(smaller_df[['text', 'id', 'title', 'sector', 'doc_vec']], smaller_df['top_cat'], test_size=0.10, random_state=2)

test_ids = x_test['id'].tolist()


print("Creating universal tokenizer.")
tokenizer = Tokenizer(nb_words=max_words)
tokenizer.fit_on_texts(x_train.text.tolist())

print("Tokenzing test data.")
y_test = np.asarray(y_test.tolist())
x_test_d2v = x_test.doc_vec.tolist()
x_test =  x_test.text.tolist()

x_test  = tokenizer.texts_to_sequences(x_test) #convert text to sequence of mapped IDs
x_test  = tokenizer.sequences_to_matrix(x_test, mode='count') #convert to matrix
x_test_d2v = np.stack(x_test_d2v, axis=0)


# y_train = y_train['top_cat']
y_train = np.asarray(y_train.tolist())
y_train = np.array(y_train)


print("Starting tokenization.")
x_train_d2v = x_train.doc_vec.tolist()
x_train = x_train.text.tolist()
x_train  = tokenizer.texts_to_sequences(x_train) #convert text to sequence of mapped IDs

print("Starting sequence to matrix process.")
x_train  = tokenizer.sequences_to_matrix(x_train, mode='count') #convert to matrix
x_train_d2v = np.stack(x_train_d2v, axis=0)
```

# Construimos la red

Puro Doc2Vec

```
In [76]: #solamente doc2vec
d2v_in = Input(shape=(200,) )
d2v_layer = Dense(20000, activation='relu')(d2v_in)
d2v_layer = Dropout(.4)(d2v_layer)
d2v_layer = Dense(10000, activation='relu')(d2v_in)
d2v_layer = Dropout(.4)(d2v_layer)
d2v_layer = Dense(5000, activation='relu')(d2v_in)
d2v_layer = Dropout(.4)(d2v_layer)
text_class = Dense(num_classes, activation='sigmoid')(d2v_layer)
d2vmodel = Model(d2v_in, text_class)
```

```
In [77]: parallel_model = multi_gpu_model(d2vmodel, gpus=4)
parallel_model.compile(loss='sparse_categorical_crossentropy',
                       optimizer='adam', metrics=['accuracy'])
res = parallel_model.fit(x_train_d2v, y_train, verbose=1, validation_data=(x_test_d2v, y_test), batch_size=512, epochs=50)
```

# Entrenamos la red

## Puro Doc2Vec

```
Train on 79773 samples, validate on 8864 samples
Epoch 1/50
79773/79773 [==============================] - 3s 42us/step - loss: 2.3817 - acc: 0.4256 - val_loss: 1.9645 - val_acc: 0.5105
Epoch 2/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.8346 - acc: 0.5300 - val_loss: 1.8849 - val_acc: 0.5245
Epoch 3/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.6955 - acc: 0.5537 - val_loss: 1.8388 - val_acc: 0.5274
Epoch 4/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.5800 - acc: 0.5748 - val_loss: 1.7969 - val_acc: 0.5351
Epoch 5/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.4687 - acc: 0.5982 - val_loss: 1.8034 - val_acc: 0.5363
Epoch 6/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.3616 - acc: 0.6227 - val_loss: 1.7929 - val_acc: 0.5405
Epoch 7/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.2613 - acc: 0.6459 - val_loss: 1.7678 - val_acc: 0.5454
Epoch 8/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.1623 - acc: 0.6716 - val_loss: 1.7786 - val_acc: 0.5452
Epoch 9/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.0689 - acc: 0.6952 - val_loss: 1.7885 - val_acc: 0.5488
Epoch 10/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.9818 - acc: 0.7180 - val_loss: 1.8052 - val_acc: 0.5499
Epoch 11/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.8962 - acc: 0.7394 - val_loss: 1.8332 - val_acc: 0.5458
Epoch 12/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.8234 - acc: 0.7607 - val_loss: 1.8339 - val_acc: 0.5493
Epoch 13/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.7553 - acc: 0.7797 - val_loss: 1.8609 - val_acc: 0.5500
Epoch 14/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.6964 - acc: 0.7954 - val_loss: 1.8859 - val_acc: 0.5459
Epoch 15/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.6404 - acc: 0.8118 - val_loss: 1.9232 - val_acc: 0.5485
Epoch 16/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.5940 - acc: 0.8247 - val_loss: 1.9358 - val_acc: 0.5486
Epoch 17/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.5504 - acc: 0.8380 - val_loss: 1.9633 - val_acc: 0.5424
Epoch 18/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.5163 - acc: 0.8480 - val_loss: 1.9883 - val_acc: 0.5439
Epoch 19/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.4844 - acc: 0.8565 - val_loss: 2.0182 - val_acc: 0.5425
Epoch 20/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.4570 - acc: 0.8652 - val_loss: 2.0375 - val_acc: 0.5435
Epoch 21/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.4289 - acc: 0.8727 - val_loss: 2.0656 - val_acc: 0.5487
Epoch 22/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.4086 - acc: 0.8796 - val_loss: 2.0985 - val_acc: 0.5481
Epoch 23/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.3844 - acc: 0.8867 - val_loss: 2.1206 - val_acc: 0.5410
Epoch 24/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.3709 - acc: 0.8900 - val_loss: 2.1628 - val_acc: 0.5416
Epoch 25/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.3564 - acc: 0.8953 - val_loss: 2.1740 - val_acc: 0.5430
Epoch 26/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.3354 - acc: 0.9003 - val_loss: 2.2019 - val_acc: 0.5398
Epoch 27/50
```

# Entrenamos la red

## Puro Doc2Vec

```
Train on 79773 samples, validate on 8864 samples
Epoch 1/50
79773/79773 [==============================] - 3s 42us/step - loss: 2.3817 - acc: 0.4256 - val_loss: 1.9645 - val_acc: 0.5105
Epoch 2/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.8346 - acc: 0.5300 - val_loss: 1.8849 - val_acc: 0.5245
Epoch 3/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.6955 - acc: 0.5537 - val_loss: 1.8388 - val_acc: 0.5274
Epoch 4/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.5800 - acc: 0.5748 - val_loss: 1.7969 - val_acc: 0.5351
Epoch 5/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.4687 - acc: 0.5982 - val_loss: 1.8034 - val_acc: 0.5363
Epoch 6/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.3616 - acc: 0.6227 - val_loss: 1.7929 - val_acc: 0.5405
Epoch 7/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.2613 - acc: 0.6459 - val_loss: 1.7678 - val_acc: 0.5454
Epoch 8/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.1623 - acc: 0.6716 - val_loss: 1.7786 - val_acc: 0.5452
Epoch 9/50
79773/79773 [==============================] - 2s 29us/step - loss: 1.0689 - acc: 0.6952 - val_loss: 1.7885 - val_acc: 0.5488
Epoch 10/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.9818 - acc: 0.7180 - val_loss: 1.8052 - val_acc: 0.5499
Epoch 11/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.8962 - acc: 0.7394 - val_loss: 1.8332 - val_acc: 0.5458
Epoch 12/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.8234 - acc: 0.7607 - val_loss: 1.8339 - val_acc: 0.5493
Epoch 13/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.7553 - acc: 0.7797 - val_loss: 1.8609 - val_acc: 0.5500
Epoch 14/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.6964 - acc: 0.7954 - val_loss: 1.8859 - val_acc: 0.5459
Epoch 15/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.6404 - acc: 0.8118 - val_loss: 1.9232 - val_acc: 0.5485
Epoch 16/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.5940 - acc: 0.8247 - val_loss: 1.9358 - val_acc: 0.5486
Epoch 17/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.5504 - acc: 0.8380 - val_loss: 1.9633 - val_acc: 0.5424
Epoch 18/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.5163 - acc: 0.8480 - val_loss: 1.9883 - val_acc: 0.5439
Epoch 19/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.4844 - acc: 0.8565 - val_loss: 2.0182 - val_acc: 0.5425
Epoch 20/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.4570 - acc: 0.8652 - val_loss: 2.0375 - val_acc: 0.5435
Epoch 21/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.4289 - acc: 0.8727 - val_loss: 2.0656 - val_acc: 0.5487
Epoch 22/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.4086 - acc: 0.8796 - val_loss: 2.0985 - val_acc: 0.5481
Epoch 23/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.3844 - acc: 0.8867 - val_loss: 2.1206 - val_acc: 0.5410
Epoch 24/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.3709 - acc: 0.8900 - val_loss: 2.1628 - val_acc: 0.5416
Epoch 25/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.3564 - acc: 0.8953 - val_loss: 2.1740 - val_acc: 0.5430
Epoch 26/50
79773/79773 [==============================] - 2s 29us/step - loss: 0.3354 - acc: 0.9003 - val_loss: 2.2019 - val_acc: 0.5398
Epoch 27/50
```

Evaluamos la red

# Redes Neuronales

In [93]:
```
fresh_docs[['title', 'topic', 'cat_1', 'cat_2', 'prob_1', 'prob_2']].sort_values('prob_1', ascending=False)
```

Out[93]:

| | title | topic | cat_1 | cat_2 | prob_1 | prob_2 |
|---|---|---|---|---|---|---|
| 7569 | Dicerna Reports Fourth Quarter and Full Year 2016 Financial and Operational Results | biz_ern | biz_guide | biz_ern | 1 | 0.016008 |
| 608861 | marriott helps make it happen for women-owned businesses and celebrates international women's day | resp_women | resp_women | resp_gen | 1 | 0.000234512 |
| 40970 | Mirati Therapeutics Reports Financial Results And Provides Business Update For The Third Quarter... | biz_ern | biz_guide | biz_ern | 1 | 0.999999 |
| 628647 | 8-K - SOHU COM INC (0001104188) (Filer) | lgl_reg_sec | lgl_reg_sec | prod_gen | 1 | 1.14492e-09 |
| 618361 | investor alert: class action lawsuit against 500.com limited announced by law offices of howard ... | lgl_cls_action | lgl_cls_action | lgl_gen | 0.999974 | 0.0619148 |
| 397280 | fuwei films celebrates commencement of trial production line installation | prod_pharm_gen | biz_prod_update | prod_pharm_gen | 0.999409 | 0.0104544 |
| 77268 | DMC Global Schedules Third Quarter Earnings Release and Conference Call | biz_ern | biz_ern_sched | biz_guide | 0.989847 | 0.0191524 |
| 277967 | jcpenney announces successful closing of real estate term loan refinance | biz_deb_loan | biz_deb_loan | biz_debt | 0.913028 | 0.00391874 |
| 296214 | lexmark's perceptive software positioned in leader's quadrant of gartner's magic quadrant for en... | biz_boast | biz_boast | jobs_hire | 0.913018 | 0.0343318 |
| 469459 | osi systems security division selected as winner of best standoff threat detection technology at... | misc_rprt | biz_expan | biz_ctrct | 0.887409 | 0.00164562 |
| 201203 | 2u, inc. chief executive officer & co-founder chip paucek to present at needham's annual interco... | evt_conf | evt_conf | misc_rprt | 0.554742 | 0.156479 |
| 262527 | developing brand identities, trading plans, strategy committees and mergers - research report on... | biz_strg | biz_strg | misc_retire | 0.404529 | 0.001072 |
| 150421 | lincoln financial launches latest "responsibility of love" advertising campaign | prod_gen | biz_advrt | misc_rprt | 0.388404 | 0.0202121 |
| 259574 | tegna board elects jennifer dulski as new director | jobs_brd | jobs_brd | biz_shrholder_action | 0.280753 | 0.00131839 |
| 613380 | gaithersburg marriott washingtonian center serves holiday dinner to homeless women | resp_women | resp_women | resp_disease | 0.15759 | 0.0164838 |
| 453075 | western asset worldwide income fund inc. announces results of annual meeting of stockholders | misc_rprt | biz_shrholder_action | misc_rprt | 0.10522 | 0.0219532 |
| 191400 | oxbridge re holdings to present at the 6th annual helios gateway conference on september 7, 2017 | evt_conf | misc_rprt | resp_gen | 0.095373 | 0.00549327 |
| 762959 | sunshine bancorp, inc. hires veteran banking executive andrew samuel | resp_gen | jobs_exc | jobs_exc_retr | 0.0259146 | 0.00166884 |
| 401806 | derma sciences acquires global long-term exclusive rights to nimbus technology from quick-med te... | lgl_ip | lgl_ip | prod_gen | 0.0194067 | 0.00256054 |
| 208394 | incomm expansion to create more than 150 jobs in georgia | jobs_hire | biz_collab | jobs_hire | 0.0184285 | 0.00941845 |
| 136563 | Bruker Corporation Announces FDA Clearance to Market the MALDI Biotyper CA System | prod_pharm_fda | prod_pharm_fda | lgl_patent | 0.0174324 | 2.80536e-05 |
| 258676 | dps instruments, inc. elects richard r. kurtz to board of directors | jobs_brd | jobs_brd | biz_dwnsize | 0.00793869 | 0.000530561 |
| 262589 | dsw designer shoe warehouse invigorates brand with new mission, strategic plans | biz_strg | prod_gen | biz_collab | 0.00223701 | 0.00174467 |
| 131873 | informatica world attendees take time out to support education programs for at-risk youth | resp_vol | resp_vol | resp_don | 0.00223203 | 0.000574176 |
| 601557 | bark at the park presented by avoderm natural pet foods, nylabone and the american pet products ... | ind_fin_bond | biz_advrt | resp_disease | 0.00205538 | 0.00177321 |
| 257908 | rcm announces election of new board members at annual meeting of stockholders | jobs_brd | jobs_exc_retr | resp_gen | 0.00190325 | 0.00110279 |

# Construimos la red

Puro texto (5000D)

```
In [81]: #solamente BOW/matriz dispersa
         bow_in = Input(shape=(5000,) )
         bow_layer = Dense(3000, activation='relu')(bow_in)
         d2v_layer = GaussianNoise(0.20)(bow_layer)
         bow_layer = Dropout(.5)(bow_layer)
         text_class = Dense(num_classes, activation='sigmoid')(bow_layer)
         bowmodel = Model(bow_in, text_class)
```

# Entrenamos la red

Puro texto (5000D)

```
In [82]: parallel_model = multi_gpu_model(bowmodel, gpus=4)
         parallel_model.compile(loss='sparse_categorical_crossentropy',
                                optimizer='adam', metrics=['accuracy'])
         alt_res = parallel_model.fit(x_train_alt, y_train, verbose=1, validation_data=(x_test_alt, y_test), batch_size=512, epochs=5)
```

```
Train on 79773 samples, validate on 8864 samples
Epoch 1/5
79773/79773 [==============================] - 18s 224us/step - loss: 2.1024 - acc: 0.5230 - val_loss: 1.3855 - val_acc: 0.6516
Epoch 2/5
79773/79773 [==============================] - 17s 208us/step - loss: 1.0674 - acc: 0.7070 - val_loss: 1.3200 - val_acc: 0.6684
Epoch 3/5
79773/79773 [==============================] - 17s 209us/step - loss: 0.7932 - acc: 0.7566 - val_loss: 1.3227 - val_acc: 0.6683
Epoch 4/5
79773/79773 [==============================] - 17s 208us/step - loss: 0.6197 - acc: 0.7945 - val_loss: 1.4341 - val_acc: 0.6723
Epoch 5/5
79773/79773 [==============================] - 17s 209us/step - loss: 0.5264 - acc: 0.8182 - val_loss: 1.4649 - val_acc: 0.6748
```

Evaluamos la red

# Redes Neuronales

```
In [96]: fresh_docs[['title', 'topic', 'cat_1', 'cat_2', 'prob_1', 'prob_2']].sort_values('prob_1', ascending=False)
```

Out[96]:

| | title | topic | cat_1 | cat_2 | prob_1 | prob_2 |
|---|---|---|---|---|---|---|
| 40970 | Mirati Therapeutics Reports Financial Results And Provides Business Update For The Third Quarter... | biz_ern | biz_guide | biz_ern | 1 | 1 |
| 608861 | marriott helps make it happen for women-owned businesses and celebrates international women's day | resp_women | resp_women | resp_lgbt | 1 | 0.319956 |
| 7569 | Dicerna Reports Fourth Quarter and Full Year 2016 Financial and Operational Results | biz_ern | biz_guide | biz_ern | 1 | 0.999997 |
| 628647 | 8-K - SOHU COM INC (0001104188) (Filer) | lgl_reg_sec | lgl_reg_sec | prod_pharm_fda | 0.999998 | 3.5234e-07 |
| 397280 | fuwei films celebrates commencement of trial production line installation | prod_pharm_gen | biz_prod_update | prod_pharm_gen | 0.99959 | 0.946358 |
| 277967 | jcpenney announces successful closing of real estate term loan refinance | biz_deb_loan | biz_debt_repricing | biz_deb_loan | 0.981668 | 0.922196 |
| 296214 | lexmark's perceptive software positioned in leader's quadrant of gartner's magic quadrant for en... | biz_boast | biz_boast | jobs_hire | 0.981554 | 0.1537 |
| 618361 | investor alert: class action lawsuit against 500.com limited announced by law offices of howard ... | lgl_cls_action | lgl_cls_action | lgl_gen | 0.968187 | 0.41085 |
| 201203 | 2u, inc. chief executive officer & co-founder chip paucek to present at needham's annual intere... | evt_conf | evt_conf | misc_rprt | 0.888602 | 0.1169 |
| 762959 | sunshine bancorp, inc. hires veteran banking executive andrew samuel | resp_gen | jobs_exc_retr | misc_retire | 0.730959 | 0.609632 |
| 77268 | DMC Global Schedules Third Quarter Earnings Release and Conference Call | biz_ern | biz_ern_sched | evt_conf | 0.676063 | 0.0700534 |
| 613380 | gaithersburg marriott washingtonian center serves holiday dinner to homeless women | resp_women | resp_women | resp_gen | 0.640837 | 0.420084 |
| 191400 | oxbridge re holdings to present at the 6th annual liolios gateway conference on september 7, 2017 | evt_conf | misc_rprt | evt_conf | 0.573823 | 0.0857405 |
| 131873 | informatica world attendees take time out to support education programs for at risk youth | resp_vol | resp_don | resp_gen | 0.429303 | 0.177793 |
| 259574 | tegna board elects jennifer dulski as new director | jobs_brd | jobs_brd | jobs_exc | 0.389172 | 0.00251958 |
| 601557 | bark at the park presented by avoderm natural pet foods, nylabone and the american pet products ... | ind_fin_bond | resp_disease | biz_advrt | 0.309008 | 0.131647 |
| 453076 | western asset worldwide income fund inc. announces results of annual meeting of stockholders | misc_rprt | misc_rprt | biz_shrholder_action | 0.273662 | 0.0924295 |
| 262527 | developing brand identities, trading plans, strategy committees and mergers - research report on... | biz_strg | biz_strg | jobs_brd | 0.136076 | 0.103029 |
| 150421 | lincoln financial launches latest "responsibility of love" advertising campaign | prod_gen | biz_advrt | jobs_brd | 0.11501 | 0.0111979 |
| 208394 | incomm expansion to create more than 150 jobs in georgia | jobs_hire | jobs_hire | lgl_ip | 0.102453 | 0.00860366 |
| 469459 | osi systems security division selected as winner of best standoff threat detection technology at... | misc_rprt | biz_expan | biz_ctrct | 0.0841566 | 0.0203793 |
| 257908 | rcm announces election of new board members at annual meeting of stockholders | jobs_brd | jobs_brd | biz_acq | 0.0813808 | 0.00509373 |
| 136563 | Bruker Corporation Announces FDA Clearance to Market the MALDI Biotyper CA System | prod_pharm_fda | prod_pharm_fda | prod_pharm_gen | 0.0647542 | 0.000411876 |
| 401806 | derma sciences acquires global long-term exclusive rights to nimbus technology from quick-med te... | lgl_ip | prod_pharm_gen | biz_expan | 0.0265843 | 0.0025167 |
| 258676 | dps instruments, inc. elects richard r. kurtz to board of directors | jobs_brd | biz_acq | jobs_brd | 0.0196316 | 0.00534062 |
| 262589 | dsw designer shoe warehouse invigorates brand with new mission, strategic plans | biz_strg | prod_gen | biz_strg | 0.000657667 | 0.000389189 |

# Construimos la red

Puro texto (50000D)

```
In [85]: #solamente BOW/matriz dispersa
         bow_in = Input(shape=(max_words,) )
         bow_layer = Dense(3000, activation='relu')(bow_in)
         d2v_layer = GaussianNoise(0.20)(bow_layer)
         bow_layer = Dropout(.5)(bow_layer)
         text_class = Dense(num_classes, activation='sigmoid')(bow_layer)
         bigmodel = Model(bow_in, text_class)
```

# Entrenamos la red



Puro texto (50000D)

```
In [86]: parallel_model = multi_gpu_model(bigmodel, gpus=4)
         parallel_model.compile(loss='sparse_categorical_crossentropy',
                                optimizer='adam', metrics=['accuracy'])
         sparse_res = parallel_model.fit(x_train, y_train, verbose=1, validation_data=(x_test, y_test), batch_size=512, epochs=5)
```

```
Train on 79773 samples, validate on 8864 samples
Epoch 1/5
79773/79773 [==============================] - 161s 2ms/step - loss: 1.9167 - acc: 0.5652 - val_loss: 1.3012 - val_acc: 0.6681
Epoch 2/5
79773/79773 [==============================] - 159s 2ms/step - loss: 0.7299 - acc: 0.7856 - val_loss: 1.3686 - val_acc: 0.6671
Epoch 3/5
79773/79773 [==============================] - 159s 2ms/step - loss: 0.4545 - acc: 0.8447 - val_loss: 1.5342 - val_acc: 0.6681
Epoch 4/5
79773/79773 [==============================] - 159s 2ms/step - loss: 0.3562 - acc: 0.8709 - val_loss: 1.6328 - val_acc: 0.6663
Epoch 5/5
79773/79773 [==============================] - 160s 2ms/step - loss: 0.3132 - acc: 0.8827 - val_loss: 1.7516 - val_acc: 0.6748
```

# Evaluamos la red



In [98]:
```
fresh_docs[['title', 'topic', 'cat_1', 'cat_2', 'prob_1', 'prob_2']].sort_values('prob_1', ascending=False)
```

Out[98]:

| | title | topic | cat_1 | cat_2 | prob_1 | prob_2 |
|---|---|---|---|---|---|---|
| 286404 | mirati therapeutics reports financial results and provides business update for the third quarter... | biz_guide | biz_guide | biz_ern | 1 | 1 |
| 608861 | marriott helps make it happen for women-owned businesses and celebrates international women's day | resp_women | resp_women | resp_lgbt | 1 | 0.319956 |
| 7569 | Dicerna Reports Fourth Quarter and Full Year 2016 Financial and Operational Results | biz_ern | biz_guide | biz_ern | 1 | 0.999997 |
| 628647 | 8-K - SOHU COM INC (0001104188) (Filer) | lgl_reg_sec | lgl_reg_sec | prod_pharm_fda | 0.999998 | 3.5234e-07 |
| 397280 | fuwei films celebrates commencement of trial production line installation | prod_pharm_gen | biz_prod_update | prod_pharm_gen | 0.99959 | 0.946358 |
| 277967 | jcpenney announces successful closing of real estate term loan refinance | biz_deb_loan | biz_debt_repricing | biz_deb_loan | 0.981668 | 0.922196 |
| 296214 | lexmark's perceptive software positioned in leader's quadrant of gartner's magic quadrant for en... | biz_boast | biz_boast | jobs_hire | 0.981554 | 0.1537 |
| 618361 | investor alert: class action lawsuit against 500.com limited announced by law offices of howard ... | lgl_cls_action | lgl_cls_action | lgl_gen | 0.968187 | 0.41085 |
| 201203 | 2u, inc. chief executive officer & co-founder chip paucek to present at needham's annual interco... | evt_conf | evt_conf | misc_rprt | 0.888602 | 0.1169 |
| 762959 | sunshine bancorp, inc. hires veteran banking executive andrew samuel | resp_gen | jobs_exc_retr | misc_retire | 0.730959 | 0.609632 |
| 77268 | DMC Global Schedules Third Quarter Earnings Release and Conference Call | biz_ern | biz_ern_sched | evt_conf | 0.676063 | 0.0700534 |
| 613380 | gaithersburg marriott washingtonian center serves holiday dinner to homeless women | resp_women | resp_women | resp_gen | 0.640837 | 0.420084 |
| 191400 | oxbridge re holdings to present at the 6th annual liolios gateway conference on september 7, 2017 | evt_conf | misc_rprt | evt_conf | 0.573823 | 0.0857405 |
| 131873 | informatica world attendees take time out to support education programs for at risk youth | resp_vol | resp_don | resp_gen | 0.429303 | 0.177793 |
| 259574 | tegna board elects jennifer dulski as new director | jobs_brd | jobs_brd | jobs_exc | 0.389172 | 0.00251958 |
| 601557 | bark at the park presented by avoderm natural pet foods, nylabone and the american pet products ... | ind_fin_bond | resp_disease | biz_advrt | 0.309008 | 0.131647 |
| 453075 | western asset worldwide income fund inc. announces results of annual meeting of stockholders | misc_rprt | misc_rprt | biz_shrholder_action | 0.273662 | 0.0924295 |
| 262527 | developing brand identities, trading plans, strategy committees and mergers - research report on... | biz_strg | biz_strg | jobs_brd | 0.136076 | 0.103029 |
| 150421 | lincoln financial launches latest "responsibility of love" advertising campaign | prod_gen | biz_advrt | jobs_brd | 0.11501 | 0.0111979 |
| 208394 | incomm expansion to create more than 150 jobs in georgia | jobs_hire | jobs_hire | lgl_ip | 0.102453 | 0.00860366 |
| 469459 | osi systems security division selected as winner of best standoff threat detection technology at... | misc_rprt | biz_expan | biz_ctrct | 0.0841566 | 0.0203793 |
| 257908 | rcm announces election of new board members at annual meeting of stockholders | jobs_brd | jobs_brd | biz_acq | 0.0813808 | 0.00509373 |
| 136563 | Bruker Corporation Announces FDA Clearance to Market the MALDI Biotyper CA System | prod_pharm_fda | prod_pharm_fda | prod_pharm_gen | 0.0647542 | 0.000411876 |
| 401806 | derma sciences acquires global long-term exclusive rights to nimbus technology from quick-med te... | lgl_ip | prod_pharm_gen | biz_expan | 0.0265843 | 0.0025167 |
| 258676 | dps instruments, inc. elects richard r. kurtz to board of directors | jobs_brd | biz_acq | jobs_brd | 0.0196316 | 0.00534062 |
| 262589 | dsw designer shoe warehouse invigorates brand with new mission, strategic plans | biz_strg | prod_gen | biz_strg | 0.000657667 | 0.000389189 |

# En fin…

Vemos que Doc2Vec aporta información que no se puede capturar con formas más tradicionales de vectorizar texto.

…en realidad, nuestro modelo incorpora ambos métodos, más un LSTM con el título, y rinde muchísimo mejor que cualquiera de los tres individualmente. :)

¿preguntas?