

Cantando el tango como ninguna.

Una aplicación de topic modeling y algunos otros experimentos en letras de tango...

Germán Rosati (CONICET-UNSAM, PIMSA)

Sobre mí

Sociología, Machine Learning, Métodos Cuantitativos

Ahora...

- Profesor “Métodos Cuantitativos” UNSAM
- Investigador Asistente CONICET

Antes...

- Coordinador Data Science / IA Digital House
- Becario doctoral (CONICET)
- Investigador invitado Freie Universität Berlin
- Analista Experto de Datos (MTEySS)
- Data Scientist FreeLance (BID, Banco Mundial, PNUD, OIT, Universidades, consultoras)

Machine learning proliferates in particle physics

Illustration by Sandbox Studio, Chicago with Corinne Mucha

06/01/18 | By Manuel Gnida

A new review in *Nature* chronicles the many ways machine learning is popping up in particle physics research.

Experiments at the Large Hadron Collider produce about a million gigabytes of data every second. Even after reduction and compression, the data amassed in just one hour at the LHC is

nature.com > nature > technology features > article

MENU ▾

nature
International Journal of science

News & Comment Research

News Opinion Research Analysis Careers Books & Culture

TECHNOLOGY FEATURE • 20 FEBRUARY 2018 • CORRECTION 07 MARCH 2018

Deep learning for biology

A popular artificial-intelligence method provides a powerful tool for surveying and classifying biological data. But for the uninitiated, the technology poses significant difficulties.

nature.com > nature > news > article

a nature

MENU ▾

nature
International Journal of science

News & Comment Research

News Opinion Research Analysis Careers Books & Culture

NEWS • 28 MARCH 2018

Need to make a molecule? Ask this AI for instructions

Artificial-intelligence tool that has digested nearly every reaction ever performed could transform chemistry.



Cambridge
Analytica



SAY BIG DATA

ONE MORE TIME

memegenerator.net

Planteo del problema

- Incorporar estas técnicas a las ciencias sociales
- ¿Cómo detectar temas en las letras de tango?

Hoja de ruta

- Enfoques habituales en análisis de texto en Ciencias Sociales
- “Nuestro” enfoque
- Pipeline de preprocesamiento
- Intuición **por (y para) sociólogos**: ¿Qué es LDA?
- Algunos resultados

Enfoque tradicional

- Problema: analizar los temas de las letras de tango
- Enfoque “hermenéutico”: analizar pocas letras en profundidad
- Temas comunes: representaciones de género, figuras del “guapo”, representaciones del arrabal, etc.



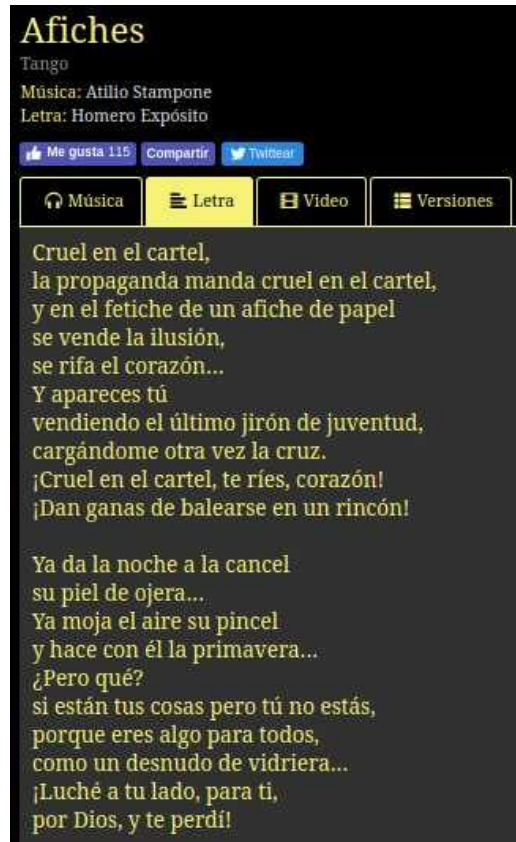
Enfoque tradicional

- Problema: analizar los temas de las letras de tango
- Enfoque “estadístico”
- Cantón (1972), analiza ciertos aspectos relevantes de las letras de los tangos cantados por Gardel



Enfoque propio

- Scrap de letras del sitio todotango.com
- Corpus: 5.700 letras
- Problema: analizar un corpus de ~5.700 letras de tango para detectar “tópicos” - Detección automática: Latent Dirichlet Allocation

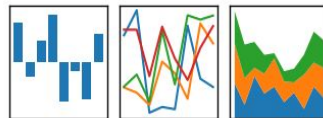


Stack



pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



StanfordNLP 0.2.0 - Python NLP Library for Many Human Languages

- Table of Contents
- [About](#)
 - [Get Started](#)
 - [License](#)
 - [Citing StanfordNLP in papers](#)
 - [Links](#)



Pipeline de preprocesamiento

1. Vectorización de texto

TANGO	agua	blanda	cartel	cruel	el	en	era	la	manda	más	propaganda	que
Cruel en el cartel, la propaganda manda cruel en el cartel,	0	0	2	2	2	2	0	1	1	0	1	0
Era más blanda que el agua que el agua blanda	2	2	0	0	2	0	1	0	0	1	0	2

Pipeline de preprocesamiento

1. Vectorización de texto
2. Eliminar stopwords (por lista)

TANGO	agua	blanda	cartel	cruel	era	manda	propaganda
Cruel en el cartel, la propaganda manda cruel en el cartel,	0	0	2	2	0	1	1
Era más blanda que el agua que el agua blanda	2	2	0	0	1	0	0

Pipeline de preprocesamiento

1. Vectorización de texto
2. Eliminar stopwords (por lista)
3. Normalización
 - Lematización
 - Stemming

Pipeline de preprocesamiento

1. Vectorización de texto
2. Eliminar stopwords (por lista)
3. Normalización
4. Eliminar stopwords (vía term-freq)

Al usar valores estándar (eliminar términos que se encuentran en más del 95% y en menos del 5% de los documentos) sobreviven pocos términos (alrededor de 150)

Subsiste en lematización y en stemming e incluso al hacerlo con los términos sin normalizar

¿Lunfardo?

Pipeline de preprocesamiento

1. Vectorización de texto
2. Eliminar stopwords (por lista)
3. Normalización
4. Eliminar stopwords (vía term-freq)
5. Ponderar matriz de términos

Term Frequency - Inverse Doc Freq

$$TF(t, d) = \frac{rc(t, d)}{\sum_{t \in d} rc(t, d)}$$

$$DF(t) = \log \frac{df(t)}{|C|}$$

$$IDF(t) = \frac{1}{DF(t)} = \log \frac{|C|}{df(t)}$$

$$TF_IDF(t) = TF(t, d) \times IDF(t)$$

Topic Modeling - Latent Dirichlet Allocation

Seeking Life's Bare (Genetic) Necessities

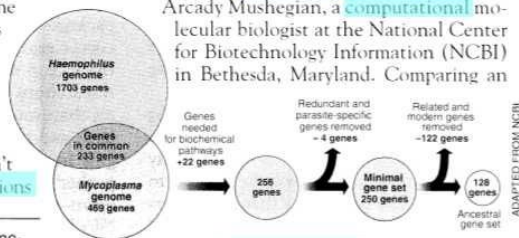
COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

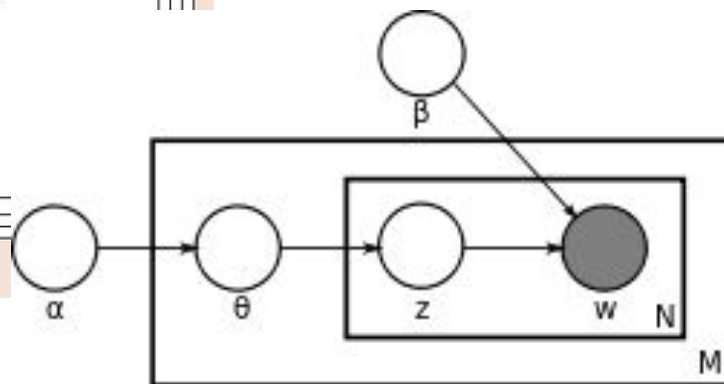
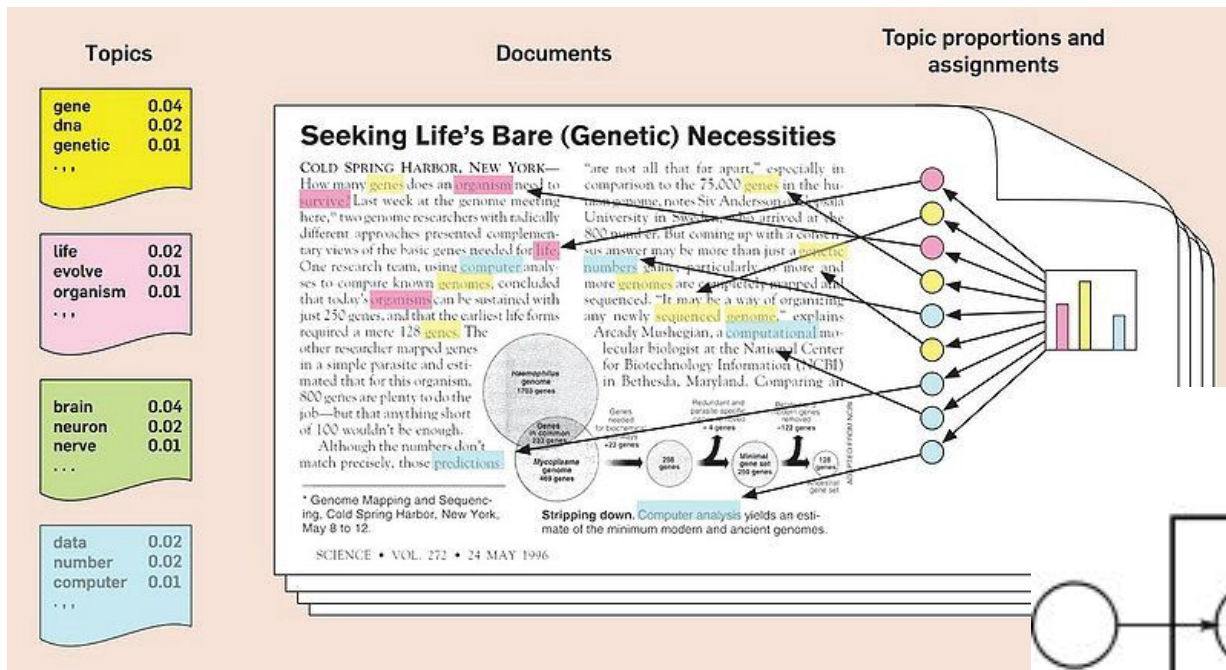


Stripping down. **Computer analysis** yields an estimate of the minimum modern and ancient genomes.

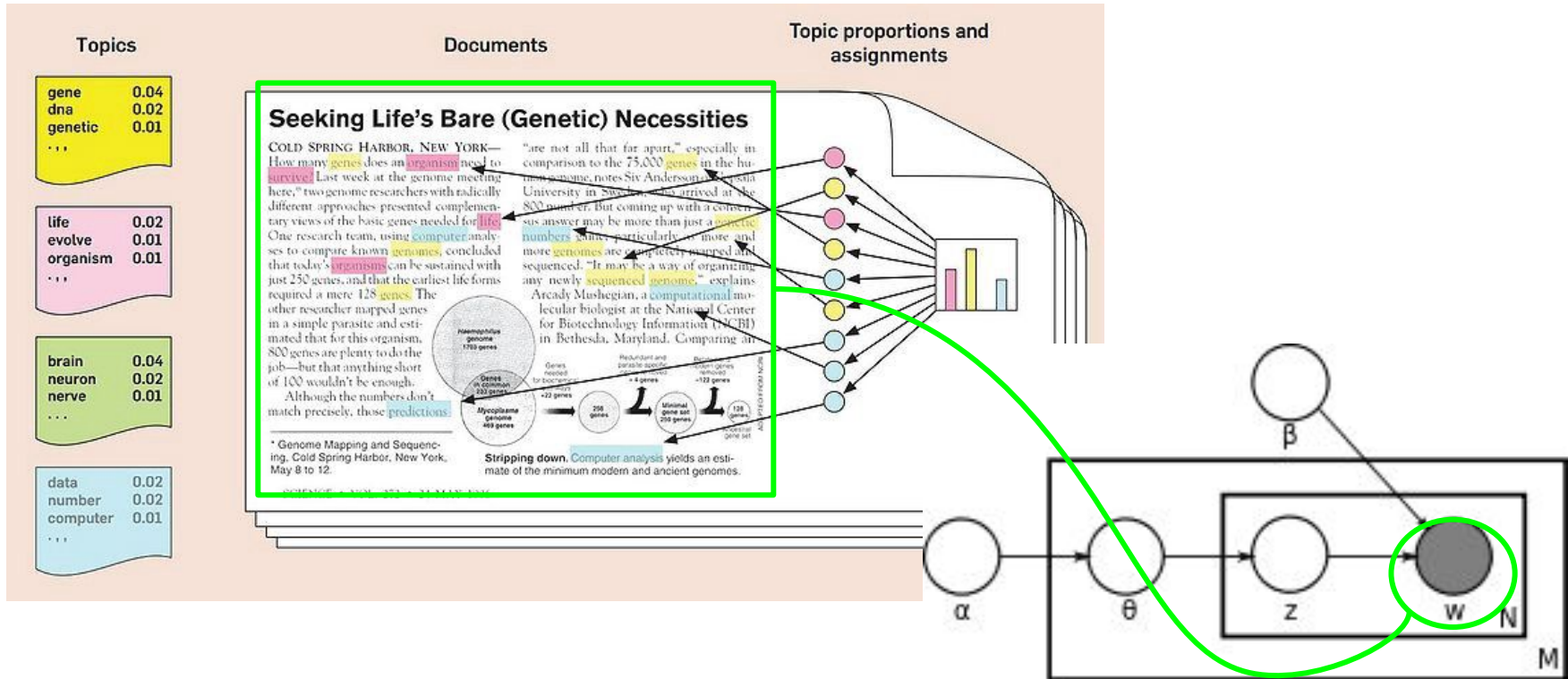
- Intuición: Un documento se compone de muchos tópicos
- Supuestos:
 - Cada tópico es una distribución de palabras con diferentes probabilidades
 - Cada documento es una mezcla de diferentes tópicos
 - Cada palabra se “extrae” de alguno de estos tópicos
- Objetivo: queremos estimar los tópicos en un corpus

[Blei, 2012]

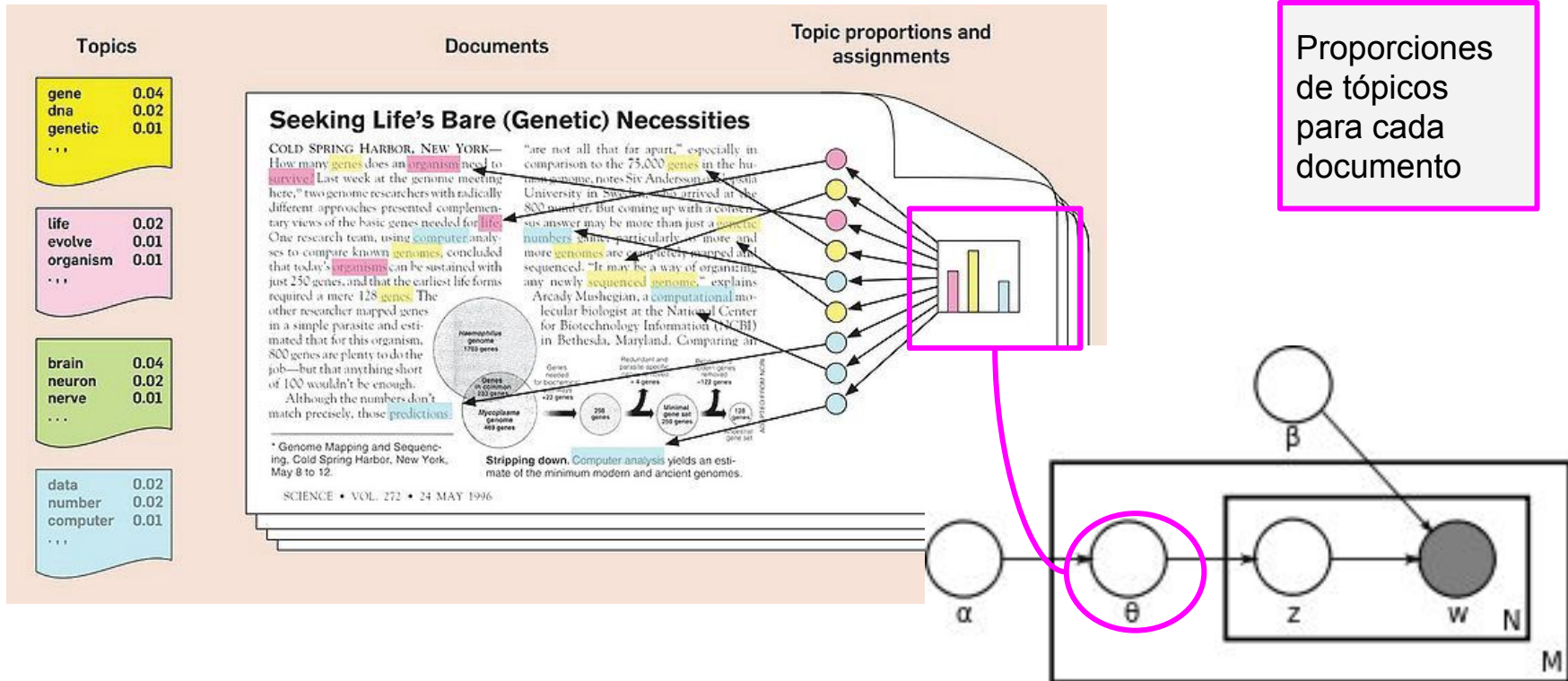
Topic Modeling - Latent Dirichlet Allocation



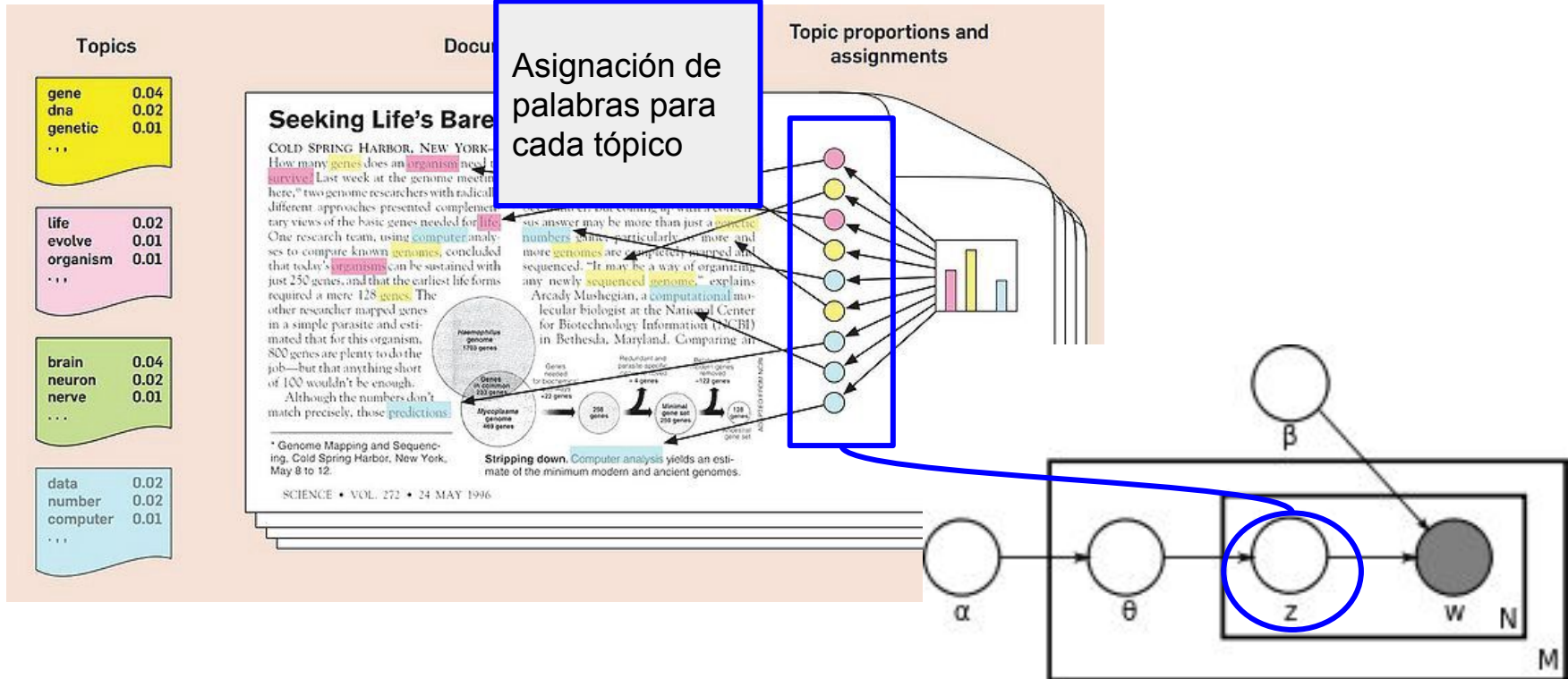
Topic Modeling - Latent Dirichlet Allocation



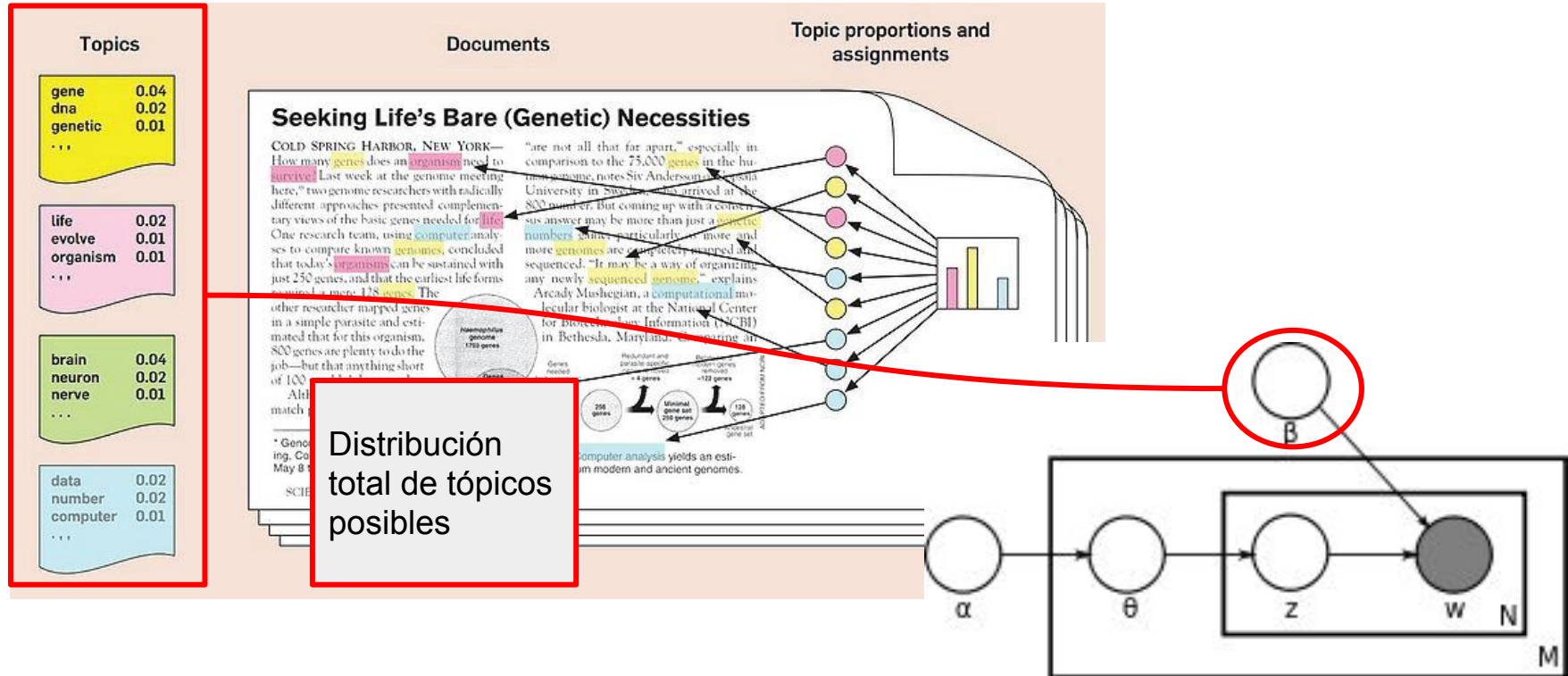
Topic Modeling - Latent Dirichlet Allocation



Topic Modeling - Latent Dirichlet Allocation



Topic Modeling - Latent Dirichlet Allocation



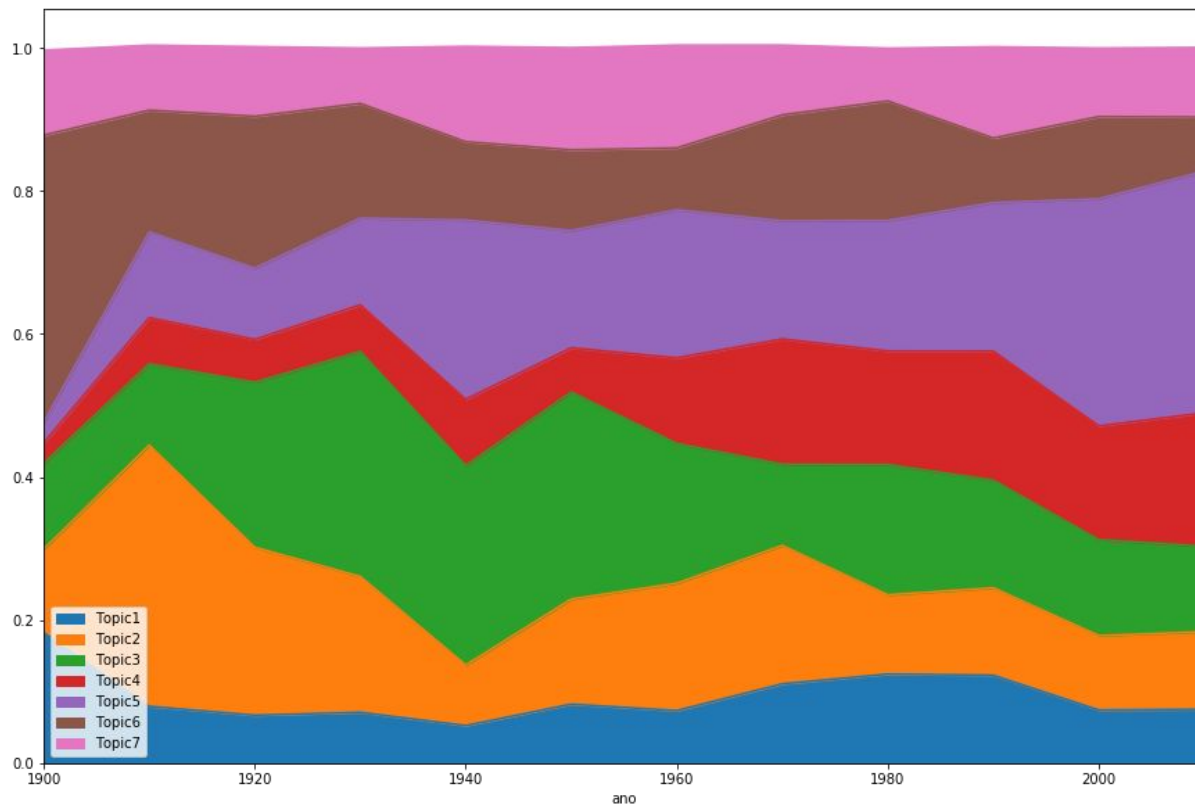
Topic Modeling - Latent Dirichlet Allocation

- Tópicos más relevantes

https://gefero.github.io/tango_scrap/

Topic Modeling - Latent Dirichlet Allocation

Evolución de los tópicos
(mediana de la
composición de las letras
de tango) 1880-2010



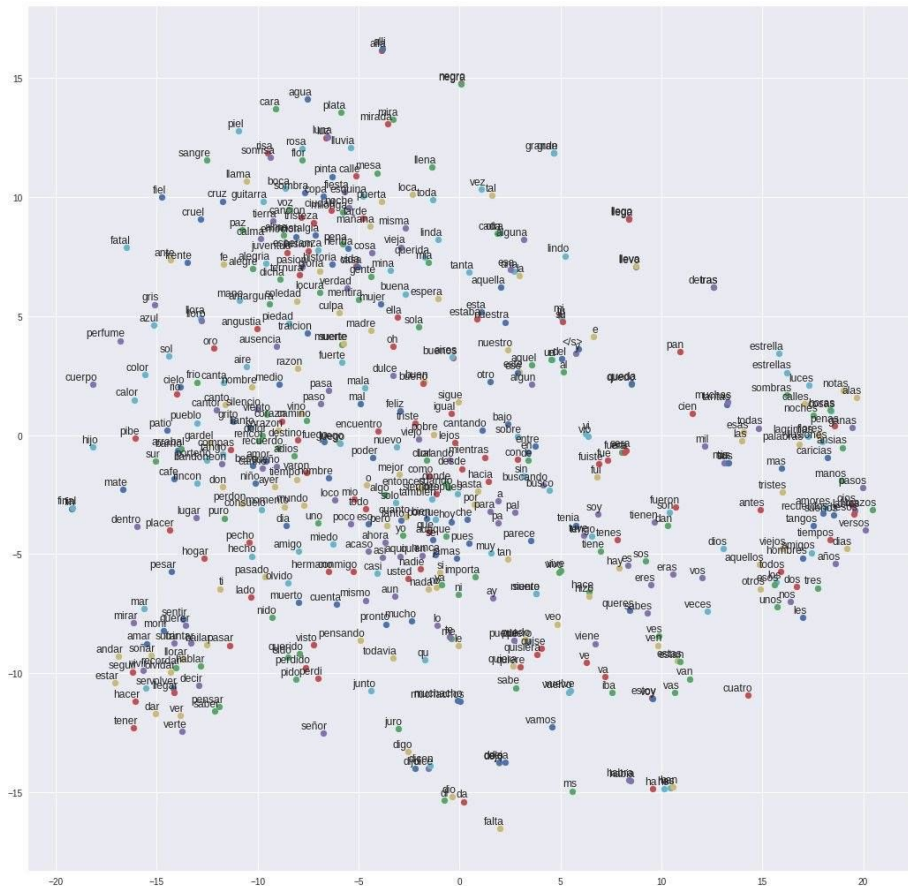
Topic Modeling - Latent Dirichlet Allocation

Composición de tópicos en algunos tangos

	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7
	Amor signo -	Imág. naturales	Amor signo +	Miscelaneo	Ciudad	Tango	Personif.
Arrabal amargo	0.02	0.02	0.02	0.02	0.85	0.02	0.02
Barrio reo	0.03	0.03	0.03	0.53	0.03	0.34	0.03
Cafetin de Buenos Aires	0.02	0.02	0.49	0.38	0.02	0.02	0.02
Garua	0.03	0.03	0.03	0.03	0.85	0.03	0.03
Lejana Tierra mía	0.03	0.03	0.03	0.03	0.84	0.03	0.03

De yapa... dos ejercicios más sobre este corpus

Estimacion de word embeddings (vía FastText)



De yapa... dos ejercicios más sobre este corpus

El próximo Discepolín...

Una red neuronal que escribe
(ojalá) tangos



CrstC @Crst_C · 18 mar. 2018

20 hs de entrenamiento... mirá lo que escribe esta red neuronal profunda de mierda...

```
In [21]: txt=""
         for char in full_string:
             txt = txt+char
         print(txt)

carta salidora
sos la banca tentadora
por la que siempre me seco
y sos el colgante fleco
de la pa pe pe pe pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa
pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa
a pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa
pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa pa
```

In []:

De yapa... dos ejercicios más sobre este corpus

El próximo Discepolín...

Una red neuronal que escribe
(ojalá) tangos



CrstC @Crst_C · 20 nov. 2017

Bueno... el estribillo tiene gancho...

```
new_model.fit(X, y, epochs=5, batch_size=50, callbacks=callbacks_list, verbose=1)
Epoch 1/5
39750/39750 [=====] - 1209s 30ms/step - loss: 1.9692
Epoch 2/5
39750/39750 [=====] - 1185s 30ms/step - loss: 1.5312
Epoch 3/5
39750/39750 [=====] - 1200s 30ms/step - loss: 1.3999
Epoch 4/5
39750/39750 [=====] - 1233s 31ms/step - loss: 1.3335
Epoch 5/5
39750/39750 [=====] - 1243s 31ms/step - loss: 1.2931

Out[11]: <keras.callbacks.History at 0x7fd8925b8c50>

In [ ]:

In [130]: generate_text(new_model, 150)
"los dos
y en la misma cortada
y en la misma despues
y en la misma cortada
y en la misma despues
y en la misma cortada
y en la misma despu

Out[130]: '"los dos \n y en la misma cortada \n y en la misma despues \n y en la misma cortada \n y en la misma despues \n y
en la misma cortada \n y en la misma despue'
```



1

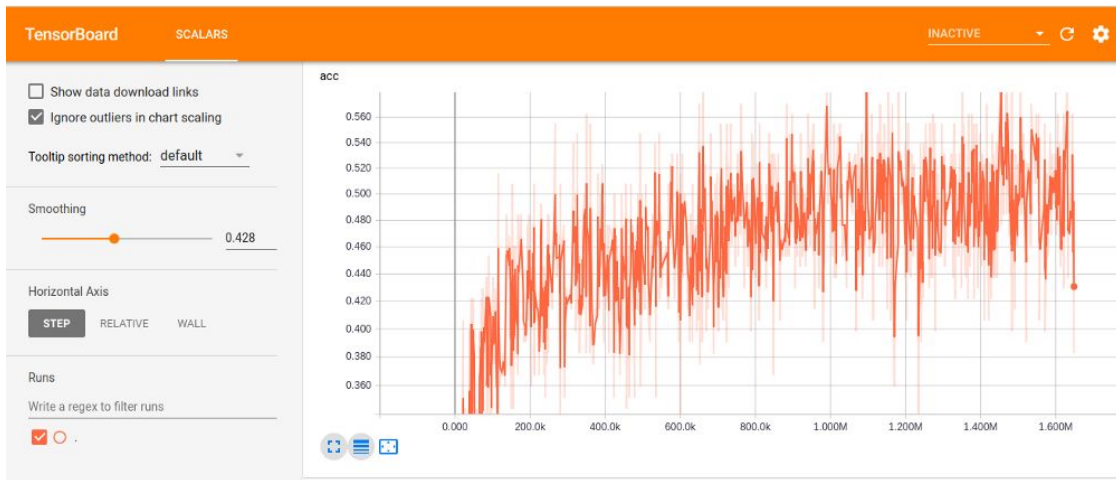


[Mostrar este hilo](#)

De yapa... dos ejercicios más sobre este corpus

El próximo Discepolín...

Una red neuronal que escribe
(ojalá) tangos



De yapa... dos ejercicios más sobre este corpus

El próximo Discepolín...

Una red neuronal que escribe
(ojalá) tangos

```
----- Generating text after Epoch: 5
----- diversity: 0.2
----- Generating with seed: "les voy a nombrar señores
las pebetas "
les voy a nombrar señores
las pebetas de la mina
y en la vida en la panta
la mas de mi corazon
y se de estan el por el mano
y el mance de mi alma
y en la vida
y en el mano
la ser
el canto de sol
por mi alma
y en el corazon
y se que no es mi mis mentinas
con el mante de mi por su sol
y en la vida de mi corazon
y en el mancio que el alma
se para la te su corazon
con el mentina con el manta
y su corazon
```

De yapa... dos ejercicios más sobre este corpus

El próximo Discepolín...

Una red neuronal que escribe
(ojalá) tangos

```
----- Generating text after Epoch: 8
----- diversity: 0.2
----- Generating with seed: "y la sagrada biblia pide la salvacion
"
y la sagrada biblia pide la salvacion
y el para el pasa
si de pasar3o
en el canto en la cariza
y el tango en la este amor
y en el perdio de amor
con tu alma de el amor
y en el cara de la misti3
si el corazon
y el cara de amor
que esta en el corazon
en el amor de la este alma
y en la para se es3co
y el cara y3ente
no se entre la cara de mi cara
y el pasar de la mana
se alegre en la tango
para el p3sio
```

De yapa... dos ejercicios más sobre este corpus

El próximo Discepolín...

Una red neuronal que escribe
(ojalá) tangos

----- Generating text after Epoch: 14

----- diversity: 0.2

----- Generating with seed: " tuco paz

sera porque me acune

en tu"

tuco paz

sera porque me acune

en tu corazon

se el desde su para

y en tu vida de tu corazon

y el perdido en la mano

con la corazon

y el canto en la para

en la canto de la vida

de mi para

te destente de tu camino

de su alma mas el corazon

y en la mano de la alma

y te de al perder

es el amor

con la para

se amor de la vida

y en el canto el corazon

y en el corazon

te viento de la perdida

y el canto el

¿Y entonces?

Dos formas de vinculación entre Cs. Sociales y Machine Learning

- **Como usuarios o consumidores**
 - \approx a la que tenemos con la estadística
 - Usuarios de métodos, APIS, etc.

¿Y entonces?

Dos formas de vinculación entre Cs. Sociales y Machine Learning

- **Como productores**
 - Planteo de nuevos problemas relevantes
 - Reformulación de nuevos métodos en base a problema

¿Preguntas?



@Crst_C



german.rosati@gmail.com



<https://gefero.github.io/>