

Text Mining

Edgar Altszyler
Bruno Bianchi

eaaltszyler@dc.uba.ar



Motivación

Disponibilidad masiva de textos



Estructura de la materia

- Clases teórico-prácticas

Estructura de la materia

- Clases teórico-prácticas
- Exámen teórico: TBA

Estructura de la materia

- **Clases teórico-prácticas**
- **Exámen teórico:** TBA
- **Proyecto final:**
 - En grupos de 4
 - Elegir una tarea clara y concisa:
 - Tarea de clasificación, regresión
 - Testeo de alguna hipótesis
 - Con un *corpus* dado en la clase o bajado por ustedes
 - Hay algunas opciones pre-establecidas
 - Entrega escrita + presentación oral

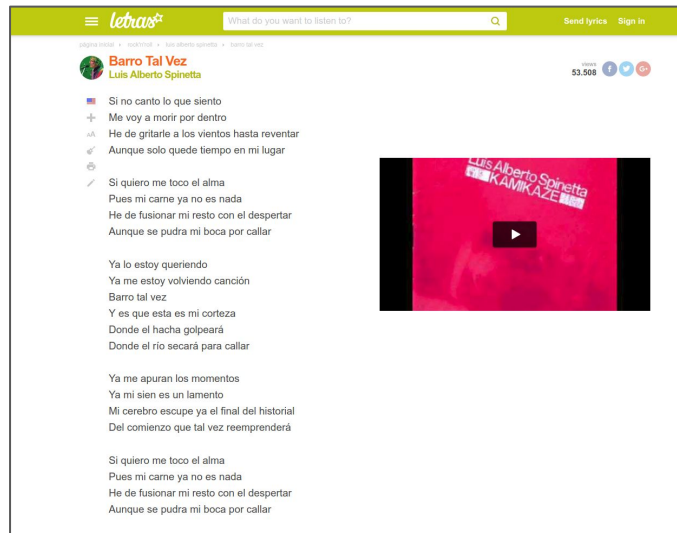
Procesamiento de texto en Python

Luis Alberto Spinetta

Luis Alberto Spinetta (Buenos Aires, 23 de enero de 1950 - Ibídem, 8 de febrero de 2012), conocido como El Flaco o simplemente por su apellido, fue un cantante, guitarrista, poeta, escritor y compositor argentino de rock, considerado uno de los más importantes y respetados músicos en Hispanoamérica. La complejidad instrumental, lírica y poética de sus obras le valió el reconocimiento en muchas partes del mundo. En 1997 la revista Billboard lo definió como «ícono del rock argentino»,¹ y en 2001 el diario Página/12 lo consideró el artista más influyente en la historia del rock argentino, tras hacer una encuesta con celebridades del rock local.² En 2014 se estableció por ley que el día de su nacimiento fuera el Día Nacional del Músico en Argentina. Spinetta fundó diversos grupos, como Almendra, Pescado Rabioso, Invisible, Spinetta Jade y Spinetta y los Socios del Desierto. En su obra hay influencia de escritores, filósofos, pensadores, psicólogos y artistas plásticos como Rimbaud, Van Gogh, Dalí, Escher, Lü Dongbin, Jung, Freud, Nietzsche, Foucault, Deleuze, Sartre, Castaneda y Artaud, así como de las culturas de los pueblos originarios americanos y de Oriente.

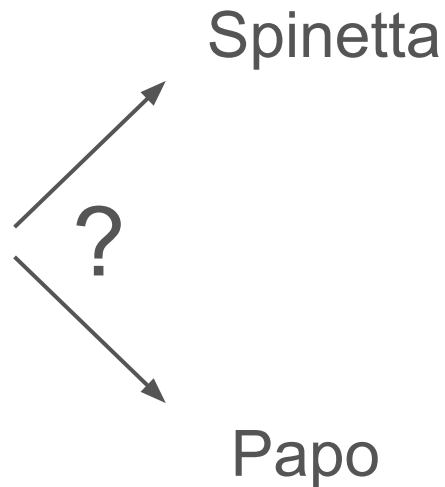


Web scraping (invitado: Ramiro Gálvez)



Clasificación de Textos

Y la espuma gira en torno a mi piel
Me han puesto manos para hablarle
A las cosas de mi.
Y al fin mi duende nació
Tiene orejas blancas
Como un soplo de pan y arroz
Y un hongo como nariz
Cuatro pelos locos y un violín que nunca calla
Solo se desprende y es igual a las guirnaldas.



Clasificación de Textos

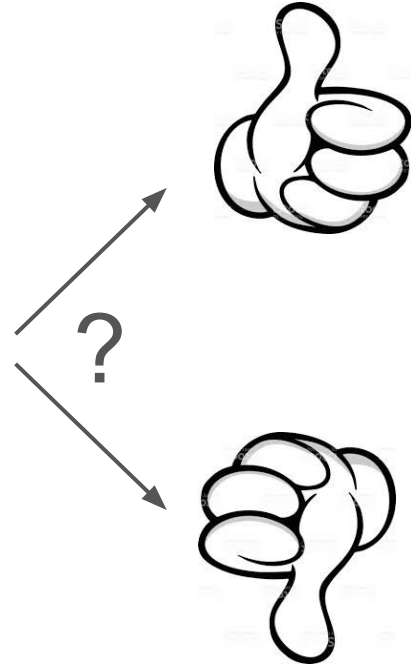
Detección de sentimiento

@juanma25



Que bien que estoy!

Empanadas + Netflix , infalible!



Sentiment / Emotion Analysis

Sentiment Analysis



Emotion Analysis



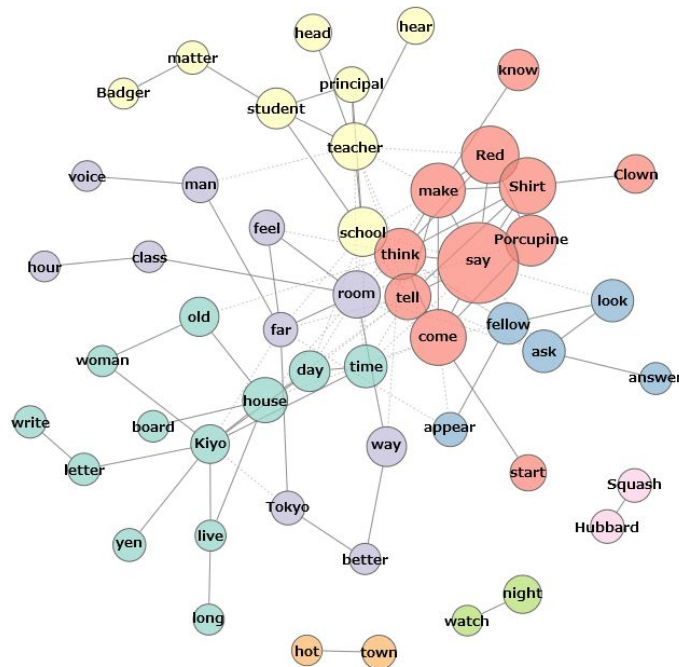
Similaridad Semántica

Estudio de asociaciones de palabras

Ispeccionando un dataset



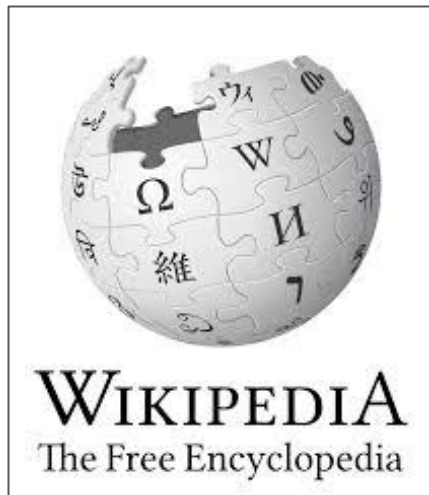
WIKIPEDIA
The Free Encyclopedia



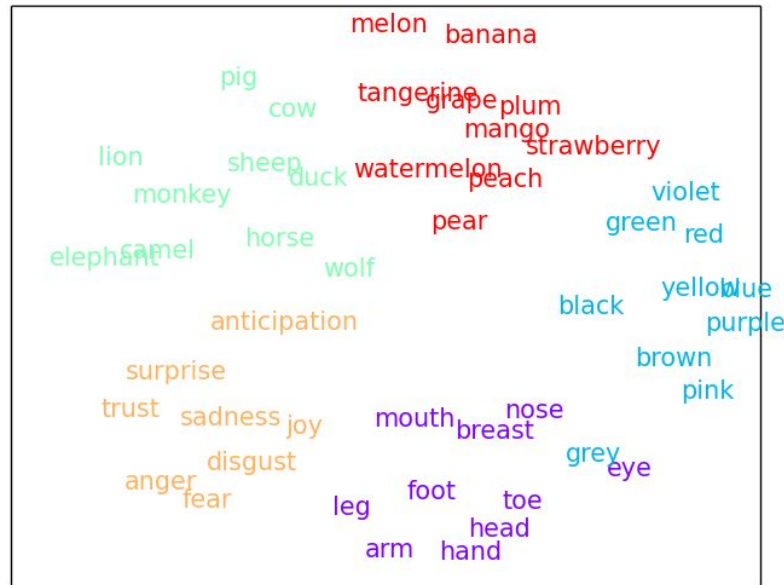
Word-Embeddings

De palabras a vectores

Corpus de textos



Word-embeddings



Language Models

Que ganas de tomar un mate **caliente**

Topic Models

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

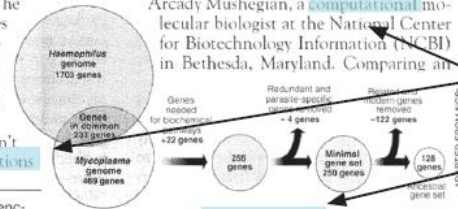
Documents

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,⁴⁰ 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**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson at 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

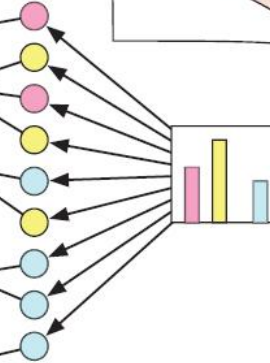


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

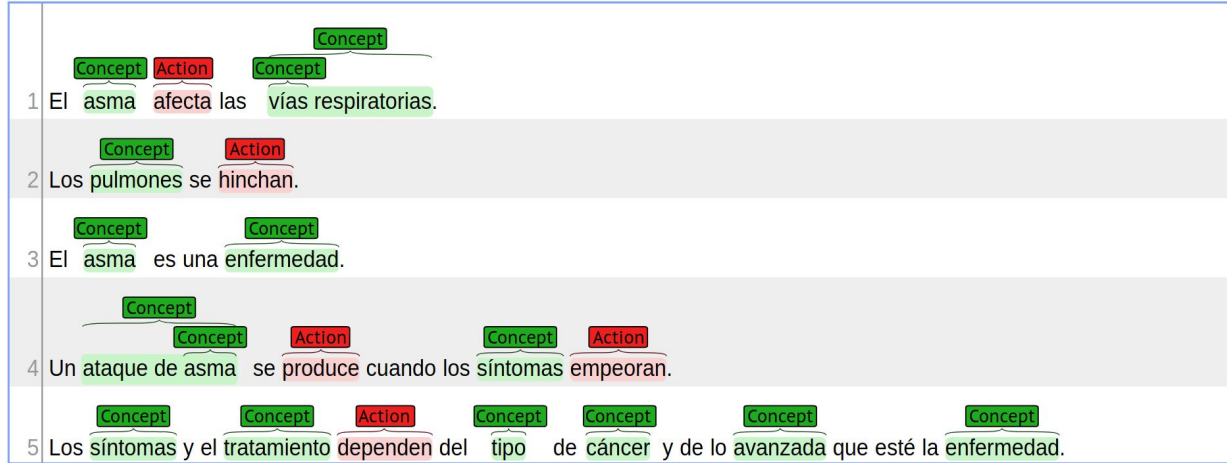
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

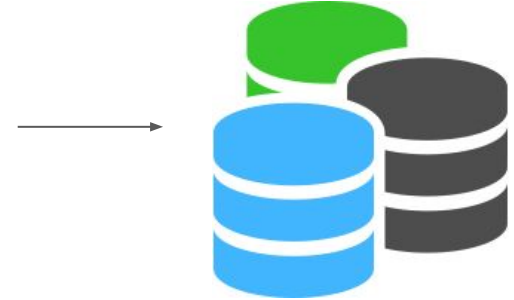
Topic proportions and assignments



Extracción de Información



Base de datos
estructurada



Comunidad NLP

Las conferencias más importantes del área son:

- ACL, NACCL, EMNLP, TACL, EACL, CoNLL,...
- Todos los paper están en ACL Anthology.
- También hay publicaciones de NLP en conferencias de ML y AI como: ICML, NIPS, ICLR, AAAI, IJCAI

Corpus

DreamBank
www.dreambank.net

Adam Schneider & G. William Domhoff
Psychology Dept., UC Santa Cruz

Más de 20.000 reportes de sueños



Query: Search [\[HELP\]](#)

Dream series:

- Barb Sanders [n=3116]
- Barb Sanders #2 [n=1138]
- Barb Sanders: baseline [n=250]
- Bay Area girls: Grades 4-6 [n=234]
- Bay Area girls: Grades 7-9 [n=154]**
- Bea 1: a high school student [n=223]
- Bea 2: a college student [n=63]
- Blind dreamers (F) [n=238]
- Blind dreamers (M) [n=143]
- Chris: a transvestite [n=100]

Filters

F M any

ind. ☐ ☐ ☐

coll. ☐ ☐ ☐

all ☐ ☐ ☐

[more info](#)

[\[ALL info\]](#)

Matching mode:

☒ AND ☐ OR ☒ subtotals

Case-sensitive?

☐ Yes ☒ No

Results display:

☐ Lists ☒ Table ☐ Contingency

Consistency chart:

	run
Bay Area girls: Grades 7-9 (n=154)	12 (7.8%)

Category	Women's Reports (N=246)	Men's Reports (N=95)
Parents/Siblings	26.8%	14.7%
Spouses/Partners	19.9%	13.7%
Other Family	10.2%	4.2%
Friends	46.7%	44.2%
Any Familiar Character	75.2%	62.1%
Travel/Vacation	16.3%	6.3
Sports	7.3%	9.5%
Entertainment/shows	7.7%	9.5%
Parties/Cafes/Bars	11.4%	6.3%
Shopping	9.8%	2.1%
Any Leisure Activity	42.3%	27.4%
School/Work/Politics	20.3%	29.5%
Dreams With No Familiar Elements	12.6%	20.0%

Corpus

50.000 reviews de IMDB
positivos/negativos

The screenshot shows the IMDb page for the movie 'The Godfather' (1972). The page layout includes a top navigation bar with the IMDb logo, a search bar, and links to 'IMDb Pro', 'IMDb Apps', and 'Help'. Below the navigation bar are tabs for 'Movies, TV & Showtimes', 'Celebs, Events & Photos', 'News & Community', and 'Watchlist'. The main content area features a large movie poster on the left and a detailed description on the right. The description includes the movie's title, year, genre, runtime, and a synopsis. It also lists the director, writers, and stars. A 'Your rating' section shows a 9.2 rating from 801,690 users. A 'Quick Links' section on the right provides links to various related content. At the bottom, there are buttons for '+ Watchlist', 'Watch Trailer', and 'Share...'. A 'Related News' section at the bottom right mentions 'Academy's Song Nominee'.

IMDb Find Movies, TV shows, Celebrities and more... All **IMDb Pro** | IMDb Apps | Help

Movies, TV & Showtimes Celebs, Events & Photos News & Community Watchlist Login

The Godfather (1972) 55

175 min - Crime | Drama - 24 March 1972 (USA)

Your rating: ★★★★★★★★ -/10
Ratings: 9.2/10 from 801,690 users Metascore: 100/100
Reviews: 1,731 user | 184 critic | 14 from Metacritic.com

The aging patriarch of an organized crime dynasty transfers control of his clandestine empire to his reluctant son.

Director: Francis Ford Coppola
Writers: Mario Puzo (screenplay), Francis Ford Coppola (screenplay), 1 more credit »
Stars: Marlon Brando, Al Pacino, James Caan | See full cast and crew »

+ Watchlist Watch Trailer Share...

Quick Links
Full Cast and Crew Plot Summary
Trivia Parents Guide
Quotes User Reviews
Awards Release Dates
Message Board Company Credits

Explore More

Like 31,995 people like this. Be the first of your friends.

Related News
Academy's Song Nominee

Corpus

50k de comentarios
Positivos/Negativos

