Text Mining

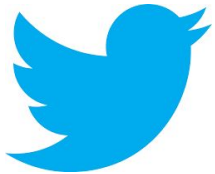
Edgar Altszyler
Bruno Bianchi

ealtszyler@dc.uba.ar

Motivación

Disponibilidad masiva de textos

Google



Estructura de la materia

- **Clases teórico-prácticas**

Estructura de la materia

- **Clases teórico-prácticas**
- **Exámen teórico**: TBA

Estructura de la materia

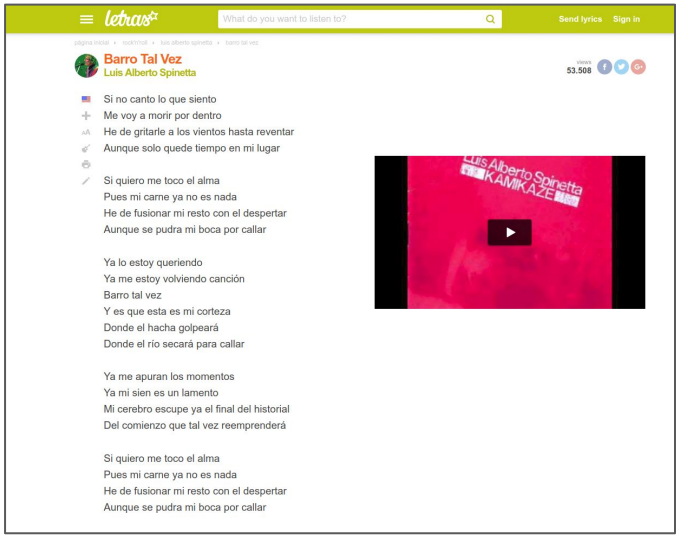
- **Clases teórico-prácticas**
- **Exámen teórico**: TBA
- **Proyecto final**:
  - En grupos de 4
  - Elegir una tarea clara y concisa:
    - Tarea de clasificación, regresión
    - Testeo de alguna hipótesis
  - Con un *corpus* dado en la clase o bajado por ustedes
  - Hay algunas opciones pre-establecidas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

Procesamiento de texto en Python

Luis Alberto Spinetta

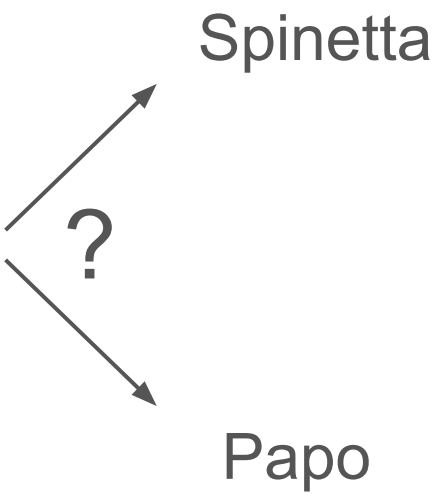
Luis Alberto Spinetta (Buenos Aires, 23 de enero de 1950 - Ibídem, 8 de febrero de 2012), conocido como El Flaco o simplemente por su apellido, fue un cantante, guitarrista, poeta, escritor y compositor argentino de rock, considerado uno de los más importantes y respetados músicos en Hispanoamérica. 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.



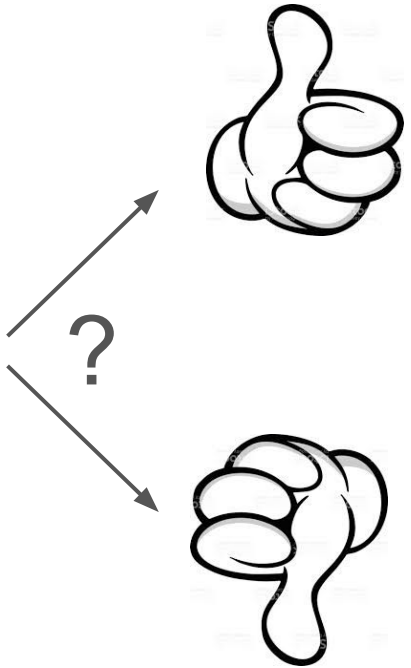
Clasificación de Textos

Detección de sentimiento

@juanma25

Que bien que estoy!

Empanadas + Netflix , infalible!



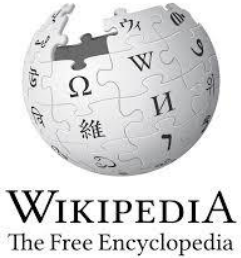
Sentiment / Emotion Analysis

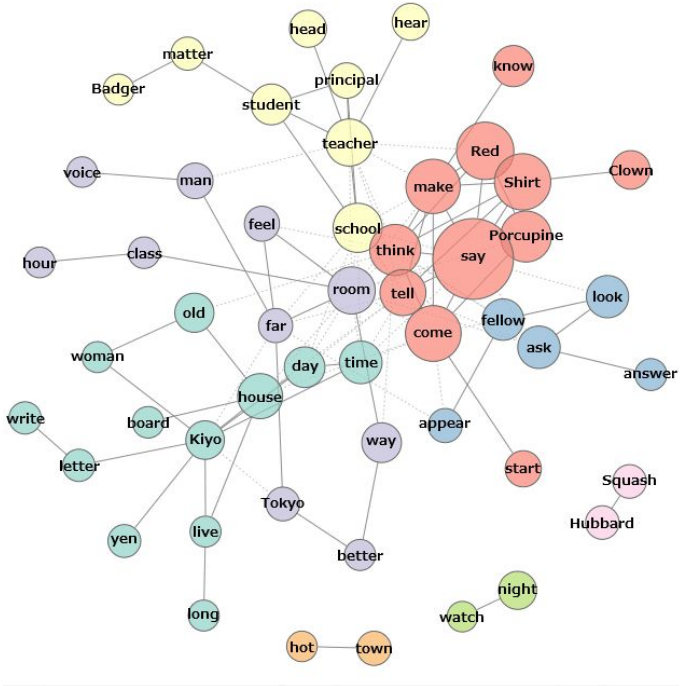
Sentiment Analysis

Emotion Analysis

Similaridad Semántica

Estudio de **asociaciones de palabras**
Inspeccionando un dataset

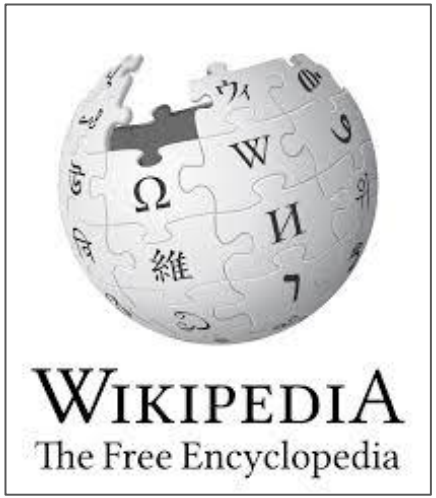




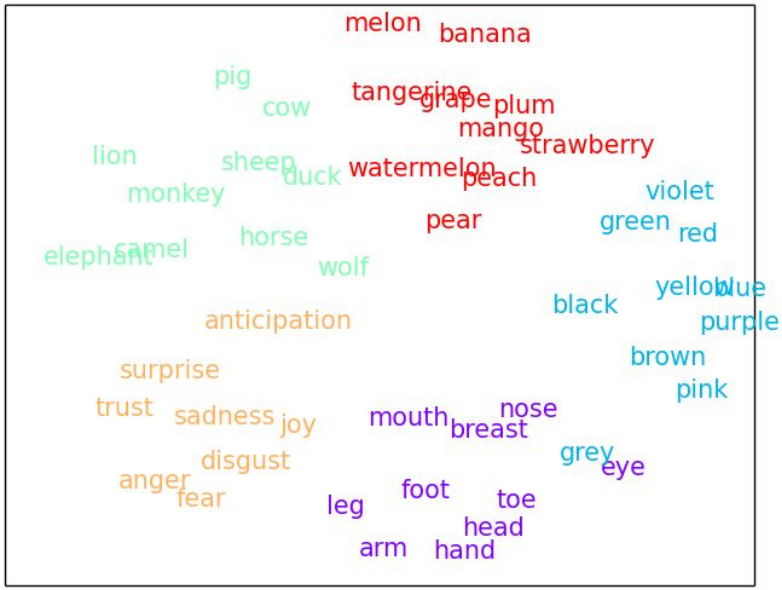
Word-Embeddings

De palabras a vectores

Corpus de textos



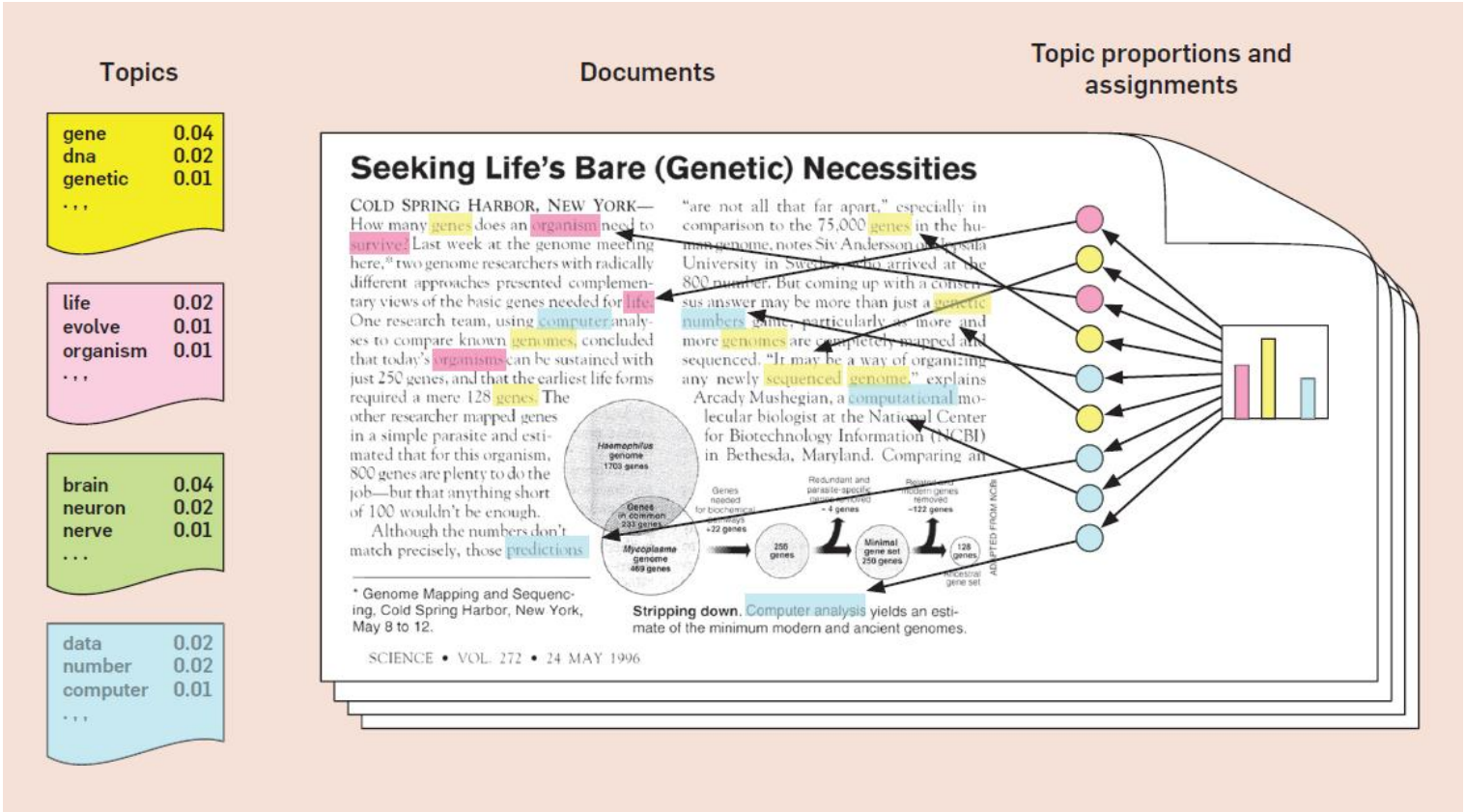
Word-embeddings



Language Models

Que ganas de tomar un mate caliente

Topic Models



Extracción de Información

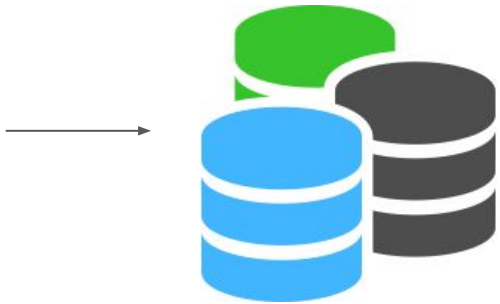
1 El asma (Concept) afecta (Action) las vías respiratorias (Concept).

2 Los pulmones (Concept) se hinchan (Action).

3 El asma (Concept) es una enfermedad (Concept).

4 Un ataque de asma (Concept) se produce (Action) cuando los síntomas (Concept) empeoran (Action).

5 Los síntomas (Concept) y el tratamiento (Concept) dependen (Action) del tipo (Concept) de cáncer (Concept) y de lo avanzada (Concept) que esté la enfermedad (Concept).



Comunidad NLP

Las conferencias más importantes del área son:

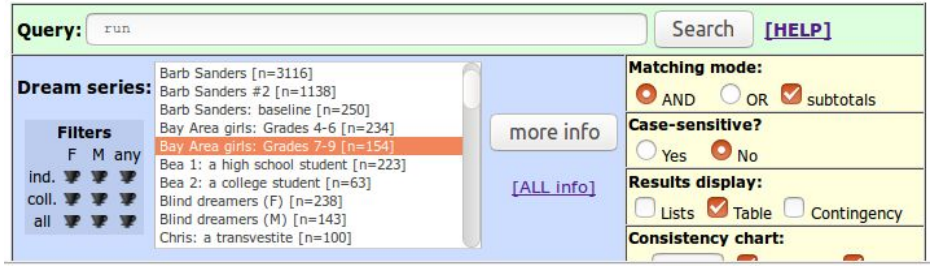
- ACL, NACCL, EMNLP, TACL, EACL, CoNLL,...
- Todos los paper están en ACL Anthology.
- También hay publicaciones de NLP en conferencias de ML y AI como: ICML, NIPS, ICLR, AAAI, IJCAI

Corpus

DreamBank
www.dreambank.net

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

Más de 20.000 reportes de sueños



	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%

Schneider, A., & Domhoff, G. W. (2016). The Quantitative Study of Dreams.

Corpus

50.000 reviews de IMDB
positivos/negativos

Corpus

50k de comentarios
Positivos/Negativos

Dubiau L. y Juan M Ale (2013) Análisis de Sentimiento sobre un Corpus en Español