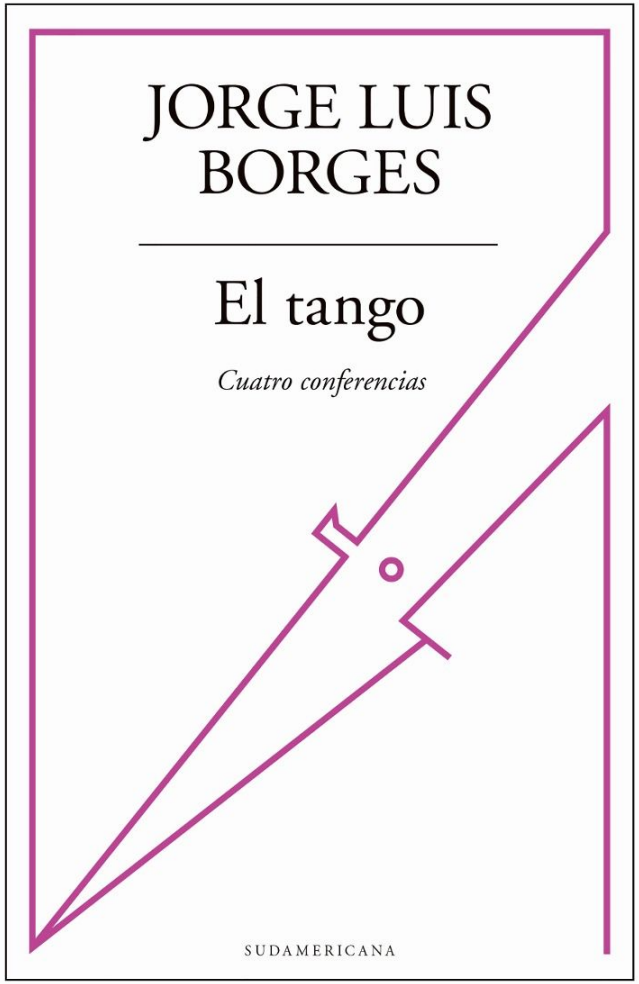
Testeando una hipótesis de Borges sobre el tango

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

Germán Rosati (CONICET-UNSAM, PIMSA)

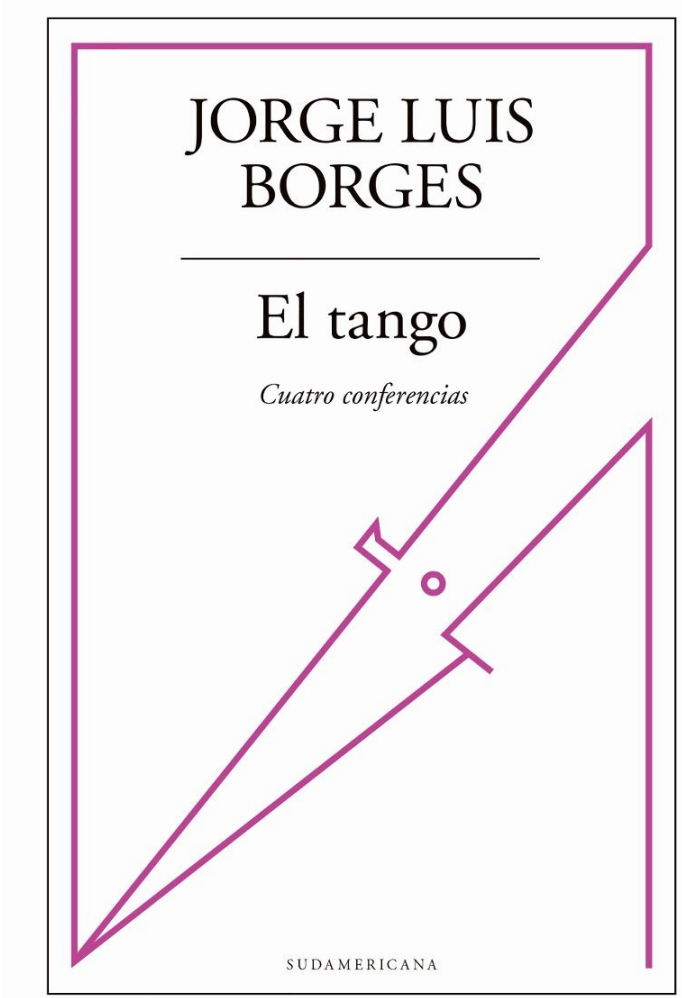
Hoja de ruta

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



“El tango, como hemos visto, empezó, surge de la milonga, y es al principio un baile valeroso y feliz. Y luego, el tango va languideciendo y entristeciéndose…”

III Conferencia, p.80-81


JORGE LUIS
BORGES
El tango
Cuatro conferencias
SUDAMERICANA

Enfoque tradicional

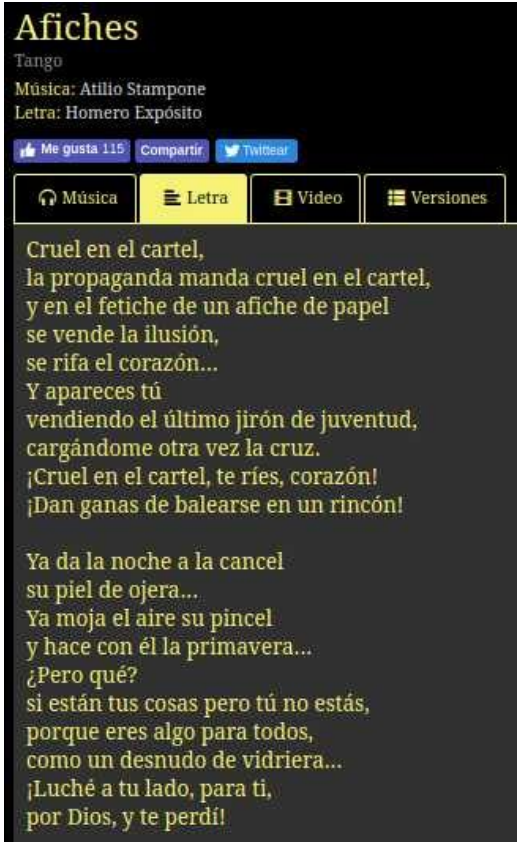
- Problema: analizar los temas de las letras de tango
- Enfoque “hermenéutico”: analizar pocas letras en profundidad
- Temas comunes: representaciones de género, figuras del “guapo”, representaciones del arrabal, etc.

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



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

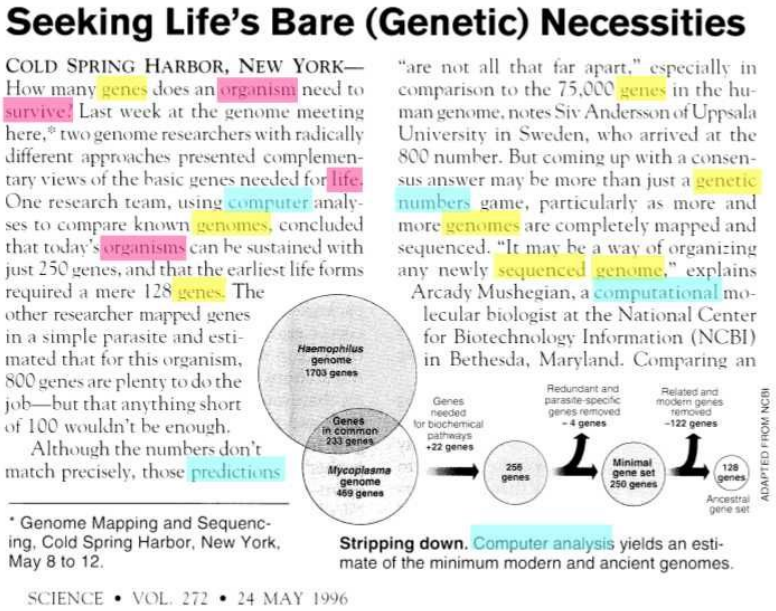
$$TF(t,d) = \frac{rc(t,d)}{\sum_{t \in d} rc(t,d)}$$

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

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

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

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 "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

* 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

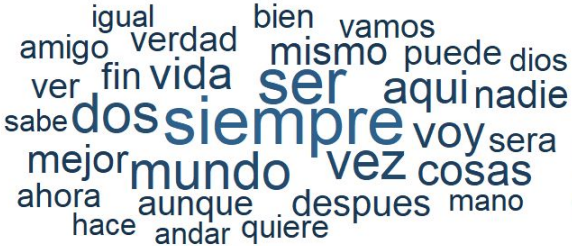
[Blei, 2012]

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

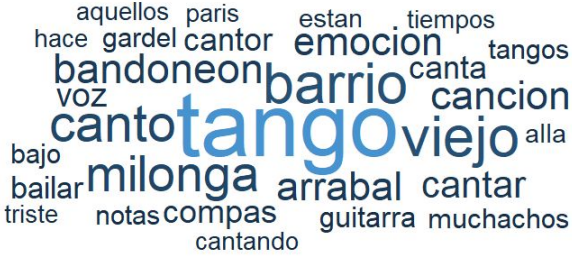
Los temas del tango: algunos resultados

Los temas del tango: algunos resultados

Misceláneo

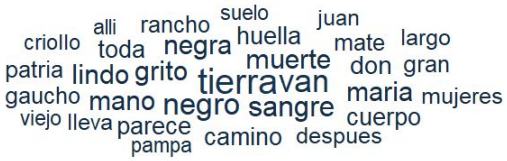


Tango + Arrabal

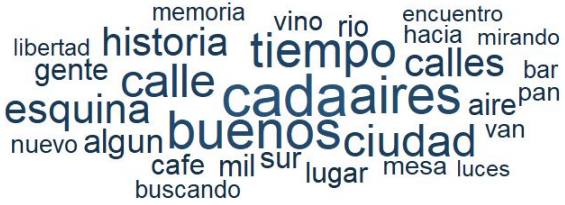


Misc. + Lunfardo

Campo, gauchesca

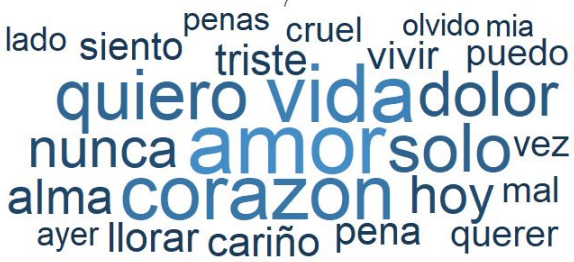


Ciudad, Buenos Aires

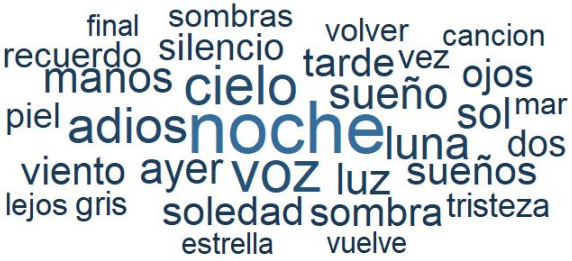


Amor signo +

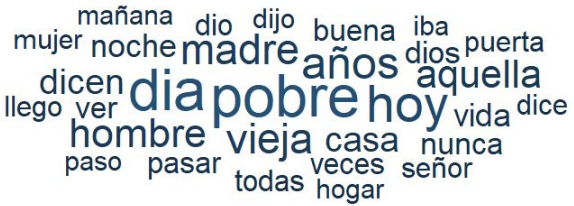
Amor signo -



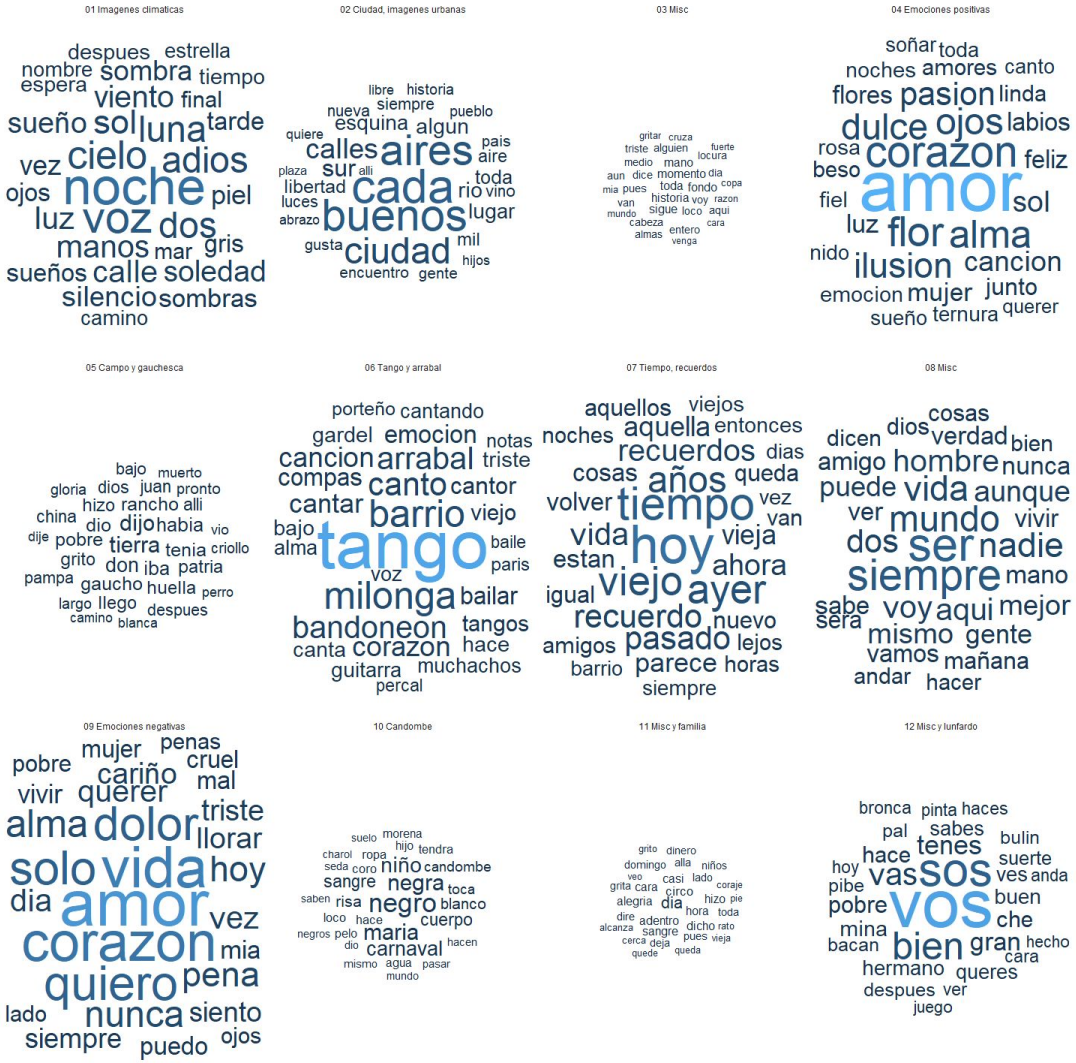
Imágenes climáticas



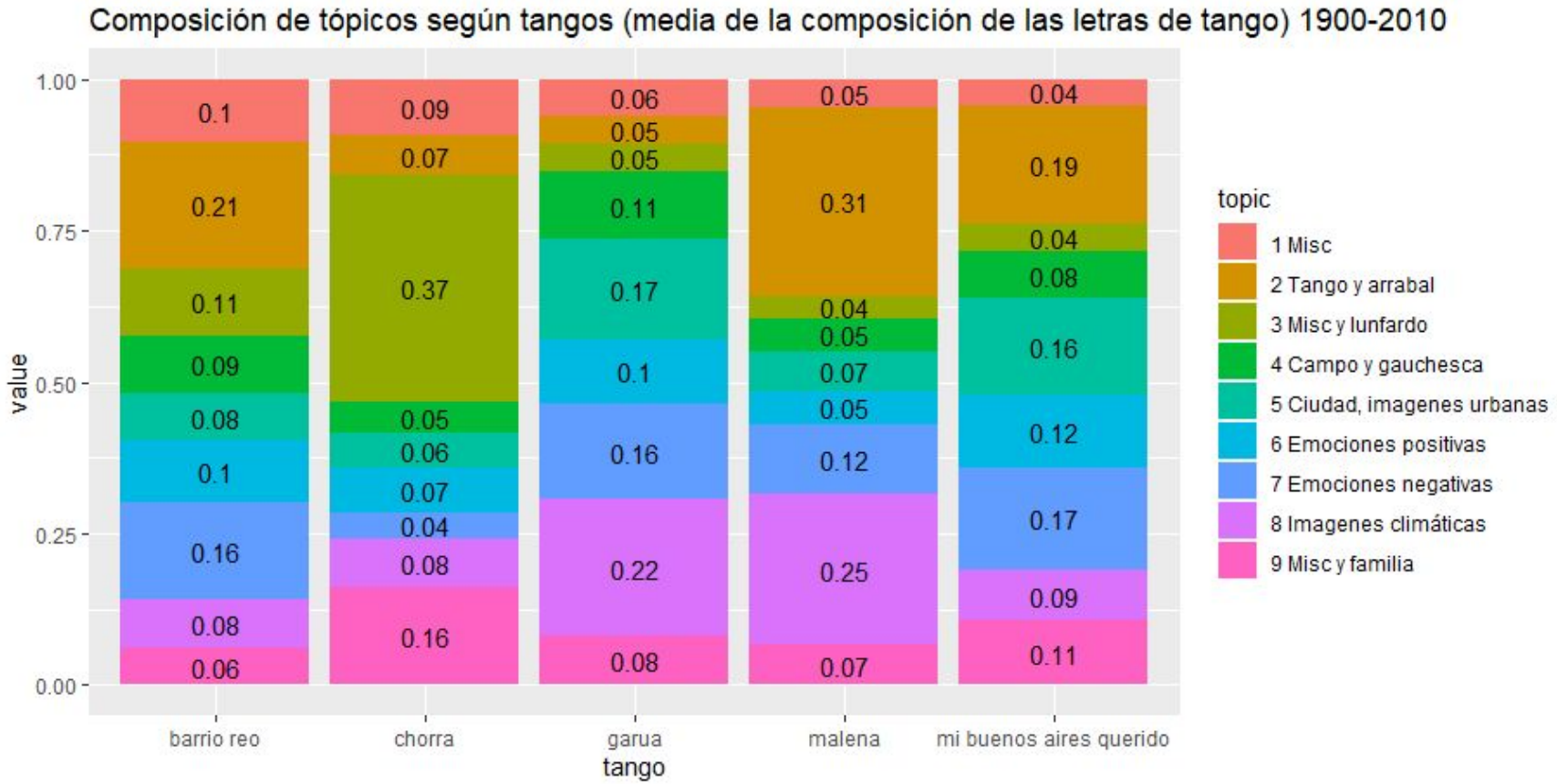
Misc. + familia



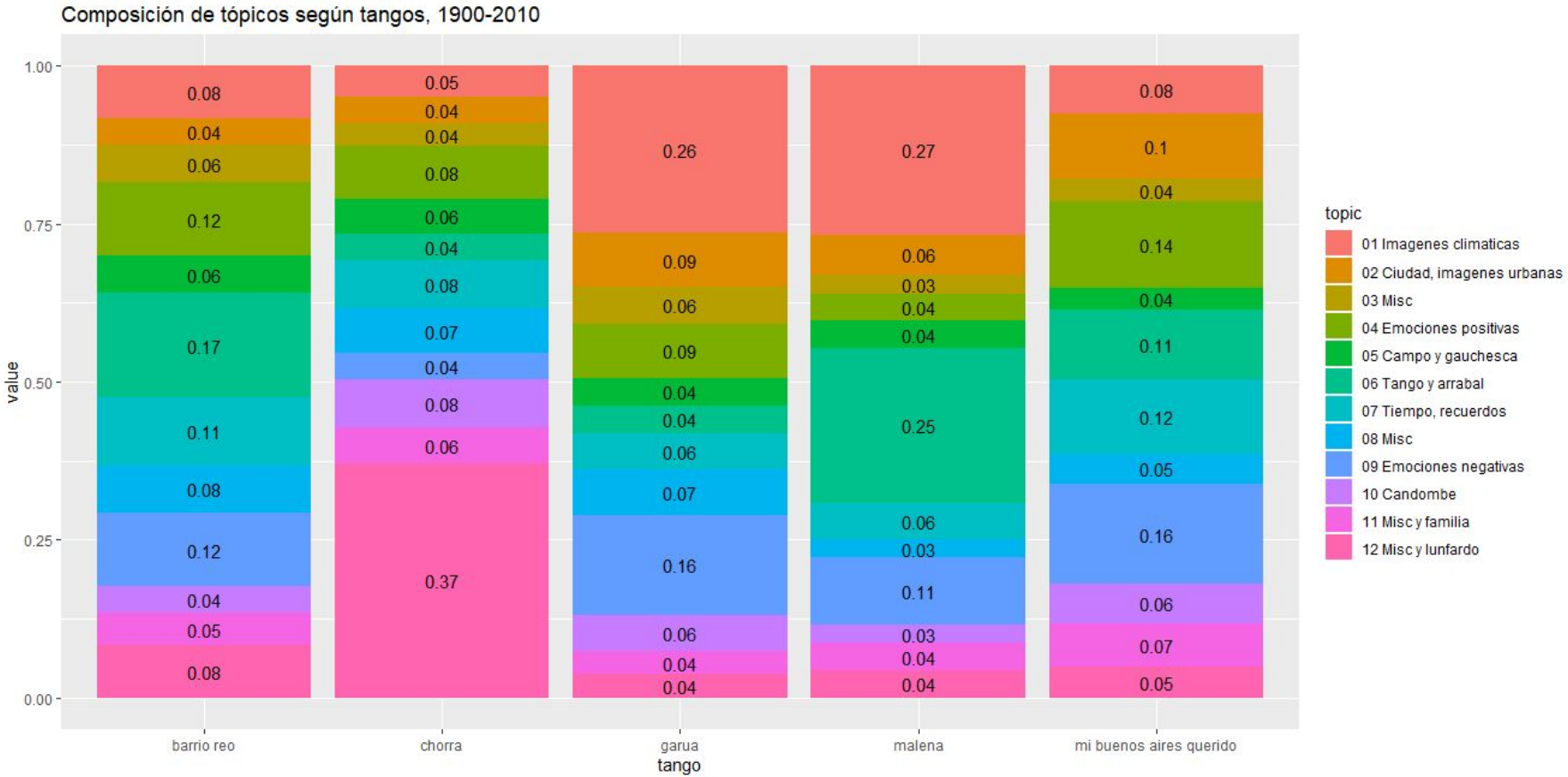
Los temas del tango:



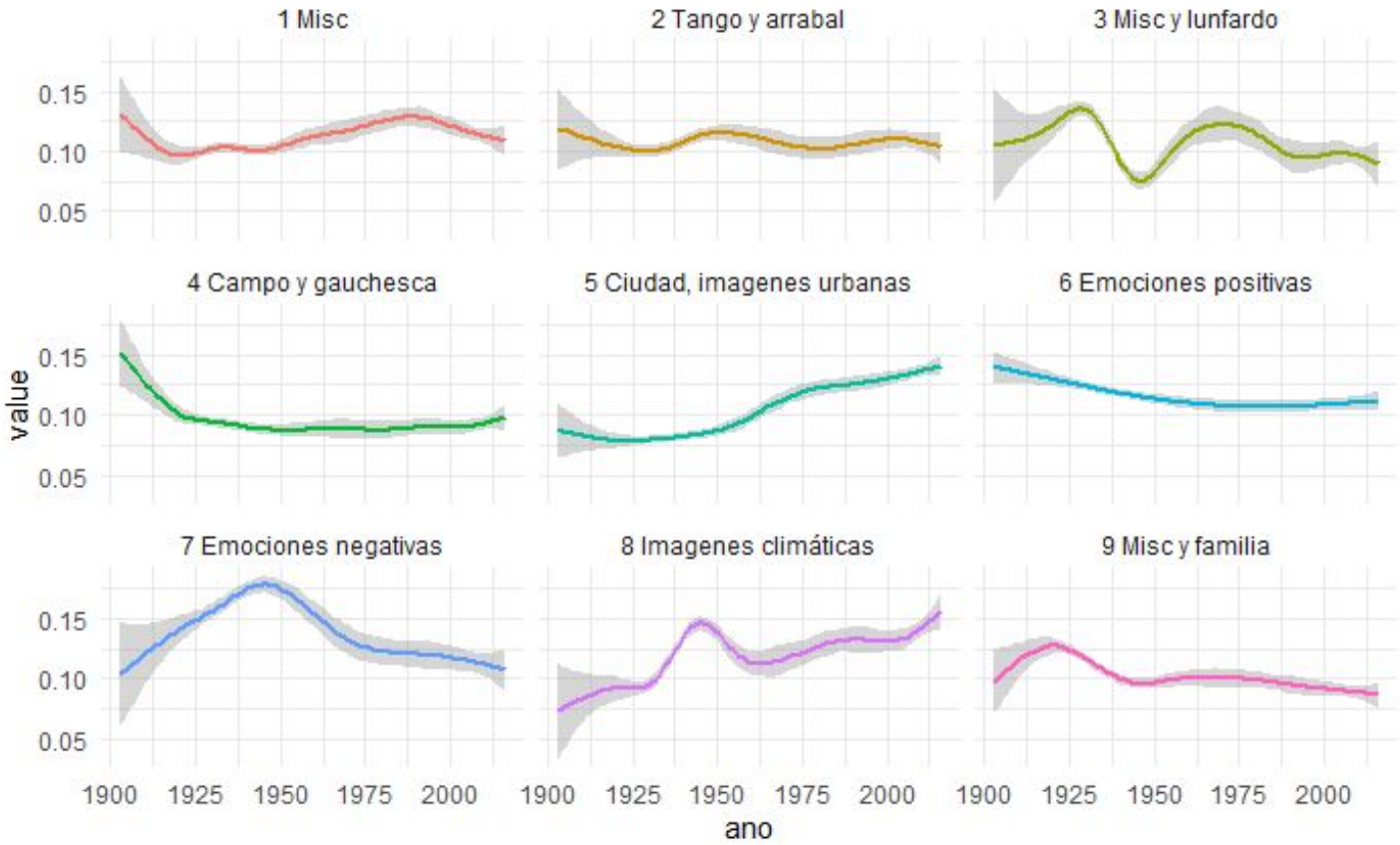
Los temas del tango: algunos resultados



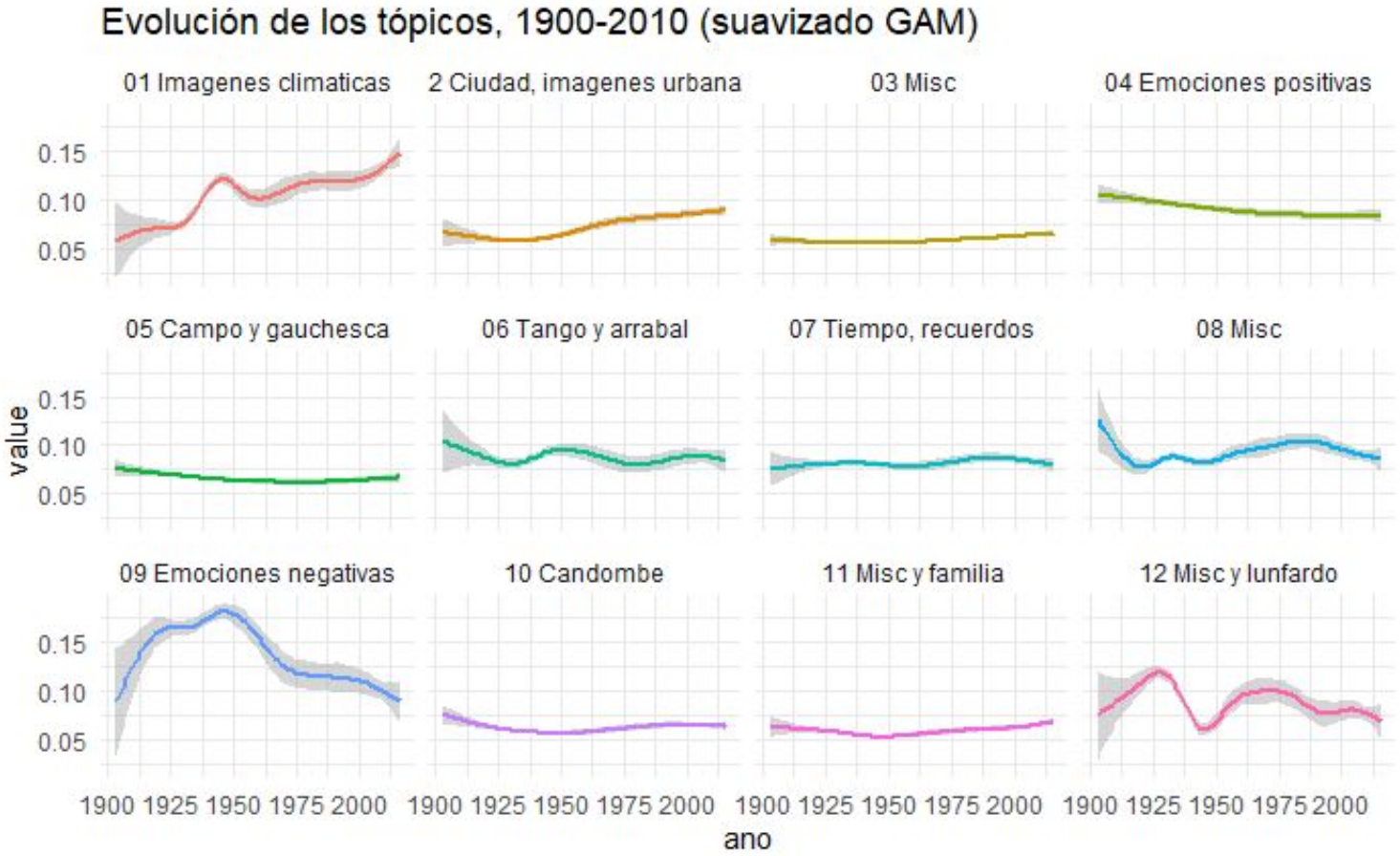
Los temas del tango: algunos resultados



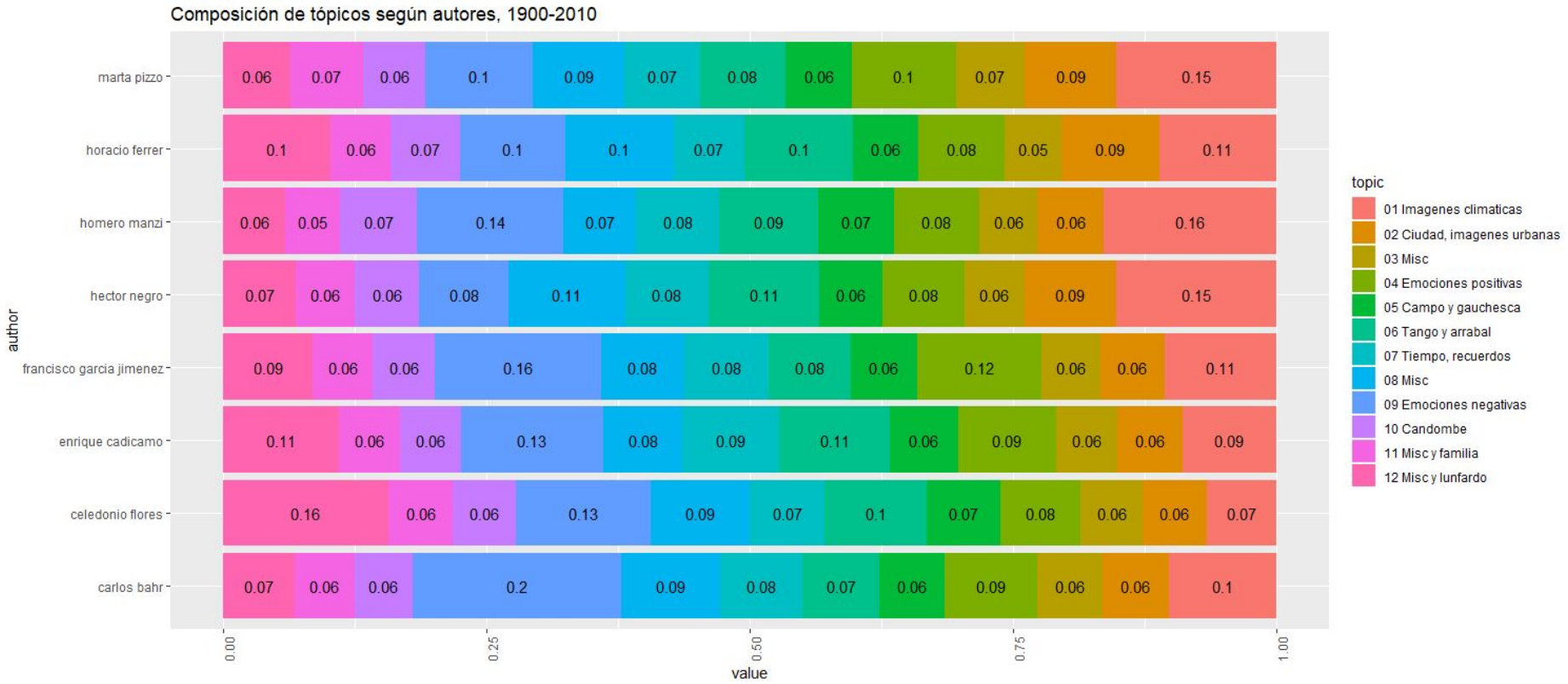
Los temas del tango: algunos resultados



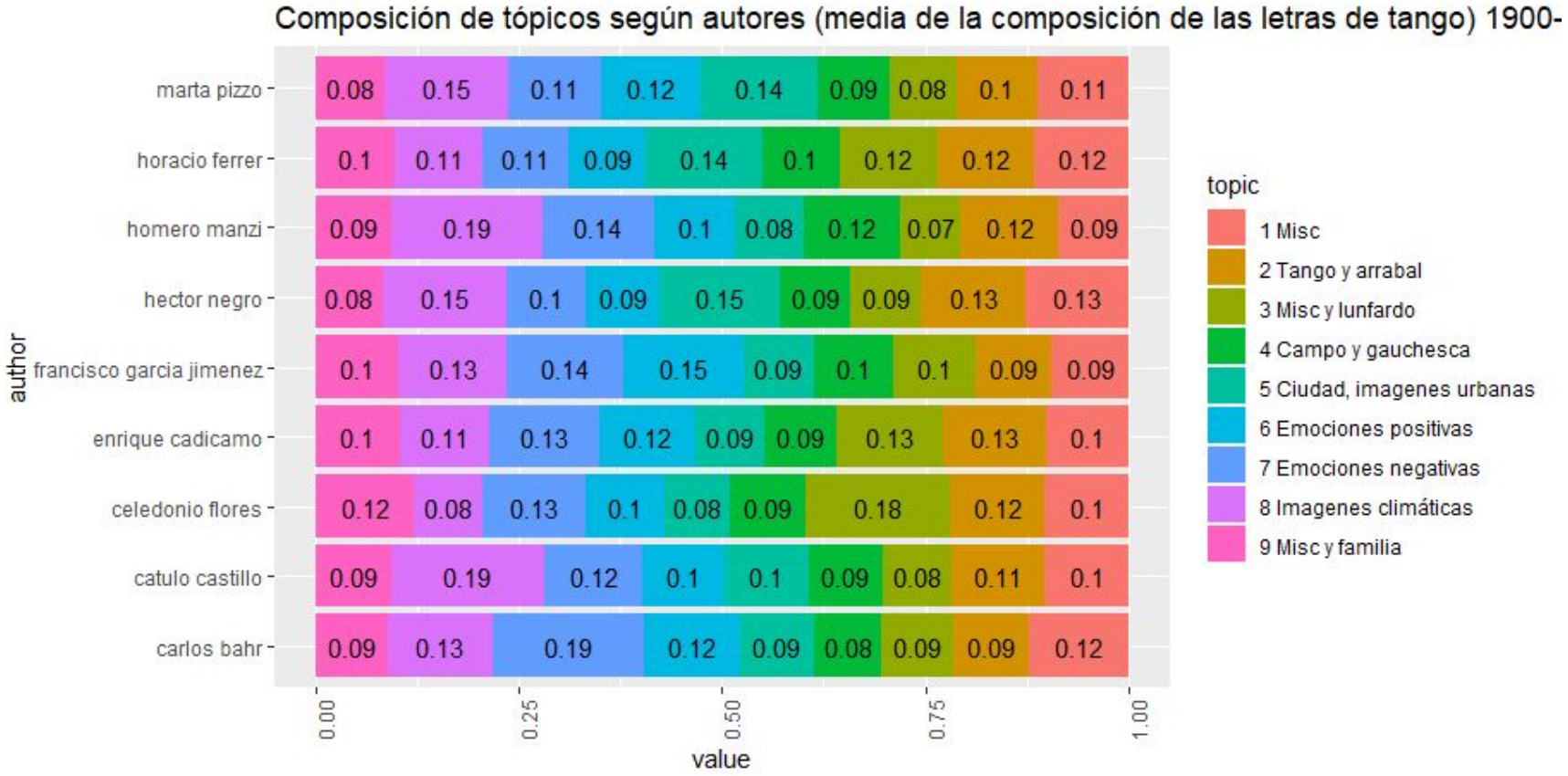
Los temas del tango: algunos resultados



Los temas del tango: algunos resultados



Los temas del tango: algunos resultados



¿Preguntas?

@Crst_C

german.rosati@gmail.com

https://gefero.github.io/