

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) sociólogos**: ¿Qué es LDA?
- Algunos resultados

JORGE LUIS
BORGES

El tango

Cuatro conferencias

SUDAMERICANA

“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...”

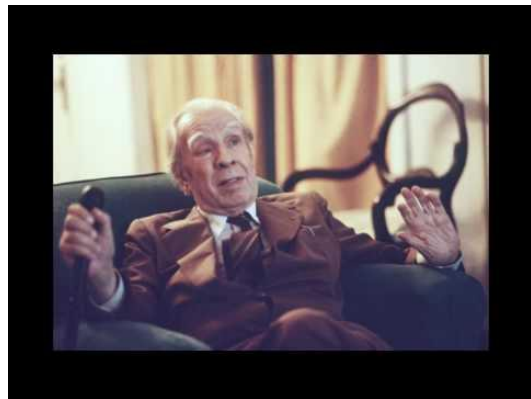
III Conferencia, p.80-81

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.



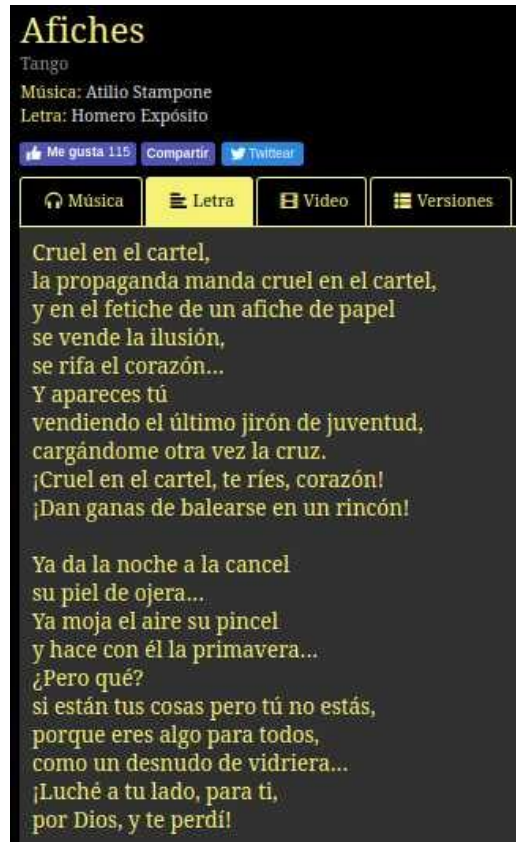
Enfoque tradicional

- 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

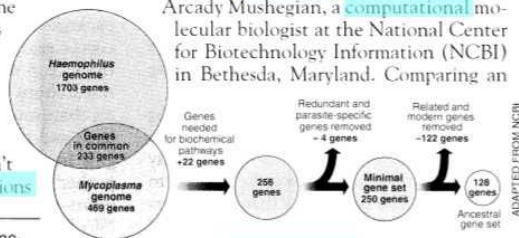
COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses 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**

* 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]

Los temas del tango: algunos resultados

Los temas del tango: algunos resultados

Misceláneo

igual bien vamos
amigo verdad mismo puede dios
ver fin vida **ser** aqui nadie
sabe **dos siempre** voy sera
mejor mundo vez cosas
ahora aunque despues mano
hace andar quiere

Tango + Arrabal

aquellos paris estan tiempos
hace gardel cantor emocion tangos
bandoneon **barrio** cancion
voz **canto tango** viejo
bajo milonga arrabal cantar
triste notas compas guitarra muchachos
cantando

Misc. + Lunfardo

hecho anda quieres bronca
sabes **hace vas** tenes bacan
bulin puro pinta **sos vos bien** suerte
juego ves gran hoy che mina pal falta
hermano haces cara

Campo, gauchesca

allí rancho suelo juan
criollo toda negra huella mate largo
patria lindo grito muerte don gran
gaucho mano negro sangre maria mujeres
viejo lleva parece pampa camino despues

Ciudad, Buenos Aires

memoria vino rio encuentro
libertad historia tiempo calles bar
gente calle cada aires aire pan
esquina buenos ciudad van
nuevo algun cafe mil sur lugar mesa luces
buscando

Amor signo +

nido risa alegría
fuego ternura beso boca flores cancion
querer mujer **corazon** vida
feliz flor **amor** labios
ilusion alma sol
linda dulce ojos pasion luz
soñar junto fiel amores rosa

Amor signo -

penas cruel olvido mia
lado siento triste vivir puedo
quiero vida dolor
nunca **amor solo** vez
alma **corazon** hoy mal
ayer llorar cariño pena querer

Imágenes climáticas

final sombras volver cancion
recuerdo silencio tarde vez ojos
manos cielo sueño sol mar
piel **adios noche** luna dos
viento ayer VOZ luz sueños
lejos gris soledad sombra tristeza
estrella vuelve

Misc. + familia

mañana dio dijo buena iba
mujer noche madre años dios puerta
dicen **dia pobre hoy** vida dice
llego ver hombre vieja casa nunca
paso pasar todas veces señor
hogar

Los temas del tango:

01 Imágenes climáticas

después estrella
nombre sombra tiempo
espera viento final
sueño sol luna tarde
vez cielo adiós
ojos noche piel
luz voz dos
manos mar gris
sueños calle soledad
silencios sombras
camino

05 Campo y gauchesca

bajo muerto
gloria dios juan pronto
china hizo rancho allí
dio dijo había vio
dije pobre tierra tenía criollo
grito don iba patria
pampa gaucho huella perro
largo llevo después
camino blanca

09 Emociones negativas

mujer penas
pobre carino cruel
vivir querer triste
alma dolor llorar
solo vida hoy
dia amor vez
corazon mia
quiero pena
lado nunca siento
siempre puedo ojos

02 Ciudad, imágenes urbanas

libre historia
nueva siempre pueblo
esquina algun pais
quiere calles aires aire
plaza sur ali toda
libertad cada rio vino
luces buenos lugar
abrazo gusta ciudad mil
encuentro gente hijos

06 Tango y arrabal

porteño cantando
gardel emocion notas
cancion arrabal triste
compas canto cantor
cantar barrio viejo
bajo baile
alma paris
voz tango
milonga bailar
bandoneon tangos
canta corazon hace
guitarra muchachos
percal

10 Candombe

sueño morena
charol ropa
seda coro niño candombe
sangre negra toca
saben risa negro blanco
loco hace cuerpo
negros pelo maria
dia carnaval hacen
mismo agua pasar
mundo

03 Misc

gitar cruza fuerte
triste alguien locura
medio mano
aun dice momento dia
mia pues toda fondo opa
van historia voy razon
mundo sigue loco aqui
cabeza entero cara
almas venga

07 Tiempo, recuerdos

aquellos viejos
noches aquella entonces
recuerdos dias
cosas años queda
volver tiempo vez van
vida hoy vieja
están ahora
igual viejo ayer
recuerdo nuevo
amigos pasado lejos
barrio parece horas
siempre

11 Misc y familia

grito dinero
domingo alla niños
veo casi lado coraje
grita cara circo hizo pie
alegría dia hora toda
dije adentro dicho rato
alcanza sangre pues vieja
cerca deja queda

04 Emociones positivas

soñar toda
noches amores canto
flores pasión linda
dulce ojos labios
corazon feliz
amor sol
luz flor alma
ilusion cancion
emocion mujer junto
sueño ternura querer

08 Misc

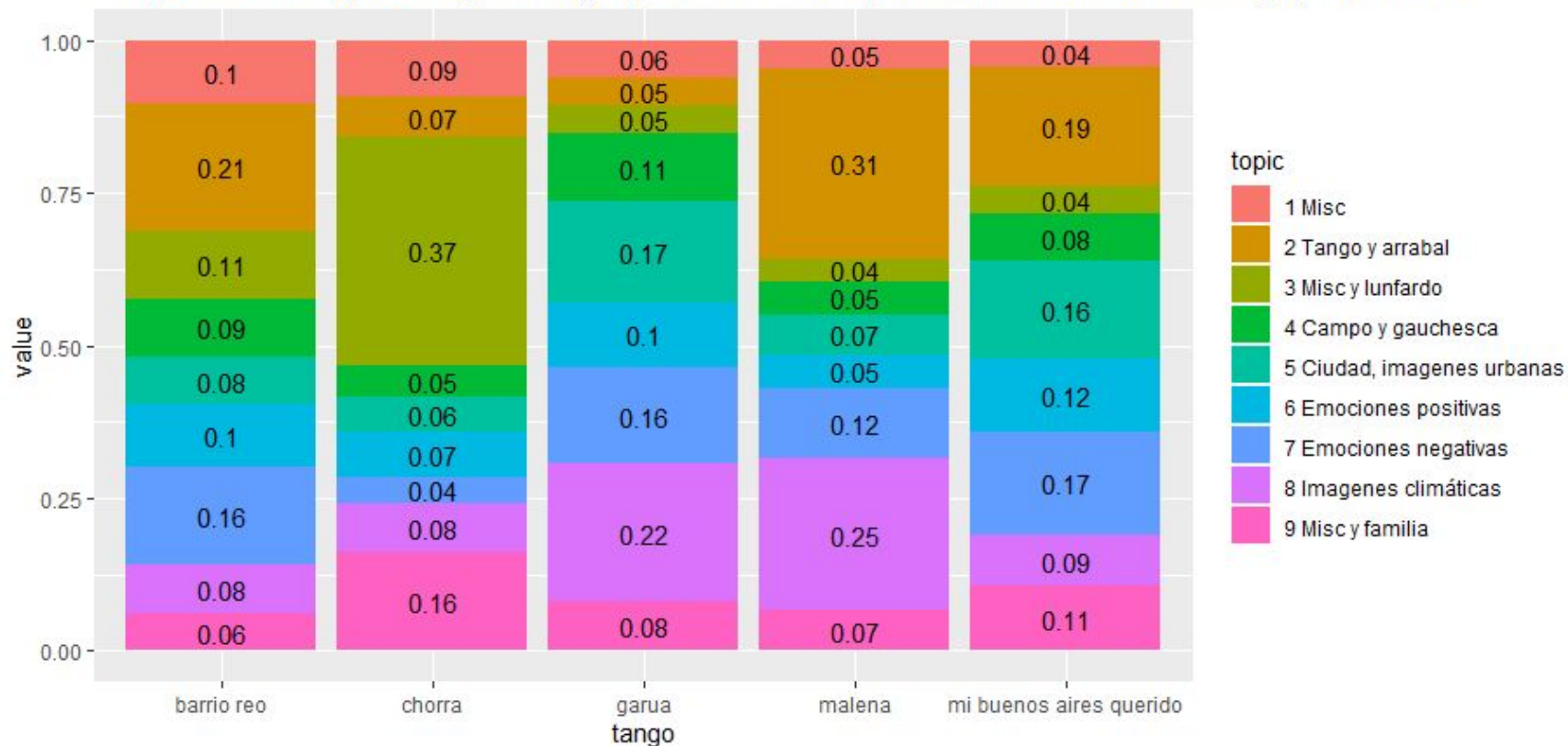
cosas
dicen dios verdad bien
amigo hombre nunca
puede vida aunque
ver mundo vivir
dos ser nadie
siempre mano
sabe voy aquí mejor
sera mismo gente
vamos mañana
andar hacer

12 Misc y lunfardo

bronca pinta haces
pal sabes bulin
hace tenes suerte
hoy vassos ves anda
pibe vos buen
pobre VOS che
mina bien gran hecho
bacan bien gran cara
hermano queres
después ver
juego

Los temas del tango: algunos resultados

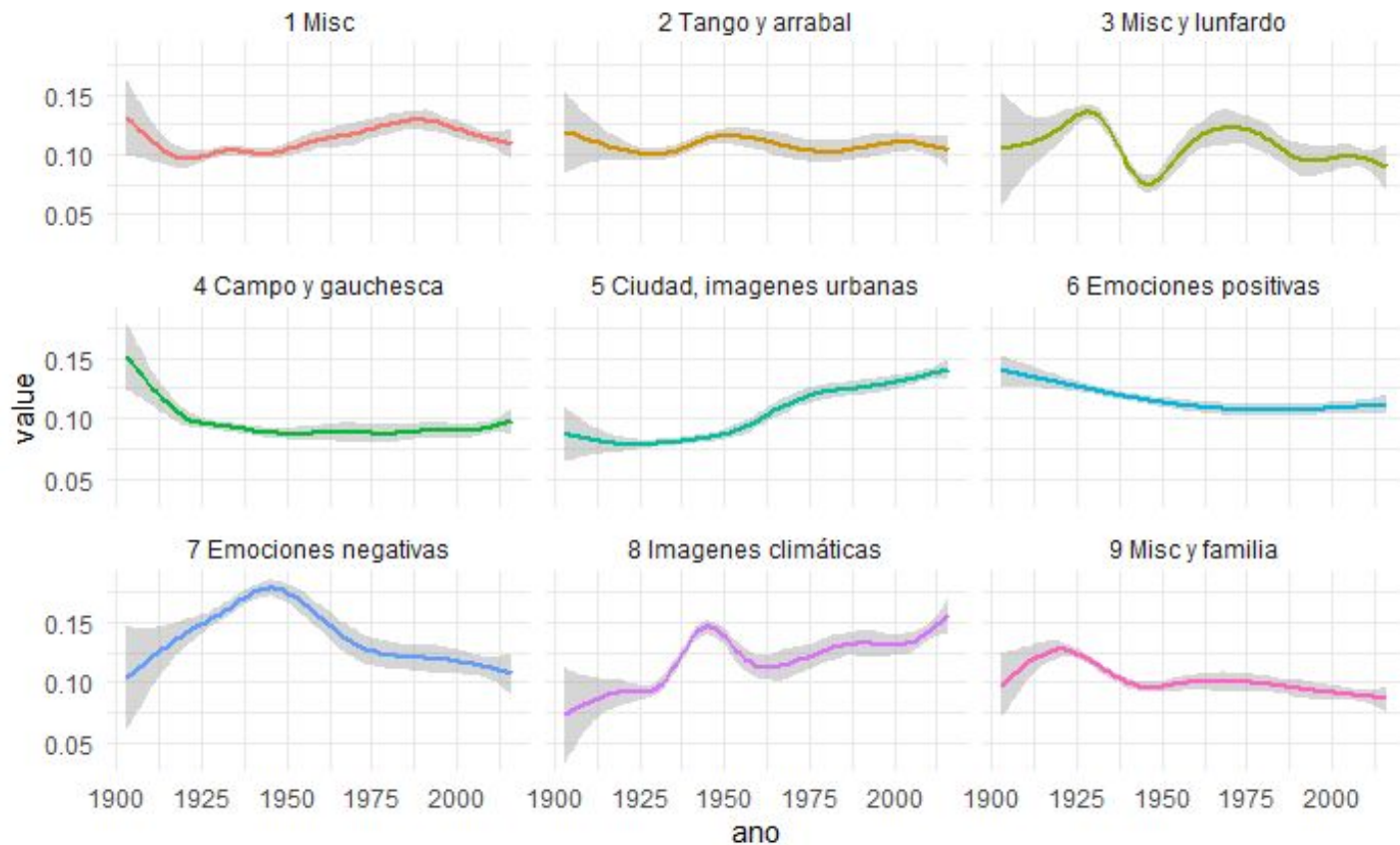
Composición de tópicos según tangos (media de la composición de las letras de tango) 1900-2010



Los temas del tango: algunos resultados

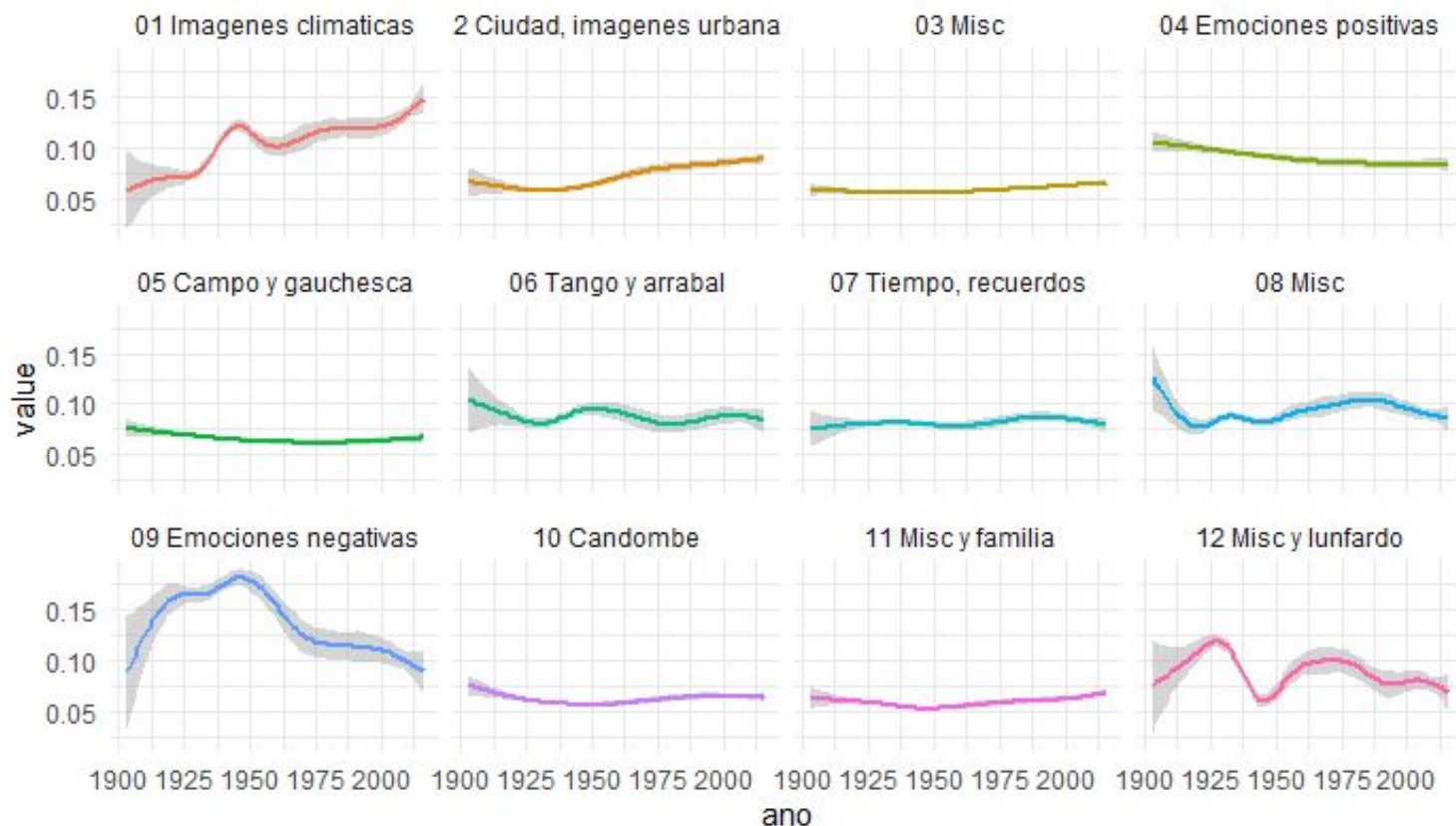


Los temas del tango: algunos resultados



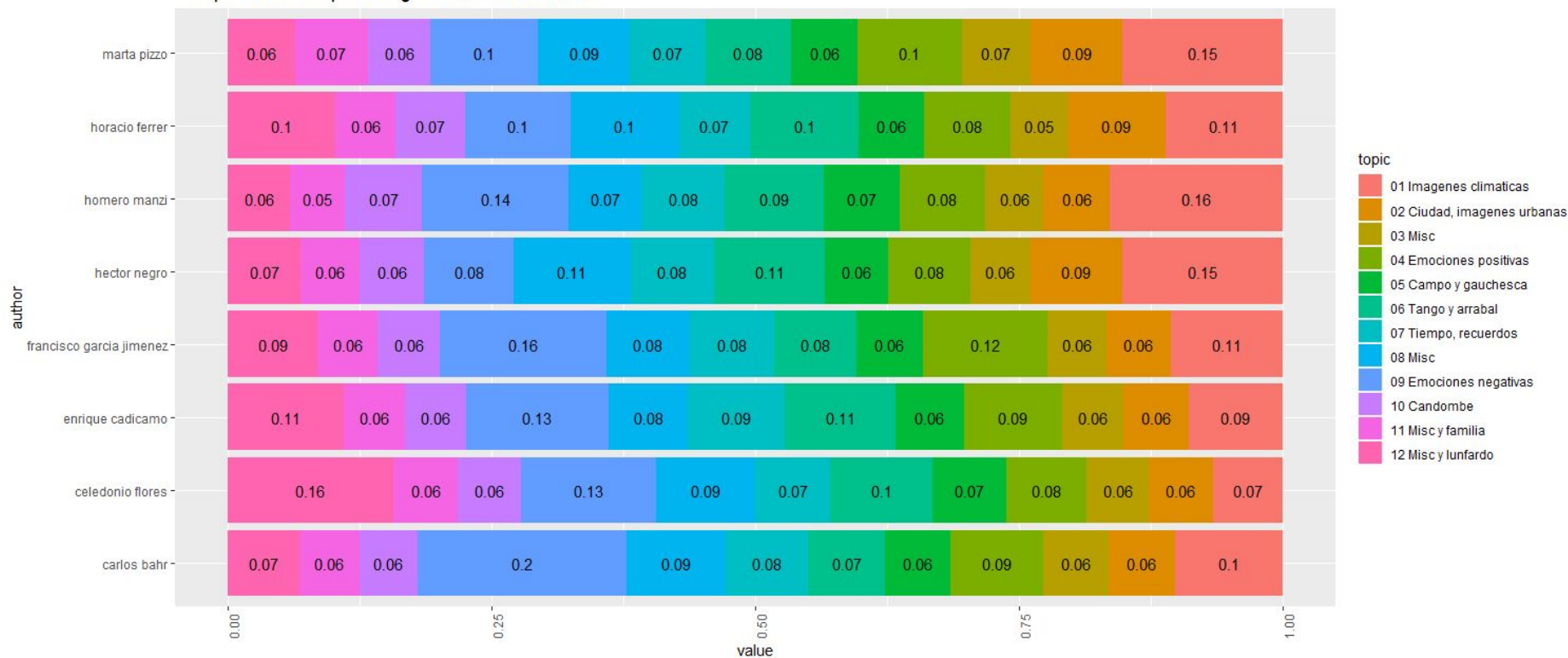
Los temas del tango: algunos resultados

Evolución de los tópicos, 1900-2010 (suavizado GAM)



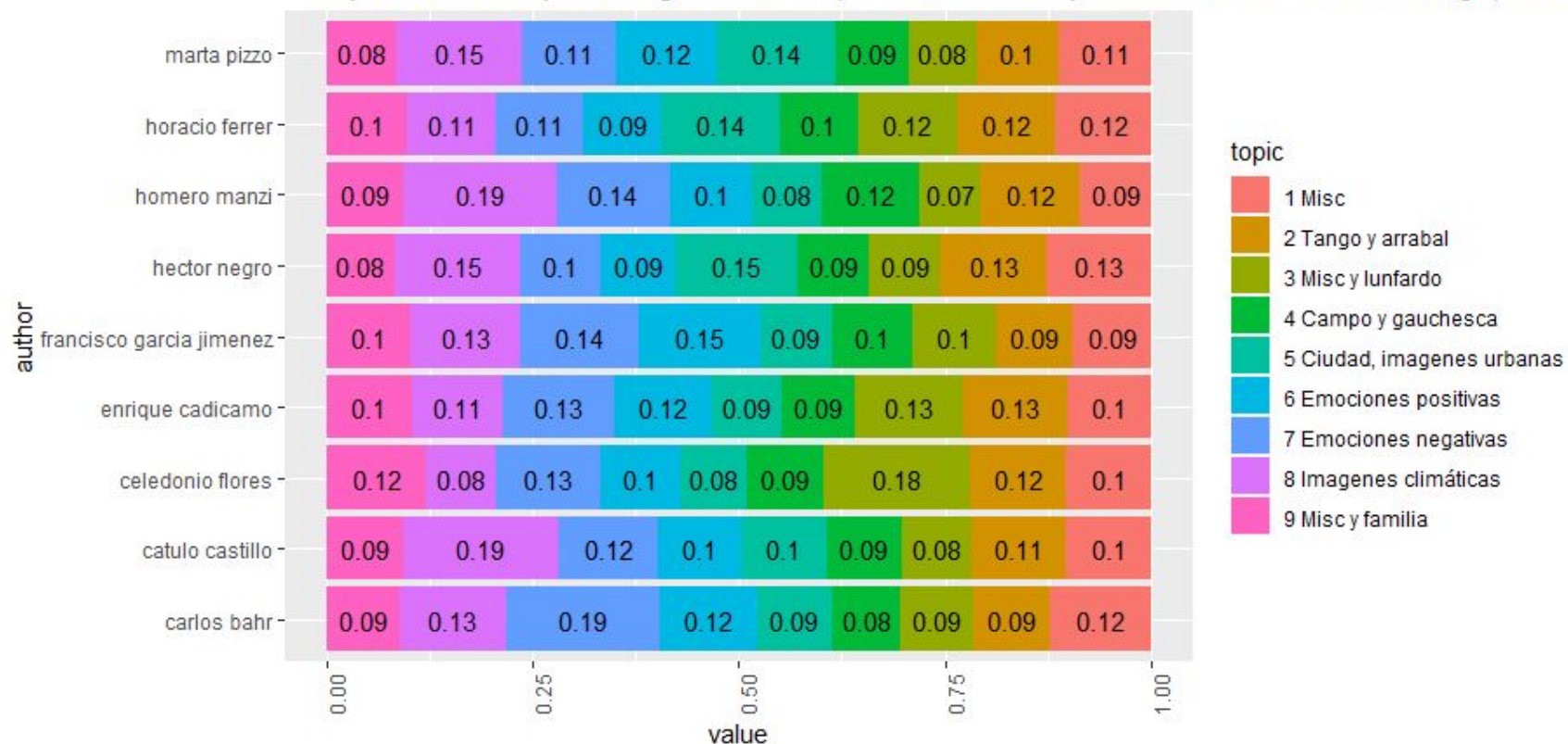
Los temas del tango: algunos resultados

Composición de tópicos según autores, 1900-2010



Los temas del tango: algunos resultados

Composición de tópicos según autores (media de la composición de las letras de tango) 1900-



¿Preguntas?



@Crst_C



german.rosati@gmail.com



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