Aplicación de Procesamiento de Lenguaje Natural en las ciencias sociales. Detección automática de tópicos en letras de tango

Germán Rosati (CONICET - IDAES/UNSAM - PIMSA)

Introducción

Independientemente de posiciones idealistas o materialistas al respecto, lo cierto es que buena parte de las interacciones a lo largo y a lo ancho de la estructura social, se encuentran mediadas por el lenguaje y tiene como producto una gran cantidad de "textos"", muchos de los cuales no son escritos, pero muchos otros sí.

Desde las ciencias sociales se ha hecho énfasis en esta idea, llegando a extremos teórico-metodológicos y afirmaciones temerarias tales como que la sociedad es un texto o que la cultura puede ser interpretada de la misma forma y con las mismas herramientas con que se aborda un texto literario. Uno de los referentes más relevantes de la antropología interpretativa escribe en uno de sus textos más famosos:

> The culture of a people is an ensemble of texts, themselves ensembles, which the anthropologist strains to read over the shoulders of those to whom they properly belong. (Geertz, 1974)[1]

Este enfoque ha ido acompañado de posiciones metodológicas más bien ligadas al análisis literario o filosófico: los métodos vinculados a la deconstrucción y a la hermenéutica constituyen algunos ejemplos. En la cita anterior resalta un requisito metodológico importante de este conjunto de enfoques: el objeto (la cultura) debe ser tratado por investigador como si fuera un texto, lo cual pone en un lugar central a la operación de "interpretación". Así, las herramientas y técnicas metodológicas de las que se suele echar mano en este tipo de enfoques se encuentran más vinculadas al análisis literario, al análisis del discurso y a métodos que buscan la comprensión del detalle y el contexto. Dichos enfoques han sido objeto de diversas críticas (Reynoso, 2007). No es objeto del presente trabajo evaluar dichas críticas y solo se remarcará un aspecto relevante a tener en cuenta: en términos generales, las posturas interpretativistas tienden a poner el eje en la capacidad subjetiva del investigador para realizar la interpretación. Esto hace que buena parte de las investigaciones realizadas bajo este tipo de marcos teóricos presenten (potencialmente) una relativa falta de sistematicidad metodológica y que el peligro de la imposibilidad de replicación de sus resultados esté siempre latente.

Un segundo punto a considerar en este tipo de abordajes se vincula al problema de la escala. Existen algunas aproximaciones basadas en la tradición hermenéutica que buscan llegar a un grado más alto de sistematización de los procesos y etapas del análisis. El Qualitative Content Analysis (Hsieh & Shannon, 2005), por ejemplo, busca sistematizar las diversas etapas y decisiones metodológicas en el proceso de análisis de datos no estructurados textuales. El problema que surge aquí es la escala: la transcripción y codificación manual de corpus textuales limita fuertemente el tamaño de los corpus a analizar.

Las técnicas *Text Mining* y *Natural Language Processing* (NLP) abordadas en este artículo (y muchas otras que no son mencionadas) pueden ser de utilidad a las ciencias sociales dado que permiten realizar una sistematización (y, eventualmente, lograr un cierto grado de automatización) de los diversos pasos de preprocesamiento de un texto y habilitan la aplicación de métodos cuantitativos de análisis para una amplia diversidad de tareas (clasificación de textos, detección de temas y tópicos, etc.). También abren la posibilidad de escalar el trabajo de forma eficiente. En lugar de leer cada uno de los textos de un corpus, tarea que rápidamente se vuelve imposible, las técnicas de minería de texto permiten analizar de forma automática corpus de escalas notablemente grandes.

El presente trabajo tiene como objetivo general discutir algunas aproximaciones metodológicas al anásis automático de textos y presentar algunas de las potencialidades que tienen en el trabajo cotidiano de las

[1] "La cultura de un pueblo es un ensamble de textos, los cuales son emsables a su vez, y que los antropólogos tratan de leer sobre los hombros de aquellos a los que les pertenecen."

ciencias sociales a partir de su aplicacion a un caso de estudio concreto: el análisis de los temas en un corpus de 5.600 letras de tango. Particularmente, nos centraremos en la discusión de algunos aspectos del flujo de trabajo para el análisis de texto y en la aplicación y discusión de una técnica específica para resolución de un problema general del Procesamiento de Lenguaje Natural: el modelo *Latent Dirichlet Allocation* (LDA) para la detección de tópicos en un corpus.

2. Antecedentes metodológicos en el análisis de letras de tango

El análisis de contenido en letras de tango no es un tema nuevo y existe una gran cantidad de literatura al respecto. Es importante tener en cuenta que el objetivo central del artículo es ilustrar la aplicación de un proceso de trabajo y de algunas técnicas de análisis vinculadas al campo del NLP. Se centra, entonces, en la dimensión metodológica del problema. Es por ello que los textos reseñados esta sección son ilustrativos de dicha dimensión.

Un abordaje común es el rastreo de un tópico o problema particular a lo largo de un corpus textual. En el caso del análisis de letras de tango un tópico habitual se vincula al problema del género. Irene López (2010) aborda cuáles son las representaciones que se hacen de las mujeres sobre un corpus de aproximadamente 10 letras de tango de diferentes épocas y autores. A su vez, Juan Gasparri (2011) busca identificar las formas en que las masculinidades (particularmente, la del "guapo" en sus diversas construcciones) aparecen en un corpus de alrededor de 16 textos. También, teniendo como eje la problemática de género, Carolina Marchese (2007) rastrea la temática amorosa y el sesgo sentimental en aproximadamente 30 tangos de fines del siglo XIX y principios del XX[2].

A su vez, otros textos tienen un enfoque más amplio y tratan de relevar una mayor cantidad de temas o problemas en las letras de tango. Así, Lucía Willenpart (2011) rastrea e identifica algunos temas comunes: el amor, el duelo amoroso, la mujer, la madre, el tango mismo, etc.

Por otro lado, Cantón (1972) se pregunta por los objetos y sujetos de los tangos cantados por Carlos Gardel. Este estudio tiene un carácter cuantitativo, por lo cual analiza un corpus significativamente más grande que los anteriores: alrededor de 100 tangos.

De esta forma, en términos metodológicos es posible identificar tres rasgos de los textos reseñados:

1. no aparecen criterios claros para la confección del corpus, con todos los potenciales segos y problemas que esto conlleva
2. el tamaño de los corpus es entre pequeño (10 tangos) y mediano (100 tangos)
3. con la excepción del texto de Cantón (1972), los textos son abordados a partir del rastreo de un tópico o pregunta particular en profundidad, privilegiándose un análisis interpretativo del contenido.

3. Construcción del corpus y flujo de trabajo

A los efectos de ilustrar un caso de aplicación de este tipo de técnicas, se presentará un flujo de trabajo clásico aplicado a la detección de tópicos en un corpus textual de letras de tango. El corpus fue construido a partir de un proceso de web scraping[3] Se descargaron todas las letras de tango disponibles en el sitio todo tango. Además de las letras se descargó información accesoria sobre el tango en cuestión. Para ello se confeccionaron dos web crawlers sencillos[4].

El corpus final consiste en alrededor de 5.600 letras de tango, agrupadas en un dataset con la siguiente estructura:

Tabla 1. Ejemplo de base de datos utilizada

[2]El tamaño de los corpus está referido a la cantidad de textos que citan o mencionan los diferentes autores en cada uno de los trabajos citados. Lógicamente, es razonable asumir que solamente se cita una fracción -desconocida- de los textos analizados.

[3]El *scraping*-literalmente, "raspado" o "rascado" consiste en la descarga y formateo de la información disponible en sitios web, información que generalmente no se encuentra en condiciones de ser trabajada de forma estadística (Mitchell, 2015).

[4]El código puede ser consultado en nuestro repositorio https://github.com/gefero/tango_scrap. A su vez, todo el material de replicación para el análisis de tópicos puede ser encontrado en el siguiente repositorio https://github.com/gefero/topic_modeling_tango

Título	Ritmo	Año	Compositor	Autor	Letra
A bailar	Tango	1943	D. Federico	H. Expósito	a bailar a bailar que la orquesta se va…
…	…	…	…	…	…
Malena	Tango	1941	L. Demare	H. Manzi	malena canta el tango como ninguna…
Zurdo	Tango	S/D	A. Pontier	F. Silva	era del tiempo lindo que siempre es antes…

Al observar el campo *Letra* (el núcleo del análisis), se nota que se trata de un caso típico de datos no estructurados: las letras constituyen texto libre y no parece respetarse la estructura tripartita del dato. Las filas sí representan una unidad (los tangos) pero no tenemos atributos, o en todo caso, solamente tenemos un atributo, la letra. El primer paso, entonces, poder pasar de esta representación no estructurada a una estructurada. Esta representación tendrá como objeto reducir la complejidad del texto, dado que "el lenguage es complejo. Pero no toda su complejidad es necesaria para analizar un texto de forma efectiva" (Grimmer & Stewart, 2013).

3.1 Modelo *BoW* (*Bag of words*)

Para llegar a esa representación estructurada[5], será necesario pensar en una estructura de datos acorde a las necesidades del análisis. La unidad de análisis serán los tangos individuales, por lo cual, cada fila en la matriz final será un tango. A su vez, la representación a utilizar será la siguiente: cada columna consistirá en un término t del vocabulario general V del corpus C. Dada celda estará constituida por el conteo crudo de ocurrencias de cada palabra (columna) en cada documento (fila). Esta representación es la que se denomina *Bag of Words* o "bolsa de palabras" y se dispone en una Matriz de Frecuencias de Términos (TFM, por sus siglas en inglés).

Por ejemplo, para dos de los tangos analizados esta matriz tomaría la siguiente forma:

Tabla 2. Ejemplo de matriz de frecuencia de términos cruda

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

Puede verse que al construir esta matriz la información sobre el orden de las palabras se ha perdido. En efecto, el orden de las columnas es ahora arbitrario (en este caso, lexicográfico) y no respeta la estructura secuencial de las palabras en un texto. Esta es una simplificación importante del modelo BoW. Esta limitación puede subsanarse parcialmente generando una TFM de bi-gramas (pares de palabras), tri-gramas (tripletas de palabras) o n-gramas. El costo es un crecimiento exponencial en la dimensionalidad de la TFM

Debe tenerse en cuenta que esta matriz se construye sobre el vocabulario V, o sea el total de términos únicos sobre el corpus C. Así, V incluiría a priori signos de puntuación, diferentes conjugaciones de verbos, sustantivos en singular o plural, etc. Esto hace que el vocabulario "crudo" de C tienda a ser demasiado grande. Es por ello que, en la etapa de preprocesamiento del texto se utilizan algunas técnicas para reducir la complejidad y la extensión de V.

Un paso simple es la eliminación de lo que suelen denominarse *stopwords*, básicamente artículos, preposiciones, conectores, etc. La lógica detrás de esta eliminación es que estas palabras se encuentran en todos los documentos d de C por lo cual aportan poca información acerca de su contenido.

Tabla 3. Ejemplo de matriz de frecuencia de términos sin stopwords

[5]Vale destacar que el flujo de trabajo descripto a continuación es uno posible -y bastante común-, pero de ninguna manera el único, ni necesariamente el "mejor" en términos absolutos. El flujo y las operaciones contenidas en el mismo deberán ser revisadas para cada problema particular (Grimmer & Stewart, 2013).

Letra	agua	blanda	cartel	cruel	manda	mas	propaganda
Cruel en el cartel la propaganda manda cruel en el cartel	0	0	2	2	1	0	1
Era más blanda que el agua que el agua blanda	2	2	0	0	0	1	0

A partir de la eliminación de las *stopwords* se obtiene en la tabla 3 una representación más resumida de la información contenida en C. No obstante, ésta no es la única operación disponible para reducir la complejidad de C.

3.2 Normalización de términos: *stemming* y lematización

El paso siguiente consiste en la reducción de la diversidad de los términos t de V manteniendo su sentido. Existen dos técnicas básicas para lograr este resultado:

1. *stemming*: remueve las declinaciones de las palabras con el objetivo de reducir la dimensionalidad de V. Aquellas palabras que remiten a un mismo concepto básico son reducidos a la misma raíz. Por ejemplo, familias, familia y familiar son reducidas a familia
2. lematización: Tiene el mismo objetivo y lógica que el *stemming*. La diferencia es que usa diccionarios, el contexto de las palabras y su función sintáctica para determinar su raíz. Así, logra discernir que mejor y mejorable remiten a la misma raíz bueno.

La diferencia entre ambas suele ser el tiempo de cómputo. En términos generales, el *stemming* tiende a ser más rápido, dado que solamente requiere de reglas para truncar las palabras. Existen diferentes algoritmos para varios idiomas (algoritmo Porter, algoritmo Snowball, etc.).

3.3 Normalización de conteos

El último paso supone normalizar los valores de las celdas de la TFM. Al momento de filtrar los stopwords se buscaba poder eliminar aquellas palabras muy frecuentes en todos los textos. Es posible extender este razonamiento para el resto de los términos de V. Así, pueden identificarse dos dimensiones de la frecuencia de dichos términos:

1. Un término t es importante si es muy frecuente en un documento d de C
2. A su vez, t es más informativo del contenido de un documento d si está presente en pocos documentos y no en todos los documentos de C.

Es decir, resulta importante analizar la frecuencia de t tanto en el documento d como en el corpus total C. Existen dos métricas para lograr este objetivo. Para la primera dimensión se parte del conteo crudo de t en d: $c(t, d)$, es decir, cada celda de la TFM "cruda". Es posible definir, entonces, una métrica llamamda *Term Frequency* (TF) es decir, el conteo crudo normalizado por la extensión del documento (el total de términos en el documento):

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

En relación a la informatividad de un término a lo largo de C, podemos definir la siguiente métrica, llamada *Document Frequency* (DF):

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

donde $df(t)$ es la cantidad de documentos en C que contienen a t; $|C|$ es el tamaño del corpus.

De esta forma, DF informa acerca de la proporción de documentos que contienen a t. Cuanto mayor es $DF(t)$ menos informativo es t. Es por ello que se usa la inversa de esta métrica:

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

$IDF(t)$ entonces, es mayor, cuanto menor es la frecuencia de t en C, es decir, cuanto más informativo es t.

Podemos combinar ambas dimensiones en una métrica resumen, llamada $TFIDF$, *Term Frequency-Inverse Document Frequency*:

$$TF - IDF(t) = \frac{c(t,d)}{\sum_{t \in d} c(t,d)} \times \log \frac{|C|}{df(t)} = TF(t,d) \times IDF(t)$$

Así, valores altos de $TF(t,d)$ y valores altos de $IDF(t)$ -o sea, valores bajos de $DF(t)$- arrojan valores altos de $TF - IDF(t)$. O sea, términos t frecuentes en d y poco frecuentes en C.

No obstante, estas operaciones mencionadas, en ciertos casos (como el del corpus en cuestión) suele suceder que el conteo crudo $c(t,c)$ de términos fucione de forma aceptable. En el caso específico de este trabajo, luego de evaluar ambas alternativas se optó por utilizar $c(t,d)$ como métrica en la TFM[6]. Para resumir, entonces, el preprocesamiento realizado para este corpus:

1. normalización a minúsculas
2. eliminación de *stopwords*
3. eliminación de puntuación y
4. eliminación de caracteres extraños y dígitos

4. Detección de tópicos

Existen varias técnicas para la detección automática de tópicos en corpus textuales, algunas de las cuales están basadas en alguna forma de descomposición de la TFM[7]. Para el presente trabajo se utilizará una de las más conocidas: *Latent Dirichlet Allocation* o LDA.

La intuición detrás de LDA (D. Blei, 2012) es que cada d de C puede exhibir varios tópicos, es decir, puede hablar de varios temas simultáneamente. Por ejemplo, al analizar un tango como "Malena" de Homero Manzi, se observa que habla de diferentes temas: del amor, del tango, del barrio, etc. La idea detrás de LDA es poder operacionalizar esta intuición a través de un modelo generativo, es decir, asume la existencia de un "proceso generador de textos".

Más formalmente[8] un tópico se define como una distribución de probabilidad a lo largo de un vocabulario V fijo. Por ejemplo, si existiera un tópico como *sentimientos o emociones* palabras como "amor", "pena", "sufrimiento", deberían tener altas probabilidades para este tópico. En cambio, palabras como "ciudad", "barrio" estarían más asociadas a un tópico que hable acerca de la *ciudad*.

Ahora bien, para cada documento d en el corpus C se generan las palabras w que lo componen en un proceso de dos etapas:

1. Se selecciona de forma aleatoria una distribución de tópicos para d
2. Para cada palabra (w) en d

i) se selecciona aleatoriamente un tópico de la distribución general de tópicos
ii) se selecciona aleatoriamente una palabra correspondiente a la distribución de todo el vocabulario V

De esta forma, cada documento d exhibe ciertos tópicos t en diferente proporción (paso 1.) cada palabra w es extraída de uno de los tópicos (paso 2.ii), donde el tópico selecionado es elegido de la distribución de tópicos de ese documento d particular (paso 2.i).

[6] Para una discusión al respecto de estas métricas, puede verse Wiedemann (2016)

[7] Existen otros métodos para la detección de tópicos, basados más bien en la descomposición de la TFM en dos componentes, una matriz de $documentos \times tpicos$ y una matriz de $trminos \times tpicos$, tales como *Non Negative Matrix Factorization* y *Latente Semantic Analysis* (Hassani, Iranmanesh, & Mansouri, 2019).

[8] Esta sección se basa en (D. Blei, 2012)

Así, el objetivo del modelado de tópicos es descubrir de forma automática los temas a los que alude un determinado conjunto de documentos. Ahora bien, como puede intuirse esa estructura de tópicos puede ser pensada como un set de variables latentes a la TFM. Lo único observado es el conjunto de documentos (preprocesado como una TFM). La estructura de tópicos (es decir, la composición de tópicos por documento y la asignación de palabras a un documento) puede ser considerada como un conjunto de variables no observadas (justamente es lo que se trata de estimar). Formalizando el razonamiento anterior, es posible ver que se trata de una probabilidad conjunta:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d)(\prod_{n=1}^{N} p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{1:K}, z_{d,n}))$$

Así, los tópicos están indexados en $\beta_{1:K}$, donde cada β_k es una distribución de probabilidad sobre el vocabulario. La proporción de tópicos para el d-ésimo documento están indexadas por θ_d, dónde $\theta_{d,K}$ es la proporción del tópico d en el documento k. La asignación de tópicos para el documento d, está dada por z_d, donde $z_{d,n}$ es la asignación de tópico para la n-ésima palabra en el documento d. Finalmente, las palabras observadas para el documento d, son w_d, donde $w_{d,n}$ es la n-ésima palabra en el documento d, que es un elemento del vocabulario fijo.

Puede verse que existen ciertas dependencias en el modelo: por ejemplo, la asignación de tópicos $z_{d,n}$ depende de la proporción de tópicos por documento θ_D. A su vez, la palabra observada $w_{d,n}$ depende de la asignación de tópicos $z_{d,n}$ y todas dependen de los tópicos β_k.

Entonces, el problema es poder estimar la estructura de tópicos a partir de los documentos observados. De esta forma, es posible formular el problema a partir del llamado "posterior":

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

El numerador es la distribución conjunta de todas las variables aleatorias del modelo y puede ser calcuado de forma simple. El problema reside en el denominador: la probabilidad marginal de las observaciones, es decir, la probabilidad de observar el corpus dado bajo cualquier modelo de tópicos. Si bien, debería (en teoría) poder calcularse la agregación de todas las distrbuciones de tópicos para cada una de las posibles estructuras de tópicos, lo cierto es que se trata de un problema computacionalmente intratable. Por ello, al igual que en muchos problemas dentro del marco bayesiano, es necesario recurrir a aproximaciones numéricas[9].

Ahora bien, el método utilizado tiene algunas supuestos que, si bien pueden deducirse de lo expuesto más arriba, será útil repasarlos:

1. Cada documento d se compone de varios tópicos
2. Un tópico, a su vez, se compone de palabras; más precisamente, un tópico es una distribución de probabilidad sobre la totalidad de palabras del vocabulario V
3. Los tópicos "preexisten" a los documentos y la distribución de probabilidad sobre V es constante
4. Dado que se basa en el modelo BoW, para la construcción de tópicos se asume que las palabras no tienen orden
5. A su vez, el orden de los documentos no es relevante. Se asume que existe una cantidad fija de tópicos (y que es un hiperparámetro del modelo). Esto puede ser un problema al analizar corpus con documetnos de épocas muy diferentes[10].

[9] No es el objetivo de este trabajo desarrollar los métodos de inferencia y aproximación. En general, se basan en métodos de inferencia variacional o en métodos basados en Markov Chain Montecarlo. Para un mayor desarrollo puede verse (Asuncion, Welling, Smyth, & Teh, 2009).

[10] (D. Blei, 2012) expone varios métodos para flexibilizar este supuesto. Particularmente, los llamados *dynamic topic modelling* son una técnica posible.

5. Resultados

Utilizando LDA se buscó detectar los tópicos más relevantes en el dataset de letras de tango.

Ahora bien, como se desprende del apartado anterior, uno de los problemas principales es determinar la cantidad de tópicos que se busca detectar. Este problema es análogo al problema de determinación de la cantidad de clusters al aplicar algoritmos de segmentación tales como *K-means*. Es más, en la etapa de preprocesamiento, también se tomaron una serie de decisiones: el tipo de TFM a utilizar y el método de normalización, etc. Es por ello, que sería posible considerar cada una de estas decisiones como un hiperparámetro a evaluar y testear todas las combinaciones posibles de estos hiperparámetros, junto con el parámetro obligatorio referido a la cantidad de tópicos a detectar en el corpus.

Existen diversas métricas que permiten cuantificar qué tan "bueno" es el número de tópicos definido en términos cuantitativos (log-likelihood, perplexity, etc.). En general el uso de estas métricas conduce a modelos que logran buena performance estadística pero no necesariamente generan tópicos que sean interpretables, es decir, que tengan algún sentido en términos semánticos. Más bien, tiende a suceder lo contrario.

En términos generales, un número de tópicos grande tiende a arrojar mejores métricas y tiende a permitir una alta resolución de la estructura latente del corpus. No obstante, se ha observado que a medida que que el número de tópicos se incrementa, la calidad de los tópicos (en términos de interpretabilidad) tiende a decrecer (Mimno, Wallach, Talley, Leenders, & McCallum, 2011, Chang, Boyd-Graber, Wang, Gerrish, & Blei (2009)). De esta forma, al igual que en muchos otros problemas, complejidad del modelo e interpretabilidad tienden a ir en direcciones contrarias[11].

En el presente ejercicio se intentó buscar k que permitiera identificar tópicos interpretables. En el anexo se presentan algunas de los principales términos para diferentes k. Un primer rasgo que puede observarse es que, independientemente de las inicializaciones, existen algunos tópicos que se mantienen:

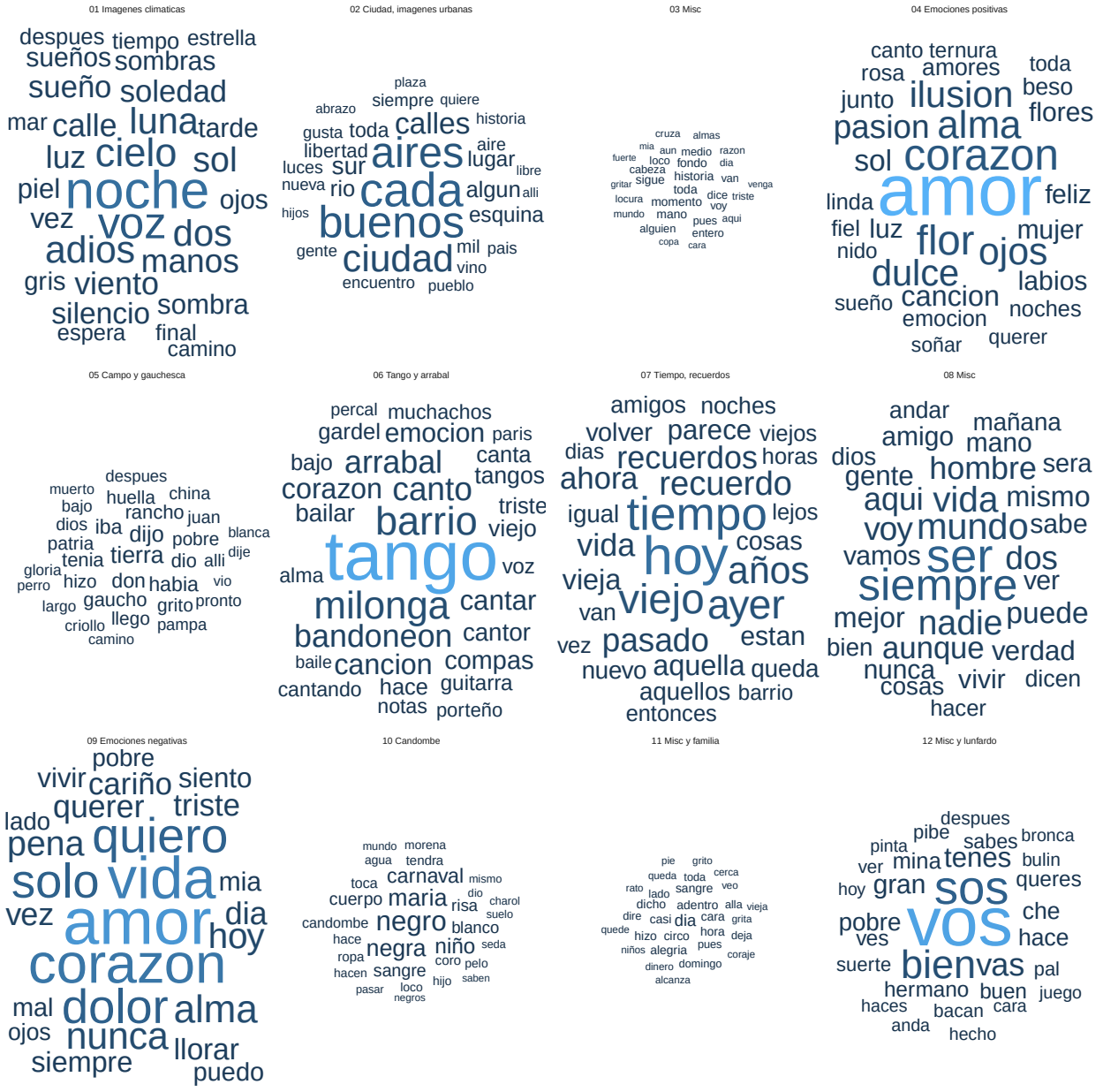
- Sentimientos y emociones con carácter positivo o negativo
- Imágenes de la noche, oscuridad y sombras asociadas a despedidas
- Imágenes que vinculan al tango y al barrio, arrabal
- Tópico sobre el tango, específicamente

Hemos llamado "misceláneos" a aquellos tópicos que presentan una distribución de palabras que no resultan interpretables.

De esta forma, seleccionamos el modelo que muestra 12 topicos. Observemos las 30 primeras palabras de cada uno de los tópicos.

Gráfico 1. Compisición de las 30 palabras de cada tópico ($k = 12$)

[11] Existe otro conjunto de métricas -coherece, topic itrusion, etc..- que se centran en la interpretabilidad (Mimno et al., 2011).



Puede verse que el primer tópico detectado tiene palabras como “noche”, “luna”, “cielo”, “sombras”, “viento”. Es decir, nos habla de *imágenes naturales o climáticas*.

A su vez, el segundo tópico capta el tema de la *ciudad y de las imágenes urbanas*: menciona términos como “buenos”, “aires”, “ciudad”, “calles”. El tópico 6 habla del *arrabal, pero sobre todo del tango mismo* (“tango”, “barrio”, “arrabal”, “canción”, “milonga”, “bandonéon”). El tópico 7 (“pasado”, “recuerdo”, “tiempo”) menciona palabras vinculadas al paso del *tiempo y a la memoria*.

Los tópicos 4 y 9 contienen palabras vinculadas a las *emociones*. El 4 (“ilusión”, “pasión”, “corazón”, “amor”) con una connotación positiva y el 9 (“amor”, “dolor”, “pena”, “triste”), negativa.

El tópico 5 y el 10 logran evidenciar tópicos de carácter “étnico”, por decirlo de alguna manera: el 5 con palabras como “china”, “gaucho”, “tierra”, “sangre” capta la cuestión de la *gauchesca y el campo*. El tópico 10, en cambio, (“carnaval”, “negro”, “morena”, “candombe”) habla sobre el *candombe y la cuestión de color*.

Por último, restan cuatro tópicos. Dos de ellos (3 y 8) tienen un carácter residual. Resultan difíciles de

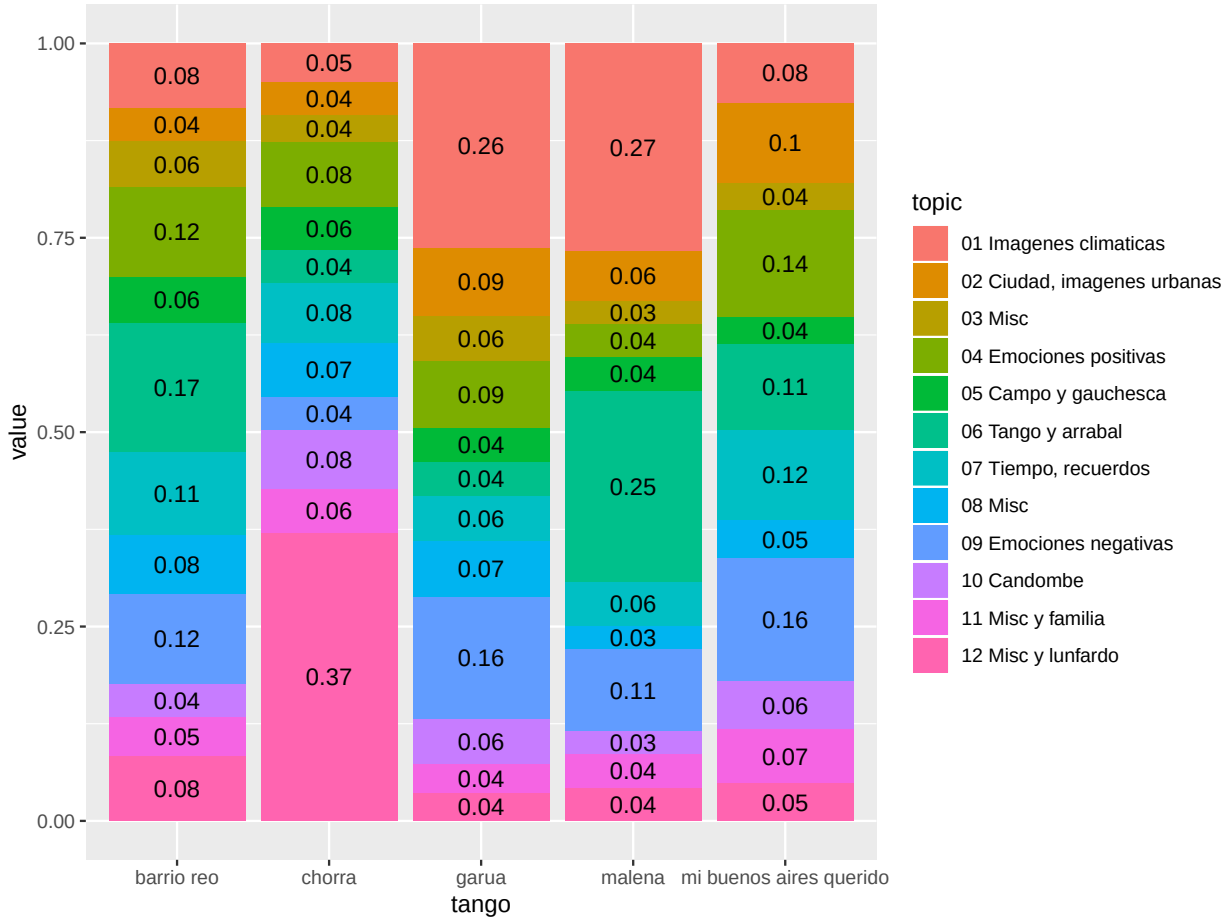
interpretar, por ello fueron etiquetados como *miscélaneos*. No obstante, el 11 y el 12, si bien contienen muchos términos que son poco interpetables, puede verse que el 11 ("vieja,"domingo","niños") contiene palabras vinculadas a la *vida familiar* y el 12, términos en *lunfardo* ("bulín, "pinta", "che", "pibe). El tópico 12, también parece tener como sus dos palabras más importantes"vos" y "sos", con lo cual, parece estar captando el hecho de que se habla de *forma directa a un interlocutor*.

Tabla 3. Identificación de los tópicos hallados

Tópico	Etiqueta
01	Imagenes climaticas
02	Ciudad, imagenes urbanas
04	Emociones positivas
05	Campo y gauchesca
06	Tango y arrabal
07	Tiempo, recuerdos
09	Emociones negativas
10	Candombe
11	Misc y familia
12	Misc y lunfardo

Una vez detectados los tópicos podemos estimar para cada documento d del corpus C qué proporción presenta de cada uno de los tópicos. A continuación se exponen, a modo de ejemplo, la composición de tópicos de cinco tangos de tres décadas diferentes:

Gráfico 2. Composición de tópicos según tangos, 1900-2010



Así, un tango como "Barrio reo", que habla de los gratos recuerdos del cantor al retornar a su barrio y de la tristeza que le produce el encuentro con su deterioro ("Hoy te encuentro envejecido"), muestra valores altos en los tópicos *emociones positivas*, *emociones negativas* y en el que habla sobre el *tango y arrabal*.

Al mismo tiempo, "Chorra", el tópico predominante es el 12. En efecto, el tango está dedicado y dirigido a una persona (la "chorra" en cuestión) y, al mismo tiempo, utiliza una buena cantidad de términos del lunfardo ("afanaste", "chorra", "cachaban", "gil").

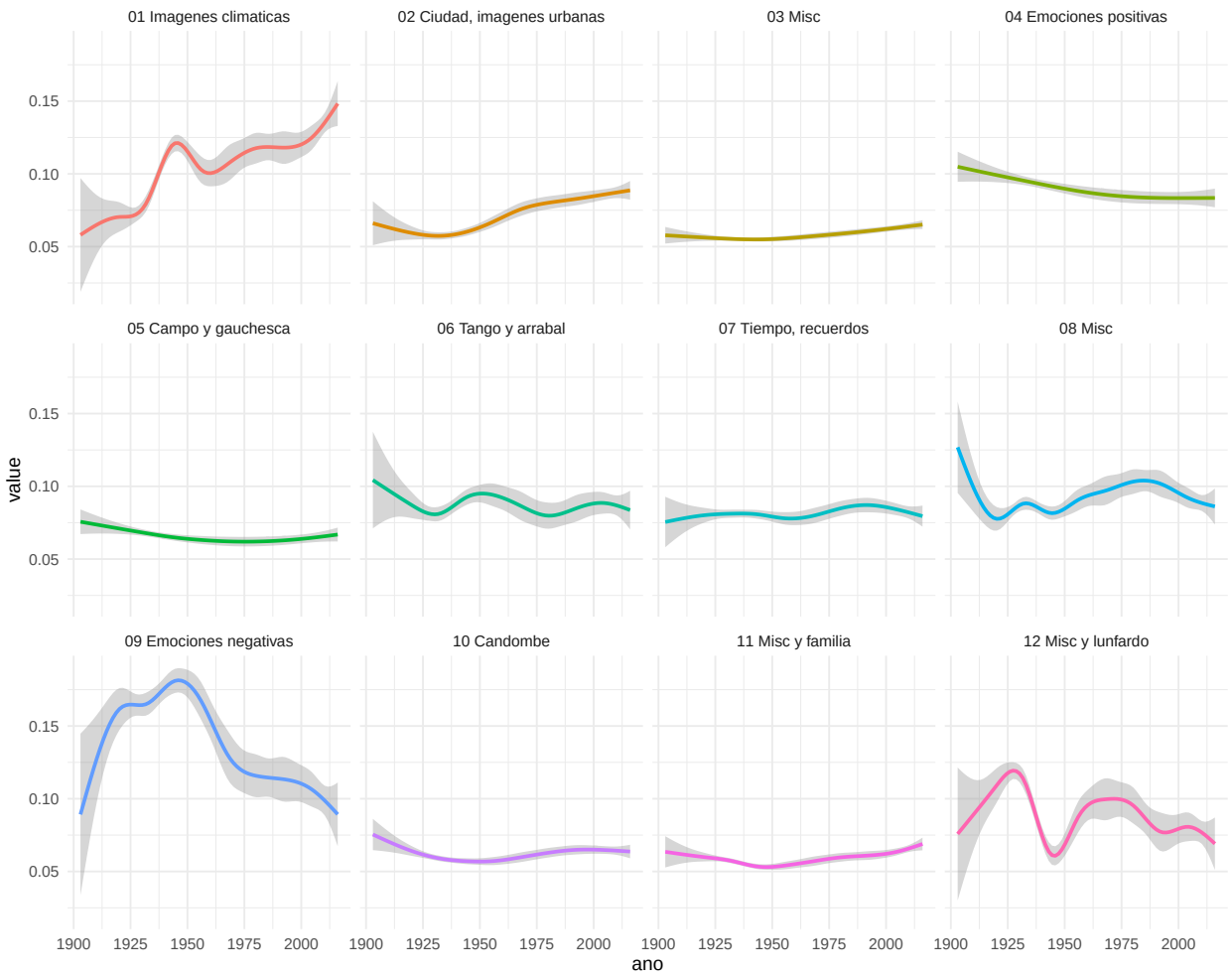
El tango "Garúa" muestra una caminata del narrador bajo la llovizna y pinta un cuadro lúgubre y oscuro ("sobre la calle la hilera de focos lustra el asfalto con luz mortecina") mientras el caminante recuerda a la mujer que, presumiblemente se fue. Es por eso que las *emociones negativas* y las *imágenes climáticas* aparecen con fuerza en este tango.

"Malena" nos habla (en tercera persona) de una cantante de tangos que pasó por desamores, que parece tener cierta predisposición a la bebida y que "canta el tango como ninguna". Esto se corresponde con la composición de tópicos detectada: *emociones negativas*, *tango y arrabal* e *imágenes climáticas* ("tono oscuro", "el frío del último encuentro", "tus manos son palomas que sienten frío").

Por úlimo, "Mi Buenos Aires querido", nos habla de la ciudad, la nostalgia del narrador y la esperanza del retorno. Esto se corresponde con los tópicos detectados: *ciudad*, *emociones negativas*, *emociones positivas* y *tiempo y recuerdos*.

Al mismo tiempo, podemos analizar la evolución de cada uno de los tópicos a lo largo de las diferentes décadas.

Gráfico 4. Evolución de los tópicos, 1900-2010 (suavizado GAM)



A partir de la evolución temporal puede verse cómo efectivamente van modificándose los temas del tango. En primer lugar, las imágenes naturales y climáticas ganan predominio de forma casi sostenida a lo largo del tiempo. En mucha menor medida, el tópico vinculado a la ciudad parece ganar cierta importancia a partir de la década del '30. También resultan interesantes las oscilaciones que parece presentar el tema del tango y el arrabal. Parece caer levemente hacia la década del '20 y vuelve a incrementarse hacia los años '50, mostrando otro valle hacia los años '70.

Pero quizás uno de los cambios más importantes es el que se observa en los tópicos 4 y 9. En efecto, puede verse que la participación de las emociones positivas es relativamente constante a lo largo del tiempo. En cambio, son las emociones negativas las que muestran diferencias fundamentales a lo largo del tiempo: muestran una tendencia al crecimiento hasta la década del '40-'50 y luego tienden a bajar de forma más suave.

Este punto es interesante porque toca algunas de las discusiones dentro del campo de los estudios del tango. Así, por ejemplo Borges (2016), planteaba que

> "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…"

De esta forma, puede verse que esta hipótesis parece ser corroborada por la información construida[12].

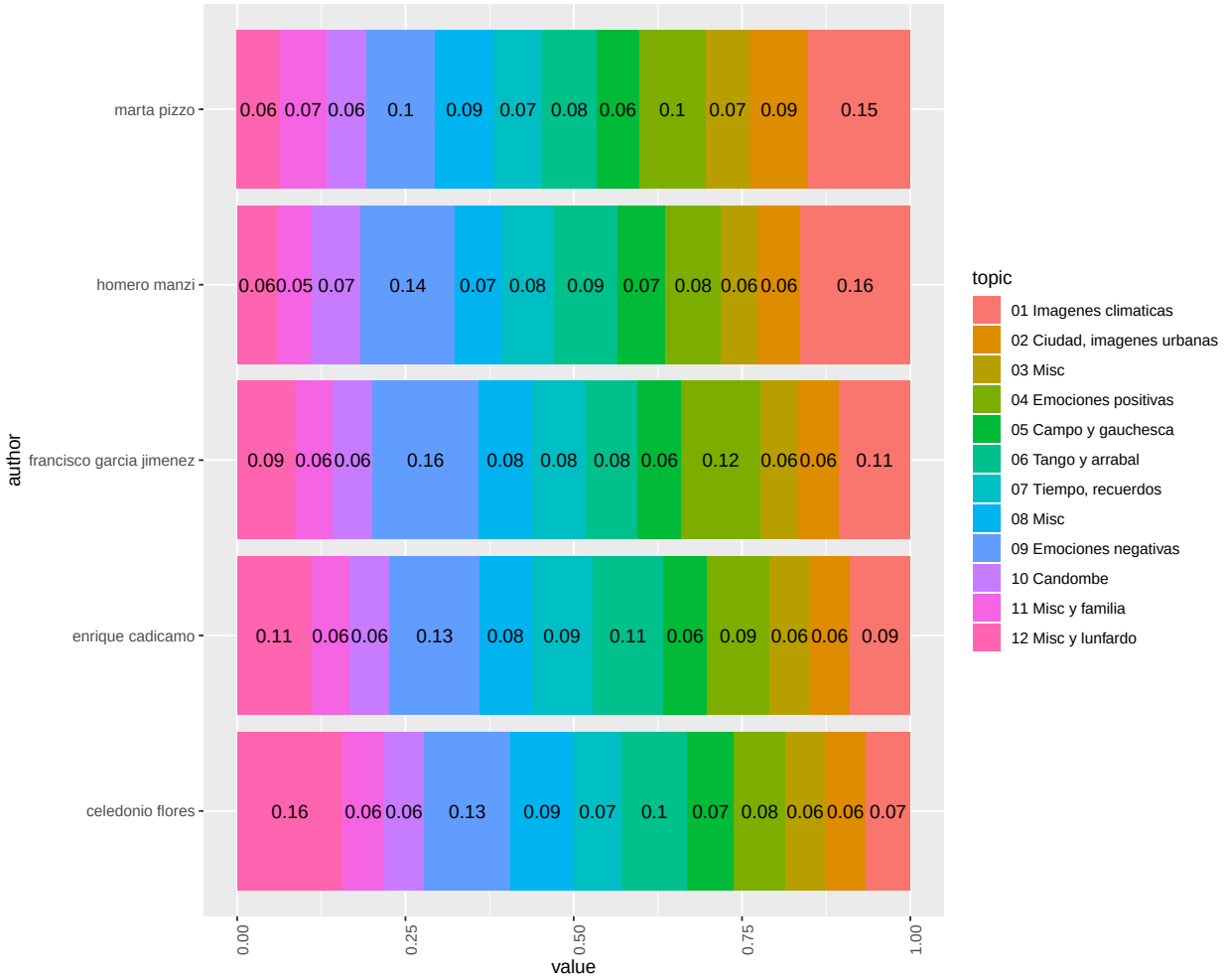
A partir de las series anteriores podrían plantearse preguntas que interroguen sobre la vinculación existente

[12]En el mismo texto, Borges (2016) discute sobre las causas de dicho cambio: explora hipótesis cuasi-sociológicas sobre la influencia negra, la italiana y otras hipótesis musicológicas, tales como la importancia de la introducción del bandoneón.

entre éstos cambios en los temas del tango u los procesos de desarrollo y expansión del capitalismo en Argentina y/o a los procesos de migración rural-urbana.

También es posible calcular la composición promedio de los diferentes tópicos en cada autor de tango. A continuación, desplegamos la composición de los cinco autores con mayor cantidad de letras en el dataset.

Gráfico 5. Composición de tópicos según autores, 1900-2010. Media de la composición de las letras de tango

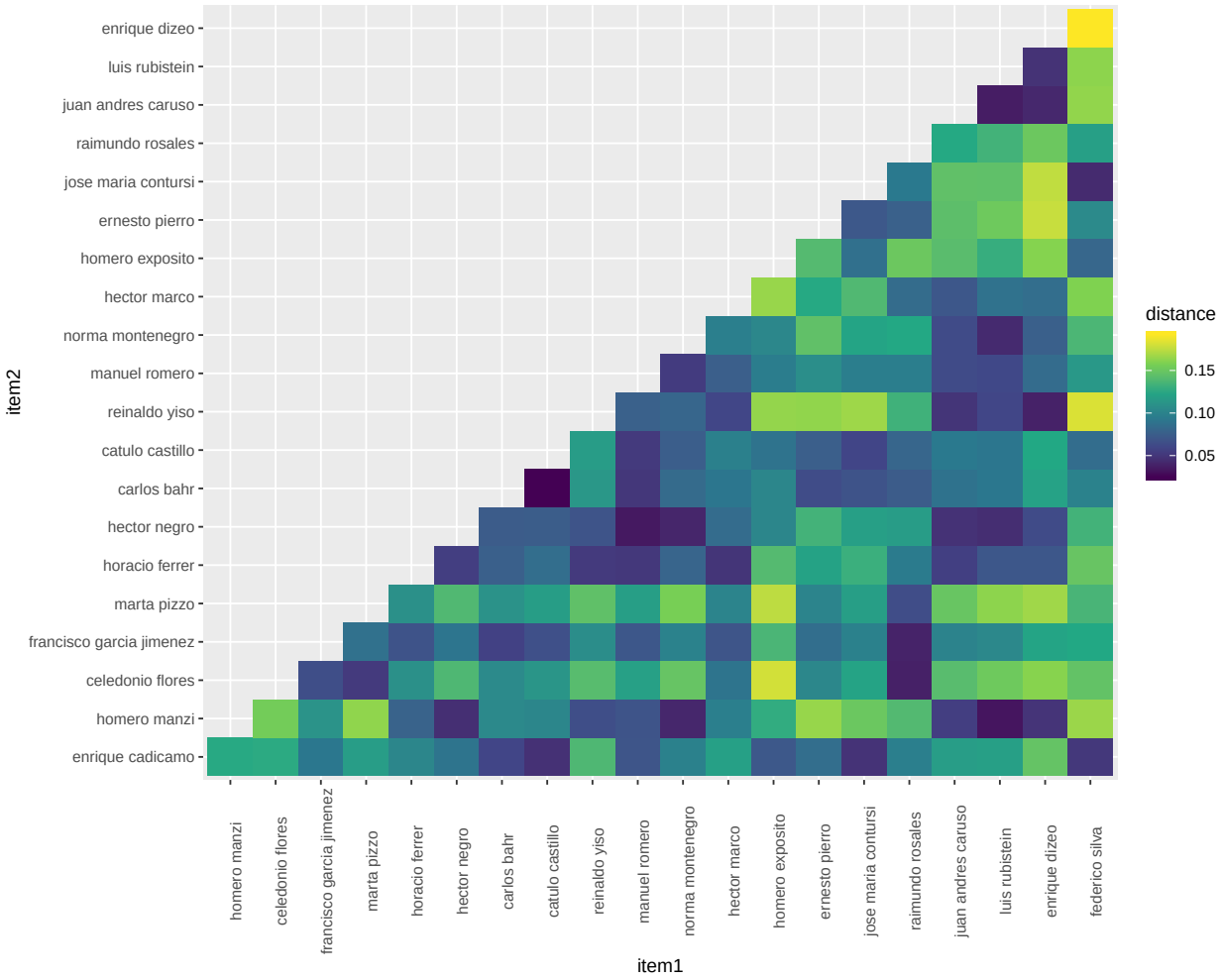


Así, por ejemplo, los temas predominantes de Cadícamo parecen ser el lunfardo y las emociones negativas. Esto contrasta con Homero Manzi, quien parece utilizar en mayor medida las imágenes climáticas (y también las emociones negativas).

Finalmente, podría pensarse en construir una matriz de distancias para cada autor en función de la composición promedio de sus tópcios, con el objetivo de encontrar autores que utilizan temas similares.[13]

Gráfico 6. Distancias en la composición de tópicos según autores, 1900-2010. Media de la composición de las letras de tango

[13] De forma análoga podría construirse una matriz a nivel de tango y buscar los tangos que hablan de tópicos similares.



De esta forma, puede verse por ejemplo, que Celedonio Flores y Homero Manzi parecen ser de los que mayor similitudes tienen en relación a los tópicos que utilizan. Algo parecido pasa con Enrique Dizeo y José María Contursi, por un lado y con Ernesto Pierro por el otro.

6. Resultados y discusión

En el presente trabajo se buscó presentar una aproximación metodológica posible para el análisis automático de textos a partir de la aplicación de una técnica de detección de tópicos (LDA) sobre un dataset de 5.600 letras de tango. A su vez, se presentó un flujo de trabajo posible para dicho análisis y se discutieron algunas técnicas para el preprocesamiento del texto.

De esta forma, fue posible estimar, mediante la técnica de topic modeling LDA, los principales temas del tango. Así, el uso de emociones positivas y negativas, imágenes de la ciudad, sobre el tango y el arrabal, sobre el campo y la gauchesca, sobre la temporalidad y la memoria, entre otros, aparecían como los más importantes. Al mismo tiempo, fue posible validar los tópicos a partir de la selección de algunos tangos y del análisis de sus letras buscando su correspondencias con los tópicos estimados.

Quizás uno de las posibilidades analíticas más interesantes fue la de poder visualizar la evolución temporal de los tópicos y, eventualmente, plantear hipótesis sobre la vinculación con procesos más generales y vinculados a otras dimensiones analíticas por fuera de las determinaciones internas del género tango (procesos vinculados a los cambios en la estructura económica, migratorios, etc.).

Finalmente, fue posible analizar las diferencias (distancias) entre los tópicos utilizados por diferentes autores.

Ahora bien, más allá de los resultados del ejercicio propuesto (que tienen como objetivo más mostrar un caso de uso de la herramienta que agotar las determinaciones del objeto en cuestión), el trabajo busca mostrar las potencialidades que este tipo de técnicas tienen para la investigación en ciencias sociales. Así, con una herramienta que permita analizar con un corpus amplio y de forma sistemática la evolución de los temas del tango (u otros géneros), sería posible vincular analítica y metodológicamente la dimensión cultural con otras esferas de la estructura social[14].

Al mismo, la detección de tópicos ha sido utilizada en los últimos tiempos para el análisis literario (Jockers & Mimno, 2013), el estudio de comunicados políticos (Grimmer, 2010), el estudio de medios (Wiedemann, 2016, DiMaggio, Nag, & Blei (2013)), el estudio de temas en leyes y proyectos (Gerrish & Blei, 2012), por nombrar algunas aplicaciones relevantes.

En efecto, sus principales ventajas radican en su escalabilidad y replicabilidad: utilizando las técnicas de análisis cualitativo de textos "tradicionales", es posible lograr gran profundidad analítica pero sobre corpus más bien pequeños o medianos y escasamente replicables. Así, los trabajos mencionados tenían una escala más bien pequeña: alrededor de 30 letras de tango. La excepción es el trabajo de Cantón (1972). El proceso de detección de tópicos encarado en el presente trabajo procesó y analizó una base de datos de alrededor de 6.200 letras de tango.

A su vez, el uso de técnicas automáticas de NLP no implica un desplazamiento de los enfoques tradicionales. Un buen ejemplo es el trabajo de Baumer y otros (2017), en el que se comparan los resultados obtenidos utilizando dos métodos de análisis sobre un mismo corpus de datos: generación de categorías utilizando la metodología de la *Grounded Theory* y detección de tópicos utilizando LDA. Los resultados sugieren tanto una coherencia como una complementariedad entre los resultados de ambos métodos.

Anexo - Tablas con los primeros 8 términos para estimación de tópicos con diferentes k.

Tabla A.1 k=15

topic	V1	V2	V3	V4	V5	V6	V7	V8
Topic 1	noche	sol	tiempo	cielo	voz	dos	luna	luz
Topic 2	vino	hombres	siempre	par	habia	dieron	flor	invierno
Topic 3	dice	mil	muerte	paz	filo	frente	gente	entero
Topic 4	amor	corazon	vida	quiero	dolor	alma	solo	hoy
Topic 5	alguien	final	cada	lejos	dio	niño	pasiones	sabia
Topic 6	siete	historia	volvio	decir	ahora	feliz	ver	viejos
Topic 7	tango	barrio	buenos	aires	milonga	canto	cancion	bandoneon
Topic 8	vamos	quieren	aunque	lleva	siente	oye	presencia	primavera
Topic 9	dias	cuatro	locos	quedo	circo	van	boca	hombres
Topic 10	lleva	gente	fuerte	rueda	sentido	quiero	romance	cuatro
Topic 11	tierra	linda	habia	gaucho	rancho	dijo	china	huella
Topic 12	negro	negra	maria	carnaval	libertad	sangre	niño	hijo
Topic 13	dice	hoy	anoche	puro	rosa	hizo	lagrimas	entro
Topic 14	mano	ahora	dire	paso	saben	traiga	media	casi
Topic 15	vos	sos	bien	hoy	ser	siempre	vida	vas

Tabla A.2 k=13

[14]Como hemos mencionado más arriba, uno de los supuestos de LDA en su versión básica, es que los tópicos preexisten a los textos y son constantes en el tiempo. Se trata de un supuesto fuerte para un análisis temporal. Es por ello que existen otras versiones de modelado de tópicos que permiten flexibilizar estos supuestos: *Dynamic topic modeling*, por ejemplo (D. Blei, 2012).

topic	V1	V2	V3	V4	V5	V6	V7	V8
Topic 1	mira	alli	dia	salio	frente	pelo	vendra	edad
Topic 2	paso	ser	llevo	partida	loco	años	bandera	fuerte
Topic 3	sangre	medio	dice	viene	tumba	vez	cuerpo	quedan
Topic 4	noche	voz	tiempo	cielo	luna	dos	sol	calle
Topic 5	tierra	dijo	don	viejo	tenia	gaucho	grito	rancho
Topic 6	tango	barrio	buenos	aires	milonga	viejo	bandoneon	arrabal
Topic 7	vamos	dias	loco	cuatro	libertad	viva	año	dale
Topic 8	amor	vida	corazon	solo	quiero	dolor	hoy	alma
Topic 9	pues	trabajo	quedo	tras	vida	pies	vio	algun
Topic 10	amor	flor	ojos	corazon	alma	dulce	cancion	ilusion
Topic 11	siempre	ser	vida	mundo	nadie	cosas	hombre	mismo
Topic 12	negro	sangre	negra	maria	niño	carnaval	risa	cuerpo
Topic 13	vos	sos	bien	hoy	vas	tenes	gran	pobre

Tabla A.1 k**=10**

topic	V1	V2	V3	V4	V5	V6	V7	V8
Topic 1	amor	corazon	alma	flor	ilusion	ojos	pasion	dulce
Topic 2	cada	buenos	aires	calle	ciudad	esquina	calles	van
Topic 3	noche	voz	cielo	dos	adios	luna	sol	manos
Topic 4	tierra	gran	don	gloria	alli	huella	grito	gaucho
Topic 5	tango	barrio	milonga	canto	bandoneon	arrabal	cancion	cantar
Topic 6	pobre	triste	dia	noche	aquella	madre	pena	dio
Topic 7	amor	vida	corazon	quiero	solo	dolor	nunca	alma
Topic 8	ser	siempre	mundo	nadie	hombre	voy	aqui	mejor
Topic 9	hoy	tiempo	ayer	vida	viejo	años	recuerdo	vez
Topic 10	vos	sos	bien	vas	tenes	gran	hace	che

Tabla A.1 k**=5**

topic	V1	V2	V3	V4	V5	V6	V7	V8
Topic 1	triste	vida	ojos	dia	pobre	alma	pena	noche
Topic 2	dos	noche	voz	tiempo	cielo	sol	adios	cada
Topic 3	tango	viejo	barrio	buenos	aires	milonga	canto	cancion
Topic 4	amor	corazon	vida	solo	quiero	dolor	hoy	nunca
Topic 5	vos	bien	sos	ser	vas	siempre	hombre	hoy

Bibliografía

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009). On smoothing and inference for topic models. *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, 27–34. Retrieved from http://dl.acm.org/citation.cfm?id=1795114.1795118

Baumer, E., Mimno, D., Guha, S., Quan, E., & Gay, G. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Assosiation for Information Science*

and Technology, *6*(68). Retrieved from https://mimno.infosci.cornell.edu/papers/baumer-jasist-2017.pdf

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4).

Borges, J. L. (2016). *El tango. cuatro conferencias*. Sudamericana.

Cantón, D. (1972). *Gardel, ¿a quién le cantás?* De la Flor.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Neural information processing systems*. Retrieved from docs/nips2009-rtl.pdf

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.S. government arts funding. *Poetics*, *41*(6). Retrieved from https://www.sciencedirect.com/science/article/pii/S0304422X13000661

Gasparri, J. (2011). Che varón, masculinidades en las letras de tango. *Revista Caracol*, (2).

Geertz, C. (1974). Deep play: Notes on the balinese cockfight. In *The interpretation of cultures* (pp. 412–453). New York: Basic Books.

Gerrish, S., & Blei, D. M. (2012). How they vote: Issue-adjusted models of legislative behavior. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 2753–2761). Retrieved from http://papers.nips.cc/paper/4715-how-they-vote-issue-adjusted-models-of-legislative-behavior.pdf

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, *18*(1).

Grimmer, J., & Stewart, M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, (2013). Retrieved from http://pan.oxfordjournals.org/

Hassani, A., Iranmanesh, A., & Mansouri, N. (2019). *Text mining using nonnegative matrix factorization and latent semantic analysis*.

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9).

Jockers, M., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, *41*(6).

López, I. (2010). Morochas, milongueras y percantas. representaciones de la mujer en las letras de tango. *Espéculo. Revista de Estudios Literarios.*, (45). Retrieved from http://webs.ucm.es/info/especulo/numero45/mutango.html

Marchese, M. (2007). Tango. el lenguaje quebrado del desarraigo. *Revista Latinoamericana de Estudios Del Discurso*, *6*(2).

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Empirical methods on natural language processing*. Retrieved from https://mimno.infosci.cornell.edu/papers/mimno-semantic-emnlp.pdf

Mitchell, R. (2015). *Web scraping with python: Collecting data from the modern web*. O'Reilly Media, Inc.

Reynoso, C. (2007). El lado oscuro de la descripción densa. *Anthropologika. Revista de Estudio E Investigaciones En Antropología*, *1*(1).

Wiedemann, G. (2016). *Text mining for qualitative data analysis in the social sciences. a study on democratic discourse in germany*. https://doi.org/10.1007/978-3-658-15309-0

Willenpart, L. (2011). El tango: Temas y motivos. *Verba Hispánica*, *9*(1).