Revista Latinoamericana de Metodología de la Investigación Social.
Nº23. Año 12. Abril – Septiembre 2022. Argentina. ISSN 1853-6190. Pp. 38-60.

Procesamiento de Lenguaje Natural aplicado a las ciencias sociales. Detección de tópicos en letras de tango

Natural Language Processing applied to Social Sciences. Topic identification in tango lyrics

Germán Rosati

Resumen

En este artículo, se presenta una aplicación de una técnica de Procesamiento de Lenguaje Natural (modelado de tópicos) sobre un *corpus* de letras de tango. Introduce un flujo de trabajo posible para el análisis textual computacional y en una técnica específica para la detección de tópicos: *Latent Dirichlet Allocation* (LDA). Se trabajará sobre un corpus de 5.617 letras buscando detectar de forma semiautomática sus temas. Los tópicos detectados abarcan desde imágenes de la ciudad, sobre el tango mismo, sobre emociones negativas y positivas, etc. Se analiza su evolución temporal y se muestra el cambio relativo de los tópicos en las letras de tango. También se valida el modelo analizando la composición de tópicos de algunos tangos canónicos. El trabajo busca ilustrar las potencialidades que estas técnicas tienen para el análisis de datos textuales en ciencias sociales: su escalabilidad y sus posibilidades de replicabilidad. Se marcan, finalmente, algunas limitaciones de este enfoque.

Palabras clave: Procesamiento de lenguaje natural; Tango; Modelado de tópicos; Web scraping; Minería de texto.

Abstract

This article presents the application of topic modeling, a natural language processing technique, in the context of tango lyrics analysis. It introduces a general workflow for computational text analysis and a specific topic detection technique: Latent Dirichlet Allocation (LDA). A semi-automated topic detection analysis of a 5,617 song *corpus* is described, finding a variety of recurring themes including urban imagery, tango itself, negative and positive emotions, etc. Further temporal analysis evidences the ebb and flow of topics in tango lyrics. The detection model is validated by examining the topic composition found in canonical song examples. This work aims to highlight the potential of these techniques for text data analysis in the social sciences, their scalability and replicability. Some limitations of this approach are also discussed.

Keywords: Natural language processing; Tango; Topic modeling; Web scraping; Text mining.

1. Introducción

El trabajo empírico en las ciencias sociales se caracteriza por una gran diversidad de fuentes de información utilizadas: desde datos altamente estructurados (cuya forma más clásica son las encuestas) hasta datos de un grado menor de estructuración. En este último grupo, los datos textuales, ya se trate de documentos, noticias, entrevistas, etc.- ocupan un lugar central.

El análisis de textos en sus diversas modalidades ha constituido un ejemplo típico de trabajo sobre datos cualitativos. Los estudios cualitativos se conforman por múltiples líneas de investigación. Una de ellas se caracteriza por tres elementos (Vasilachis, 2006: 6): 1) una posición epistemológica interpretativista, interesada en la comprensión y el análisis de los modos en que el mundo es interpretado por los sujetos; 2) el uso de métodos de recolección de información flexibles, es decir poco estructurados y 3) el empleo de métodos y procesos analíticos centrados en la comprensión del detalle, la complejidad y el contexto. Otros autores (Maxwell, 2004) marcan la importancia que el significado y la interpretación tienen en la investigación cualitativa. Coexisten en esta perspectiva diferentes escuelas sumamente diversas: la etnometodología; el análisis de la conversación, del discurso y de género; el análisis narrativo; la hermenéutica objetiva; la sociología del conocimiento hermenéutica; la fenomenología; el análisis de pequeños mundos de la vida; la etnografía y los estudios culturales (Vasilachis, 2006: 24). Así, las herramientas y técnicas metodológicas de las que se suele hacer uso en este tipo de enfoques se encuentran más vinculadas al análisis literario o del discurso y a métodos que buscan la comprensión profunda de los *corpus*.[1] Dichos enfoques han sido objeto de diversas críticas (Reynoso, 2007). No es objeto del presente trabajo evaluar dichas críticas y solo se marcará un aspecto de carácter metodológico: las posturas interpretativistas tienden a poner el eje en la capacidad subjetiva del investigador para realizar la interpretación o el análisis. Esto hace que parte de las investigaciones puedan presentar una relativa falta de sistematicidad metodológica. Al mismo tiempo, el peligro de la imposibilidad de replicación de sus resultados suele estar latente.

Tales problemas han sido abordados por otros autores y perspectivas que han buscado avanzar hacia un incremento de la replicabilidad de los estudios basados en datos cualitativos. Así, el llamado *Qualitative Comparative Analysis* elaborado en la década del '80 por Charles Ragin (1987) intentó estandarizar aspectos del análisis comparativo en las ciencias sociales. Lo hizo a partir de la sistematización y el ordenamiento de las unidades o casos. El objetivo era identificar diferencias y semejanzas entre las mismas a través del uso de diversos procedimientos inferenciales y comparativos basados en el álgebra booleana.[2]

El análisis de contenido abarca diferentes enfoques que van desde análisis impresionistas, interpretativos y escasamente estructurados hasta análisis estrictamente textuales (Hsieh & Shannon, 2005). La llamada *Grounded Theory* (Tie, Birks y Francis, 2019) busca la construcción de teoría social mediante procedimientos inductivos generados a partir de los datos producidos en la investigación. Una vez que tales datos han sido recolectados se siguen algunas etapas analíticas básicas: la codificación (que busca identificar los puntos de anclaje conceptuales en las fuentes); la conceptualización (los códigos con contenidos similares son agregados); la categorización (los conceptos similares son agrupados en categorías teóricas relevantes) y la teorización (una teoría surge del agrupamiento de categorías teóricas).

De forma similar, el *Qualitative Content Analysis* (Mayring, 2000) busca sistematizar las etapas y decisiones metodológicas en el proceso de análisis de textos. Procura sistematizar el proceso de investigación (preprocesamiento, análisis e interpretación) y desarrollar procedimientos de inferencia metodológicamente controlados (desarrollo inductivo de categorías, resúmenes, análisis de contexto, aplicación deductiva de categorías).

---

[1] 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: 452).

[2] Se ha utilizado como método de análisis de entrevistas en profundidad para diversos propósitos analíticos (Rosati y Chazarreta, 2017).

Este tipo de enfoques suele utilizar softwares como herramientas auxiliares del proceso. Paquetes como *Atlas-Ti* (Murh, 1997) o *Nvivo* (Richards, 1999) permiten estandarizar parte del procesamiento de la información mediante búsquedas en base a diferentes métodos textuales o basadas en métricas de similitud, codificación/anotación del *corpus*, generación de mapas o redes conceptuales y otros procedimientos similares (Casanova Correa y Pavón, 2002).

El uso de procedimientos de control metodológico, ya sean analógicos o mediatizados por algún software, permite avanzar en la estandarización de muchas de las operaciones analíticas mencionadas más arriba. No obstante, el proceso sigue teniendo un carácter fuertemente manual y continúa residiendo en el investigador. Dicha base manual, el uso de interfaces gráficas y la consecuente inexistencia de scripts o sintaxis que codifiquen dichas operaciones muestran que los procesos tienen un grado medio o bajo de estandarización. Las posibilidades de replicar tanto el procesamiento de los datos como los análisis se encuentran atadas a la decisión del investigador de documentar de forma minuciosa cada una de las decisiones tomadas.

A su vez, aparece un problema vinculado a la escala: la transcripción y codificación manual de *corpus* textuales (incluso contando con la mediación de algún software que asista dicha operación) limita fuertemente el tamaño de los *corpus* a analizar. Salvo que se cuente con recursos suficientes, con un equipo altamente capacitado y con disponibilidad de tiempo, el carácter manual de estos procesos puede llegar a limitar la escala de los *corpus* construidos y analizados.

La combinación de técnicas de *web scraping*, *Text Mining* y *Natural Language Processing* (NLP) puede ser de utilidad a las ciencias sociales atendiendo a estos problemas. Por un lado, permiten realizar una sistematización (y, eventualmente, lograr un cierto grado de automatización) de las diversas etapas del proceso de investigación, desde la recolección de datos y construcción del *corpus* hasta el preprocesamiento de un texto y su análisis. Específicamente, las técnicas de NLP habilitan la aplicación de métodos cuantitativos de análisis para una amplia diversidad de tareas (clasificación de textos, detección de temas y tópicos, detección de estructuras semánticas, 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 gran escala.

El presente trabajo tiene como objetivo discutir algunas aproximaciones metodológicas al análisis computacional de textos y presentar algunas de las potencialidades que tienen para las ciencias sociales. Esto se realizará a partir de su aplicación a un caso de estudio concreto: el análisis de los temas en un *corpus* de 5.617 letras de tango. Se concentrará en la discusión del flujo de trabajo y en la aplicación de una técnica específica para la detección de tópicos: el modelo *Latent Dirichlet Allocation* (LDA). Estos aspectos permitirán abordar los dos problemas mencionados: la escalabilidad y replicabilidad. A su vez, se propone realizar una ilustración del tipo de análisis, problemas y preguntas de investigación que tales técnicas habilitan.

2. Antecedentes metodológicos en el análisis de letras de tango

Existe una profusa literatura científica vinculada al análisis discursivo y de contenido de la poética del tango. El presente artículo no pretende abordar dicho campo problemático en la totalidad de sus determinaciones. Es por ello que los documentos reseñados en esta sección son ilustrativos de los aspectos metodológicos mencionados previamente.

Un abordaje común es el rastreo de un tópico o problema particular a lo largo de un *corpus* textual. En el análisis de letras de tango un tópico habitual es la problemática de género: representaciones de la mujer (López, 2010), masculinidades -por ejemplo, la figura del guapo- (Gasparri, 2011) o el sesgo sentimental de las letras (Marchese, 2007). Otros textos tienen un enfoque más amplio e identifican una mayor cantidad de temas. Así, Lucía Willenpart (2011) trabaja algunos temas comunes: el amor, el duelo amoroso, la mujer, la madre, el tango mismo,

etc. Un rasgo común en todos estos textos es el tamaño pequeño de los *corpus*:[3] el más grande cuenta con 30 letras (Marchese, 2007).

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. los criterios para la confección del corpus no aparecen explicitados
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 corpus son abordados a partir del rastreo de un conjunto de tópicos o preguntas particulares, privilegiando un análisis interpretativo del contenido.

3. Construcción del *corpus*

Se presentará un flujo de trabajo clásico aplicado a la detección de tópicos en un corpus de letras de tango. El mismo fue construido a partir de un proceso de *web scraping*. Se descargaron todas las letras de tango, milongas, valses, etc., disponibles en el sitio Todo Tango.[4] Además de las letras se descargó información accesoria sobre cada letra particular. Para ello, se confeccionaron dos *web crawlers* (es decir, programas que recorren automáticamente la estructura de un sitio *web*).[5]

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 cuantitativa (Mitchell, 2015). El uso de este tipo de técnicas hace posible abordar los dos problemas mencionados más arriba en la etapa de construcción de la información. Por un lado, permite sistematizar y hacer replicable la construcción del corpus. Esto se logra a partir de la ejecución de una rutina (con su correspondiente código) mediante la cual cualquier investigador puede reproducir los procedimientos de descarga de datos y construcción del corpus. Por otro lado, permite incrementar la capacidad de captura de información: el tamaño total del corpus construido en este trabajo es de una escala sensiblemente mayor que los utilizados en los estudios reseñados en la sección anterior.

Ahora bien, además de estas ventajas deben tenerse en cuenta algunas limitaciones. En primer lugar, existe una primera fuente de sesgo: dado que sería notablemente complejo construir un relevamiento completo de letras de tango, el conjunto que conforma el corpus final -pese a su amplitud- ha sido curado y seleccionado por los administradores del sitio. En este sentido, los problemas son similares a los que podrían producirse al utilizar compilaciones realizadas por diversos estudiosos.[6]

A su vez, los metadatos recabados (autor, fecha de composición, etc.) presentan diversos grados de calidad, cobertura e integridad. El atributo correspondiente a la fecha es uno de los que mayores cantidades de no respuesta presentan (alrededor de un 64%). Una vez más, en este caso, prevalece el sesgo presente en la fuente.

[3] 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.

[4] https://www.todotango.com/

[5] El código puede ser consultado en uno de los repositorios del proyecto https://github.com/gefero/tango_scrap.

[6] Véase por ejemplo Russo y Marpegán (2000).

En ambos casos, estos sesgos podrían mitigarse a partir del scrapeo de otras fuentes o sitios.[7] Dado que el objetivo de este trabajo es mostrar la aplicación de ciertas técnicas analíticas y no la extracción de conclusiones definitivas respecto a la poética del tango se utilizaron estas variables a título ilustrativo.

El *corpus* final consiste en de 5.617 letras de tango, agrupadas en un dataset con la siguiente estructura:

Tabla 1. Ejemplo de base de datos utilizada

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...

Fuente: Elaboración propia.

3.1. Análisis descriptivo de los metadatos

En este apartado se presenta un análisis sintético de los metadatos de las letras analizadas.

Tabla 2. Distribución de las letras de tango por ritmo

Ritmo	f	%
Tango	3964	70.6%
Milonga	477	8.5%
Vals	473	8.4%
Poema lunfardo	273	4.9%
Canción	156	2.8%
Candombe	35	0.6%
Otros	239	4.2%
Total	5617	100%

Fuente: Elaboración propia en base a datos recolectados de todotango.com

El 87.5% se concentran dentro de ritmos del tipo tangos, milongas y vals.

[7] Se intentó complementar la información de fecha a partir del srapeo de otro sitio (www.el-recodo.com). Sin embargo, no se lograron mejoras significativas en la no respuesta a la variable fecha. Queda pendiente para sucesivas aproximaciones al problema evaluar otros sitios o eventualmente completar dicho campo mediante la automatización de búsquedas en internet de las fechas faltantes. Todo el código que efectúa el scraping de ambos sitios y los *datasets* en cuestión pueden ser encontrados en el repositorio mencionado en la nota 4.

Tabla 3. Distribución de las letras de tango por autor

Autor	f	%
Enrique Cadícamo	168	3.0%
Homero Manzi	105	1.9%
Celedonio Flores	102	1.8%
Francisco G. Jiménez	80	1.4%
Marta Pizzo	80	1.4%
Horacio Ferrer	79	1.4%
Héctor Negro	77	1.4%
Carlos Bahr	74	1.3%
Cátulo Castillo	73	1.3%
Otros	4779	85.1%
Total	5617	100.0%

Fuente: Elaboración propia en base a datos recolectados de todotango.com

A su vez, se observa que los primeros 9 autores acumulan el 15% de las letras del conjunto de datos. Los principales autores son Cadícamo, Manzi y Flores.

Tabla 4. Distribución de las letras de tango por década

Década	f	%
1900	9	0.1%
1910	24	0.4%
1920	365	6.4%
1930	375	6.7%
1940	398	7.1%
1950	140	2.5%
1960	67	1.2%
1970	65	1.1%
1980	120	2.1%
1990	79	1.4%
2000	179	3.2%
2010	153	2.7%
Sin dato	3643	64.8%
Total	5617	100.0%

Fuente: Elaboración propia en base a datos recolectados de todotango.com

Finalmente, al analizar la distribución de letras por década, se observa la gran cantidad de datos faltantes (casi un 65%), lo cual obligará a tomar con precaución, y a título ilustrativo, análisis posteriores. A su vez, (y como era de esperarse) es el período que va de los años '20 a los '40 el que mayor cantidad de letras encuentra.

4. Preprocesamiento del texto

Al observar en la tabla 1 el campo *Letra* (el núcleo del análisis), puede advertirse 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 aparecen atributos, o en todo caso, solamente es visible un atributo, la letra. El primer paso, entonces, es transformar esta representación no estructurada en una estructurada. La misma tendrá como objeto reducir la complejidad del texto, dado que "el lenguaje es complejo. Pero no toda su complejidad es necesaria para analizar un texto de forma efectiva" (Grimmer & Stewart, 2013: 72).

4.1. Modelo BoW (Bag of words)

Para llegar a esa representación estructurada,[8] será necesario pensar en un formato 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, cada columna consistirá en un término t del vocabulario general V del *corpus* C. Cada 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 5. Ejemplo de matriz de frecuencia de términos cruda

Letra	agua	blanda	cartel	cruel	el	en	era	la	manda	mas	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

Fuente: Elaboración propia en base a datos recolectados de todotango.com

Puede verse que al construir esta matriz la información sobre el orden de las palabras se ha perdido. 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 del *corpus* C. Así, V incluiría a priori pronombres, preposiciones, diferentes conjugaciones de verbos, sustantivos en singular o plural, etc. Esto hace que el vocabulario "crudo"

[8] Vale destacar que el flujo de trabajo descrito a continuación es uno de los más comunes 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).

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 aportan poca o nula información acerca del contenido de cada documento.

Tabla 6. Ejemplo de matriz de frecuencia de términos sin stopwords

Letra	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

Fuente: Elaboración propia en base a datos recolectados de todotango.com

A partir de la eliminación de las *stopwords*[9] y de los signos de puntuación se obtiene en la tabla 6 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.

4.2. Normalización de conteos

El último paso[10] supone normalizar los valores de las celdas de la TFM. Al momento de filtrar los *stopwords* se buscaba poder eliminar aquellas palabras con poca información acerca del

[9] En el caso de los pronombres personales, los mismos fueron eliminados en base a una lista. Sin embargo, sobrevivió a esta eliminación el pronombre "vos", dado que no se encontraba en la lista original. No obstante, resulta interesante ver cómo en los resultados del ejercicio de detección de tópicos este pronombre aparece en un tópico propio junto con verbos conjugados en segunda persona. En aras de la facilidad de lectura no se explicita de forma detallada el tratamiento realizado en cada etapa del preprocesamiento y, particularmente, el realizado para las diferentes *stopwords*, signos de puntuación, etc. Puede consultarse todo el material de replicación para el análisis de tópicos en el siguiente repositorio https://github.com/gefero/topic_modeling_tango.

[10] Suele ser habitual, a los efectos de reducir la diversidad de términos, realizar algún proceso de normalización de los mismos. Se busca llevar las palabras flexionadas a una forma normal que represente a toda una clase de palabras. Esta forma normal, llamada lema, puede ser pensada típicamente como la palabra utilizada como entrada en los diccionarios de lengua: el infinitivo para las conjugaciones verbales, el masculino singular para adjetivos, etc. Esto suele requerir la detección de la función sintáctica de la palabra, su contexto y la construcción de diccionarios de palabras que mapean formas flexionadas a lemas. También pueden estimarse los lemas a partir de algoritmos y modelos probabilísticos (grafos de asociación, *clustering*, redes neuronales, etc.). A esta operación de reducción se la llama lematización. Existe, además, una segunda forma de reducción -*stemming*- más simple (si bien puede considerarse una variante de la lematización). La misma opera sobre las declinaciones de las palabras. Aquellas palabras que remiten a un mismo concepto básico son reducidas a la misma raíz. Por ejemplo, "familias", "familia" y "familiar" son reducidas a "familia". No se entra en mayores especificidades porque dichos procedimientos no fueron utilizados para el presente trabajo. Para mayor detalle puede consultarse (Jurafsky y Martin, 2006).

contenido 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 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". Se define una métrica llamada *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, puede definirse 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 $TF-IDF$, *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, en ciertos casos (como el del *corpus* que se analiza en este trabajo) suele suceder que el conteo crudo $c(t,c)$ de términos funciona 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.[11]

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
5. construcción de la TFM basada en $c(t, c)$

5. Detección de tópicos

Existen muchas técnicas para la detección automática de tópicos en *corpus* textuales. Algunas de ellas están basadas en ciertas formas de descomposición de la TFM.[12] Para el presente trabajo se utilizará uno de los métodos más conocidos: *Latent Dirichlet Allocation* o LDA.

La intuición detrás de LDA (Blei, 2012: 79) 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, asumiendo la existencia de un "proceso generador de textos": un proceso aleatorio imaginario por el cual un documento es producido.

Más formalmente,[13] 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* sería esperable que palabras como "amor", "pena", "sufrimiento", tuvieran altas probabilidades de inclusión en el mismo. En cambio, palabras como "ciudad", "barrio" estarían más asociadas a un tópico que hable acerca de la *ciudad*.

El proceso que genera cada texto del *corpus* funciona bajo los siguientes supuestos. 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 seleccionado es elegido de la distribución de tópicos de ese documento d particular (paso 2.i).

---

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

[12] Existen métodos que dividen en dos componentes en la TFM, una matriz de *documentos x tópicos* y otra de *términos x tópicos*, tales como *Non Negative Matrix Factorization* y *Latent Semantic Analysis* (Hassani, Iranmanesh, & Mansouri, 2019).

[13] Esta sección se basa en Blei (2012).

Así, el objetivo del modelado de tópicos es descubrir los temas a los que alude un determinado conjunto de documentos. 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 lo que se trata de estimar).

Formalizando el razonamiento anterior, es posible ver que se trata de una probabilidad conjunta:

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

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. Las palabras observadas en el documento d, son w_d. Aquí, $w_{d,n}$ es la n-ésima palabra en el documento d, que es un elemento del vocabulario.

Puede verse que existen ciertas dependencias en el modelo: por ejemplo, la asignación de tópicos en un documento ($z_{d,n}$)depende de la proporción de tópicos por documento θ_D. A su vez, la palabra observada $w_{d,n}$ depende tanto de la asignación de tópicos $z_{d,n}$ como de todos 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 estimado de forma relativamente 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 distribuciones 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.[14]

Ahora bien, el método utilizado tiene algunos supuestos que, si bien pueden deducirse de lo expuesto más arriba, es útil hacerlos explícitos:

1. cada documento *d* se compone devarios 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.

[14] 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 Asunción, Welling, Smyth, y Teh (2009).

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. el orden de los documentos no es relevante.
6. 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 documentos de épocas muy diferentes.[15]

Utilizando LDA se buscó detectar los tópicos más relevantes en el *corpus* de letras de tango. Ahora bien, como se desprende del apartado anterior, uno de los problemas principales es determinar la cantidad de tópicos a estimar. Este problema es análogo al problema de determinación de la cantidad de clusters al aplicar algoritmos de *clustering* tales como *K-means*. Es más, en la etapa de preprocesamiento, también se tomaron una serie de decisiones: el tipo de TFM a utilizar, por ejemplo. 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 el ajuste (es decir, qué tan "bueno" es) del 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 que no necesariamente generan tópicos interpretables.

En términos generales, un número de tópicos más grande tiende a arrojar mejores métricas y a permitir una alta resolución de la estructura latente del *corpus*. No obstante, se ha observado que al aumentar el número de tópicos la calidad de los tópicos (en términos de interpretabilidad) tiende a decrecer (Mimno et al., 2011; Chang et al., 2009). De esta forma, al igual que en muchos otros problemas, complejidad del modelo e interpretabilidad tienden a ir en direcciones contrarias.[16]

Párrafo aparte merece la problemática de ciertos procedimientos retóricos y estilísticos en textos literarios. Dado que el modelado de tópicos tiende a operar sobre la co-ocurrencia de palabras para la definición de tales tópicos, en algunos casos, recursos como las metáforas pueden generar problemas de identificación. En este sentido, las metáforas no son tematizadas en este trabajo. No obstante, algunos estudios han intentado utilizar el modelado de tópicos (y, particularmente, LDA) como detectores de metáforas.[17]

6. Resultados

En el presente ejercicio se intentó buscar un k (número de tópicos) que permitiera identificar temas interpretables. En el anexo se presentan algunos de los principales términos para diferentes k. Un primer rasgo que puede observarse es que, independientemente del número de tópicos que se elija, existen algunos temas 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 o al arrabal

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

[16] Existe otro conjunto de métricas -*coherence*, por ejemplo.- que se centran en la interpretabilidad (Mimno et al., 2011).

[17] Ver, por ejemplo, Navarro-Colorado et al (2011) y Heintz et al (2013).

- Tópico sobre el tango, específicamente

Se han llamado "misceláneos" a aquellos tópicos que presentan una distribución de palabras que no resulta interpretable. Constituyen en este sentido tópicos que no parecen tener un significado atribuible y resultan de carácter residual. Por este motivo no serán objeto de análisis.

6.1. Identificación de tópicos

De esta forma, el modelo seleccionado es el que muestra un k=12 tópicos. Es importante recordar que, dado que los tópicos se definen como una distribución de probabilidad sobre las palabras del vocabulario, una misma palabra tiene una cierta probabilidad de pertenencia a todos los tópicos. Las diferencias, entonces, son de carácter relativo. Ciertas palabras pertenecen a ciertos tópicos *con mayor probabilidad que otras*. Esto hace que pueda existir un cierta superposición (*overlapping*) entre términos a lo largo de los tópicos.

En el gráfico siguiente se exponen las primeras 30 palabras de cada uno de los tópicos, es decir, las 30 palabras con mayor probabilidad de pertenencia. A su vez, el tamaño en la nube de cada término es proporcional a dicha probabilidad. Puede verse, entonces, que salvo algunas excepciones como "amor", las principales palabras tienden a no superponerse.

Gráfico 1. Composición de las 30 palabras de cada tópico (*k=12*)

Fuente: Elaboración propia en base a datos recolectados de todotango.com

Es también importante remarcar que, al igual que en otros métodos estadísticos,[18] la interpretación de los tópicos (es decir, su etiquetado) recae sobre el equipo de investigación. En ese sentido, la identificación y etiquetado se basa en el análisis de sentido que se presentan en las palabras con mayor probabilidad de pertenencia a cada tópico.

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", (los cuales remiten, claramente a "Buenos Aires"), "ciudad", "calles". El tópico 6 habla del *arrabal, pero sobre todo del tango mismo* ("tango", "barrio", "arrabal", "canción", "milonga", "bandoneón"). 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 más positiva y el 9 ("amor", "dolor", "pena", "triste"), aparentemente negativa. Resulta interesante notar cómo un mismo término ("amor") aparece vinculado tanto a un tópico con signo negativo como a otro con connotación positiva. Es decir, que el modelo parece haber sido capaz de detectar ciertos matices y usos diferentes asociados al término "amor".

El tópico 5 y el 10 logran evidenciar temas 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, la cuestión étnica y la negritud*.

Por último, restan cuatro tópicos. Dos de ellos (3 y 8) tienen un carácter residual. Resultan difíciles de interpretar, por ello fueron etiquetados como misceláneos. No obstante, el 11 y el 12, si bien contienen muchos términos que son poco interpretables, puede verse que el 11 ("vieja", "domingo", "niños") contiene algunas 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", junto con verbos conjugados en la segunda persona del singular, con lo cual, parece estar captando el hecho de que se habla de *forma directa a un interlocutor*.

Tabla 7. Identificación de los tópicos hallados

Tópico	Etiqueta
01	Imágenes climáticas
02	Ciudad, imágenes 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

Fuente: Elaboración propia en base a datos recolectados de todotango.com

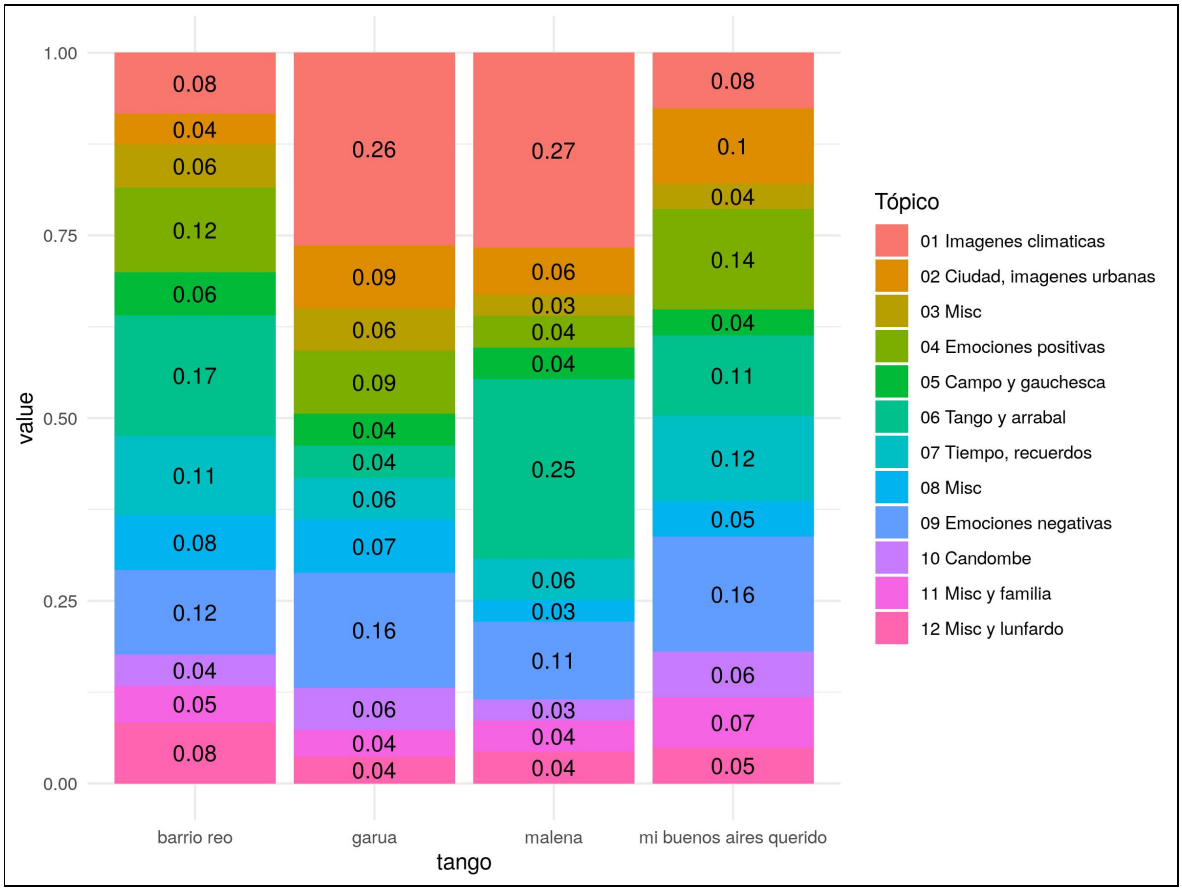
[18] Así, por ejemplo, en otros modelos no supervisados como el análisis de componentes principales (PCA, por sus siglas en inglés) o en las diversas formas de *clustering*, suele resultar necesaria la intervención del investigador para hacer inteligibles algunos de los *outputs* del modelo.

6.2. *Algunos análisis posibles*

Una vez detectados los tópicos es posible estimar para cada documento d del *corpus* C en qué proporción presenta cada uno de los tópicos. Es decir, para cada letra de tango será posible calcular qué composición muestra de cada uno de los 12 tópicos detectados. Una forma útil de validación de los resultados es explorar si el análisis manual de los tangos coincide con los tópicos que fueron detectados de forma automática.

A continuación, se exponen, a modo de ejemplo, la composición de tópicos de cuatro tangos de tres décadas diferentes:

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



Fuente: Elaboración propia en base a datos recolectados de todotango.com

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*.

El tango "Garúa" tematiza 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.

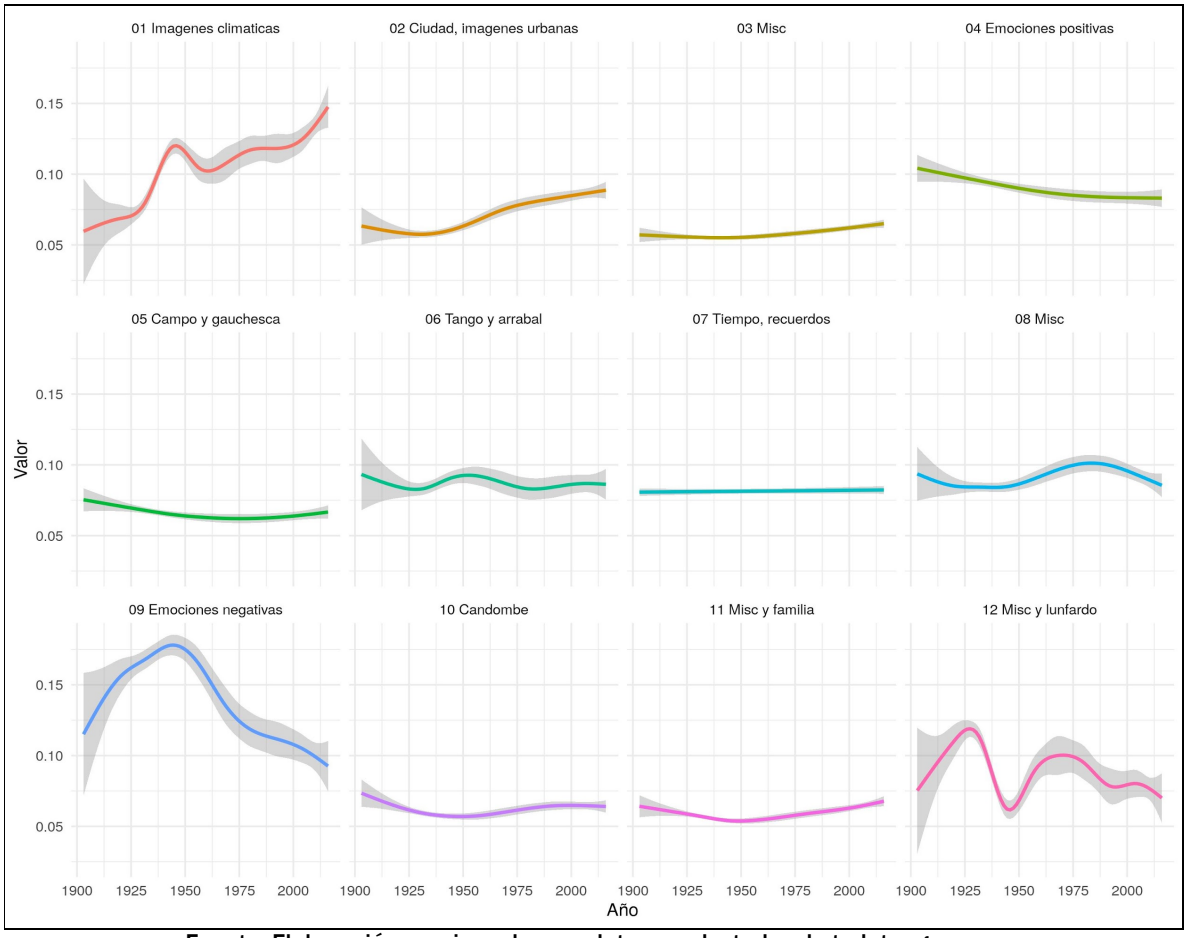
"Malena" habla (en tercera persona) de una cantante de tangos que pasó por desamores, que parece tener cierta predisposición a la bebida y que "canta el tango como ninguna". Esto se corresponde con la composición de tópicos detectada: *emociones negativas*, *tango y arrabal* e

imágenes climáticas (“tono oscuro”, “el frío del último encuentro”, “tus manos son palomas que sienten frío”).

Por último, en “Mi Buenos Aires querido”, se menciona a 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, sería posible analizar la evolución de cada uno de los tópicos a lo largo de las diferentes décadas.

Gráfico 3. Evolución de los tópicos, 1900-2010 (suavizado GAM)



Fuente: Elaboración propia en base a datos recolectados de todotango.com

En primer lugar, las imágenes naturales y climáticas ganan predominio de forma sostenida a lo largo del tiempo. En mucha menor medida, el tópico vinculado a la ciudad parece cobrar cierta importancia a partir de la década del 1930. También resultan interesantes las oscilaciones que parece presentar el tema del tango y el arrabal. Parece caer levemente hacia la década del 1920 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 más marcadas: se ve una tendencia al crecimiento hasta la década del 1940 y 1950. Posteriormente, tiende a descender su importancia.

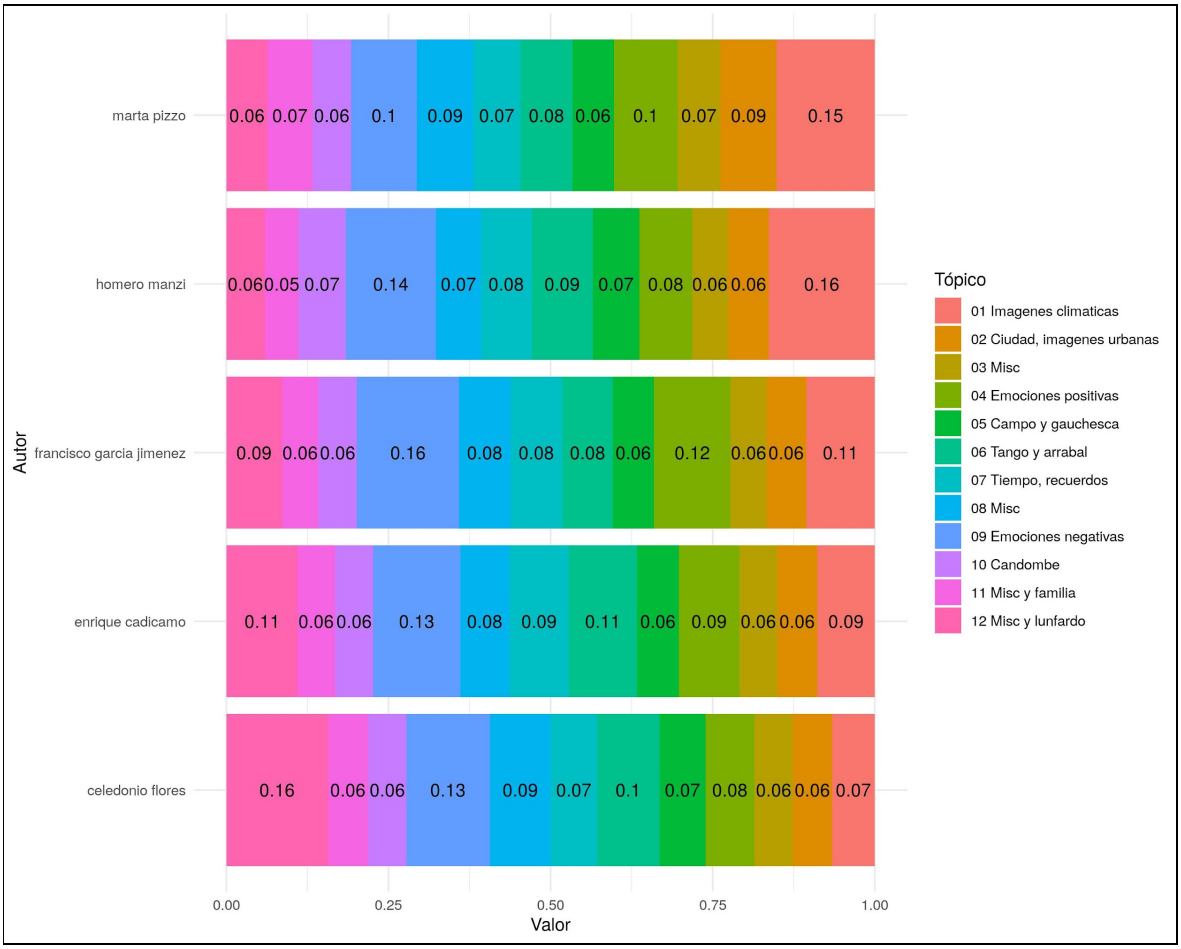
Este punto es interesante porque se vincula con algunas de las discusiones dentro del campo de los estudios del tango. Así, por ejemplo, Borges (2016: 80-81), planteaba hacia la

década del 1960 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 tener cierto sustento y habilita el estudio más profundo de esta problemática.[19]

A partir de las series anteriores podrían plantearse preguntas que interroguen sobre la vinculación existente entre tales cambios en los temas del tango y los procesos de desarrollo y expansión del capitalismo en Argentina y/o a los movimientos de migración rural-urbana.

También, es posible calcular la composición promedio de los diferentes tópicos en cada autor de tango.

Gráfico 4. Composición de tópicos según autores, 1900-2010.
Media de la composición de las letras de tango



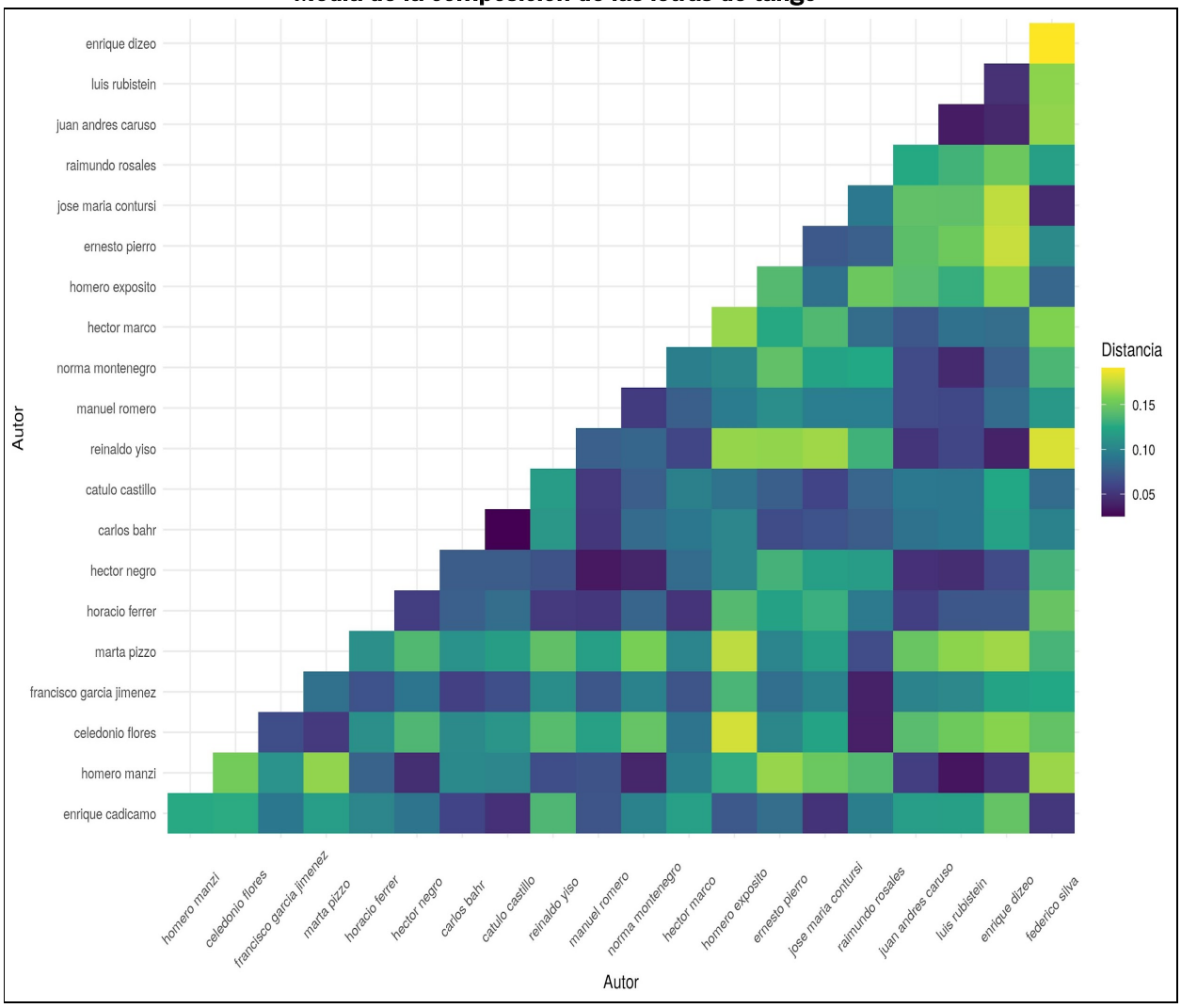
Fuente: Elaboración propia en base a datos recolectados de todotango.com

Si se analizan tales métricas para los cinco autores con mayor cantidad de letras en el *corpus*, se hace manifiesto que 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).

[19] 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.

Finalmente, puede construirse una matriz de distancias para cada autor en función de la composición promedio de sus tópicos, con el objetivo de encontrar autores que utilizan temas similares.[20]

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



Fuente: Elaboración propia en base a datos recolectados de todotango.com

Celedonio Flores y Homero Manzi parecen ser los que mayores 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.

7. Discusión y comentarios finales

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 *corpus* grande de letras de tango. A su vez, se presentó un flujo de trabajo posible para dicho análisis y se discutieron algunas técnicas para el pre-procesamiento del texto (eliminación de

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

stopwords y otros signos, normalización de conteos, construcción de una TFM mediante el modelo *Bag of Words*).

De esta forma, fue posible identificar los principales temas del tango. 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, aparecen 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 evaluando su correspondencia con los tópicos estimados.

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

Ahora bien, más allá de los resultados del ejercicio propuesto (acotados a mostrar un caso de uso de las herramientas), el trabajo busca ilustrar las potencialidades que este tipo de técnicas tiene para la investigación sobre datos textuales en ciencias sociales. Así, con una herramienta que permita analizar con un *corpus* amplio y de forma más 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.[21] Quizás una de las posibilidades analíticas más interesantes fue la de poder visualizar la evolución temporal de los tópicos y la eventual posibilidad en trabajos más sistemáticos de plantear hipótesis sobre la vinculación con procesos más generales de la sociedad argentina (los cambios en la estructura económica, movimientos migratorios, etc.).

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), de medios (Wiedemann, 2016; DiMaggio, Nag, y Blei, 2013) y de temas en leyes y proyectos (Gerrish & Blei, 2012), por nombrar algunas aplicaciones relevantes.

En efecto, sus principales ventajas radican en su escalabilidad y en sus mayores posibilidades de 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, en muchos casos, con resultados poco o medianamente replicables. Los antecedentes mencionados tenían una escala más bien pequeña: alrededor de 30 letras de tango, con la excepción del texto de Cantón (1972). El proceso de detección de tópicos encarado en el presente trabajo procesó y analizó 5.617 letras de tango.

No obstante, en este tipo de enfoques existen algunas limitaciones que es importante remarcar. En relación a la construcción del *corpus*, la dependencia de la fuente de datos a utilizar resultó una primera desventaja (si bien, no tan diferente al uso de compilaciones de letras de tango realizadas por diferentes autores). En la etapa de análisis, los modelos de detección de tópicos presentan problemas para el tratamiento adecuado de algunas figuras retóricas (como, por ejemplo, las metáforas). También se presentan como problemas potenciales a resolver, la correcta determinación del número de tópicos y el etiquetado de los mismos. La complementación con análisis en profundidad de los documentos del *corpus* resulta, en este punto, de suma utilidad.

Por ello, es importante remarcar que el uso de técnicas de NLP no implica un desplazamiento de los enfoques basados en la interpretación manual de los documentos. Un trabajo (Baumer et al, 2017) compara 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. Las conclusiones de dicho trabajo sugieren tanto coherencia como complementariedad entre los resultados de ambos métodos.

[21] 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 (Blei, 2012).

8. Bibliografía

ASUNCIÓN, A., WELLING, M., SMYTH, P., y TEH, Y. W. (2009). On smoothing and inference for topic models. *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pp-pp. 27-34. Disponible en http://dl.acm.org/citation.cfm?id=1795114.1795118. Fecha de consulta: 20/08/2019.

BAUMER, E., MIMMO, D., GUHA, S., QUAN, E., & GAY, G. (2017) Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?" *Journal of the Association for Information Science and Technology* 68 (6). Dispoible en https://mimno.infosci.cornell.edu/papers/baumer-jasist-2017.pdf. Fecha de consulta: 12/07/2019.

BLEI, D. (2012). Probabilistic topic models. *Communications of the ACM* 55 (4), 77-84.

BORGES, J. L. (2016). *El tango. Cuatro conferencias*. Buenos Aires: Sudamericana.

CANTÓN, D. (1972). *Gardel, ¿a quién le cantás?* Buenos Aires: De la Flor.

CASANOVA CORREA, J. y PAVON RABASCO, F. (2002). Nuevas herramientas para el procesamiento de datos cualitativos. *Ágora Digital* 3, 1-13. Disponible en: http://rabida.uhu.es/dspace/bitstream/handle/10272/6616/Nuevas_herramientas.pdf;jsessionid=000A8A1862E6EDF881EA03D14ABCC096?sequence=2

CHANG, J., BOYD-GRABER, J., WANG, C., GERRISH, S., y BLEI, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems,* 32:288-296. Disponible en: https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.

DIMAGGIO, P., NAG, M., y 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), 570-606.

GASPARRI, J. (2011). Che varón, masculinidades en las letras de tango. *Revista Caracol* 2, 175-215.

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

GERRISH, S., y BLEI, D. (2012). How they vote: Issue-adjusted models of legislative behavior. F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* 25, 2753-2761. DIsponible en 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), 1-35.

GRIMMER, J., y STEWART, M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21 (3), 267-297.

HASSANI, A., IRANMANESH, A., y MANSOURI, N. (2019). *Text mining using nonnegative matrix factorization and latent semantic analysis*. Disponible en: https://arxiv.org/abs/1911.04705.

HEINTZ, I., GABBARD, R., SRINIVASAN, M., BARNER, D., BLACK, D., FREEDMAN, M. y WEISCHEDEL, R. (2013). Automatic extraction of linguistic metaphors with lda topic modeling. *Proceedings of the First Workshop on Metaphor in NLP*. Disponible en: https://www.aclweb.org/anthology/W13-0908.pdf.

HSIEH, H.-F., y SHANNON, S. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research* 15 (9), 1277–1288.

JOCKERS, M., y MIMNO, D. (2013). Significant themes in 19th-century literature. *Poetics*, 41 (6), 750-769.

JURAFSKY, D. Y MARTÍN, J. (2008). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New Jersey: Prentice Hall.

LÓPEZ, I. (2010). Morochas, milongueras y percantas. Representaciones de la mujer en las letras de tango. *Espéculo. Revista de Estudios Literarios* 45. Disponible en: https://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), 45-60.

MAYRING, P. (2000). Qualitative Content Analysis. *Forum: Qualitative Social Research* 2(1). Disponible en: https://www.qualitative-research.net/index.php/fqs/article/view/1089/2385

MAXWELL, J. (2004). Reemergent scientism, postmodernism, and dialogue across differences. *Qualitative Inquiry* 10 (1), 35-41.

MIMNO, D., WALLACH, H., TALLEY, E., LEENDERS, M., y MCCALLUM, A. (2011). Optimizing semantic coherence in topic models. *Empirical methods on natural language processing*. Disponible en: https://mimno.infosci.cornell.edu/papers/mimno-semantic-emnlp.pdf

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

MURH, T. (1997). *Atlas.Ti-Visual Qualitative Data Analysis-Management-Model Building-Release 4.1*. Berlín. Consultado de https://personalpages.manchester.ac.uk/staff/andrew.j.howes/manshort.pdf

NAVARRO-COLORADO, B., TOMÁS, D., VÁZQUEZ, S., MOREDA, P., IZQUIERDO, R., SAQUETE, E., LLOPIS, F. (2011). Procesamiento automático de metáforas con métodos no supervisados. *Procesamiento del Lenguaje Natura,* 47, 345-346.

RAGIN, C. (1987). *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. California: University of California Press.

REYNOSO, C. (2007). El lado oscuro de la descripción densa. *Anthropologika. Revista de Estudio e Investigaciones en Antropología*, 1(1), 136-196.

RICHARDS, L. (1999). *Using Nvivo In Qualitative Research*. London: Sage.

ROSATI, G. y CHAZARRETA, A. (2017). El Qualitative Comparative Analysis (QCA) como herramienta analítica. Dos aplicaciones para el análisis de entrevistas. *Revista Latinoamericana de Metodología de las Ciencias Sociales* 7 (1) https://doi.org/10.24215/18537863e018

RUSSO. J. Y MARPEGÁN, D. (2000). *Letras de tango*. Buenos Aires: Basílico.

TIE, Y., BIRKS, M., y FRANCIS, K. (2019). Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine*, 7. Doi: 10.1177/2050312118822927

VASILACHIS, I. (2006). La investigación cualitativa. En Vasilachis I. (comp). *Estrategias de investigación cualitativa (*pp-pp. 23-6). Barcelona: Gedisa.

WIEDEMANN, G. (2016). *Text mining for qualitative data analysis in the social sciences. A study on democratic discourse in germany*. Berlín: Springer.

WILLENPART, L. (2011). El tango: Temas y motivos. *Verba Hispánica*, 9 (1), 219-230.

Anexos:

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

Fuente: Elaboración propia en base a datos recolectados de todotango.com

Tabla A.2 *k*=13

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

Fuente: Elaboración propia en base a datos recolectados de todotango.com

Tabla A.3 *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

Fuente: Elaboración propia en base a datos recolectados de todotango.com

Tabla A.4 *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

Fuente: Elaboración propia en base a datos recolectados de todotango.com

Autor.

Germán Rosati

Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET); Universidad Nacional de San Martín (UNSAM); Instituto de Altos Estudios Sociales (IDAES); Programa de Investigaciones sobre el Movimiento de la Sociedad Argentina (PIMSA), Argentina

Investigador Asistente del CONICET. Doctor en Ciencias Sociales por la Universidad de Buenos Aires (UBA). Magíster en Generación y Análisis de Información Estadística por la Universidad Nacional de Tres de Febrero (UNTREF). Licenciado en Sociología por la Universidad de Buenos Aires (UBA). Docente en la carrera de Sociología y Antropología (IDAES-UNSAM).

E-mail: german.rosati@gmail.com

Citado.

ROSATI, Germán (2022). "Procesamiento de Lenguaje Natural aplicado a las ciencias sociales. Detección de tópicos en letras de tango". *Revista Latinoamericana de Metodología de la Investigación Social - ReLMIS*. N°23, Año 12, pp. 38-60.

Plazos.

Recibido: 23/12/2019. Aceptado: 20/10/2020.