

# **Los temas del rock nacional. Una aproximación mediante técnicas de minería de texto.**

Germán Rosati y Laia Domenech.

Cita:

Germán Rosati y Laia Domenech (2021). *Los temas del rock nacional. Una aproximación mediante técnicas de minería de texto. XIV Jornadas de Sociología. Facultad de Ciencias Sociales, Universidad de Buenos Aires, Buenos Aires.*

Dirección estable: <https://www.aacademica.org/000-074/715>

# Los temas del rock nacional. Una aproximación mediante técnicas de minería de texto

German Rosati (CONICET/IDAES-UNSAM/PIMSA)

Laia Domenech (IDAES-UNSAM)

## 1. Introducción

El rock argentino ha sido objeto de una gran cantidad de estudios que lo han abordado a partir de diversas dimensiones. Analizado tanto como fenómeno cultural que condensa diversos aspectos de resistencia y determinaciones generacionales (Alabarces 1995), como espacio de una ideología libertaria y antiautoritaria (Vila 1985), como expresión de culturas populares juveniles (Alabarces 1995), a partir de su ideario (Pujol 2007). El análisis del contenido de las letras de rock juega un rol importante, aunque en muchos casos subordinado metodológicamente a otros indicadores o dimensiones: entrevistas con los autores/músicos, con las audiencias, etc. Ahora bien, ¿cuáles son los temas del rock? ¿Es posible caracterizar las etapas del rock nacional según los temas que aborda su lírica a lo largo del tiempo?

Existen relativamente pocos trabajos que aborden de forma central la “lírica” del rock, sus poéticas y particularmente, los temas de los que el género habla. Como plantea Conde (2007), el rock aún parece estar esperando su “santificación” por parte del mundo académico. En este sentido, los dos volúmenes de “Poéticas de Rock” (Conde 2007 y 2008) constituyen un hito en el análisis de diversos aspectos formales en la poesía del rock argentino. En esta compilación se abordan los estilos, temas y poéticas de diversos autores del rock nacional: el rock previo a La Balsa, las canciones de Lito Nebbia, Luis Alberto Spinetta, Arco Iris y muchos otros. También Conde (2020), ha encarado un análisis formal de los temas y recursos literarios en la obra de Charly García, con especial énfasis en el disco Clics Modernos del año 1983. Rodríguez Lemos y Secul Giusti (2011) realizan un análisis discursivo de un corpus de letras de rock correspondiente a la llamada “primavera democrática” (Clics Modernos, Soda Stereo y Oktubre). Un ejercicio similar, es realizado por Abot, Campolongo y Suredan (2006) sobre León Gieco.

Tanto en los trabajos que toman como eje central las poéticas o los estilos como en aquellos en los que el análisis de las letras no es el eje fundamental, se observan algunos rasgos comunes. Por un lado, se basan en un corpus relativamente acotado de letras, ya sea circunscrito a algunos discos o a algunos autores. Por otro, aplican un enfoque que podríamos llamar “hermenéutico” o “interpretativo” a partir del cual buscan reconstruir ciertos temas o constantes en el estilo de las letras analizadas. De esta forma, centran su análisis en lo que podríamos llamar “el canon” del rock: aquellos autores y bandas paradigmáticos o representativos de una forma de producir letra y música en un período determinado. Este

enfoque permite, efectivamente, un análisis de gran profundidad conceptual de los diferentes corpus mencionados.

Así, Alabarces (1995: 58) detecta algunos temas relevantes en el rock de los 60-70: "... el tópico de la ciudad, el sistema, la gran sociedad como metáforas condensadoras de los significados negativos están omnipresentes". Conde (2002), por su parte, define la soledad, el fracaso, la muerte como temas fundamentales y relativamente transversales en la obra de Charly García. Rodríguez Lemos y Secul Giusti (2011) destacan el sexo y el erotismo así como la experiencia con drogas y estimulantes como dos temas recurrentes del rock entre 1983 y 1986.

No obstante, en el presente trabajo, intentamos un enfoque diferente (y vale la pena aclararlo: complementario). Realizamos una primera aproximación a las preguntas planteadas más arriba a partir del análisis de un corpus extensivo de unas 20.000 letras del rock argentino (construido a partir de técnicas de web scraping). La escala del corpus constituye, al mismo tiempo, una ventaja y una limitación.

Por un lado, se trata de un corpus significativamente más extenso del que habitualmente utilizan los estudios sobre la temática. Esto nos permitirá, en cierta forma, correr del canon y realizar un análisis de "todas" las letras del rock, en busca de sus temas, en lugar de centrar el análisis en ciertos autores clásicos. Las comillas son precisas: efectivamente, no se trata de la totalidad de las letras producidas por el rock nacional. De hecho, como hemos mencionado en otra parte (Rosati 2022), "... 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". No obstante, resulta relevante marcar que sí implica desdibujar el concepto de canon y no analizar exclusivamente las obras "representativas" de los diversos períodos del rock argentino<sup>1</sup>.

Por otro lado, esa escala implica la imposibilidad de abordar una lectura en profundidad de la totalidad de los textos que conforman el corpus. Esto supone la necesidad de aproximarnos a una estrategia analítica que tiene puntos de contacto con lo que Moretti (2015) llama "lectura distante"<sup>2</sup>. Para esto, utilizaremos algunas técnicas de procesamiento

---

<sup>1</sup> Tal y como plantea Moretti (2015: 79:108) en "El matadero de la literatura" podríamos preguntarnos qué sucede con la "enormidad de lo no leído": aquellos textos que quedan fuera del canon (generalmente validado por el mercado y los consumidores) pero que, en muchos casos, constituyen elementos decisivos para su conformación. Moretti analiza el uso del recurso de los "indicios" en los cuentos policiales en Inglaterra de fines del siglo XIX, incorporando a su estudio a cuentos y autores olvidados (los que llama "rivales") como consecuencia de la influencia de Conan Doyle y Jane Austen.

<sup>2</sup> "... la historia literaria se convertirá rápidamente en algo muy distinto a lo que conocemos hoy: pasará a ser de segunda mano, un collage de investigaciones ajenas sin una sola lectura textual directa. Aún ambiciosa (...) pero con una ambición directamente proporcional a la distancia con respecto al texto: cuanto más ambicioso sea proyecto, mayor deberá ser la distancia (...) Y si queremos ver más allá del canon (...) la lectura cercana no servirá (...) ya sabemos leer textos; ahora tenemos que aprender a no leerlos. Lectura distante, donde la distancia, cabe repetirlo, es condición

de lenguaje natural<sup>3</sup>, particularmente vinculadas a la detección y al modelado de tópicos. Se trata de los primeros resultados exploratorios y preliminares aún de una investigación en curso. Presentaremos una primera especificación (aún en elaboración) de uno de los modelos más utilizados y más simples para la detección de tópicos en corpus textuales: Latent Dirichlet Analysis.

La ponencia se organiza del siguiente modo: en el apartado siguiente exponemos el modo de construcción del corpus y delineamos algunas características de la técnica utilizada para el análisis de tópicos. En el apartado 3 pasamos revista a algunos de los principales resultados del ejercicio de modelado de tópicos: primero, realizamos una caracterización de los mismos en función de las palabras más probables; en segundo término, validamos el modelo evaluando su funcionamiento en algunas letras de canciones de gran popularidad; en tercer lugar, analizamos la composición de los tópicos y su evolución a lo largo del tiempo en las 5 bandas más escuchadas en el último año según Spotify; cerramos este apartado presentando la evolución temporal de los tópicos detectados en todo el corpus y planteando algunas preguntas al respecto. La ponencia se cierra con el planteo y la especificación de nuevas preguntas y futuros pasos en la investigación.

## 2. Metodología

La recolección de datos se realizó a partir de un “scrapeo” de todas las letras de canciones que se encuentran en la página <https://rock.com.ar/> con Python. Esto dio como resultado una base de datos con 20.271 canciones de más de 700 artistas<sup>4</sup>. Las técnicas de *scraping* -“raspado” o “rascado”- suponen la captura de información disponible en sitios web, generalmente, con bajos grados de estructuración (Mitchell, 2015).<sup>5</sup> Además de la letra de cada canción, se recabó información complementaria (llamada, en la jerga, metadata): el nombre de la banda/artista, el nombre del disco en el que fue incluida y la fecha de lanzamiento del disco en cuestión. Un punto importante a tener en cuenta es que la fecha con la que contamos es la de edición del disco en el que se encuentra la letra analizada. Esto plantea una limitación al análisis, en tanto la fecha de composición/escritura de una canción no necesariamente coincide con la de la edición en un disco.

Luego de una limpieza de texto del corpus de estas canciones (limpieza de stopwords, pasaje a minúsculas, remoción de símbolos y tildes), se aplicó uno de los

---

del conocimiento, es lo que permite colocar el foco en unidades mucho más pequeñas o mucho más grandes que el texto: recursos, temas, tropos...” (Moretti 2015: 62-63).

<sup>3</sup> Un antecedente fundamental en el caso argentino es el trabajo pionero de Darío Cantón (1972). En el mismo realiza un análisis cuantitativo de las letras de 99 tangos cantados por Carlos Gardel.

<sup>4</sup> El código utilizado tanto para el web scraping como para el análisis del corpus puede encontrarse en [https://github.com/gefero/factor\\_data\\_rock\\_nacional](https://github.com/gefero/factor_data_rock_nacional)

<sup>5</sup> Hemos desarrollado con más detalle algunas de las ventajas y desventajas de la utilización de estas técnicas en Rosati (2022).

métodos más conocidos para la detección automática de tópicos en corpus textuales *Latent Dirichlet Allocation* o LDA.

La intuición detrás de LDA (Blei, 2012: 79-80) es que cada documento ( $d$ ) del corpus ( $C$ ) puede exhibir varios tópicos: puede hablar de varios temas simultáneamente. En este sentido<sup>6</sup>, un tópico es definido formalmente como una distribución de probabilidad a lo largo de un vocabulario. Por ejemplo, si existiera un tópico como *amor* sería esperable que palabras como “amor”, “pena”, “sufrimiento”, tuvieran altas probabilidades de inclusión en el mismo. En cambio, palabras como “casa”, “hermano”, “barrio” estarían más asociadas a un tema que hablara acerca de la *vida cotidiana*. LDA asume que el texto que se “genera” es resultado de un proceso dos etapas:

1. se selecciona aleatoriamente 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  $V$

El resultado es que cada documento ( $d$ ) muestra cierta composición diferenciada de tópicos (paso 1). A su vez, cada palabra ( $w$ ) del texto es “sorteada” de una de las “bolsas de palabras” que conforman un tópico (paso 2.ii) donde el tópico seleccionado es elegido de la distribución de tópicos de ese documento (paso 2.i).

El objetivo central es descubrir los temas a los que alude un determinado conjunto de documentos. Lo único observado es el conjunto de documentos. La estructura de tópicos puede ser considerada como un conjunto de variables no observadas. 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  $\beta_k$  es una distribución de probabilidad sobre el vocabulario. La proporción de tópicos para el  $d$ -ésimo documento están indexadas por  $\theta_d$ , donde  $\theta_{d,k}$  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

---

<sup>6</sup> Esta sección se basa en Blei (2012).

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.

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<sup>7</sup>. Ahora bien, el método utilizado tiene algunas supuestos:

1. cada documento  $d$  se compone de varios tópicos, es una mixtura
2. un tópico se compone de palabras: un tópico ES una distribución de probabilidad sobre la totalidad de palabras del vocabulario
3. los tópicos son previos a los documentos, los preexisten y la distribución de probabilidad sobre el vocabulario es constante
4. las palabras no tienen orden y el orden de los documentos no es relevante
5. existe una cantidad fija de tópicos. Esto puede ser un problema al analizar corpus con documentos de épocas muy diferentes<sup>8</sup>.

Utilizando LDA se buscó detectar los tópicos más relevantes en el corpus de letras de rock. 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*. 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.

Puede decirse que un número de tópicos más grande tiende a hacer posible una “alta resolución” de la estructura latente del corpus. Ahora bien, una gran cantidad de tópicos

---

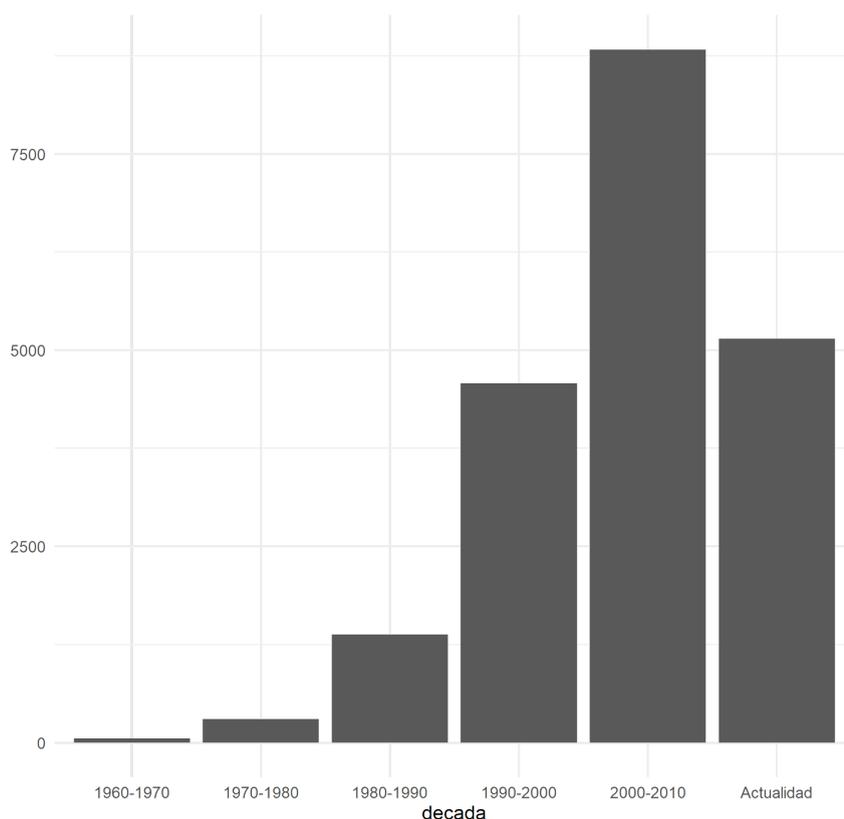
<sup>7</sup> En términos generales, se trata de un problema computacionalmente intratable por lo cual es ineludible recurrir a aproximaciones numéricas.

<sup>8</sup> Blei (2012) expone varios métodos para flexibilizar este supuesto.

puede llevar a la existencia de tópicos de “baja calidad”, es decir, tópicos que no son interpretables (Mimno et al., 2011; Chang et al., 2009): como suele ser habitual, la complejidad del modelo y la posibilidad de interpretar sus resultados de forma parsimoniosa se encuentran en tensión permanente<sup>9</sup>.

### 3. Análisis descriptivo del corpus

**Gráfico 1. Frecuencia de letras según década.**



**Fuente:** elaboración propia en base a scraping de sitio rock.com.ar

La frecuencia de las canciones nos permite ver aquello que mencionamos previamente: no estamos haciendo un análisis de “todas” las canciones del rock nacional, sino que estamos recopilando la información que encontramos en un sitio donde ya hubo una curaduría previa. En este sentido, vemos que el grueso de las canciones se encuentra

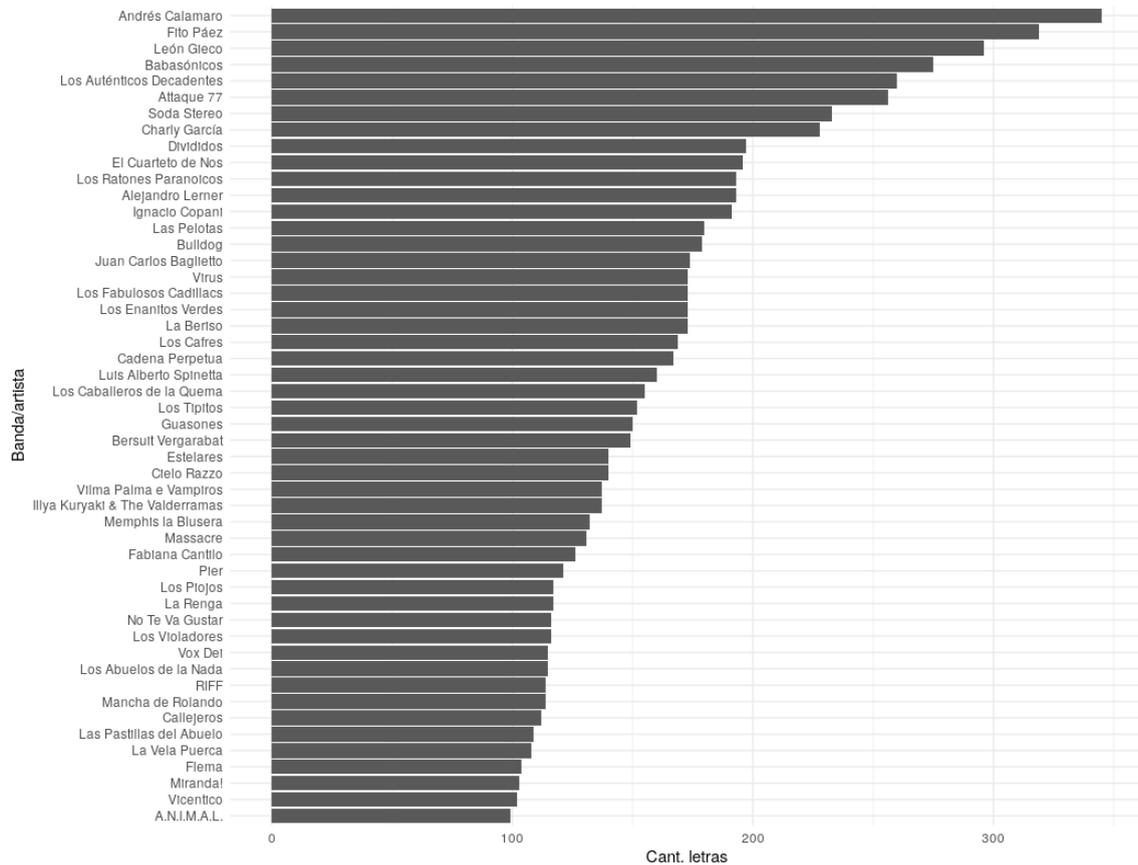
---

<sup>9</sup> Existe otro conjunto de métricas *-coherence*, por ejemplo.- que se centran en la interpretabilidad (Mimno et al., 2011). A su vez, es necesario mencionar la problemática de ciertos procedimientos retóricos y estilísticos en textos literarios. El modelado de tópicos opera sobre la co-ocurrencia de palabras para la definición de tales tópicos. Así, la presencia de recursos y tropos como las metáforas pueden generar problemas de identificación. No obstante, algunos estudios han intentado utilizar el modelado de tópicos (y, particularmente, LDA) como detectores de metáforas. Navarro-Colorado et al (2011).

en la década de los 2000-2010. El “pico” de canciones está en 2004, donde se registraron 1.212 canciones.

Podemos realizar también un análisis descriptivo de los artistas con mayor cantidad de letras en la base de datos

**Gráfico 2. Frecuencia de letras de los 50 artistas con mayor cantidad de letras**



**Fuente:** elaboración propia en base a scraping de sitio rock.com.ar

Podemos observar una distribución de frecuencias sumamente heterogénea entre los artistas. Se observa como los primeros 50 artistas (en cantidad de letras), que representan el 6% de los artistas registrados, concentran unas 8200 letras, es decir, aproximadamente un 40% del total de letras en la base de datos utilizada.

#### 4. Resultados principales

Luego de iterar sobre modelos con distintas cantidades de tópicos, definimos el modelo de 10 tópicos. En él, hallamos 9 temas conceptualmente relevantes: el amor, el rock y sexo, la angustia, la violencia, imágenes de la naturaleza, el desamor o los fantasmas, la música, imágenes de la vida cotidiana y el amor o desamor. Uno de los tópicos fue

caracterizado como misceláneo. Pueden verse, a continuación, las 10 palabras de mayor probabilidad de pertenencia a cada tópico

**Tabla 1. Tópicos y las palabras de mayor probabilidad de pertenencia**

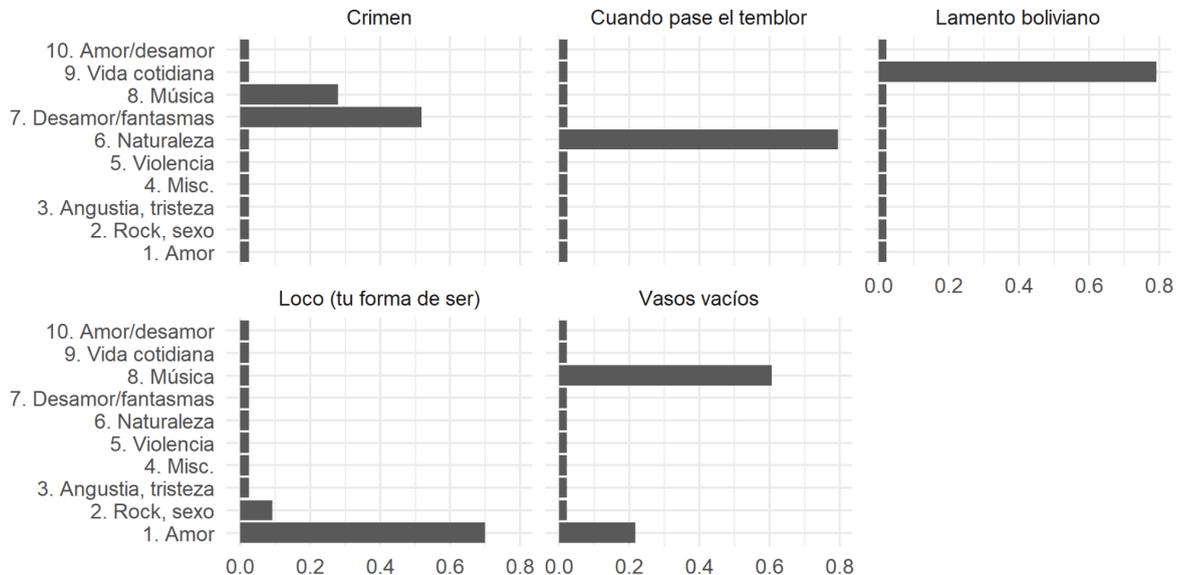
Tópico	Palabras
1. Amor	vuelvo esperar noches aire vivir sueños luna pasa buscarte amores
2. Rock/sexo	gusta roll nena bailar parar angeles digo caliente alcohol ritmo
3. Angustia/tristeza	abismo angustia digan gato pasa olvida hombre crees ciudad llores
4. Misc.	vuelvas dando perdi vienen mente pibe argentina tenerte tierra rios
5. Violencia	hombre pasara tierra fuego sangre tira gritos viene falta alguien
6. Naturaleza	puedes veras nube vuelve caminos desierto seguir jardin nombre cambia
7. Desamor/fantasmas	nena dejame dificil importante debes vivir busques necesita tonto mueve
8. Música	musica mirar vivir pensar espejo vengo control realidad real ritmo
9. Vida cotidiana	vuelvas revolucion historia casa calle mujer dime alguien años nene
10. Amor/desamor	verte recuerda dame conozco necesito amigo suerte empezo arde fuego

#### 4.1 Análisis de canciones

Una forma habitual de validar modelos de detección de tópicos consiste en analizar ciertos documentos de los cuales se conoce el contenido y analizar si los resultados del modelo (la proporción de tópicos que detecta en cada documento) es adecuada.

Seleccionamos las canciones de rock nacional más escuchadas en Spotify<sup>10</sup>: *Crimen* de Gustavo Cerati, *Cuando pase el temblor* de Soda Stereo, *Lamento boliviano* de Los Enanitos Verdes, *Loco (Tu forma de ser)* de Los Auténticos Decadentes y *Vasos Vacíos* de Los Fabulosos Cadillacs.

**Gráfico 3. Composición de tópicos de 5 canciones escuchadas en Spotify**



Fuente: elaboración propia en base a scraping de sitio rock.com.ar

*“La espera me agotó  
no se nada de vos  
dejaste tanto en mí.  
En llamas me acosté  
y en un lento degradé  
supe que te perdí”*

Este fragmento de *Crimen* ilustra el abordaje del amor que tiene la canción. El modelo detecta un valor de 0,51 en ese tópico. A su vez encontramos prevalencia alta del tópico de música, la cual se explica por algunas palabras que aparecen también en el tópico: como pensar, vivir.

*Cuando pase el temblor* de Soda Stereo hace alusión a múltiples imágenes de la naturaleza:

<sup>10</sup> Fuente: <http://www.che-loco.com/>

*“Yo  
caminaré entre las piedras  
hasta sentir el temblor en mis piernas (...)  
Estoy sentado en un cráter desierto  
sigo aguardando el temblor  
en mi cuerpo”.*

Las cuales el modelado identifica exitosamente, alcanzando un valor de casi un 0,8 en ese tópico.

*Lamento boliviano* está caracterizado principalmente por el tópico de imágenes de la vida cotidiana. Esto se debe a que el modelo identifica las partes de la canción donde describe estar *“Borracho y loco”* o dice *“Nena, no te peines en la cama”*.

Cuando analizamos la letra de *Loco (tu forma de ser)*, podemos entender que nos habla explícitamente de un enamoramiento en una fiesta, por lo que el modelo da altos puntajes en el tópico de amor y rock y sexo.

*“Te vi bailar, brillando con tu ausencia  
sin sentir piedad  
chocando con las mesas.  
Te burlaste de todos,  
te reíste de mí  
Y tus amigos escaparon de vos.  
Y a mí me volvió loco tu forma de ser.”*

Finalmente, en *Vasos vacíos* el modelo logra detectar los fragmentos que hablan de música:

*“Levanta los brazos mujer  
Y ponte esta noche a bailar”*

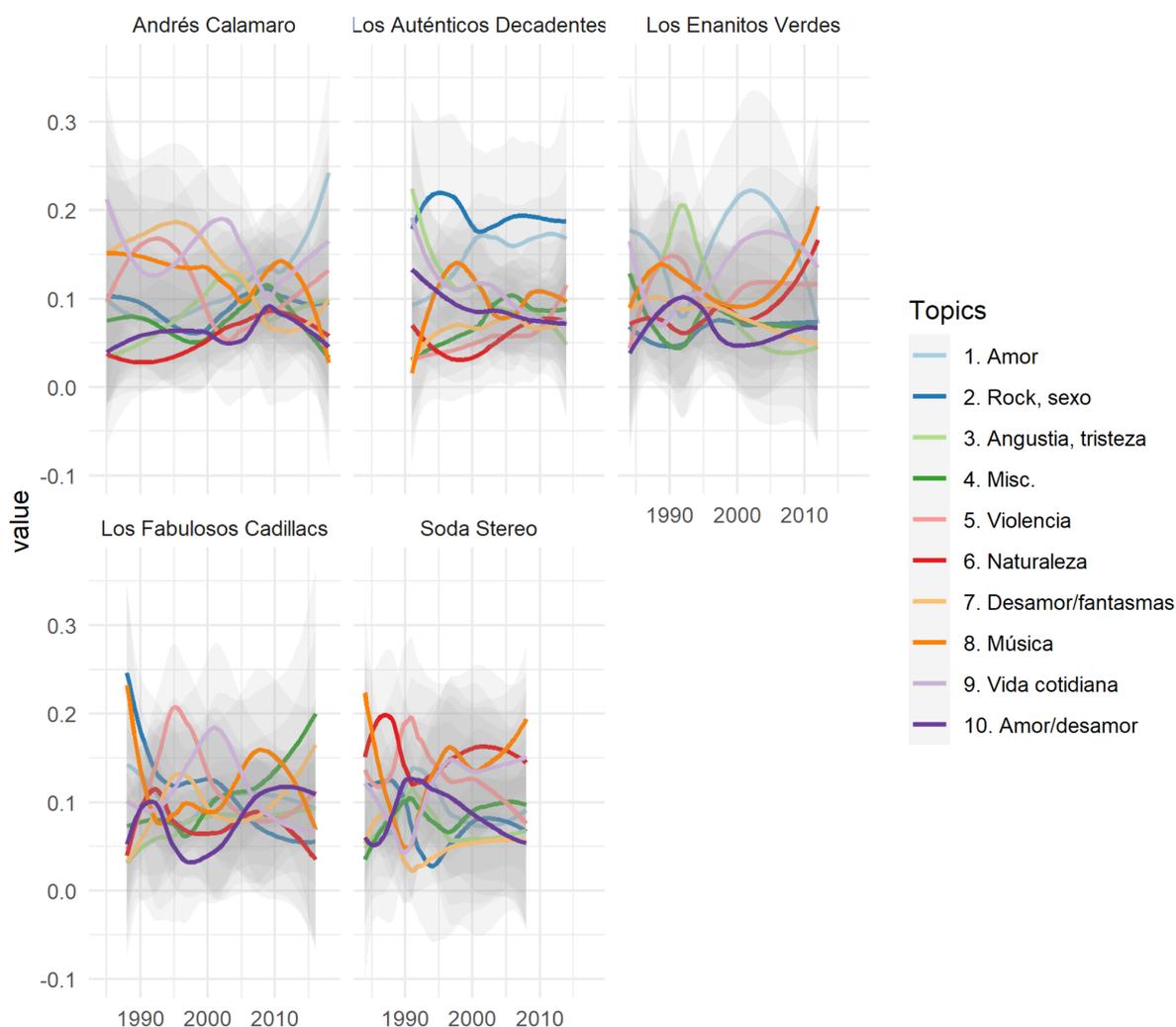
Así como los que tocan el tema amoroso:

*“A ti te quiero decir  
no te preocupes mi amor  
que yo te voy a entender  
que yo te voy a querer.”*

## 4.2 Análisis de bandas

Ahora bien, ¿pueden encontrarse diferencias entre los tópicos de diferentes bandas? Para abordar este problema seleccionamos las bandas de rock argentino más escuchadas en Spotify (con el mismo criterio que el apartado anterior). A fines de reducir el efecto de los tópicos predominantes en las canciones, realizamos el análisis a partir de la frecuencia de los tópicos máximos: cada canción fue clasificada seleccionando el tópico que concentra el mayor valor. Por ejemplo, si una canción mostraba una distribución de 80% tópico *amor*, 5% tópico *rock/sexo* y 15% *violencia*, consideramos que la misma tenía como tema predominante el “amor”.

**Gráfico 4. Evolución en el tiempo de la composición de tópicos de 5 artistas más escuchados en Spotify**



**Fuente:** elaboración propia en base a scraping de sitio rock.com.ar

Encontramos que los temas varían en distintos niveles entre los distintos artistas. Para Andrés Calamaro encontramos tres picos de tópicos mayoritarios: las imágenes de desamor y fantasmas durante los '90-2000, la vida cotidiana a comienzos del nuevo milenio y el amor a medida que tendemos hacia la actualidad. También aparece con bastante prevalencia el tópico de la música.

En el caso de Los Auténticos Decadentes, se mantiene a lo largo del tiempo como mayoritario el tópico que habla de rock y sexo. A partir del '2000, toma relevancia también el tópico de amor.

Los Enanitos Verdes tienen dos picos de tópicos mayoritarios: la angustia a principios de los '90 y el amor una vez que este otro desciende. En menor medida aparecen también las imágenes sobre la vida cotidiana.

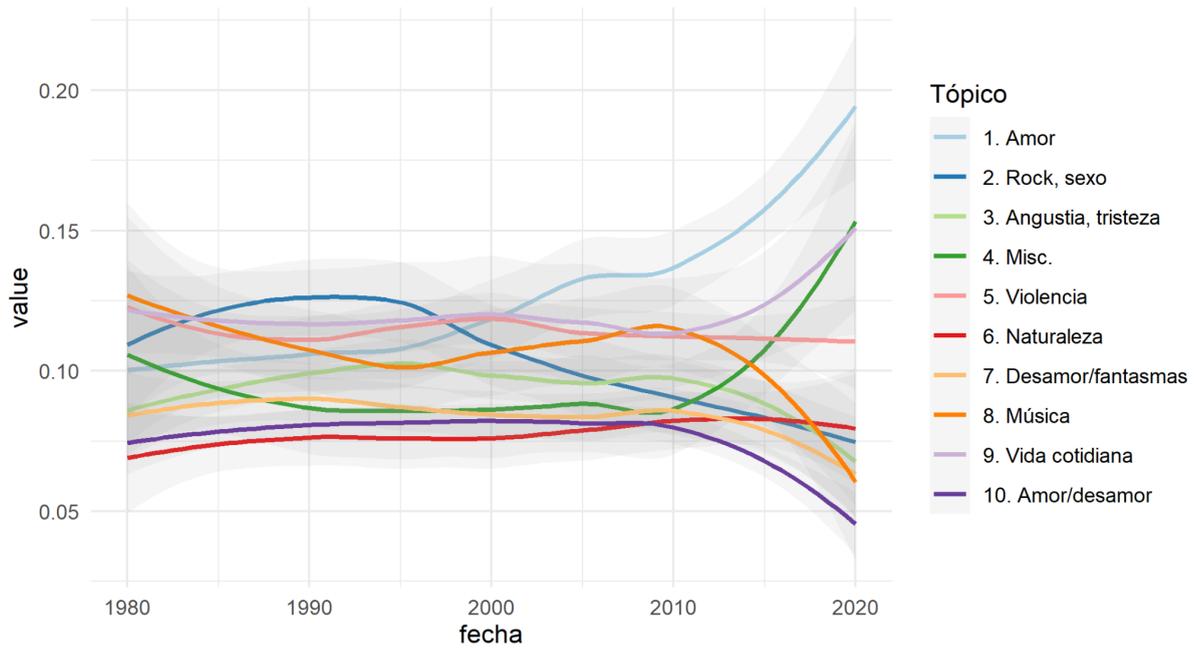
Los Fabulosos Cadillacs tienen una distribución más irregular: comienzan con una alta prevalencia del rock/sexo y la música, que una vez que desciende es reemplazada por el tópico de violencia. A partir de los 2000 aparecen las imágenes de la vida cotidiana, y a mediados de la década reaparece el tópico de la música. A medida que tendemos hacia la actualidad, toma relevancia el tópico misceláneo, lo que nos da la idea de la emergencia de un nuevo (o nuevos) temas a detectar en el rock nacional.

Por último, Soda Stereo presenta dos tópicos que aparecen con relevancia en distintos períodos del tiempo: la música e imágenes de la naturaleza. Ambas mantienen una prevalencia alta durante gran parte del período, excepto en los '90 que alcanza un pico mayoritario el tópico de la violencia.

#### **4.3 Evolución de tópicos**

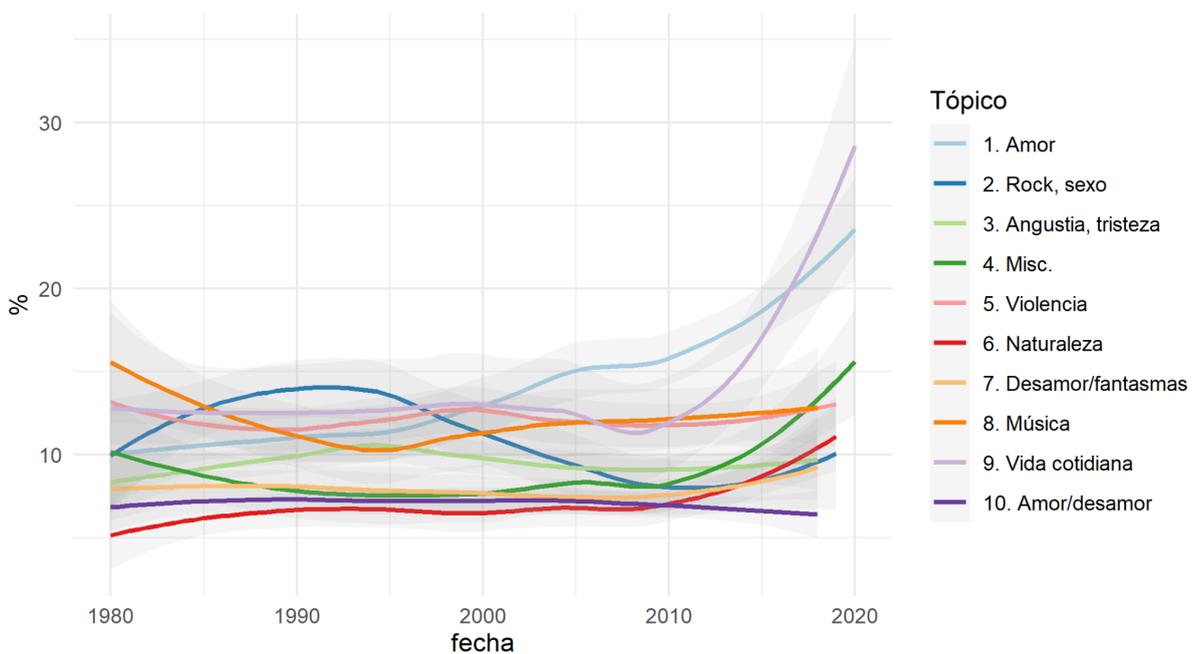
Para calcular la variación de los temas de los que se habló en el rock a lo largo del tiempo, se realizó un cálculo de la media de tópicos y el análisis de la frecuencia de los tópicos mayoritarios entre las distintas décadas.

**Gráfico 5. Evolución en el tiempo de la media de tópicos (calculada sobre el total de canciones del corpus)**



**Fuente:** elaboración propia en base a scraping de sitio rock.com.ar

**Gráfico 6. Evolución en el tiempo del tópico mayoritario de cada canción del corpus**



**Fuente:** elaboración propia en base a scraping de sitio rock.com.ar

En el primer gráfico encontramos la prevalencia del tópico de amor sobre los demás a partir de la década del 2000 en adelante. Entre la mitad de la década del '80 y hasta

finales de los '90, el tópico de rock y sexo tiene la mayor prevalencia, y previo a eso, el bastante similar tópico que habla sobre música. Los segundos tópicos que mayor valor tienen y se mantienen a valores relativamente estables a lo largo del tiempo son aquellos que hablan de violencia y de imágenes de la vida cotidiana.

Vemos resultados similares en el gráfico de tópicos mayoritarios. Prevalece la tendencia mayoritaria de música-rock-amor con puntos de cambio en mediados de los '80, finales de los '90 y a partir de los 2000. Sin embargo, aparece con mayor prevalencia a partir de la década de 2010 el tópico de imágenes de la vida cotidiana, superando al tópico de amor a medida que nos acercamos a la actualidad.

## 5. Comentarios finales y nuevas preguntas

En el presente trabajo hemos presentado los primeros resultados (aún descriptivos y exploratorios) de la aplicación de un modelo de detección de tópicos a un corpus extenso de unas 20.000 letras de rock argentino.

En este sentido, parece posible ver que el uso de este tipo de análisis permite escalar la cantidad de textos incorporados a los corpus analíticos, lo cual trae una doble consecuencia. Por un lado, implica la necesidad (y la imposibilidad) de leer *todos* los textos de un corpus. Esto supone encarar una aproximación al análisis textual de documentos a partir de una "lectura distante" (Moretti 2015). Por otro lado, este enfoque debería permitir complementar y ampliar el análisis en profundidad de pocos documentos, particularmente en tanto brinda la posibilidad de incorporar documentos que no suelen estar incluidos en los "cánones" oficiales (en este caso, los autores clásicos del rock, como Charly García, Spinetta, etc.).

A su vez, nos fue posible realizar un primer análisis sobre la diversidad de los temas en diferentes autores, tomando como ejemplo cinco bandas populares, la comparación sobre las temáticas de canciones individuales.

Finalmente, fue posible analizar la evolución de los tópicos a lo largo del tiempo. En este sentido, se hicieron visibles algunas tendencias que es interesante marcar. En el primer período (toda la década del 1980 hasta mediados de los años noventa) se observa una prevalencia similar en todos los tópicos. No obstante, el tema vinculado a la música y al rock y al sexo toma cierta preeminencia. ¿En qué medida se vincula esta tendencia con la importancia que adquirieron grupos como Virus, su "estrategia de la alegría" y el destape cultural que se daba como consecuencia de la salida de la dictadura militar?

Este tópico alcanza su máximo valor a principios de la década de 1990 y luego comienza a desvanecerse. Como contracara, los temas que tocan aspectos del amor y la vida cotidiana ganan peso a partir de los años 2000. En este sentido, podría plantearse si

este proceso es coherente con el surgimiento de las llamadas “literaturas del yo” (por ejemplo, en narrativa o en poesía).

Por último, el tópico que hemos llamado “Misceláneo” y que no resulta fácilmente interpretable parece mostrar un incremento bastante veloz hacia el final del periodo. ¿Hasta qué punto sería posible interpretar este movimiento como la emergencia de un tópico nuevo que aún se encuentra en un estado embrionario y sin una delimitación clara?

Estas preguntas muestran que será necesario avanzar en la flexibilización de algunos de los supuestos de la versión estándar de LDA. Particularmente problemático es el supuesto que asume que los tópicos no pueden cambiar con el tiempo. Si bien fue posible detectar algunos cambios interesantes en la evolución de los tópicos a lo largo del tiempo, será necesario incorporar la posibilidad de permitir que las palabras que forman parte de los tópicos, cambien a lo largo del tiempo. En este sentido, las técnicas de *dynamic topic modeling* (Blei 2012) y/o de *structural topic modeling* (Roberts, Stewart y Tigley 2019) pueden ser de gran utilidad.

Al mismo tiempo, es importante mencionar que los tópicos detectados apuntan a constituir una primera aproximación al problema. Como tales, constituyen una primera “capa” de analítica que debería ser susceptible de mejoras. Así, un segundo paso será el análisis del “tono” de las letras de rock. ¿Se trata de un tono positivo o negativo? ¿Cuáles son las emociones predominantes en las letras analizadas? ¿Cuál es su evolución temporal? En una palabra: realizar análisis de sentimiento sobre el corpus.

Algo similar podemos decir sobre los “objetos” de las letras de rock: ¿cuáles son los lugares más mentados? ¿Qué tipo de personas y objetos aparecen en las letras? ¿Cómo se vinculan a su vez con los tópicos que hemos detectado? Estos problemas podrían ser abordados mediante técnicas de reconocimiento de entidades (Named Entity Recognition).

## 6. Bibliografía citada

Abot, Marcos, Campolongo, Pablo y Suredan, María Jorgelina, *El Noticiero del rock. León Gieco y el imaginario social argentino*, Facultad de Periodismo y Comunicación Social, Universidad Nacional de La Plata, Tesis de grado.

Alabarces, Pablo (1995). *Entre gatos y violadores. El rock nacional en la cultura argentina contemporánea*. Buenos Aires: Colihue.

Blei, David (2012) “Probabilistic topic models”. *Communications of the ACM* Vol. 55, N°4, p. 77-84.

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

Chang, Jonathan, Boyd-Garber, Jordan, Wang, Chong, Gerrish, Sean, y Blei, David (2009) “Reading tea leaves: How humans interpret topic models”. *Neural information*

*processing systems.* Disponible en:  
<https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models>.

Conde, Oscar (2007). *Poéticas del rock. Vol 1.* Buenos Aires: Marcelo Héctor Oliveri Editor.

Conde, Oscar (2008). *Poéticas del rock. Vol 2.* Buenos Aires: Marcelo Héctor Oliveri Editor.

Conde, Oscar (2020). *Charly García, 1983. Acerca de Clicks modernos.* Buenos Aires: UNIPE Editorial Universitaria.

Mimno, David; Wallach, Hannah; Talley, Edmund; Leenders, Miriam y McCallum, Andrew (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, Ryan (2015). *Web scraping with python: Collecting data from the modern web.* California: O'Reilly.

Moretti, Franco (2015). *Lectura distante.* Buenos Aires: Fondo de Cultura Económica.

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 Natural*, n°47, p. 345-346.

Pujol, Sergio (2007). *Las ideas del rock. Genealogía de la música rebelde.* Rosario: Homo Sapiens.

Roberts, Margaret; Stewart, Brandon; Tingley, Dustin (2019). "stm: An R Package for Structural Topic Models", *Journal of Statistical Software*, Vol. 91, n°2.

Rodríguez Lemos, Federico y Secul Giusti, Cristian (2011). *Si tienes voz, tienes palabra. Análisis discursivo de las líricas del rock argentina en la "primavera democrática" (1983-1986)*, Facultad de Periodismo y Comunicación Social, Universidad Nacional de La Plata, Tesis de grado.

Rosati, Germán (2022). "Procesamiento de Lenguaje Natural aplicado a las ciencias sociales. Detección de tópicos en letras de tango", en *Revista Latinoamericana de Metodología de la Investigación Social*, aceptado y en prensa.

Vila, Pablo (1985). "Rock nacional, crónicas de la resistencia juvenil", en Jelin, Elizabeth (comp.). *Los nuevos movimientos sociales. Vol. 1*, Buenos Aires: CEAL.