

Tesis de Maestría

Descubrimiento de patrones temporales en un corpus de letras de música folklórica y del rock rioplatense

Bach, Ana Josefina

2016-10-07

Este documento forma parte de la colección de tesis doctorales y de maestría de la Biblioteca Central Dr. Luis Federico Leloir, disponible en digital.bl.fcen.uba.ar. Su utilización debe ser acompañada por la cita bibliográfica con reconocimiento de la fuente.

This document is part of the doctoral theses collection of the Central Library Dr. Luis Federico Leloir, available in digital.bl.fcen.uba.ar. It should be used accompanied by the corresponding citation acknowledging the source.

Cita tipo APA:

Bach, Ana Josefina. (2016-10-07). Descubrimiento de patrones temporales en un corpus de letras de música folklórica y del rock rioplatense. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires.

Cita tipo Chicago:

Bach, Ana Josefina. "Descubrimiento de patrones temporales en un corpus de letras de música folklórica y del rock rioplatense". Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. 2016-10-07.

EXACTAS UBA

Facultad de Ciencias Exactas y Naturales



UBA

Universidad de Buenos Aires



UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Descubrimiento de patrones temporales en un corpus de letras de música folklórica y del rock rioplatense

Tesis presentada para optar al título de Magíster en Explotación de Datos y
Descubrimiento de Conocimiento

Ana Josefina Bach

Director de tesis: Jose Castaño

Buenos Aires, 2016

Resumen

Culturomics es la aplicación de recopilación y análisis de datos para el estudio de la cultura humana. La minería de textos temporal se presenta como una herramienta para alcanzar los objetivos de Culturomics, mediante el procesamiento automático y el establecimiento de patrones para explicar la historia. El objetivo de este trabajo fue utilizar técnicas de explotación de datos y aprendizaje automático para detectar la existencia patrones temporales en las letras del rock y del folklore argentino. Para ello se armó un corpus de letras de rock y folklore comprendidas entre 1960 y 2014. A este corpus se le aplicaron técnicas de agrupamiento de tópicos y de clasificación para determinar la existencia de una relación entre los tópicos y los hitos históricos.

Palabras claves: Culturomics, Minería de textos temporal, LDA, Factorización matricial no negativa dinámica, VowpalWabbit, Word2Vec.

Abstract

Culturomics is the application of high-throughput data collection and analysis to the study of human culture. Text mining is presented as a useful methodology to achieve the goals of Culturomics, performing automatic processing, and setting patterns to explain history. The aim of this study is to use data mining techniques and machine learning to detect any patterns in Argentine rock and folklore songs throughout history. For this purpose, a corpus of rock and folk song lyrics was built considering the periods between the year 1960 and 2014. This corpus was subject to clustering and classification of topics techniques to determine the presence of a relationship between topics and milestones.

Keywords: Culturomics, Temporal Text Mining, LDA, Dynamic Non-negative Matrix Factorization, VowpalWabbit, Word2Vec.

Índice general

| | |
|---|-----------|
| 1. Introducción | 1 |
| 1.1. Culturomics | 1 |
| 1.2. Minería de textos temporal | 2 |
| 1.3. Trabajo anterior | 3 |
| 1.4. Objetivo de la tesis | 4 |
| 1.5. Organización de la tesis | 5 |
| 2. Técnicas utilizadas para el análisis y agrupamiento de textos | 6 |
| 2.1. Latent Dirichlet Allocation (LDA) | 6 |
| 2.2. Modelado dinámico de tópicos (<i>Dynamic Topic Modeling</i>) | 9 |
| 2.3. Factorización matricial no negativa (<i>Non-negative Matrix Factorization</i>) | 11 |
| 2.4. Word2vec | 13 |
| 2.5. Topics a través del tiempo (<i>Topics over the time</i>) | 17 |
| 3. Materiales y métodos | 20 |
| 3.1. Materiales | 20 |
| 3.2. Métodos | 26 |
| 3.2.1. Análisis exploratorio | 27 |
| 3.2.2. Clasificación | 51 |
| 4. Resultados | 56 |
| 4.1. Experimentos Exploratorios | 56 |
| 4.1.1. Ventanas de tiempo | 56 |
| 4.1.2. Corpus sin separar ventanas | 67 |
| 4.2. Clasificación | 75 |

| | |
|---|-----------|
| 4.2.1. Métricas de evaluación | 75 |
| 5. Conclusión | 81 |
| 5.1. Conclusión | 81 |
| 5.2. Trabajos a futuro | 82 |
| A. Lista de Stopwords | 83 |
| B. Gráficos de uso de las palabras a través del tiempo | 85 |

CAPÍTULO 1

Introducción

En este capítulo se introduce el concepto de *Culturomics* para la interpretación de las tendencias culturales y la relación de Culturomics con Minería de texto. Se presenta un trabajo anterior realizado sobre un corpus de letras de rock argentino, en el cual se aplicaron técnicas de minería de texto para detectar patrones de comportamiento temporales. Finalmente se define el objetivo de esta tesis.

1.1. Culturomics

Culturomics es un neologismo creado por investigadores de Harvard para referirse a una forma de lexicología computacional que estudia el comportamiento humano y las tendencias culturales reflejadas en el lenguaje y en el uso de palabras [Michel et al., 2011]. Inicialmente se utilizaron libros para su estudio, pero también se planea incorporar periódicos, manuscritos, mapas, obras de arte, y otras creaciones humanas. La clave de este enfoque es el procesamiento automático de millones de textos.

Los resultados de Culturomics son un nuevo tipo de evidencia en las humanidades. Al igual que con los fósiles de criaturas antiguas, el reto de Culturomics reside en la interpretación de esta evidencia.

Las características de una sociedad se pueden determinar a través de fuentes no tradicionales, como la música. Las obras musicales, como objeto susceptible de estudio en el campo de Culturomics, poseen información temporal y proveen un archivo de la evolución de la cultura.

“El análisis estilístico no tiene por qué considerar parámetros externos a la música tales como la ideología, las circunstancias políticas y sociales. Sin embargo, la historia

del estilo, según yo la veo, no puede explicarse sin referirse a aspectos de la cultura externos a la música” (Meyer, 2000, citado por [Astor, 2008] en “Música para la historia, historia para la música”).

En Culturomics es importante mirar el pasado, tratando de entender el fenómeno lingüístico y cultural a lo largo del tiempo. Un interrogante que surge es si el cambio del lenguaje está correlacionado a eventos importantes, si es que existen tales eventos, o a cambios culturales. También es útil encontrar qué tipo de cambio es propenso a ser más estable y qué tipo de cambio es más efímero [Tahmasebi et al., 2015].

Los desafíos técnicos de Culturomics descansan en la falta de conjuntos de datos y métodos para pruebas automáticas y evaluación a través de largos períodos de tiempo. El resumen automático de textos es un proceso de producción de un número limitado de oraciones que representan un conjunto de documentos. Debido a los largos períodos de tiempo que se analizan en Culturomics hay una necesidad de ir más allá del resumen tradicional, agregando la variable temporal. El resumen temporal semántico puede ayudar a los usuarios a trabajar con el exceso de información, y seguir el desarrollo de los temas de interés a lo largo de períodos más amplios de tiempo.

Según Suchanek y Preda [Suchanek y Preda, 2014] hasta el momento Culturomics sólo se limitó al estudio estadístico de palabras clave. Los autores proponen utilizar artículos de noticias para descubrir tendencias en la historia y en la cultura, explicándolas a través de reglas lógicas explícitas. Este campo lo denominan *Semantic Culturomics*, donde las bases de conocimiento ayudan a los humanos a entender la sociedad. Semantic Culturomics es un análisis de gran escala de documentos con la ayuda de bases de conocimiento, con el objetivo de descubrir, explicar y predecir las tendencias y eventos en la historia y la sociedad.

El puntapié inicial dado por Culturomics para el análisis automático de la literatura en inglés, constituye una fuente de motivación para aplicar algún tipo de análisis automático a otras fuentes cultura, específicamente, en letras de rock y folklore argentinas, que son las que mejor expresan los estados de ánimo de la sociedad.

Para obtener datos relevantes de un texto es necesario sistematizar el conjunto de la información contenida en el mismo.

1.2. Minería de textos temporal

Recopilar datos, organizarlos y analizarlos se puede hacer automáticamente, pero identificar, estructurar y utilizar la información, requiere aplicar cierta intuición y conocimiento. La minería de textos temporal se presenta como una metodología útil para alcanzar los objetivos de Semantic Culturomics, en cuanto al procesamiento automático, y al establecimiento de patrones para explicar la historia.

La minería de textos temporal (*Temporal Text Mining-TTM*) se refiere al descubrimiento de patrones temporales en textos recolectados a lo largo del tiempo. Debido a

que la mayoría de los textos poseen identificadores temporales, TTM tiene varias aplicaciones en múltiples dominios, como el resumen de eventos en artículos de noticias y el descubrimiento de tendencias de investigación en la literatura científica. Esta metodología se puede aplicar para interpretar los cambios culturales a lo largo de la historia de forma automática. Las aplicaciones de TTM se pueden agrupar en descubrimiento de tendencias y resumen de eventos.

Descubrimiento de tendencias: Se presentaron metodologías para la detección de tendencias emergentes en textos [Roy et al., 2002]. Se introdujo un marco de trabajo para la detección de anomalías o novedades en secuencias temporales [Ma, 2003]. En [Morinaga y Yamanishi, 2004] se consideraron tres tareas en el análisis de tendencias de tópicos: la identificación de una estructura de tópico, la detección de tópicos emergentes, y la caracterización de los tópicos. En [Perkio et al., 2004] aplicaron un modelo basado en una versión discreta multinomial de análisis de componentes principales, con el objetivo de explorar el comportamiento temporal de los tópicos. En [Wang y McCallum, 2006] presentaron el modelo *Topics over time*, que captura la popularidad de un tópico a través de la distribución beta. [Blei y Lafferty, 2006] propusieron *Dynamical topic model*, donde la distribución de las palabras en los tópicos y la popularidad se relaciona a través de las épocas, usando modelos de estado espaciales. En [Mølgaard et al., 2009] se estudiaron métodos temporales de minería de texto para recuperación de información musical. [Ahmed y Xing, 2012] propusieron *Infinite dynamic topic model (iDTM)*. *iDTM* se aplica para analizar la evolución de tendencias, distribución y número de tópicos a través del tiempo. En [Kleinberg, 2003] se busca modelar la explosión de tópicos emergentes como transiciones de estado.

Resumen de eventos: Mei y Zhai [Mei y Zhai, 2005] estudiaron una tarea particular de TTM, el descubrimiento y resumen de la evolución de patrones en los textos. En [Handey y Potey, 2015] el objetivo es la generación automática de resúmenes para capítulos de series, donde el usuario pueda especificar el período de análisis.

1.3. Trabajo anterior

En un trabajo anterior de la Especialización en Explotación de Datos y Descubrimiento del Conocimiento, se realizó un análisis de 2.167 letras del rock nacional.

El objetivo fue utilizar técnicas de explotación de datos y aprendizaje automático para detectar la existencia patrones en las letras del rock argentino a lo largo de la historia.

Se descargaron letras de canciones entre 1967 y 2012 de la página <http://www.rock.com.ar/> con el módulo Beautiful Soup de Python. Se utilizó el paquete koRpus [Michalke, 2015] del programa R [R Core Team, 2013] para ver cuán similares son distintos textos entre sí, implementando fórmulas para legibilidad de los textos y de diversidad lexicográfica. En el cuadro 1.1 se observa la composición temporal de la muestra.

Cuadro 1.1: Distribución temporal de temas

| Etapas | Casos |
|---------------|--------------|
| 1967-1974 | 227 |
| 1975-1982 | 217 |
| 1983-1990 | 333 |
| 1991-1997 | 563 |
| 1998-2003 | 498 |
| desde 2004 | 329 |

De la base compuesta por 2.167 casos y 126 variables creadas con la información de frecuencias y los índices lexicográficos, se extrajeron los datos faltantes, resultando en una base de 1.814 casos. Se aplicó Latent Dirichlet Allocation, calculando la participación de cada texto en 10 tópicos. Se aplicaron técnicas de análisis de componentes principales, análisis discriminante, agrupamiento jerárquico y no jerárquico (k-means), partición alrededor de medioides (PAM), árboles CHAID y Random Forest. El grado de instancias correctamente clasificadas por los distintos métodos no fue satisfactorio.

Alguna de las razones que se supone que motivaron este bajo rendimiento fue el tamaño de la muestra y las variables elegidas para realizar los experimentos. Los índices extraídos en forma automática, así como los tópicos, no extrajeron suficiente información relevante.

1.4. Objetivo de la tesis

“Para Claudio Díaz, autor de Libro de viajes y extravíos: un recorrido por el rock argentino (1965-1985), el fenómeno forma parte de un aplanamiento general. Me parece que hay un empobrecimiento bastante fuerte en el plano musical como en lo poético. Si pasás de los últimos trabajos de los Redondos de Ricota para adelante, notás ese aplanamiento: las letras son cada vez más tontas, la música es cada vez más tonta” [Schilling, 2006].

Tomando como punto de partida el artículo de Schilling [Schilling, 2006] sobre la mutación de la metáfora a lo textual en las letras del rock nacional, el objetivo de este trabajo es definir un marco metodológico que presente en forma sistemática la integración de las técnicas estadísticas de análisis lexicográfico y de exploración multivariada, así como técnicas específicas de minería de texto temporal. Se busca detectar y predecir patrones temporales de tópicos textuales en un corpus.

El objetivo particular es desarrollar un modelo predictivo de patrones temporales para aplicar en un corpus de letras de música folklórica y del rock argentinas, escritas entre 1960 y 2014, disponibles en la web.

Se busca la detección de tópicos generados de forma automática que representen un concepto que se relacione con el período histórico (dictadura, democracia, crisis).

1.5. Organización de la tesis

Luego de haber introducido el concepto de Culturomics y su relación con el objetivo de esta tesis, en el **Capítulo 2** se realiza una revisión de algunas metodologías utilizadas en la literatura para la minería temporal de textos. En el **Capítulo 3** se presenta el armado del corpus y los experimentos realizados. En el **Capítulo 4** se exponen los resultados. En el **Capítulo 5** se muestra la conclusión y los trabajos a futuro.

Técnicas utilizadas para el análisis y agrupamiento de textos

En este capítulo se detallan algunas de las principales técnicas utilizadas para el modelado de tópicos: Latent Dirichlet Allocation, Modelado dinámico de tópicos (*Dynamic Topic Modeling*), Factorización matricial no negativa (*Non-negative Matrix Factorization*), Word2vec y Tópicos a través del tiempo (*Topics Over Time*).

2.1. Latent Dirichlet Allocation (LDA)

El modelado de tópicos es una técnica que suele aplicarse para el agrupamiento y clasificación de datos en una colección de textos [Blei, 2012].

Previamente al desarrollo conceptual del algoritmo de Latent Dirichlet Allocation (LDA), se definen los conceptos de *palabra*, *documento*, *corpus* y *tópico*.

Una *palabra* es una unidad básica de información discreta, que en este contexto se define como un elemento de un vocabulario indexado V .

Un *documento* es una secuencia de N palabras denotadas por $w = (w_1, w_2, \dots, w_N)$, donde w_n es la n -ésima palabra en la secuencia.

Un *corpus* es una colección de M documentos denotada por: $D = \{w_1, w_2, \dots, w_M\}$.

Un *tópico* es una distribución de probabilidad sobre un vocabulario fijo.

El proceso de generación de cada documento en la colección se desarrolla como se describe a continuación:

1. Elegir una distribución aleatoria sobre los tópicos.
2. Para cada palabra en el documento:

- a) Elegir aleatoriamente un t3pico.
- b) Dado ese t3pico, elegir una palabra probable (generada en el paso 1).

Este modelo refleja la intuici3n que los documentos contienen m3ltiples t3picos. Cada documento muestra los t3picos en distinta proporci3n (paso 1), cada palabra en cada documento se extrae de uno de los t3picos (paso 2b), donde el t3pico seleccionado es elegido de la distribuci3n por documento sobre los t3picos (paso 2a).

Todo proceso generativo basado en probabilidades se basa en la existencia de variables no observables en la colecci3n. Para obtener informaci3n sobre estas es necesario inferir la distribuci3n conjunta entre eventos conocidos y eventos latentes. Se puede obtener esta informaci3n a trav3s del uso de distribuciones condicionales de eventos ocultos, dado que ya se conocen las distribuciones de eventos observables. En el modelo LDA, los eventos observables son la aparici3n de palabras en los documentos; y las variables ocultas son todas aquellas que caracterizan la estructura de t3picos de una colecci3n de documentos.

En modelado probabilístico generativo, se tratan los datos como provenientes de un proceso generativo que incluye variables ocultas. Se ejecuta el an3lisis de los datos utilizando una distribuci3n conjunta para computar la distribuci3n condicional de las variables ocultas, dadas las variables observadas.

El proceso generativo se define como:

1. Para cada t3pico k , definir una distribuci3n sobre las palabras $\phi_k \sim Dir(\alpha)$.
2. Para cada documento d ,
 - a) Definir un vector con proporciones de t3pico $\theta_d \sim Dir(\beta)$.
 - b) Para cada palabra i
 - 1) Definir una palabra $w_{d,i} \sim Mult(\theta_d), z_{d,n} \in \{1, \dots, K\}$.
 - 2) Definir una palabra $w_{d,i} \sim Mult(\phi_{z_{d,i}}), w_{d,i} \in \{1, \dots, V\}$.

Dada la siguiente notaci3n:

- Dir es una distribuci3n Dirichlet
- $Mult$ es una distribuci3n Multinomial
- $\beta_{1:K}$, son los t3picos
- β_k es una distribuci3n de palabras para el t3pico k
- θ_d , es la proporci3n de t3picos para el documento d -3simo
- $\theta_{d,k}$ es la proporci3n del t3pico k en el documento d

- z_d es la asignación de tópicos para el documento d -ésimo
- $z_{d,n}$ es la asignación de tópicos para la n -ésima palabra en el documento d
- w_d son las palabras observadas para el documento d
- $w_{d,n}$ es la n -ésima palabra en el documento d

A partir de esta notación, se puede definir el proceso generativo de documentos a través de la distribución conjunta de variables observables y ocultas a continuación:

$$\begin{aligned}
p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) \\
= \prod_{i=1}^k p(\beta_i) \prod_{d=1}^D p(\theta_d) \left(\prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)
\end{aligned} \tag{2.1}$$

Esta distribución especifica una serie de dependencias. La asignación del tópico $z_{d,n}$ depende de la proporción del tópico θ_d por documento.

Las palabras observadas $w_{d,n}$ dependen de la asignación de tópico $z_{d,n}$ y de todos los tópicos $\beta_{1:K}$. Estas dependencias definen el modelo LDA.

Una vez definido el modelo que representa las relaciones entre los tópicos, los documentos y las palabras existentes en un corpus, para que este sea de utilidad es necesario calcular las distribuciones condicionales de la estructura de los tópicos, dada la colección de documentos. Esta distribución es lo que se llama como posterior. La definición de posterior se desprende de la ecuación 2.1 y se detalla a continuación:

$$\begin{aligned}
p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) \\
= \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}
\end{aligned} \tag{2.2}$$

El numerador es la distribución conjunta de todas las variables aleatorias, que pueden ser fácilmente calculadas para cualquier conjunto de variables ocultas. El denominador es la probabilidad marginal de las observaciones, que es la probabilidad de ver el corpus observado bajo cualquier modelo de tópicos. En teoría, puede computarse sumando la distribución conjunta de cualquier posible combinación de la estructura de tópicos oculta. Como para muchos problemas bayesianos, no se puede computar la probabilidad posterior por el denominador, conocido como “evidencia”. Los algoritmos de modelado de tópicos producen una aproximación de la ecuación 2.2 formando una distribución alternativa a lo largo de la estructura latente de tópicos que se adapta para acercarse a la real posterior.

2.2. Modelado dinámico de tópicos (*Dynamic Topic Modeling*)

Mientras que el modelado de series temporales se focaliza en datos continuos, el modelado de tópicos fue diseñado para datos categóricos [Blei y Lafferty, 2006].

El proceso LDA asume que los documentos provienen de forma intercambiable del mismo conjunto de tópicos. Para algunas colecciones de documentos, el orden de los mismos refleja un conjunto de tópicos que evolucionan. En el modelado dinámico de tópicos (*Dynamic Topic Modeling-DTM*), se supone que los datos están divididos por un intervalo de tiempo, por ejemplo, un año.

Se modelan los documentos para cada intervalo con un k tópico, donde los tópicos se asocian con un intervalo t evolucionado de los tópicos asociados con un intervalo $t - 1$.

Para un modelo K compuesto con V términos, $\beta_{t,k}$ es el V -vector de parámetros naturales para el tópico k en el intervalo de tiempo t .

La representación más común de una distribución multinomial es su parametrización promedio. Si se designa el parámetro medio de V por π , el i -ésimo componente del parámetro natural está dado por $\beta_i = \log \left(\frac{\pi_i}{\pi_V} \right)$. Dirichlet no es susceptible de aplicarse para el modelado secuencial, por lo tanto se cambian los parámetros naturales de cada tópico $\beta_{t,k}$ en un modelo de espacio de estados que evoluciona con ruido gaussiano.

La versión más simple de ese modelo es:

$$\beta_{t,k} \mid \beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 I) \quad (2.3)$$

donde $\beta_{t,k}$ es la distribución de palabras del tópico k en el tiempo t , N es una distribución normal logística, σ^2 es la varianza, e I es la matriz identidad.

El enfoque es para modelar secuencias de variables compuestas aleatorias, encajando distribuciones gaussianas en un modelo dinámico y mapeando de los valores emitidos a una simplex.

En LDA, las proporciones específicas θ provienen de una distribución Dirichlet. En DTM, se usa una distribución normal logística con media α para expresar incertidumbre sobre las proporciones. La estructura secuencial entre modelos se captura con un modelo dinámico simple:

$$\alpha_t \mid \alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I) \quad (2.4)$$

Dónde α_t es la distribución de tópicos por documento en el tiempo t .

El proceso generativo para el intervalo t de un corpus secuencial es:

1. Generar tópicos $\beta_t \mid \beta_{t-1} \sim N(\beta_{t-1,k}, \sigma^2 I)$.
2. Generar $\alpha_t \mid \alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I)$

3. Para cada documento:

- a) Generar $\eta \sim N(\alpha_t, \alpha^2 I)$.
- b) Para cada palabra:
 - 1) Generar $Z \sim Mult(\pi(\eta))$.
 - 2) Generar $W_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$.

Donde:

- $W_{t,d,n}$ es una palabra específica
- π mapea a los parámetros naturales multinomiales a los parámetros promedio
- $\pi(\beta_{k,t})_w = \frac{\exp(\beta_{k,t,w})}{\sum_w \exp(\beta_{k,t,w})}$

Trabajar con series de tiempo sobre los parámetros naturales permite el uso de modelos gaussianos para las dinámicas temporales. Sin embargo, debido a la no conjugabilidad de los modelos gaussianos y multinomiales, las inferencias posteriores no son tratables.

Para manejar los datos se utilizan métodos variacionales como alternativas determinísticas a una simulación estocástica.

La idea bajo los métodos variacionales es optimizar los parámetros libres de la distribución sobre las variables latentes, de manera que la distribución sea cercana en la medida de divergencia Kullback-Liebler (KL) sobre la verdadera posterior. La medida Kullback-Liebler calcula la distancia entre dos distribuciones de probabilidad definidas sobre la misma variable aleatoria.

Esta distribución puede usarse como un sustituto de la verdadera distribución posterior. En el modelo dinámico, las variables latentes son los tópicos $\beta_{t,k}$ (distribución de palabras del tópico k en el tiempo t), la mezcla de proporciones $\theta_{t,d}$ (distribución de tópicos para el documento d en el tiempo t), y los indicadores de tópicos $z_{t,d,n}$ (tópico para n -ésima palabra en el documento d en el tiempo t).

La distribución variacional refleja la estructura de grupo de las variables latentes. Hay parámetros variacionales para cada secuencia de tópicos de parámetros multinomiales, y parámetros variacionales para cada una de las variables latentes a nivel de documento.

En la aproximación promedio, cada variable latente se considera independientemente del resto. En la distribución variacional de $\{\beta_{k,1}, \dots, \beta_{k,T}\}$, se retiene la estructura secuencial del tópico postulando un modelo dinámico con observaciones variacionales gaussianas $\{\hat{\beta}_{k,1}, \dots, \hat{\beta}_{k,T}\}$.

Estos parámetros se ajustan para minimizar la divergencia KL entre la resultante distribución posterior, que es gaussiana, y la verdadera distribución posterior, que no

es gaussiana. La distribución variacional de las variables latentes a nivel de documento siguen la misma forma que en Blei [Blei et al., 2003]. Cada vector de proporción $\theta_{t,d}$ está dotado con un parámetro libre Dirichlet $\gamma_{t,d}$, cada indicador de tópico $z_{t,d,n}$ está dotado con un parámetro libre multinomial $\varphi_{t,d,n}$. Se propone el uso de dos aproximaciones variacionales, una basada en un filtro de Kalman, y una segunda basada en una regresión wavelet.

2.3. Factorización matricial no negativa (*Non-negative Matrix Factorization*)

Las técnicas de factorización matricial en data mining pertenecen a la categoría de métodos de espacios vectoriales. Estas técnicas permiten escribir una matriz como producto de dos matrices con una estructura especial. En este trabajo no se explicarán los métodos de factorización matricial. Para ver en detalle los distintos tipos de descomposición de matrices, ver [Eldén, 2007], [Koren et al., 2009].

La factorización matricial no negativa (*Non-negative matrix factorization-NMF*) es una familia de algoritmos no supervisados que simultáneamente realizan reducción de dimensiones y agrupamiento. NMF se volvió popular como herramienta de exploración de datos en bioinformática, y fue utilizado para agrupamiento de textos [Gaujoux y Seoighe, 2010].

La motivación detrás de NMF es que además de la reducción de dimensiones, el resultado es una matriz no negativa que puede ser mejor modelada e interpretada. En minería de textos bajo el modelo de espacio vectorial, las colecciones de documentos se guardan como matrices término-documento de elementos no negativos, donde cada columna representa un documento.

NMF es un algoritmo de factorización matricial que encuentra la factorización positiva de una matriz positiva [Xu y Gong, 2003].

Se utiliza el vector término-frecuencias para representar cada documento. Se define $W = \{f_1, f_2, \dots, f_n\}$ como el vocabulario completo de los documentos de un corpus después de la remoción de las stopwords y de las operaciones de stemming. X_i es el vector término frecuencia del documento d_i , definido como:

$$X_i = [x_{1i}, x_{2i}, \dots, x_{ni}]$$

$$x_{ji} = t_{ji} \times \log \left(\frac{n}{idf_j} \right)$$

Donde t_{ji} , idf_j, n , representan la frecuencia-término de la palabra $f_j \in W$, el número de documentos conteniendo la palabra f_j , y el número total de documentos en el corpus, respectivamente. Adicionalmente, X_i se normaliza a la unidad de longitud euclídea.

Utilizando X_i como la i -ésima columna, se construye la matriz término-documento $m \times n$ denominada X . Esta matriz se utilizará para calcular la factorización matricial no negativa. El resultado de agrupamiento de los documentos se obtendrá directamente del resultado de la factorización.

Un corpus consiste en k grupos de documentos. El objetivo es factorizar X en la matriz no negativa U ($m \times k$) y la matriz no negativa V^T ($k \times n$) que minimiza la siguiente función objetivo:

$$J = \frac{1}{2} \| X - UV^T \|^2 \quad (2.5)$$

Donde $\| \cdot \|^2$ representa la suma elevada al cuadrado de todos los elementos en la matriz. La función J puede reescribirse:

$$\begin{aligned} J &= \frac{1}{2} \text{tr} \left((X - UV^T) (X - UV^T)^T \right) \\ &= \frac{1}{2} \text{tr} (XX^T - 2XVU^T + UV^T VU^T) \\ &= \frac{1}{2} (\text{tr} (XX^T) - 2\text{tr} (XVU^T) + \text{tr} (UV^T VU^T)) \end{aligned} \quad (2.6)$$

Sea $U = [u_{ij}]$, $V = [v_{ij}]$, $U = [U_1, U_2, \dots, U_k]$, el anterior problema de minimización se puede reformular como: minimizar J con respecto a U y V bajo las siguientes restricciones: $u_{ij} \geq 0$, $v_{ij} \geq 0$, donde $0 \leq i \leq m$, $0 \leq j \leq k$, $0 \leq x \leq n$ y $0 \leq y \leq k$.

Este es un problema de optimización con restricciones que se puede resolver con el método del multiplicador de Lagrange. Sean α_{ij} y β_{ij} los multiplicadores de Lagrange para la restricción $u_{ij} \geq 0$ y $v_{ij} \geq 0$, respectivamente, y $\alpha = [\alpha_{ij}]$, $\beta = [\beta_{ij}]$, la L de Lagrange es:

$$L = J + \text{tr}(\alpha U^T) + \text{tr}(\beta V^T) \quad (2.7)$$

Las derivadas de L con respecto a U y V son:

$$\frac{\partial L}{\partial U} = -XV + UV^T V + \alpha \quad (2.8)$$

$$\frac{\partial L}{\partial V} = -X^T U + VU^T U + \beta \quad (2.9)$$

Utilizando la condición Kuhn-Tucker donde $\alpha_{ij} u_{ij} = 0$ y $\beta_{ij} v_{ij} = 0$, se obtienen las siguientes ecuaciones para u_{ij} y v_{ij} :

$$(XV)_{ij} u_{ij} - (UV^T V)_{ij} u_{ij} = 0 \quad (2.10)$$

$$(X^T U)_{ij} v_{ij} - (VU^T U)_{ij} v_{ij} = 0 \quad (2.11)$$

Estas ecuaciones concluyen en la siguiente actualización:

$$u_{ij} \leftarrow u_{ij} \frac{(XV)_{ij}}{(UV^T V)_{ij}} \quad (2.12)$$

$$v_{ij} \leftarrow v_{ij} \frac{X^T U_{ij}}{(V U^T U)_{ij}} \quad (2.13)$$

La solución de minimizar la función J no es única. Si U y V son las soluciones a J , luego UD y VD^{-1} también formarán una solución para cualquier matriz diagonal positiva D .

Para hacer la solución única, se requiere que el largo euclídeo de la columna vector en la matriz U sea 1. Este requerimiento de normalizar U se puede alcanzar con:

$$v_{ij} \leftarrow v_{ij} \sqrt{\sum_i u_{ij}^2} \quad (2.14)$$

$$u_{ij} \leftarrow \frac{u_{ij}}{\sqrt{\sum_i u_{ij}^2}} \quad (2.15)$$

En resumen, el algoritmo de agrupamiento se compone de los siguientes pasos:
Dado un corpus,

- Construir la matriz de términos-documentos X en la cual la columna i representa la frecuencia ponderada del vector de términos del documento d_i .
- Realizar NMF en X para obtener dos matrices no negativas U y V utilizando las ecuaciones 2.12 y 2.13.
- Normalizar U y V usando las ecuaciones 2.14 y 2.15.
- Usar V para determinar el grupo de cada punto de datos. Examinar cada fila i de la matriz V . Asignar el documento d_i al cluster x si $x = \operatorname{argmax}_j v_{ij}$.

Este método difiere de Latent Semantic Indexing (LSI) basado en Singular Vector Decomposition (SVD) y los métodos relacionados de agrupamiento espectral en que el espacio latente semántico derivado por NMF no necesita ser ortogonal. Se garantiza que cada documento tome sólo valores no negativos en todas las direcciones latentes semánticas.

2.4. Word2vec

Con el objetivo de la automatización del resumen de textos, se puede utilizar una variedad de metodologías, como Hidden Markov Models, técnicas de grafos, y acercamientos sobre distribución de probabilidades [Heuer, 2015]. Según lo expuesto por

Sahlgren [Sahlgren, 2005] (hipótesis distributiva) las palabras con un significado similar ocurren en similares contextos. Esto implica que el significado de las palabras se puede inferir por su distribución contextual. Bruni y otros [Bruni et al., 2014] demuestran que esta teoría tiene múltiples raíces teóricas en psicología, lingüística, lexicografía y filosofía. El objetivo de la semántica distributiva es encontrar una representación, por ejemplo, un vector, que aproxime el significado de una palabra. La hipótesis distributiva propone que los términos con propiedades de distribución similares tienen un significado similar [Sahlgren, 2005].

Uno de los desafíos de la semántica distributiva es computar vectores de palabras que sean representaciones adecuadas de las palabras. Se utilizan varias arquitecturas para computar tales vectores de palabras. Tradicionalmente, los vectores de palabras se entrenan como parte de un lenguaje de modelado de redes neuronales. De acuerdo con Mikolov, las redes neuronales artificiales pueden pensarse como una proyección no lineal de los datos [Mikolov et al., 2013]. Las redes neuronales artificiales (RNA) son un sistema para el tratamiento de la información, cuya unidad de procesamiento básica está inspirada en la neurona humana [Herrero y Bote, 1998]. Los principios de funcionamiento de las RNA son:

- Aprendizaje adaptativo.
- Autoorganización.
- Tolerancia a fallos.
- Operación en tiempo real.
- Fácil inserción en la tecnología existente.

Las RNA están formadas por tres grupos de neuronas:

- Neuronas de entrada: reciben información del exterior.
- Neuronas de salida: transmiten información al exterior.
- Neuronas ocultas: no tienen contacto con el exterior, y solamente intercambian información con otras neuronas de la red.

En cualquier tipo de RNA las neuronas se encuentran interconectadas entre sí, organizándose en capas, y formando diferentes topologías. Las topologías se pueden clasificar de acuerdo con tres criterios:

- número de niveles o capas.
- número de neuronas por nivel.

- formas de conexión.

En este trabajo no se explicarán en detalle las distintas arquitecturas, para ampliar ver [Yegnanarayana, 2009].

Word2vec fue desarrollado por Mikolov, Sutskever, Chen, Corrado y Dean en Google y publicado en 2013 [Mikolov et al., 2013]. Word2vec es una herramienta que implementa dos formas de computar representaciones de palabras: continuous bag-of-words (CBOW) y continuous skip-gram (CSG).

Word2vec toma un corpus como entrada y produce vectores de palabras como resultado. Primero construye un vocabulario desde los textos de entrenamiento y luego aprende una representación vectorial de palabras. El vector de palabras resultante puede usarse como variables en aplicaciones de lenguaje natural y aprendizaje automático. La ventaja que presenta Word2vec es su utilización de modelos de redes neuronales que entienden el significado semántico de las palabras.

Las arquitecturas que se utilizaron previamente a CBOW y CSG fueron RNA alimentadas hacia adelante (Feedforward Neural Net Language Model -NNLM), y Redes neuronales con conexiones recurrentes (Recurrent Neural Net Language Model -RNNLM).

El modelo de lenguaje probabilístico de red neuronal feedforward (NNLM) se compone de capas de entrada, capa de proyección, capa oculta y capas de salida. En la capa de entrada, N palabras anteriores se codifican utilizando una de-codificación V , donde V es el tamaño del vocabulario. La capa de entrada es entonces proyectada a una capa P proyección que tiene dimensión $N \times D$, utilizando una matriz de proyección compartida, donde D es la dimensionalidad de las palabras en el espacio. Como sólo N entradas están activas en un momento dado, la composición de la capa de proyección es una operación relativamente barata.

El modelo RNNLM no tiene una capa de proyección, sólo entrada, oculta y salida. Lo que tiene especial este modelo es la matriz recurrente que conecta la capa oculta a sí misma, usando conexiones demoradas en el tiempo. Esto permite al modelo recurrente formar una memoria de corto plazo.

Continuous Bag-of-Words Model es similar a feedforward NNLM, donde la capa oculta no lineal se remueve y la capa proyectada se comparte para todas las palabras, por lo tanto, todas las palabras se proyectan en la misma posición (sus vectores se promedian). Se llama a esta arquitectura modelo bag-of-words, porque el orden de las palabras en la historia no influencia la proyección.

Continuous Skip-gram Model es similar a CBOW, pero en vez de predecir la palabra basada en el contexto, trata de maximizar la clasificación de una palabra basada en otra palabra en la misma oración. Se utiliza cada palabra como una entrada en un clasificador log-lineal con una capa continua proyectada, y predice palabras con un cierto rango antes y después de la actual palabra.

Aumentando el rango mejora la calidad de los vectores de palabras resultantes, pero

también incrementa la complejidad computacional. Desde que las palabras más distantes están usualmente menos relacionadas con la palabra actual que aquellas cercanas a ella, se les da menos peso a las palabras distantes muestreando menos de esas palabras en los ejemplos de entrenamiento.

Para reducir el número de categorías y mejorar la eficiencia, la primera opción de los investigadores fue el agrupamiento (clustering). La idea principal fue agrupar las palabras similares en una clase y luego utilizarla para representar cada palabra. Los vectores computados por Word2vec contienen información semántica. En word2vec la mayor contribución es que se puede con exactitud calcular la similitud entre palabras utilizando esos vectores [Yuan et al., 2014].

Para ello, se pueden obtener las palabras similares calculando la similitud coseno como medida de distancia. Se pueden agrupar estas palabras similares en la misma clase computando la similitud coseno luego de definir el número de grupos. Luego de obtener los grupos, se pueden contar las palabras que aparecen en él y sumar el número como un valor de la dimensión correspondiente. El vector obtenido es la variable del documento. Este vector integra la información semántica de las palabras similares. Los vectores de Word2vec se pueden utilizar como atributos para tareas de procesamiento de lenguaje natural supervisadas, como clasificación de documentos, reconocimiento de entidades nombradas y análisis de sentimiento.

Por ejemplo, en [Yuan et al., 2014] se utilizaron dos conjuntos de datos. Luego del pre-procesamiento (eliminación de puntuación, números y stopwords), se armó un corpus. Cada cinco palabras contiguas se tomó como una ventana, y se introdujo en Word2vec. Se definió la dimensión del vector en 200. Se pudieron obtener $n \times 200$ vectores después de entrenar la red neuronal, donde n es el número de palabras. Tomando estos vectores, se utilizaron para calcular la similitud coseno cada dos palabras. Luego, las palabras más similares se agruparon en un cluster. En el diseño de los experimentos, se definió el número de clusters en 50. Luego se etiquetaron los clusters en cada palabra y se computaron los *features* (atributos) de los documentos a través de estos clusters. Se inicializaron los *features* en un vector de dimensión $d \times 50$, donde d es el número de documentos y cada columna representa un cluster. Si una palabra pertenece al i -ésimo cluster aparece en el documento, el correspondiente *feature* sumará +1 en la i -ésima dimensión. Con este método, se obtuvieron todos los vectores de *features*. Se dividieron los datos en entrenamiento y testing, y se entrenó un clasificador SVM (Support Vector Machine, ver en [Tong y Koller, 2002]). Se compararon los resultados de este método versus LDA y TF-IDF (ver ecuación en 3.1). En ambos conjuntos de datos, el resultado utilizando word2vec fue superior al resto.

2.5. Tópicos a través del tiempo (*Topics over the time*)

El algoritmo Tópicos a través del tiempo (*Topics over the time-TOT*) modela el tiempo conjuntamente con los patrones de co-ocurrencia de palabras [Wang y McCallum, 2006]. El modelo no discretiza el tiempo, y no realiza asunciones de Markov sobre los estados de transiciones en el tiempo.

TOT parametriza una distribución continua en el tiempo asociado con cada tópico, y los tópicos son responsables de la generación de los tiempos como las palabras. La estimación de parámetros es conducida para descubrir tópicos que capturen simultáneamente la co-ocurrencia y localidad de aquellos patrones en el tiempo. Cuando un patrón fuerte de co-ocurrencia de palabras aparece por un pequeño espacio en el tiempo y luego desaparece, TOT crea un tópico con una estrecha distribución temporal.

Cuando un patrón de co-ocurrencia de palabras permanece consistente a través de un espacio de tiempo, TOT creará un tópico con una distribución temporal ancha. En Wang y McCallum [Wang y McCallum, 2006], se utilizó una distribución Beta sobre un espacio de tiempo normalizado cubriendo todos los datos. Una variable aleatoria tiene distribución Beta de parámetros α, β en $[0, 1]$ (se representa como $X \sim Be(\alpha, \beta)$), con $\alpha, \beta > 0$ si su función de densidad es

$$f(x|\alpha, \beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} & \text{para } 0 < x < 1 \\ 0 & \text{resto} \end{cases}$$

La función Γ se define como:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

En TOT se ajustan los parámetros del modelo de acuerdo con un modelo generativo, en el cual la distribución multinomial por documento sobre los tópicos es muestreada desde una distribución Dirichlet. Por cada ocurrencia de una palabra, se muestrea un tópico, luego una multinomial por tópico genera la palabra, y una distribución Beta por tópico genera la marca de tiempo.

La marca temporal, que en la práctica es siempre observada y constante a lo largo del documento, se asocia con cada palabra en el documento.

Los tópicos de TOT y su significado se modelan como una constante a lo largo del tiempo. TOT captura los cambios en la ocurrencia (y co-ocurrencia condicionada al tiempo) de los propios tópicos, no cambios en la distribución de palabras en cada tópico. El descubrimiento de tópicos está influenciado no solo por la co-ocurrencia de palabras, sino también por información temporal. Más que modelar una secuencia de cambios de estados con una asunción Markov en la dinámica, TOT modela valores absolutos de marcas de tiempo normalizados. Esto permite que TOT visualice

grandes rangos de dependencias temporales, para predecir valores temporales absolutos dado un documento sin marca temporal, y predecir la distribución de tópicos dada una marca temporal. También ayuda a evitar el riesgo de un modelo Markov de dividir un tópico inapropiadamente en dos cuando hay una pequeña brecha en su aparición. TOT evita la discretización asociando con cada tópico una distribución continua a lo largo del tiempo. Se pueden aplicar muchas distribuciones. Todos los resultados en [Wang y McCallum, 2006] utilizan la distribución Beta, para la que el rango de tiempo utilizado para la estimación de parámetros se normalizó de 0 a 1.

TOT es un modelo generativo de marcas de tiempo y palabras en documentos en los documentos con marca temporal. Se puede describir al proceso generativo de la siguiente manera:

1. Generar una multinomial $T\phi_z$ de una Dirichlet previa β , una para cada tópico z ;
2. Para cada documento d , generar una multinomial θ_d de una Dirichlet previa α ; luego para cada palabra w_{di} en el documento d :
 - a) Generar un tópico z_{di} de una multinomial θ_d ;
 - b) Generar una palabra w_{di} de una multinomial $\phi_{z_{di}}$;
 - c) Generar una marca de tiempo t_{di} de Beta $\psi_{z_{di}}$.

Donde:

- T es el número de tópicos
- D es el número de documentos
- V es el número de palabras únicas
- N_d es el número de tokens en el documento d
- θ_d es la distribución multinomial de tópicos específicos del documento d
- ϕ_z es la distribución multinomial de palabras específicas del tópico z
- ψ_z es la distribución beta de tiempo específica del tópico z
- z_{di} es el tópico asociado con el i -ésimo token en el documento d
- w_{di} es el i -ésimo token en el documento d
- t_{di} es la marca de tiempo asociada con el i -ésimo token en el documento d

En el proceso generativo anterior, una marca de tiempo se genera para cada palabra. Todas las marcas de tiempo de las palabras en un documento son observadas, como lo es la marca de tiempo del documento. La distribución posterior de los tópicos depende de la información de dos modalidades, el tiempo y el texto. La inferencia no se puede realizar con exactitud en este modelo. Se utiliza muestreo de Gibbs para aproximar la inferencia. Por simplicidad y velocidad se estiman las distribuciones de Beta ψ_z por el método de momentos, una vez por iteración de muestreo Gibbs. Se pueden estimar los valores de los hiperparámetros α y β usando un algoritmo Gibbs EM.

Aunque un documento se modela como una mezcla de tópicos, hay típicamente sólo una marca de tiempo asociada con un documento. El proceso generativo describe los datos donde hay una marca de tiempo asociada con cada palabra. Cuando se ajusta el modelo de datos típicos, cada marca de tiempo de los documentos de entrenamiento se copia a todas las palabras en el documento. Sin embargo, después de ajustar, si se corre como un modelo generativo, el proceso generará distintas marcas de tiempo para las palabras en el mismo documento.

Usando este modelo se puede predecir una marca de tiempo dadas las palabras en un documento. Para facilitar la comparación con LDA, se pueden discretizar las marcas de tiempo. Dado un documento, se predice su marca de tiempo eligiendo la marca de tiempo discreta que maximiza la posterior, la cual es calculada multiplicando la probabilidad de la marca de tiempo de todos los tokens de su correspondiente distribución Beta.

CAPÍTULO 3

Materiales y métodos

En este capítulo se describe el proceso de armado de un corpus de letras del rock y del folklore argentinos entre 1960 y 2014. Luego se detallan los experimentos realizados con el objetivo de identificar en el corpus patrones temporales.

3.1. Materiales

Para el armado del corpus, se utilizaron las siguientes fuentes:

1. <http://www.rock.com.ar/> La página posee varios nodos (discos, letras, bios, enciclopedia, blog, imágenes, notas, publicidad).
2. <http://www.folkloredelnorte.com.ar/> La página no cuenta con el año de los temas, y en algunos casos, tampoco con el autor. Para completar el año se utilizó como fuente adicional la página de SADAIC (<http://www.sadaic.org.ar/>).
3. <http://www.cmtv.com.ar/> Contiene letras en castellano de España y América Latina, dentro de las cuales hay letras de rock argentino y de folklore.

Se realizaron pruebas para la descarga de las letras con las siguientes herramientas: R [R Core Team, 2013], Wget (<https://www.gnu.org/software/wget/>) y HTTrack Website Copier (<https://www.httrack.com/>).

En la figura 3.1 a continuación se muestra el flujo de trabajo para el armado del corpus.

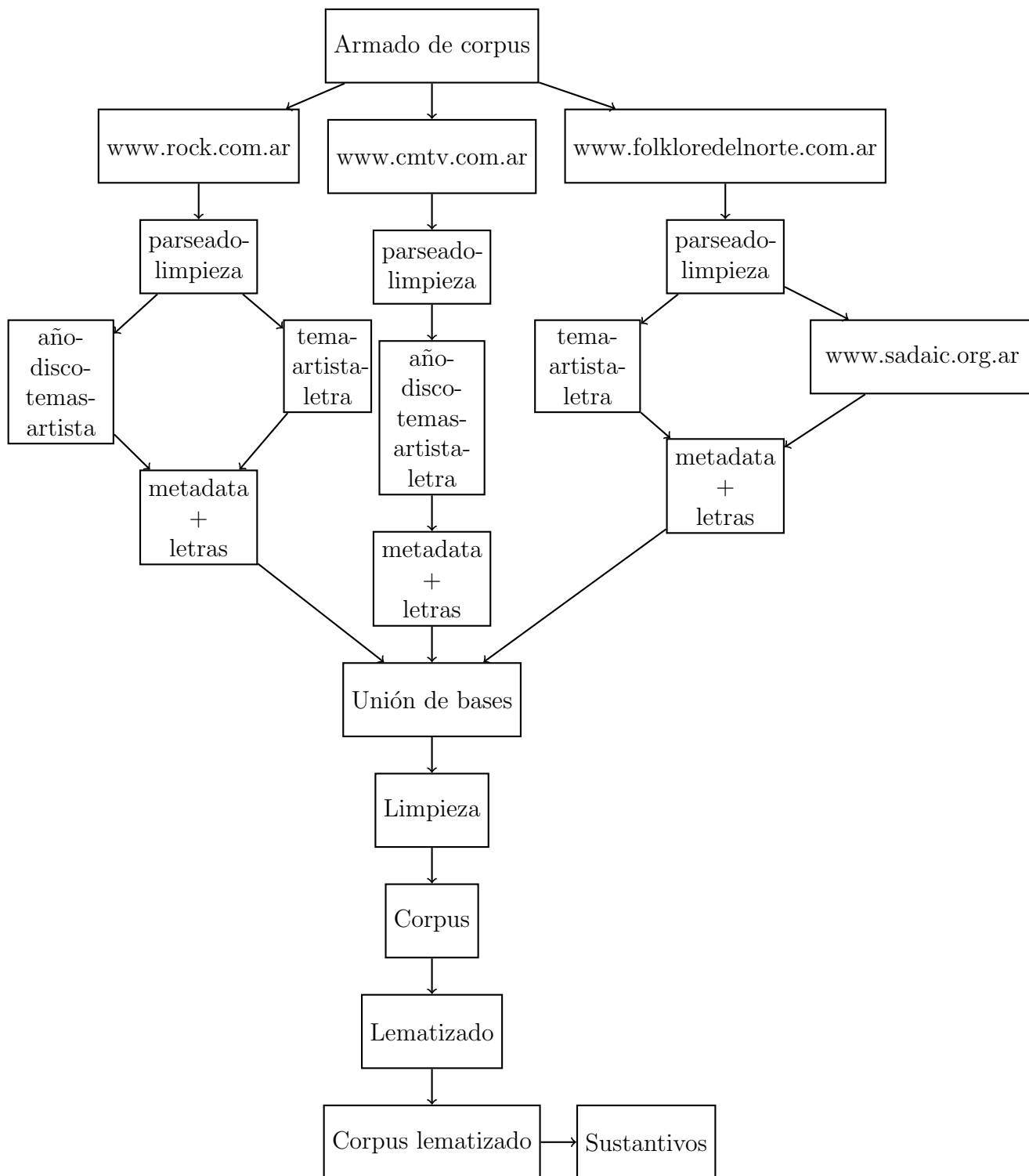


Figura 3.1: Flujo de trabajo para el armado del corpus

El software que se utilizó para la lectura, análisis sintáctico (*parsing*), extracción de nodos, limpieza y armado de corpus fue R. Los paquetes que se utilizaron para estos procesos fueron: XML [Lang, 2013b], RCurl [Lang, 2013a], stringR [Wickham, 2015], plyr [Wickham, 2011], tm [Meyer et al., 2008] y data.table [Dowle et al., 2014]. Para el lematizado se utilizó el software freeling (<http://nlp.lsi.upc.edu/freeling/>).

El árbol de la página <http://www.rock.com.ar/> se descargó totalmente en carpetas con la herramienta Httrack. Las carpetas que se utilizaron fueron discos y temas. En la carpeta discos, se recorrieron las subcarpetas y se analizó el texto (*parsing*), para extraer los nodos que contienen el nombre del disco, el intérprete, la fecha de edición, y la lista de temas (solo los nombres). Esta información se guardó en un vector compuesto por temas, disco, año, intérprete. Se ordenó la combinación intérprete-letra por año y se seleccionó el mínimo de los años para aquellos casos donde la combinación no era única. El resultado fue la tabla de discos con los temas una única vez.

La carpeta letras contiene 19.116 letras en formato html. Al igual que con la carpeta discos, se recorrieron las subcarpetas y se analizó el texto, separando el título de la canción, el intérprete y la letra propiamente dicha.

El resultado fue un vector con título del tema, intérprete, y letra de la canción (ver figura 3.2).

```
> head(dt.lettras)
```

| | temas | interprete |
|----|------------------------|------------------------|
| 1: | 1 segundo es demasiado | 1 segundo es demasiado |
| 2: | Cadenas | 1 segundo es demasiado |
| 3: | Caminar | 1 segundo es demasiado |
| 4: | Casa de ahorcados | 1 segundo es demasiado |
| 5: | El rescate | 1 segundo es demasiado |
| 6: | En busca del sol | 1 segundo es demasiado |


```
letra
```

```
1:
\n\nha pasado un cuarto, de mi vida\nhan pasado varios años ya\nhe tenido tristezas, al
egrias\ncometí el pecado de amar.\nmirá por la ventana... descubrirás\nLas palomas blan
cas, la libertad.\nno puedo cambiar, la historia\nNo puedo encontrar remedio ya\ny ahor
a estoy pagando el precio, en soledad\nEncerrado entre rejas, a oscuras\nno sé que sign
ifica: libertad\nsí no podré ver palomas blancas... nunca más.\nveo pasar sonriendo al
tiempo \ndesde un rincón\nPrivado de ver el cielo, sin una razón\nLa tormenta se acerca
\nA mis brazos hoy\nya llega la sentencia, a mi corazón\nMira por la ventana, descubrir
ás\nLas palomas blancas, la libertad.\nno puedo tapar el muro\nno puedo escaparle a la
verdad\ny ahora estoy pagando el precio, en soledad\nse aproxima, sonriente, la muerte
a mi vida\nno sé que significa: libertad\nsí no podré ver palomas blancas... nunca más.
2:
\n\nmira cuando vos ya no calles,\nmira que el dolor no te va a esquivar; \nse prende f
uego la mentira y ésta farsa ya no tiene un lugar. \nmira que sensación se siente volar
, \nes crudo éste dolor pero te hace andar; \nno es tan difícil que los sueños se estro
peen \ntoda ésta historia sin despegar. \nse rompen las cadenas que por tantos siglos a
taron las fieras, \nexplotan las ventanas de esta Patria oscura que tanto me ciega.\nmí
ra que ésta vez ya no hay vuelta atrás, \nrompiste con el Dios de la inmensidad; \nesto
esta claro toco el techo la balanza \nno es tiempo de callar... \nse rompen las cadenas
que por tantos siglos ataron las fieras, \nexplotan las ventanas de esta Patria oscura
que tanto me ciega. \nmira cuando vos... ya no calles.
```

Figura 3.2: Matriz de letras

La tabla de discos se vinculó con la tabla de letras. De esta manera se obtuvo la metadata necesaria para completar el corpus (disco, intérprete, nombre del tema, año) – ver figura 3.3 –.

```
> head(bd)
```

| | temas | interprete | año | disco |
|---|------------------------|------------------------|------|---------------------------------|
| 1 | | Agente Naranja | 2003 | "Frágil_cero" |
| 2 | 0351 4236802 | Los Cuervos | 2003 | "Nunca" |
| 3 | 1 Corazón | Séptima ola | 2004 | "Un vivo bárbaro" |
| 4 | 1 segundo es demasiado | 1 segundo es demasiado | 2001 | "1 segundo es demasiado (demo)" |
| 5 | 10 años después | Miranda | 2012 | "Luna magistral" |
| 6 | 10 mandamientos | La 25 | 2005 | "Ruta 25" |

Figura 3.3: Metadata de letras del rock nacional

En el caso de la página de folklore <http://www.folkloredelnorte.com.ar/>, la herramienta Htrack presentó problemas con la codificación de las letras, por lo que se utilizaron las librerías XML y RCurl de R para la descarga del cancionero.

A diferencia de la página de rock, la sintaxis no es la misma para cada canción, por lo tanto, el proceso de limpieza fue más trabajoso.

Para obtener el año de cada canción, se armó una consulta para completar el formulario de la página de SADAIC de forma automática. En la figura 3.4 se puede observar el resultado de la consulta para la canción Abrazado a tu cintura.

```
[1] " Título | "
```

```
[2] " | subtítulo"
```

```
[3] "#1196749 | ISWC T-037412366-6"
```

```
[4] "Registrada el 01/02/2007"
```

```
[5] "ABRAZADO A TU CINTURA"
```

```
[6] "BELLO PATRICIA ELISABETBALMACEDA VICTOR MIGUEL"
```

```
[7] " | "
```

Figura 3.4: Resultado de consulta de fecha de registro en SADAIC de letras de folklore

En algunos casos, aparece más de un título repetido, y por lo tanto distintos años y autores. El autor no está separado de la letra ni del título, y aparece de distintas formas en los documentos descargados, por lo que se extrajeron las primeras 10 palabras de cada texto, para compararlo con el resultado de la página, y seleccionar el autor que correspondía. En los casos que no se encontró el autor, no se incluyó la canción en el corpus.

Para la página <http://www.cmtv.com.ar/>, se realizó la descarga de las letras con wget. En esta página se encuentran contenidos de música en castellano (de España y América Latina). Luego del análisis sintáctico y la extracción de los autores, discos, años, temas y letras, se eliminaron los intérpretes no argentinos e intérpretes argentinos de otros géneros musicales.

Con las tres bases obtenidas (rock, folklore y cmtv), con la misma estructura (título, letra, autor, año y disco), se armó una base única y se homogeneizó el texto. Se ordenó la base por título/autor/disco y se eliminaron los registros duplicados. De la base resultante, se buscaron palabras clave (grandes éxitos, unplugged, en concierto,

oro, obras cumbres, colección, remix, acústico y en vivo) y se eliminaron aquellos discos cuyo título las incluía.

De las letras restantes, se ordenaron los temas por autor, y se revisaron aquellos autores con más de 100 canciones. Para cada uno de ellos, se ordenaron las letras alfabéticamente, y se revisó que no hubiera letras repetidas con errores de tipeo, eliminando aquellos temas que eran en colaboración con otro intérprete. Como resultado de esta base, se obtuvieron 32.703 letras, de las cuales 32.291 corresponden al período bajo análisis.

Con el objetivo de eliminar letras duplicadas, se ordenaron los temas por fecha y por autor, y se calcularon las distancias entre las letras y los títulos. En aquellos casos en que la combinación autor/título/letra tuviera una distancia igual a cero caracteres, se eligió la letra de menor fecha, y se eliminó el resto.

Con esa base, se lematizaron las letras en una librería de código abierto para el procesamiento multilingüe automático (Freeling). Se obtuvo como resultado una salida con la palabra original, el lema, la etiqueta EAGLE (codificación para la representación morfológica de las palabras) y la probabilidad. Se agregaron a la base anterior los lemas, y se filtraron los sustantivos.

Con el objetivo de descartar algún caso en que la letra sea igual, pero el título distinto, se utilizaron los sustantivos. En base a los sustantivos, se creó una matriz de documentos/términos, pesada con la medida *Term frequency-inverse document frequency (Tf-idf)*.

Tf-idf es una medida que expresa la relevancia de una palabra para un documento en un corpus. Se define $Tf_{i,j}$, como la frecuencia del término t_i en el documento d_j . Debido a la presencia de documentos de distinto largo, se normaliza por la frecuencia máxima del término en el documento.

$$Tf_{i,j} = \frac{f_{i,j}}{\max f_{i,j}} \quad (3.1)$$

Sea Q el número de documentos en la colección y n_i el número de documentos donde aparece el término t_i , se define el *idf* del término t_i como

$$idf_{t_i} = \log_{10} \frac{Q}{n_i} \quad (3.2)$$

El modelo $tf - idf$ combina ambos modelos, quedando el peso w de un documento como

$$w(t_i, d_j) = Tf_i \times \log_{10} \frac{Q}{n_i} \quad (3.3)$$

Luego se normalizaron los vectores y se aplicó k-medias (*k-means*), con $k = 15$, para agrupar letras similares. Para los agrupamientos generados, se convirtieron las letras en una matriz de documentos-términos, y se calculó la distancia euclídea entre ellos. Al

inspeccionar los primeros elementos, se encontraron las letras duplicadas. A modo de ejemplo, se muestra el agrupamiento jerárquico de uno de los 15 grupos de k-medias (ver figura 3.5).

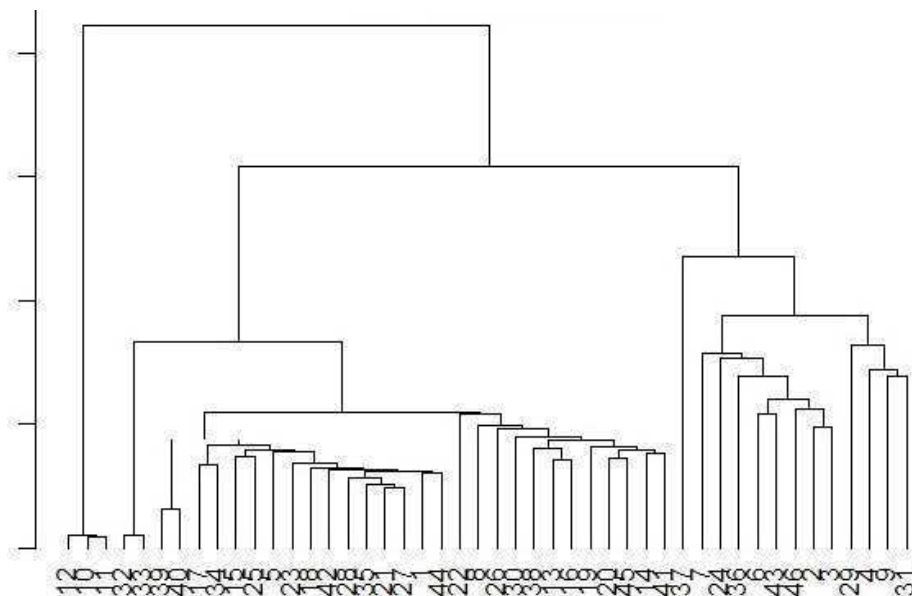


Figura 3.5: Agrupamiento jerárquico de documentos de un cluster de k-means

Se inspeccionaron las letras 10, 11, 12. Tienen la misma letra, pero distintas versiones, por eso en la limpieza anterior no fueron eliminadas (figura 3.6):

| | titulo | artista |
|------|-----------------------------|-------------|
| 7440 | loco de atar | la zimbabwe |
| 7441 | loco de atar radio mix | la zimbabwe |
| 7442 | loco de atar tribal version | la zimbabwe |

| letra |
|--|
| 7440 \n\t\t\tLOCO DE ATAR\n\t\t\tPasan las noches y yo sig |
| o aquí escondido \nsoy parte de algún plan en otro lugar \ |
| nNo tengo nada perdido, por perdido \nJugando fuerte se qu |
| e puedo ganar \nY se que hay muchos que se sorprenderán \n |
| Y se que hay otros que se arrepentirán \nPorque dejaron un |

Figura 3.6: Letras repetidas, eliminadas luego del agrupamiento k-means

El resultado de esta limpieza es una base de dimensión 30.969 filas x 8 columnas (título, artista, letra, fecha, disco, lemas, sustantivos, id base de origen –rock o folklore-).

3.2. Métodos

Con este corpus de 30.969 letras se realizaron dos tipos de experimentos para probar el cambio de tópicos a lo largo del tiempo. Estos experimentos se pueden clasificar en exploratorios y de clasificación. En este capítulo se explicarán los experimentos, los resultados se analizan en el [Capítulo 4](#).

Los experimentos exploratorios consistieron en la división del corpus en ventanas de tiempo para generar matrices de similitud, con el objetivo de detectar tópicos emergentes y tópicos en decadencia. Luego sobre estas ventanas se aplicó también factorización matricial no negativa dinámica, con el mismo objetivo del experimento anterior. Sobre ese mismo corpus, sin dividirlo en ventanas, se eligieron distintos números de tópicos. El objetivo fue comparar la evolución de esos tópicos en toda la línea de tiempo, así como el uso de las palabras en cada etapa temporal. Finalmente se experimentó con el modelado dinámico de tópicos.

Los experimentos de clasificación consistieron en la prueba de distintos algoritmos para predecir la época a la que pertenecen las letras del corpus.

En la figura [3.7](#) se resumen los experimentos realizados agrupados por tipo:

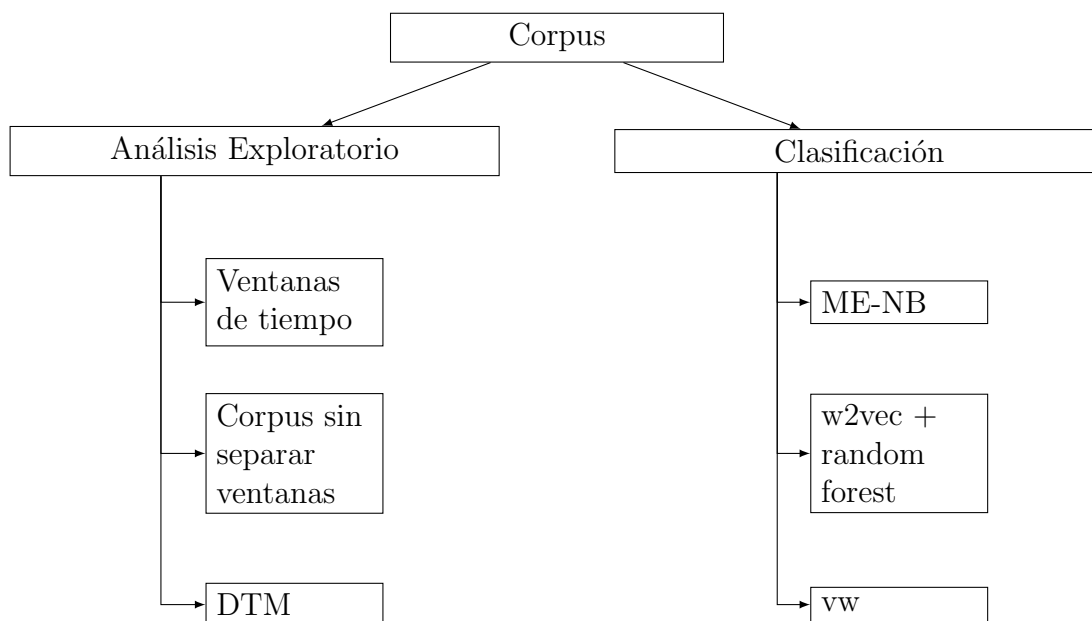


Figura 3.7: Esquema de experimentos. DTM: Modelado dinámico de tópicos, ME: Máxima entropía, NB: Naive Bayes, w2v: Word2vec, vw: VowpalWabbit

En la figura [3.8](#) se reproduce el histograma de los documentos. La mayoría de los documentos se agrupan a partir del año 2000, debido a la mayor cantidad de letras digitalizadas en internet.

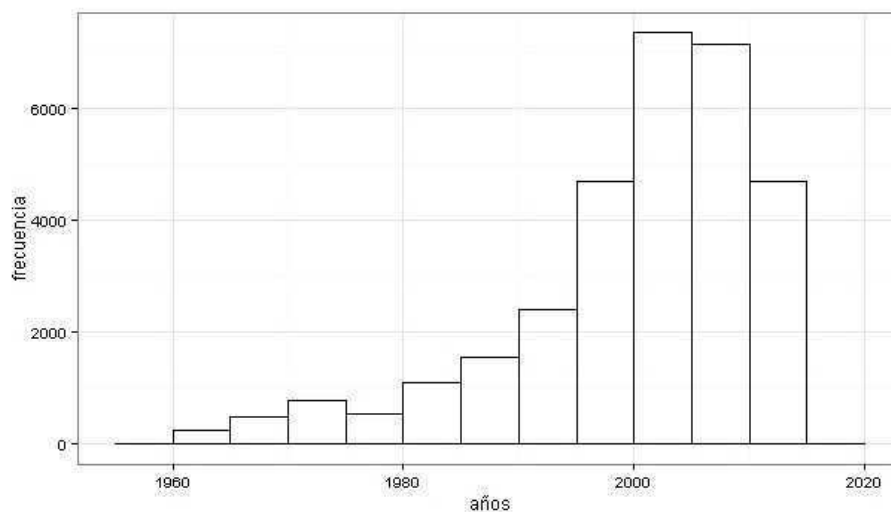


Figura 3.8: Histograma de documentos por fecha

3.2.1. Análisis exploratorio

Para analizar un corpus con marcas de tiempo, se puede separar el corpus en ventanas temporales de n períodos, o bien realizar un análisis integral del corpus, sin separar en períodos.

3.2.1.1. Ventanas de tiempo

Se busca probar que para encontrar un tópico emergente, a ventanas de tiempo más pequeñas, más específicos serán los temas, y más fácil la detección de novedades o de tópicos en decadencia [Mei y Zhai, 2005]. Con las ventanas de tiempo se realizaron dos tipos de experimentos, uno basado en una matriz de similitud, y otro aplicando el modelo de Factorización matricial no negativa dinámica (DNMF) [Greene y Cross, 2015].

Matriz de similitud

Para crear las ventanas temporales, se seleccionaron las fechas de las canciones, y se dividió la base en ventanas superpuestas de ancho = 5 años, siguiendo lo propuesto en [Mei y Zhai, 2005]. La elección del ancho = 5 años fue experimental. El objetivo fue buscar en la superposición algún momento específico donde se produjera un cambio en la tendencia de tópicos.

En la figura 3.9 se observa cómo se componen las cuatro primeras ventanas. Por ejemplo, la ventana 1 (w_1), abarca los períodos 1960-1964, la ventana 2 (w_2), abarca

los períodos 1961-1965, y así sucesivamente hasta el año 2014 inclusive. Se obtuvieron 51 ventanas.

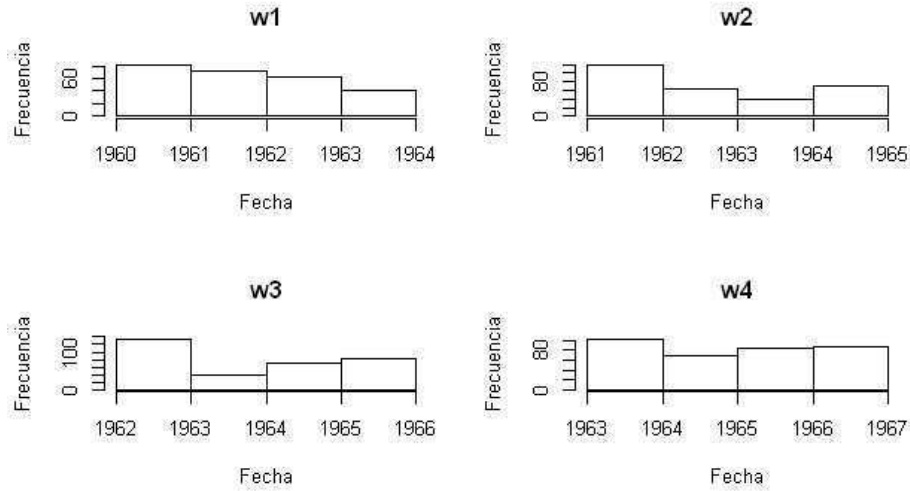


Figura 3.9: Histograma de cantidad de documentos por ventanas w_1 (1960-1964) a w_4 (1963-1967)

Se generó una lista inicial de *stopwords*, con las palabras más frecuentes, y se ejecutó un modelo LDA con 100 tópicos, antes de realizar los experimentos definitivos. En base a los resultados, se agregaron más palabras a la lista de *stopwords*. Luego se generaron los diccionarios para cada ventana, tomando o bien sólo los lemas, o sólo los sustantivos.

Para decidir el número de tópicos óptimo para este experimento, se utilizó la medida Kullback-Leibler [Arun et al., 2010]. Para calcular esta medida se define LDA como un mecanismo de factorización de matrices, donde un corpus C se divide en dos matrices: M_1 de orden $T \times W$, y M_2 de orden $D \times T$, donde D es el número de documentos, T es el número de tópicos y W es el tamaño del vocabulario del corpus. Ambas son matrices estocásticas donde la fila k -ésima en M_1 es una distribución de palabras en el tópico k -ésimo y la fila n -ésima en M_2 es la distribución de tópicos en el n -ésimo documento. Si no fueran matrices estocásticas, pero sólo representaran cuentas, si el elemento $(i, j)^{-ésimo}$ en la matriz M_1 indicara las veces que la palabra j fue asignada al tópico i y el elemento $(i, j)^{-ésimo}$ en la matriz M_2 indicara el número de veces que el tópico j es asignado a una palabra en el documento i , luego:

$$\sum_{v=1}^W M_1(t, v) = \sum_{d=1}^D M_2(d, t) \forall t = 1, \dots, T \quad (3.4)$$

El número de palabras asignadas a cada tópico se observa de dos maneras: una como suma de filas de las palabras y otra como la suma de columnas de los documentos. Sin embargo, cuando ambas matrices se normalizan por fila (como en LDA), esta igualdad no se mantiene. La idea de la medida es tomar ventaja de que ambas sumas representan la proporción de tópicos asignados al corpus y pueden compararse.

Primero se crea la medida de divergencia KL. Sean p y q distribuciones de probabilidad de una variable aleatoria discreta, como esta medida no es simétrica, sino una medida de entropía, se toma la entropía de (p, q) y se suma a la entropía de (q, p) [Grainger, 2014].

Para cada k número de tópicos, se calcula C_{M_1} , que es la distribución de valores singulares para M_1 (matriz documento-palabra). Luego se calcula C_{M_2} , que es la distribución obtenida normalizando el vector $L \times M_2$, donde L es el largo del corpus y M_2 la matriz documento-tópico.

El cálculo es el siguiente:

$$Divergencia(M_1, M_2) = KL(C_{M_1} \parallel (C_{M_2}) + KL(C_{M_2} \parallel (C_{M_1})) \quad (3.5)$$

En la figura 3.10 se observan los valores de la medida KL simétrica para la primera ventana $w1$: 1960-1964, ejecutando LDA con tópicos $k = \{1, 2, 3, \dots, 50\}$. El mínimo valor representa el valor óptimo de k (número de tópicos). En este caso el menor valor (tomando $k > 1$) se alcanza para $k = 2$.

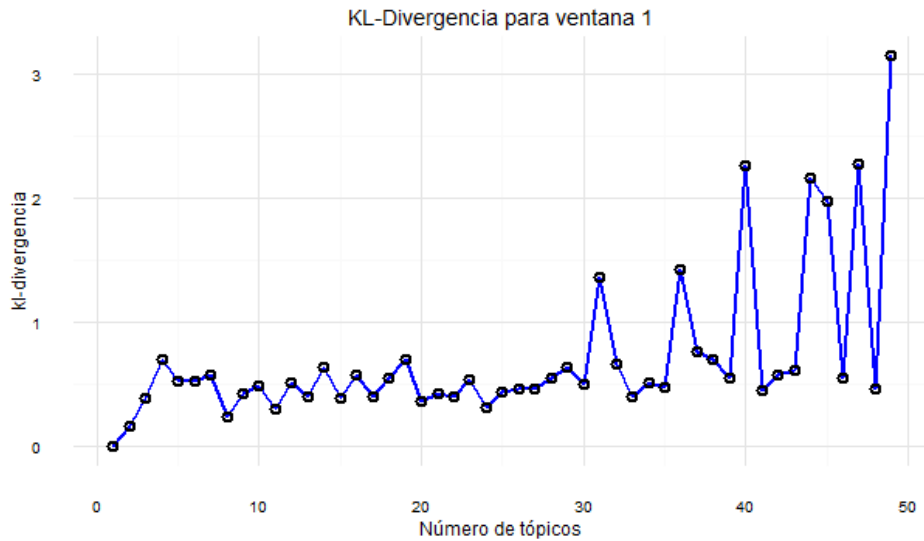


Figura 3.10: KL Simétrica para $w1$ (1960-1964)

En la figura 3.11 se observa la participación de cada tópico en el corpus. La distribución es similar en ambos casos.

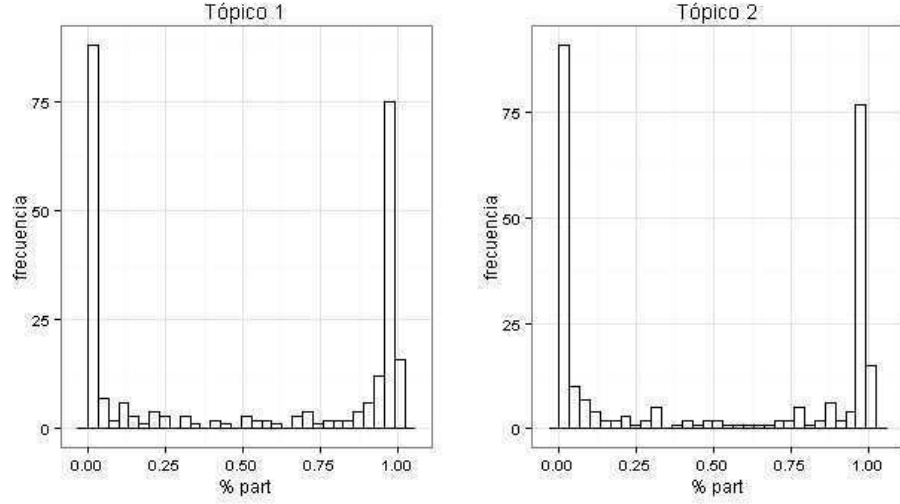


Figura 3.11: Histograma de participación de cada tópico de la ventana 1 (1960-1964)

El siguiente paso es determinar el nacimiento o decadencia de los tópicos a través de las ventanas. Para ello, se puede calcular la similitud entre los tópicos en distintas ventanas (estabilidad de tópicos). Para determinar la similitud entre dos conjuntos de tópicos, se calcula la medida KL entre los tópicos de la ventana t y los tópicos de la ventana $t + 1$, por ejemplo. En [Steyvers y Griffiths, 2007] la medida KL entre los tópicos j_1 y j_2 se calcula como:

$$KL(j_1, j_2) = \frac{1}{2} \frac{\sum_{k=1}^W \theta_k'^{j_1} \log_2 \theta_k'^{j_1}}{\theta_k''^{j_2}} + \frac{1}{2} \frac{\sum_{k=1}^W \theta_k''^{j_2} \log_2 \theta_k''^{j_2}}{\theta_k'^{j_1}} \quad (3.6)$$

Donde θ' y θ'' corresponden a la distribución de tópicos de dos ventanas consecutivas. Se toman las estructuras de la ventana t y $t + 1$:

- Proporciones de tópicos: una matriz $K \times V$, conteniendo las probabilidades para cada tópico, donde V es el largo del vocabulario.
- Vocabulario: una matriz $V \times 1$, representando el vocabulario.

Dadas dos instancias de tópicos, con diferentes vocabularios y orden de palabras, se calcula la medida KL entre ellas, dando como resultado una matriz de similitud. Si un tópico en t tiene baja divergencia con un tópico en $t + 1$ (son similares) y alta divergencia con el resto de tópicos de $t + 1$ (distintos), se puede decir que el conjunto de tópicos es similar (baja divergencia = similares, alta divergencia = distintos).

Cada combinación en la matriz muestra la divergencia KL entre dos tópicos. Una combinación con un tono más claro en $x = i$ e $y = j$ significa que los tópicos i y j

son más similares (baja divergencia), mientras que una combinación con un tono de color más oscuro significa que son menos similares (alta divergencia). En la figura 3.12 se muestra la matriz de similitud entre los tópicos de la ventana 1 y los tópicos de la ventana 2. El tópico menos divergente es el tópico 3 de la ventana 2 haber-cantar-querer-tener-noche-si-pa. Los términos más relevantes del tópico 1 de la ventana 2 (más divergente) son carnaval-ver-haber-bailar-querer-cantar-pa.

| | W2-t1 | W2-t2 | W2-t3 | W2-t4 | W2-t5 |
|-------|-------|-------|-------|-------|-------|
| W1-t1 | 0.5 | 0.4 | 0.22 | 0.44 | 0.48 |
| W1-t2 | 0.72 | 0.46 | 0.19 | 0.46 | 0.54 |

Figura 3.12: Matriz de similitud $w1(1960-1964)$ versus $w2(1961-1965)$

Para analizar el comportamiento de los tópicos divergentes en su propia ventana se adaptó el código de la librería LDAVis de R [Sievert y Shirley, 2014]. El objetivo es comparar los tópicos dentro de una misma ventana temporal de forma visual. Esta representación tiene dos componentes, el modelo global de los tópicos, y luego la frecuencia y relevancia de los términos.

Para representar los tópicos, se muestra cuán prevalente es cada tópico, y cómo se relacionan los tópicos entre ellos. Las variables de entrada son:

- ϕ : matriz $K \times W$ contiene la función de densidad de probabilidad estimada sobre W términos en el vocabulario para cada uno K tópicos en el modelo
- θ : matriz $D \times K$ contiene la función de densidad de probabilidad estimada sobre K tópicos en el modelo para cada uno de los documentos D en el corpus
- nd : número de tokens observados en el documento d , donde nd tiene que ser un entero mayor a cero
- $vocab$: un vector de largo W conteniendo los términos en el vocabulario, listado en el mismo orden que las columnas de ϕ

- Mw : frecuencia del término w en todo el corpus, donde Mw debe ser un entero mayor a cero para cada término $w = 1 \dots W$

Se computan las cuentas de tokens en los K tópicos:

- Frecuencia de tópicos = suma de columnas ($\theta \times nd$)
- Proporción de tópicos = Frecuencia de tópicos / suma (Frecuencia de tópicos)

Cada tópico se representa con una figura, y su prevalencia se representa por el tamaño de un círculo. Luego se calcula la matriz de distancia entre tópicos, utilizando una versión simétrica de la medida KL, la divergencia Jensen Shannon.

La matriz de entrada es la matriz ϕ , donde cada fila contiene la distribución de los términos para cada tópico, con tantas filas como tópicos, y tantas columnas como términos en el vocabulario. La matriz de distancia tiene una dimensión $K \times K$, que se reduce a $K \times 2$, utilizando componentes principales.

Para la segunda ventana (1961-1965), se representaron los 5 tópicos en sus dos componentes principales (figura 3.13).

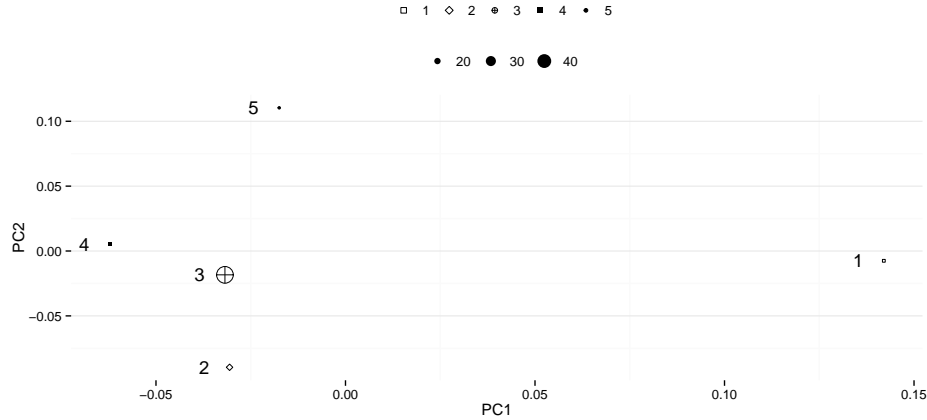


Figura 3.13: Representación de tópicos en dos primeros componentes principales de la ventana 2 (w_2 : 1961-1965)

En la figura 3.13 se representan los 5 tópicos de la ventana 2, y se ve la distancia entre el tópico 3 (el menos divergente de la ventana 2 con respecto a la ventana 1 -figura 3.12 -) y el resto de los tópicos.

Con estos gráficos (matriz de similitud y representación en componentes principales) se analizó la relación de los tópicos entre ventanas y dentro de su propia ventana respectivamente. Si un tópico se caracteriza como emergente, pero tiene una distancia muy cercana con el resto de los tópicos de su ventana, no se lo toma como emergente.

Factorización matricial dinámica no negativa (*Dynamic Non-negative Matrix Factorization*)

En “Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis” [Greene y Cross, 2015] se analizan las interacciones políticas en el parlamento europeo considerando como la agenda política de las sesiones evolucionó a lo largo del tiempo. Para detectar temas latentes en los discursos legislativos a lo largo del tiempo, analizan el contenido de los discursos usando un método de modelado dinámico de tópicos basado en dos capas de NMF. Primero dividen los discursos en distintas ventanas temporales, en los cuales bajos niveles de tópicos por ventana se identifican aplicando NMF. Luego los tópicos de cada ventana se representan como una matriz combinada de tópicos-documentos. Aplicando NMF a esta nueva representación, se pueden identificar tópicos dinámicos de alto nivel que potencialmente abarcan varias ventanas de tiempo.

Para este experimento se utilizaron **sólo los sustantivos**. El proceso tiene tres pasos:

- Pre-procesamiento: se tokenizan los documentos, se remueven las *stopwords* (las mismas utilizadas para las ventanas del experimento en python), y se construye una matriz de documentos-términos para cada ventana de tiempo. Se seleccionan las opciones de peso tfidf y normalizado de largo de documentos.
- Modelado de tópico para cada ventana: se generan ventanas temporales de tópicos, teniendo en cuenta hitos históricos como separadores de cada ventana. Como resultado el corpus se dividió en seis ventanas. Se aplica NMF para los resultados del paso anterior (pre-procesamiento). Para cada ventana se eligieron 7 tópicos.
- Modelado dinámico de tópicos: luego que se crearon las ventanas de tópicos temporales, se combinan los resultados de las ventanas para generar tópicos dinámicos que abarcan varias ventanas de tiempo.

A los resultados producidos para cada ventana se les asignó un color, donde cada color representa una temática global (naturaleza, folklore/campo, sentimientos, términos en inglés, rock).

El modelo de factorización matricial no negativa dinámica utiliza también una medida de coherencia conocida como $TC-W2V$, que se utiliza para comparar diferentes modelos de tópicos y elegir un modelo con una cantidad de tópicos adecuada [O’Callaghan et al., 2015]. Esta medida evalúa la relación de un conjunto de términos principales que describen un tópico, basada en la similitud de sus representaciones en un espacio Word2vec. La coherencia de un tópico t_h se representa por sus términos más sobresalientes dada la similitud coseno entre todos los términos relevantes en el espacio Word2vec.

$$coh(t_h) = \frac{1}{\binom{h}{2}} \sum_{j=2}^t \sum_{i=1}^{j-1} \cos(wv_i, wv_j) \quad (3.7)$$

Un puntaje general para un modelo de t3pico T consistiendo en K t3picos est3 dado por la media de los puntajes de coherencia de t3picos.

$$coh(T) = \frac{1}{k} \sum_{h=1}^k coh(t_h) \quad (3.8)$$

Un valor adecuado para k es aquel donde alcanza el m3ximo valor de coherencia. Se utiliz3 la medida de coherencia con un n3mero de k t3picos distinto, para elegir el n3mero 3ptimo. En este paso se eligi3 un n3mero de t3picos entre 4 y 20. Para cada uno de estos valores, el modelo prueba el valor de coherencia y recomienda una cantidad 3ptima de t3picos por ventana. Para el modelado din3mico, tambi3n calcula la misma medida, y muestra el n3mero de t3picos a calcular para los t3picos din3micos. En este caso, los valores de k seleccionados fueron entre 4 y 10.

3.2.1.2. Corpus sin separar ventanas

El objetivo de este experimento es facilitar la comparabilidad de t3picos en los distintos per3odos hist3ricos. Se parte del supuesto que para analizar la evoluci3n temporal de los t3picos, la cantidad de t3picos y los t3picos en s3 deben ser los mismos para todo el corpus. En esta secci3n se exponen los indicadores que se tomaron para la elecci3n del n3mero 3ptimo de t3picos, la evoluci3n temporal de esos t3picos, la evoluci3n temporal de palabras de esos t3picos y la obtenci3n de t3picos emergentes y en decadencia.

Elecci3n de n3mero de t3picos

Para realizar estos experimentos, se utiliz3 Mallet [McCallum, 2002], con los siguientes par3metros para el entrenamiento de t3picos (cuadro 3.1):

Cuadro 3.1: Par3metros para Latent Dirichlet Allocations

| t3picos | α |
|---------|----------|
| 10 | 5.00 |
| 20 | 2.50 |
| 50 | 1.00 |
| 100 | 0.50 |
| 150 | 0.33 |
| 200 | 0.25 |
| 250 | 0.20 |

La relación entre el número de tópicos y α surge de $\alpha = \frac{50}{t}$, donde t es el número de tópicos [Griffiths y Steyvers, 2004]. Un alto valor de α significa que cada documento es probable que contenga una mezcla de la mayoría de los temas, y no específicamente un tema. Un valor bajo de α significa que es más probable que un documento pueda contener una mezcla de sólo unos pocos, o incluso sólo uno de los temas.

Parámetros adicionales:

- *keepsequence* preserva el documento como una secuencia de *features* de palabras, más que un vector de cuentas de *features* de palabras.
- lista de *stopwords*.
- *optimize – interval* = 10: esta opción activa una optimización de hiperparámetros, que permite al modelo ajustarse mejor a los datos permitiendo que algunos tópicos sean más prominentes que otros. Una optimización cada 10 iteraciones es razonable.
- número de iteraciones = 1000.
- $\beta = 0,1$. Un alto valor de β significa que cada tema es probable que contenga una mezcla de la mayor parte de las palabras, y no una palabra específica, mientras que un valor bajo significa que un tema puede contener una mezcla de sólo algunas de las palabras.

Para determinar la calidad de los tópicos, y poder elegir una de las distintas opciones parametrizadas, se tomaron en cuenta medidas que surgen como resultado del procesamiento en Mallet.

En “Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements” [Boyd-Graber et al., 2014] se analizan medidas de los tópicos para definirlos como “buenos” o “malos”. Para este trabajo se utilizaron algunas de esas medidas para determinar el número óptimo de tópicos para realizar los experimentos.

Número de palabras: número total de tokens asignados al tópico. Tópicos pequeños suelen ser ilógicos, tópicos grandes son demasiado generalizados. Una explicación para esta relación es que los tópicos más comunes estén bien representados en muchos documentos. Para los tópicos menos frecuentes, el modelo debe estimar la distribución de palabras para una muestra menor. Los tópicos más pequeños son más vulnerables a mezclarse con otros tópicos porque no tienen su propia distribución (figura 3.14).

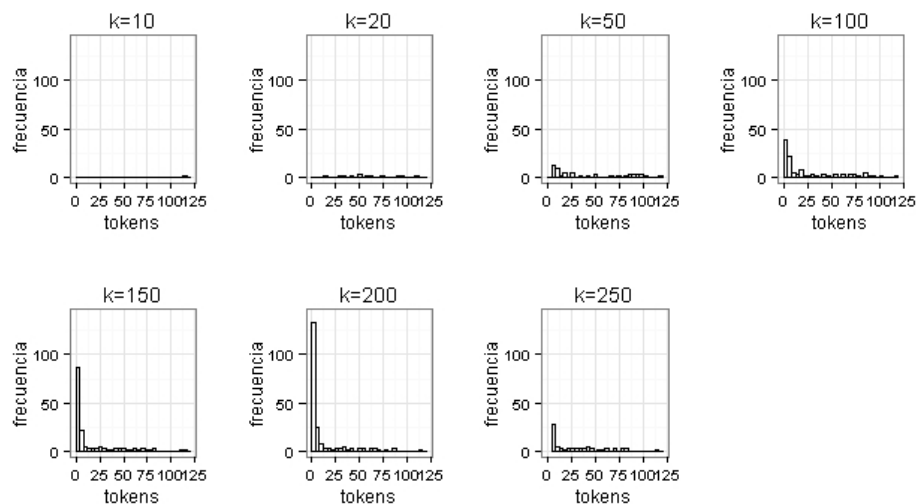


Figura 3.14: Cantidad de tokens (en miles) según número de tópicos

Largo de palabras: para cada palabra, cuenta el número de caracteres. Tópicos con muchas palabras cortas tienden a ser problemáticos. Esta métrica normaliza el largo de las palabras contra el largo promedio de las principales palabras en los tópicos. Números negativos significan palabras cortas, números positivos significan palabras largas. Esta métrica es útil para encontrar tópicos anómalos. La intuición es que las palabras con significado más específico tienden a ser más largas, y viceversa. El largo de las palabras no necesariamente significa que el tópico no sea interpretable, pero indica que es un tipo diferente de agrupamiento de palabras que aparecen juntas (figura 3.15).

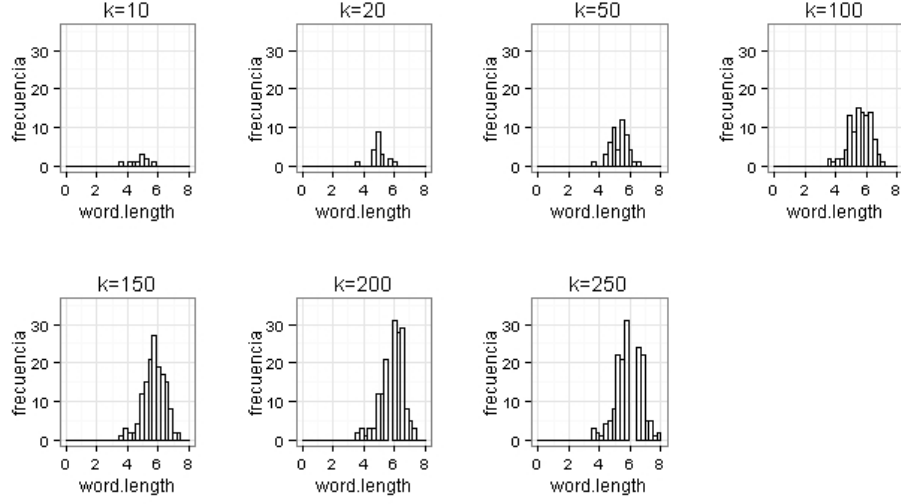


Figura 3.15: Largo de palabras según número de tópicos

Coherencia: probabilidad de palabras dadas las palabras más altamente clasificadas. Esta métrica extrae combinaciones ilógicas. Sea $D(v)$ la frecuencia de documentos para la palabra de tipo v (por ejemplo, el número de documentos con al menos un token del tipo v) y $D(v, v')$ la frecuencia co-documento del tipo de palabras v y v' , se defina coherencia de tópicos como:

$$C(t; V^t) = \sum_{m=2}^M \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \quad (3.9)$$

Donde $V^{(t)} = (v_1^{(t)}, \dots, v_M^{(t)})$ es una lista de las M palabras más probables en el tópico t . Una cuenta suavizada de 1 se incluye para evitar tomar el logaritmo de 0.

Esta medida descansa en las estadísticas de co-ocurrencia de palabras extraídas del corpus que está siendo modelado, y no depende de un corpus externo de referencia. Los números cercanos a 0 indican mayor coherencia [Mimno et al., 2011].

Para este conjunto de datos, a mayor número de tópicos, la medida se aleja de cero.

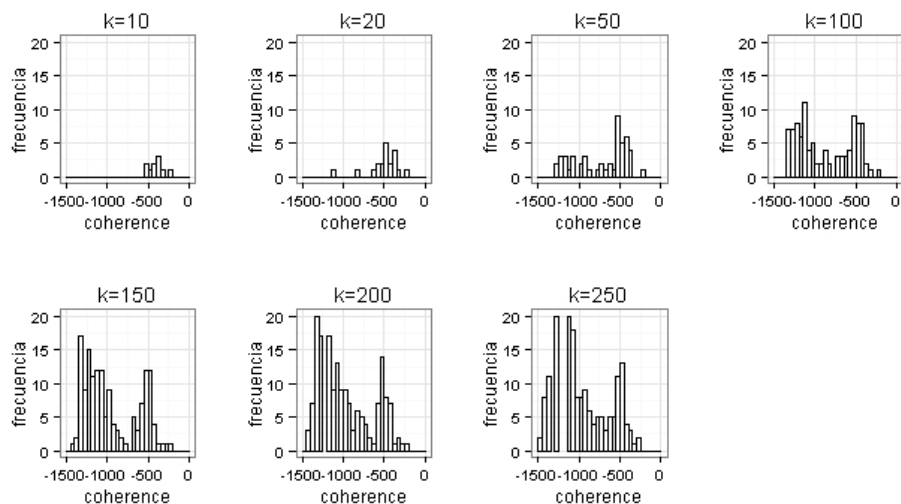


Figura 3.16: Coherencia según número de tópicos

Distancia de corpus: Se puede definir un tópico global contando el número de veces que cada palabra se utiliza en todos los documentos y normalizando esas cuentas. Las distribuciones de tópicos que son similares a esta distribución a nivel del corpus, de acuerdo a una medida de similitud entre distribuciones, como la distancia Jensen-Shannon o la distancia de Hellinger, consisten en las palabras más comunes en el corpus. Estos tópicos se perciben como demasiado generales. Tener un pequeño número de estos tópicos generales puede ayudar a mejorar la calidad de los otros tópicos. Valores más altos significan temas más específicos, temas con los valores más bajos serían lo que se obtendría mediante el recuento de todas las palabras, independientemente del tópico. A medida que aumenta el valor de k , mayores son los valores de la medida corpus-distancia (figura 3.17).

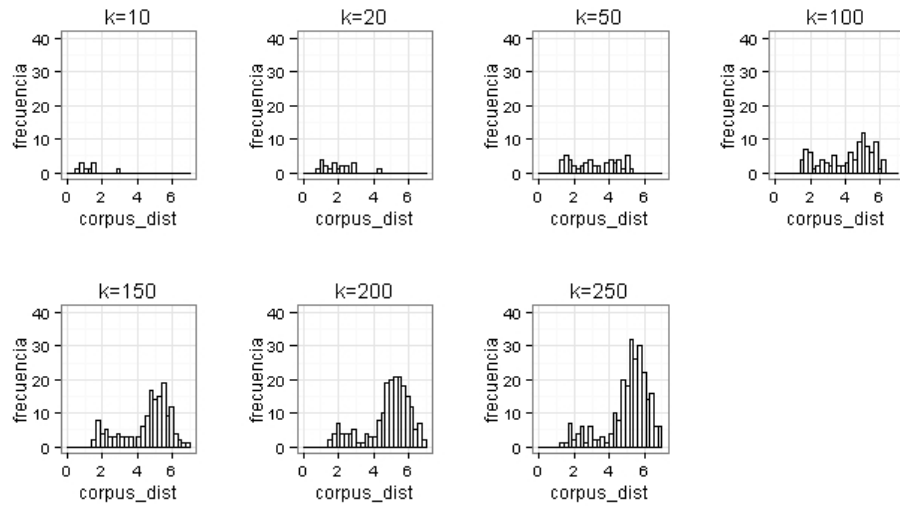


Figura 3.17: Distancia según número de tópicos

Diferencia token/documento: Esta métrica compara el número de veces que una palabra ocurre en un tópico (medido en tokens) y el número de documentos en el que la palabra aparece como parte de ese tópico (instancias de la palabra asignadas a otro tópico no se cuentan) -ver figura 3.18-.

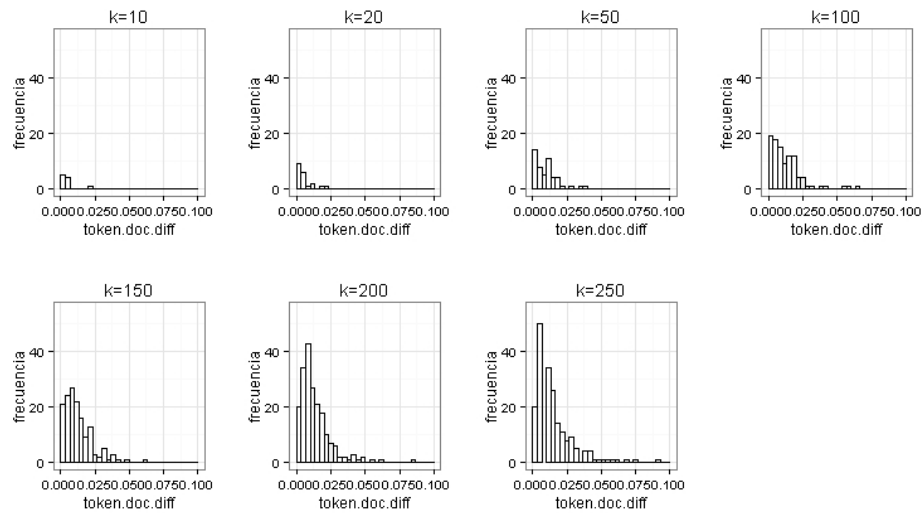


Figura 3.18: Diferencia token/documento según número de tópicos

Documentos de rango 1: temas vacuos o demasiado generales ocurren en poca medida en muchos documentos. Esta métrica cuenta fuera de los documentos que contienen un

tópico dado, cuantas veces ese tópico es el único más común en un documento. Números bajos indican posiblemente tópicos no interesantes (figura 3.19).

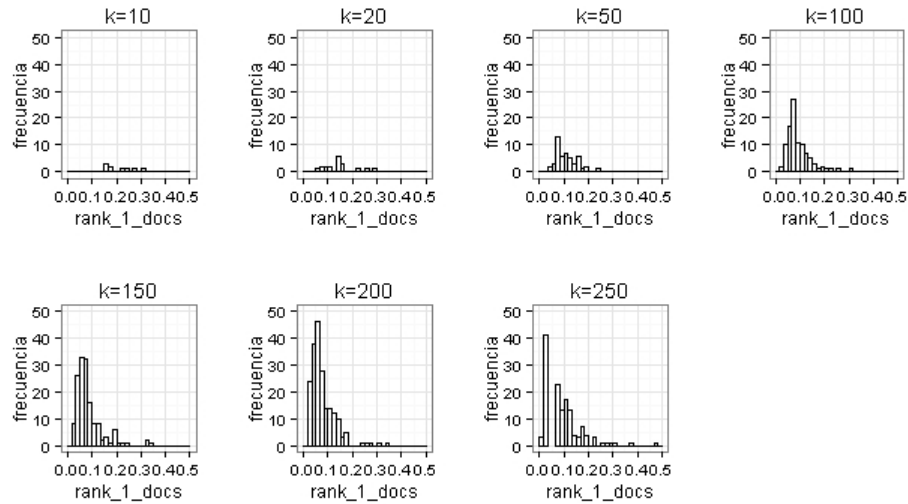


Figura 3.19: Documentos de rango 1 según número de tópicos

Entropía de documentos: medida de dispersión de los tópicos a lo largo de los documentos. Esta medida parece correlacionarse con el logaritmo del número de tokens en el tópico (figura 3.20).

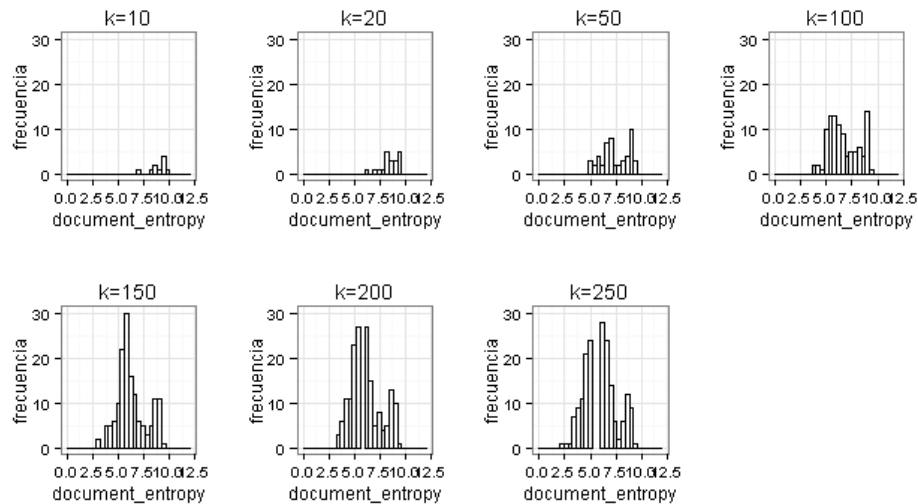


Figura 3.20: Entropía de documentos según número de tópicos

Teniendo en cuenta estas medidas, se decidió que el número de tópicos que mejor representaba el corpus era entre 10 y 50 tópicos. Para elegir el número óptimo de tópicos se graficaron los tópicos en ejes temporales para $k = 10$, $k = 20$ y $k = 50$.

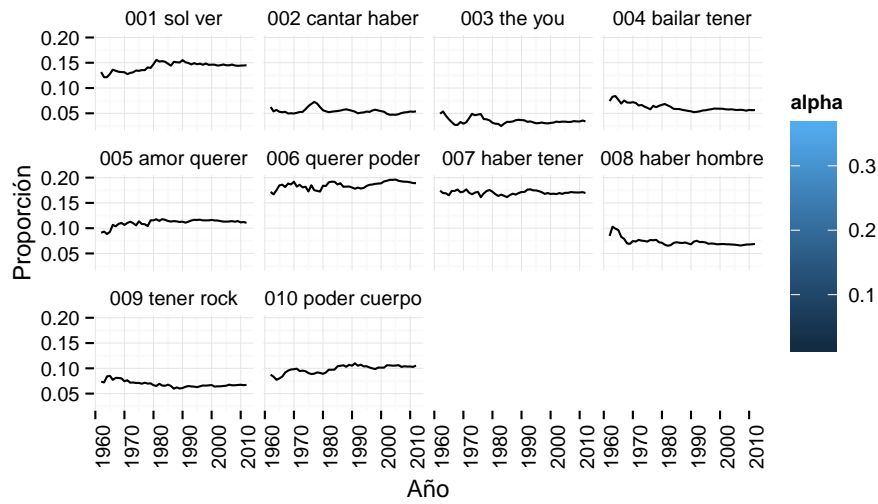


Figura 3.21: Tópicos a través del tiempo para $k = 10$

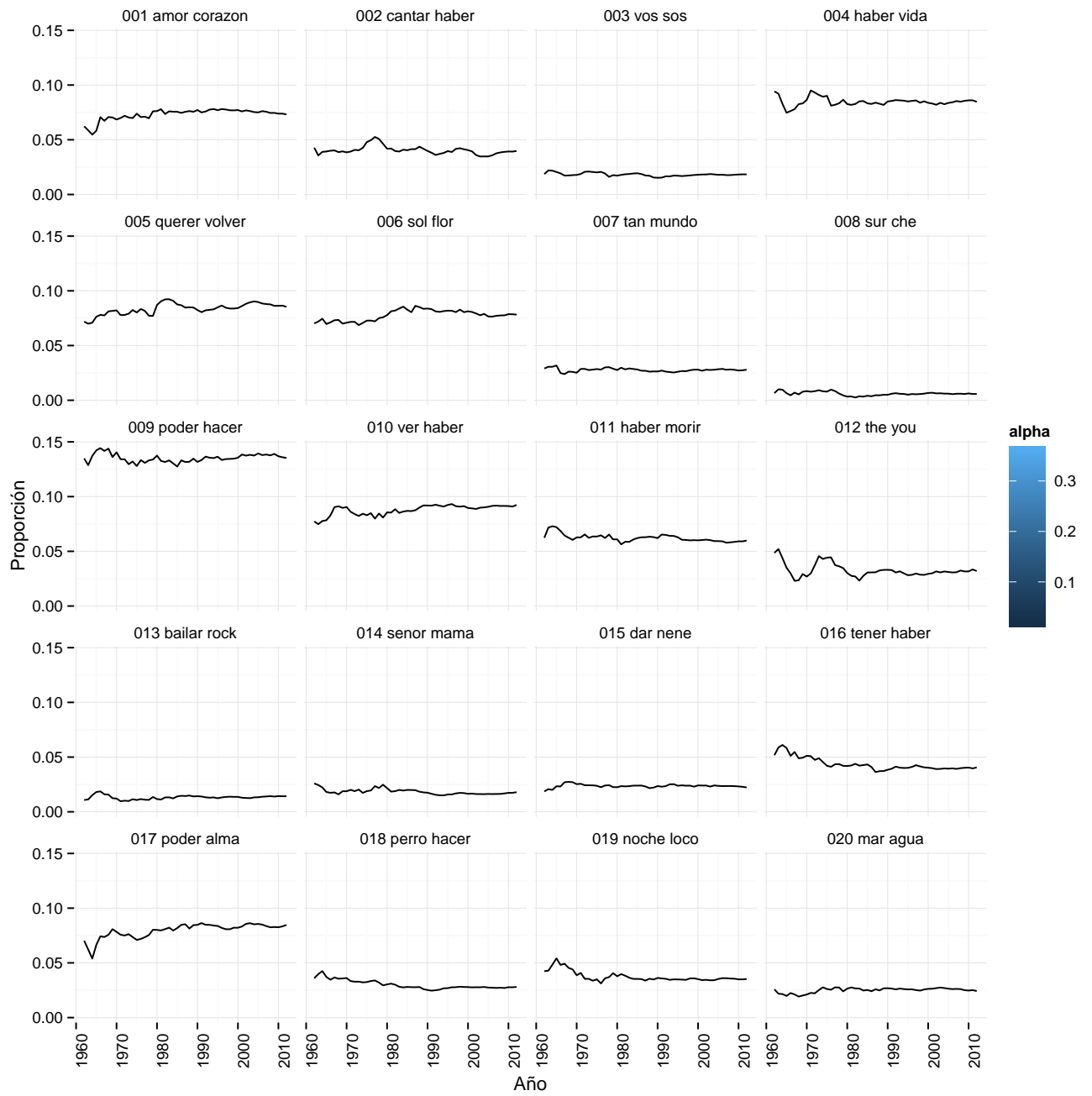


Figura 3.22: Tópicos a través del tiempo para $k = 20$

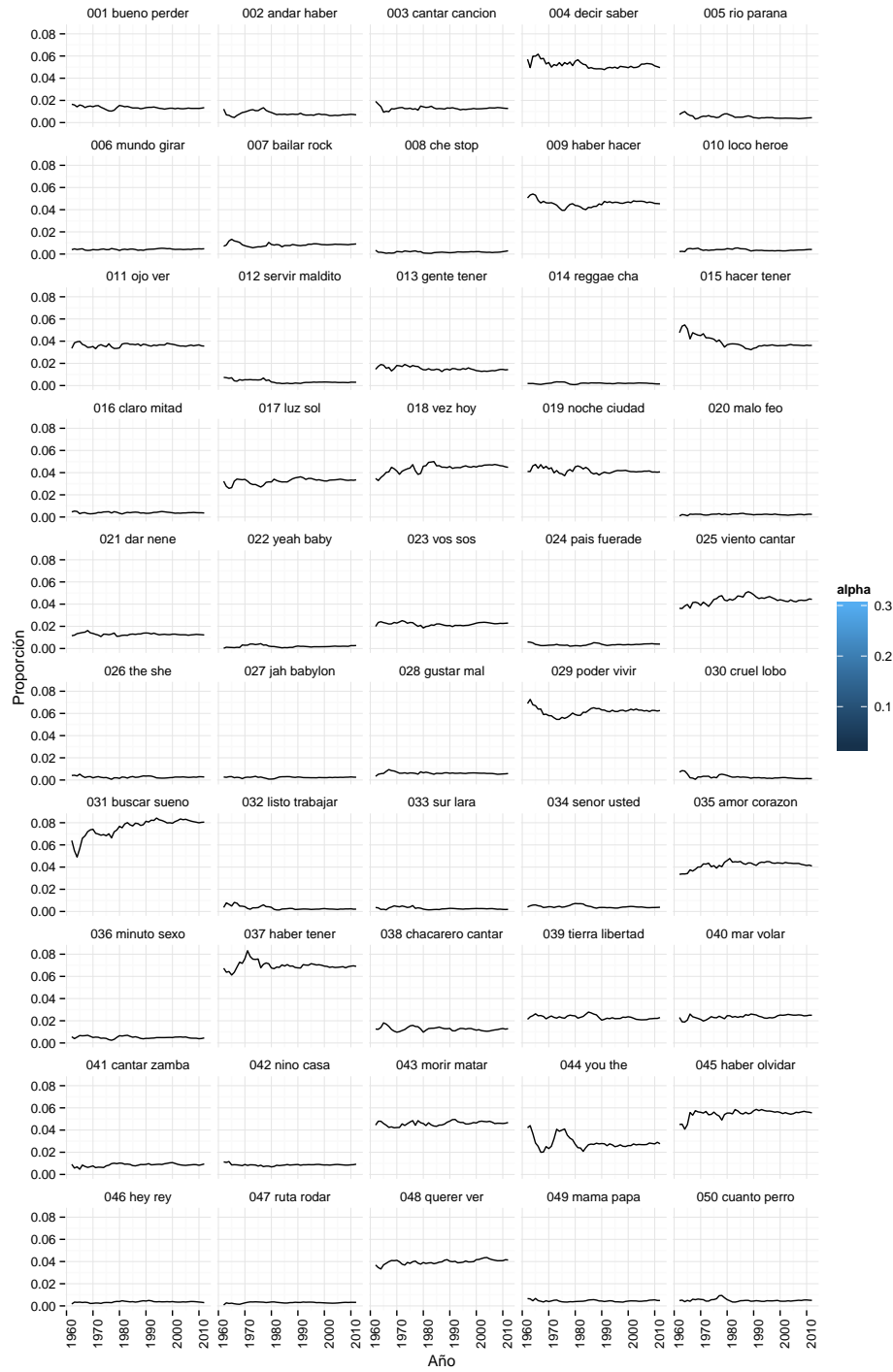


Figura 3.23: Tópicos a través del tiempo para $k = 50$

Para $k = 10$, se graficaron todos los tópicos en el mismo eje temporal (figura 3.24).

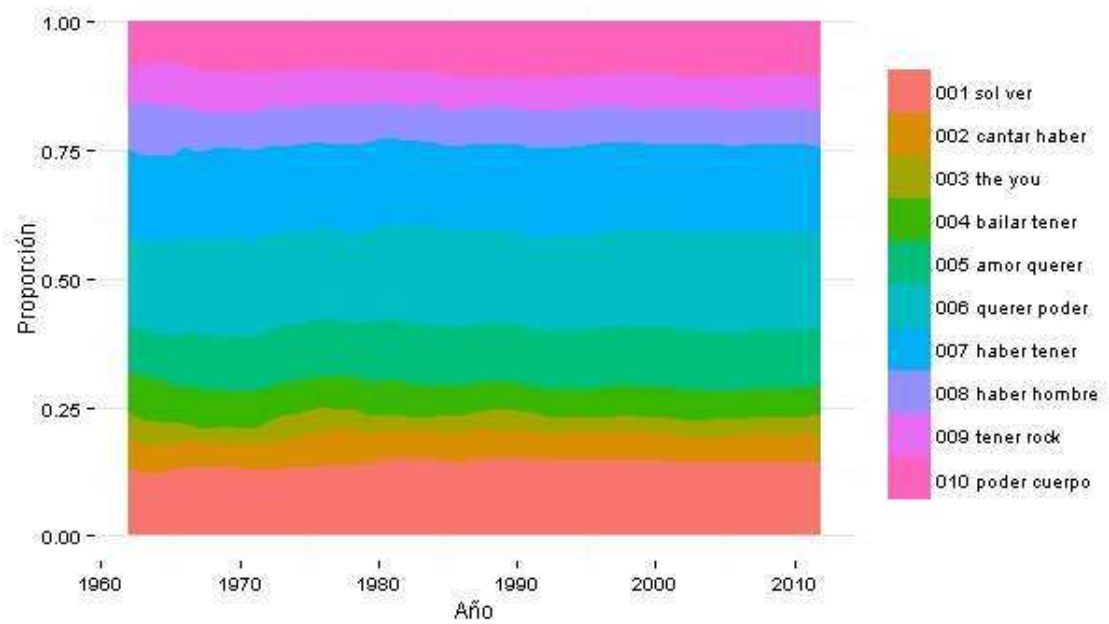


Figura 3.24: Evolución temporal de tópicos para $k=10$

Para $k = 20$ y $k = 50$ se graficaron sólo los tópicos con mejor medida de coherencia (figuras 3.25 y 3.26).

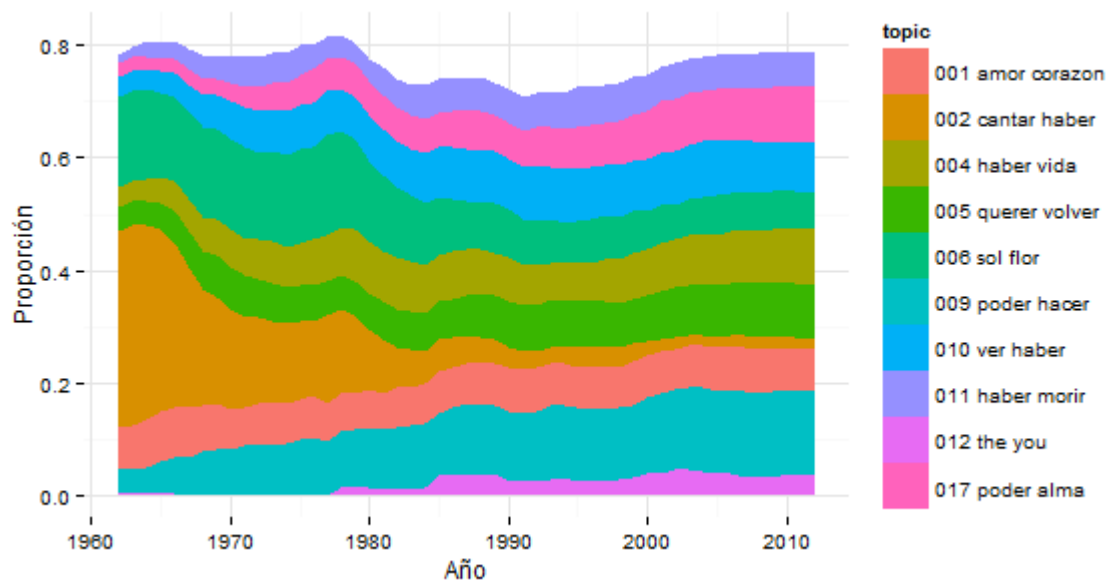


Figura 3.25: Evolución temporal de tópicos para $k=20$

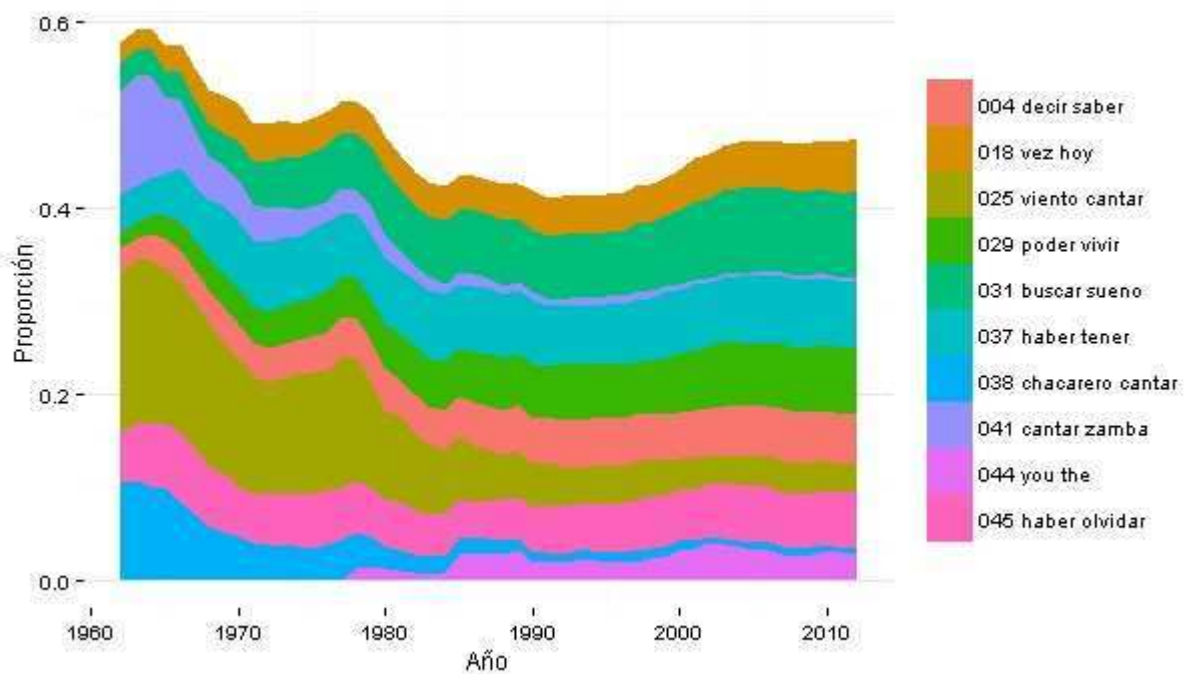


Figura 3.26: Evolución temporal de tópicos para $k=50$

En las figuras 3.21, 3.22 y 3.23 se presentaron los tópicos en gráficos separados, para ver el comportamiento individual de cada tópico en el eje temporal. Esto permite analizar no sólo el cambio de comportamiento temporal, sino la proporción de cada uno de los tópicos. Por ejemplo, en la figura 3.21 hay tópicos con cambio de comportamiento en el tiempo (tópico 009 tener rock), pero con poca participación en el corpus. En las figuras 3.24, 3.25 y 3.26 se presentaron los tópicos en el mismo gráfico. En la figura 3.24 no se observa un cambio de proporción según la línea temporal. En cambio, en las figuras 3.25 y 3.26 se puede ver una variación en la proporción de acuerdo con la época. Por ejemplo, en la figura 3.22 el tópico 002 presenta mayor proporción antes de 1980, mientras que luego de esa época, decae. En base a estos gráficos, se decidió que la mejor representación temporal estaba dada por $k = 20$. Para $k = 10$ no se observaron cambios de la participación de los tópicos en el tiempo, y para $k = 50$ hay pocos tópicos que representen un cambio temporal.

Uso de las palabras a través del tiempo

En “Words alone: Dismantling topic models in the humanities” [Schmidt, 2012], se busca saber que palabras caracterizan un autor o género. Si se compara la diferencia entre frecuencias de distintos corpus, se obtiene la lista de palabras comunes. Se utiliza la fórmula de Dunning’s log likelihood:

$$\text{Dunning's ll} = O_i \times \ln \left(\frac{O_i}{E_i} \right) \quad (3.10)$$

Donde O es la frecuencia observada, y $\frac{O}{E}$ es la frecuencia observada dividida por la frecuencia esperada.

En “Accurate methods for the statistics of surprise and coincidence” [Dunning, 1993], Dunning realizó un análisis de una muestra de palabras con el objetivo de encontrar palabras que aparezcan juntas con una frecuencia mayor a la esperada, basado sólo en las frecuencias. Creó una tabla de contingencia que contenía las siguientes cuentas de cada bigrama que aparecía en el texto:

Cuadro 3.2: Tabla de contingencia

| | |
|-----------------------|----------------------------|
| $K(AB)(k_{.11})$ | $K(\neg AB)(k_{.12})$ |
| $K(A\neg B)(k_{.21})$ | $K(\neg A\neg B)(k_{.22})$ |

Donde $\neg AB$ representa el bigrama en el cual la primera palabra no es la palabra A y la segunda es la palabra B . Si las palabras A y B ocurren independientemente, se puede esperar $p(AB) = p(A)p(B)$ donde $p(AB)$ es la probabilidad de A y B ocurriendo en una secuencia, $p(A)$ es la probabilidad de A apareciendo en la primera posición, y $p(B)$ es la probabilidad de B apareciendo en la segunda posición. La hipótesis nula sería que A

y B son independientes, $p(A | B) = p(A | \neg B) = p(A)$. Esto significa que para probar la independencia de A y B se puede buscar si la distribución de A dado B (primera fila de la tabla) es la misma que la distribución de A dado que B no está presente (segunda fila de la tabla). No realiza un test para probar que A y B son independientes, sino que utiliza un test estadístico como una medida para destacar los casos en que A y B están fuertemente asociados en un texto. Con estos datos se computa el log-likelihood ratio score (LLR). LLR compara dos hipótesis sobre la probabilidad de ocurrencia de una palabra en un corpus y la probabilidad de la misma palabra en otro corpus. Se iguala el conteo de palabras en un corpus a un experimento de tipo Bernoulli. Si la probabilidad que la próxima palabra concuerde con un prototipo es p , luego el número de coincidencias generadas en las próximas n palabras es una variable aleatoria K con distribución binomial

$$p(K = k) = p^k(1 - p)^{n-k} \binom{n}{k} \quad (3.11)$$

cuya media es np y su varianza es $np(1 - p)$. Si $np(1 - p) > 5$, la distribución de esta variable será aproximadamente normal. Hay otra clase de tests que no dependen del supuesto de normalidad. Estos tests asumen que el modelo es conocido, pero que los parámetros de los modelos son desconocidos. La probabilidad que una salida experimental descrita por k_1, \dots, k_n se observará para un modelo dado descrito por un número de parámetros p_1, p_2, \dots se denomina función de verosimilitud para el modelo y se escribe como

$$H(p_1, p_2, \dots; k_1, \dots, k_m) \quad (3.12)$$

donde todos los argumentos a la izquierda del punto y coma son parámetros, y los argumentos a la derecha son valores observados. Para experimentos Bernoulli repetidos, $m = 2$ porque se observan el número de experimentos y el número de salidas positivas y sólo hay una p . La fórmula explícita de la función de verosimilitud es

$$H(p : n, k) = p^k(1 - p)^{n-k} \binom{n}{k} \quad (3.13)$$

Para resumir la notación, los parámetros del modelo se pueden expresar en un solo parámetro, así como los valores observados. Luego la función de verosimilitud puede escribirse como

$$H(\omega; k) \quad (3.14)$$

donde ω se considera como un punto en el espacio de parámetros Ω , y k un punto en el espacio de observaciones K . La razón de verosimilitud para una hipótesis es la relación entre el máximo valor de la función de verosimilitud sobre el subespacio representado por la hipótesis y el máximo valor de la función de verosimilitud en todo el espacio de parámetros.

$$\lambda = \frac{\max_{\omega \in \Omega_0} H(\omega, k)}{\max_{\omega \in \Omega} H(\omega, k)} \quad (3.15)$$

Donde Ω es el espacio de parámetros completo y Ω_0 es la hipótesis que está siendo testeada. La característica particular de las razones de verosimilitud es que la cantidad $-2 \log \lambda$ está asintóticamente distribuida como una distribución χ_2 con grados de libertad iguales a la diferencia en dimensión entre Ω y Ω_0 . La comparación de dos procesos binomiales o multinomiales se puede llevar a cabo usando razones de verosimilitud. En el caso de dos distribuciones binomiales

$$H(p_1, p_2; k_1, n_1, k_2, n_2) = p_1^{k_1} (1 - p_1)^{n_1 - k_1} \binom{n_1}{k_1} p_2^{k_2} (1 - p_2)^{n_2 - k_2} \binom{n_2}{k_2} \quad (3.16)$$

La hipótesis que las dos distribuciones tienen el mismo parámetro subyacente está representada por el conjunto $\{(p_1, p_2) \mid p_1 = p_2\}$. La verosimilitud para este test es:

$$\lambda = \frac{\max_p H(p, p; k_1, n_1, k_2, n_2)}{\max_{p_1, p_2} H(p_1, p_2; k_1, n_1, k_2, n_2)} \quad (3.17)$$

El valor máximo se alcanza con $p_1 = \frac{k_1}{n_1}$ y $p_2 = \frac{k_2}{n_2}$ para el denominador, y $p = \frac{k_1 + k_2}{n_1 + n_2}$ para el numerador. Esto reduce la razón a

$$= \frac{\max_p L(p, k_1, n_1) L(p, k_2, n_2)}{\max_{p_1, p_2} L(p_1, k_1, n_1) L(p_2, k_2, n_2)} \quad (3.18)$$

donde

$$L(p, q, n) = p^k (1 - p)^{n - k} \quad (3.19)$$

Tomando el algoritmo de la razón de verosimilitud esto es igual a

$$-2 \log \lambda = 2[\log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) - \log L(p, k_1, n_1) - \log L(p, k_2, n_2)] \quad (3.20)$$

Esta metodología es similar a otras mencionadas en [Sanderson, 2010] que se utilizan para selección de atributos, como Información mutua y Chi cuadrado.

Se decidió utilizar la medida de Dunning, por su sencillez de cálculo, adaptando el código del artículo de Schmidt "Words alone: Dismantling topic models in the humanities" [Schmidt, 2012].

Se dividió el tiempo en cuartiles, en función de la fecha, y se contaron las frecuencias en cada intervalo temporal. Luego se calculó la media de la medida Dunning log-likelihood. Esto permite ver los tópicos no como una única lista ordenada, sino como cuatro listas con un orden distinto dependiendo del período. Se seleccionaron los términos más relevantes por tópicos, y se graficaron en cada época, mostrando la diferencia de posición en el tópico según el período (ver resultados en la Subsubsección 4.1.2.2).

Tópicos emergentes y tópicos en decadencia (*Hot and Cold Topics*)

Los tópicos emergentes son tópicos con una correlación positiva con el momento de publicación. Los tópicos en decadencia son tópicos que tienen una correlación negativa con el momento de publicación.

Para calcular los tópicos emergentes y en decadencia, se adaptó la función del paquete jstor [Marwick, 2013].

Se tomaron como entrada las proporciones documento-tópico por fecha, para calcular la proporción promedio por fecha y una media móvil de 5 años para suavizar la curva. Se calculó la correlación de Pearson entre tópico y año y el p-valor para esta correlación ($p = 0,05$).

Se extrajeron las 5 correlaciones más positivas, y las 5 correlaciones más negativas, para el número de tópicos elegido. Los resultados se analizan en la [Subsubsección 4.1.2.3](#).

3.2.1.3. Modelado dinámico de tópicos (*Dynamic Topic Modeling*)

Se utilizó el modelo de Blei implementado en C/C++ [Blei y Lafferty, 2006]. La implementación toma como entrada:

- Un documento por línea, donde cada línea contiene un índice entero que corresponde a una sola palabra y la cuenta de palabras únicas para cada índice.
- Un archivo con las secuencias temporales, que tiene el número de documentos para cada ventana de tiempo.
- Un archivo con todas las palabras en el vocabulario, ordenadas de la misma manera que los índices de palabras.
- Un archivo con la información de cada documento.

Este modelo da como resultado:

- Las distribuciones de palabras para cada tópico para todas las ventanas.
- Las gamas asociadas con cada documento. Si se divide esto por la suma de cada documento se obtiene la mezcla esperada de tópicos.

Los parámetros utilizados fueron:

- Número de tópicos: 20.
- α : 0,01.
- Mínimo de iteraciones 6, máximo de iteraciones 20, máximo en iteraciones: 10.

Los archivos de salida se procesaron con un script en R, que dió como resultado las palabras que componen los tópicos según la ventana temporal y la participación de los tópico en cada documento.

La salida del modelo DTM es el orden de cada palabra perteneciente al tópico según la ventana de tiempo.

Se reproducen algunos ejemplos:

Cuadro 3.3: DTM Ejemplo 1

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----|---------|--------|-----|---------|--------|-------|---------|---------|---------|----------|
| W1 | canción | cantar | voz | alegría | gracia | canto | navidad | piedra | brindar | claro |
| W2 | canción | cantar | voz | alegría | gracia | canto | navidad | piedra | brindar | claro |
| W3 | canción | cantar | voz | alegría | gracia | canto | navidad | claro | brindar | piedra |
| W4 | canción | cantar | voz | alegría | gracia | canto | claro | navidad | brindar | pájaro |
| W5 | canción | cantar | voz | alegría | gracia | canto | claro | navidad | brindar | borracho |
| W6 | canción | cantar | voz | alegría | gracia | canto | navidad | claro | brindar | borracho |

Cuadro 3.4: DTM Ejemplo 2

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----|-------|-------|--------|--------|--------|--------|-----|--------|--------|----------|
| W1 | matar | tu | guerra | de | muerte | morir | el | sangre | ley | bastar |
| W2 | matar | tu | guerra | de | muerte | morir | el | ley | sangre | bastar |
| W3 | matar | tu | guerra | de | muerte | morir | el | ley | sangre | bastar |
| W4 | matar | tu | guerra | morir | de | muerte | ley | el | sangre | arma |
| W5 | matar | morir | guerra | de | tu | muerte | el | ley | sangre | bastar |
| W6 | matar | morir | de | guerra | muerte | tu | ley | el | sangre | infierno |

La salida si bien muestra un cambio en el orden de las palabras según la ventana, no arrojó mejores resultados que LDA, con los parámetros utilizados. Aparecen en el primer ejemplo para las ventanas 5 y 6 el término borracho. En "Propuestas para una antropología argentina" [Berbeglia, 2007], se menciona la existencia de una apología del consumo del alcohol en las tribus juveniles, que muestra a estos grupos como contraculturas. En el segundo caso, la palabra morir tiene mayor importancia en las dos últimas ventanas.

Esta técnica no fue útil para el objetivo buscado, por lo que no se profundizará en los resultados.

3.2.2. Clasificación

Para los experimentos de clasificación, se utilizaron Mallet, Python, R y Wowpal-Wabbit [Langford et al., 2007a]. Los objetivos de estos experimentos fueron:

- Analizar el grado de precisión de clasificación de las letras en períodos temporales. Si el grado de precisión fuese superior al azar, se concluiría que hay características de las palabras o su combinación que permiten identificar un período histórico.
- Comparar la precisión en la clasificación entre métodos clásicos como Máxima Entropía y Naive Bayes, versus Word2vec y Wowpal Wabbit.

3.2.2.1. Naive Bayes y Máxima Entropía

Con el corpus de **lemas completo y el de sustantivos** se ejecutaron los algoritmos de Naive Bayes y Máxima entropía en Mallet. Los parámetros utilizados fueron:

- Evaluación: Para la evaluación de la clasificación, se utilizaron validación cruzada con $k = 10$ y división en entrenamiento y testing.
- Tokens: se exportaron los lemas y los sustantivos desde R, previa limpieza de las tildes, ñ, ü, ya que generan problemas con la codificación.
- Años: se generaron dos tipos de agrupamiento por fecha
 - Por décadas
 - Por hitos históricos
- Balanceo: Para mejorar la performance de los algoritmos, se submuestrearon las clases mayoritarias. Se respetó la proporción de cada año en cada ventana para realizar el submuestreo. En el caso del corpus por décadas, se decidió tomar una muestra de 300 temas para cada año para las ventanas con mayor número de casos (ventanas 4 y 5, ver tabla 3.7). Para el corpus balanceado por hitos se submuestrearon también las ventanas mayoritarias, teniendo en cuenta el número mínimo de documentos por año. En la ventana 4 se tomaron 350 documentos por año, al igual que en la ventana 5, mientras que en la ventana 6 el número de muestras por año fue 200, ya que había años que no superaban ese número (ver tabla 3.8). El número de documentos por muestra fue una decisión experimental, podría haber sido un número menor.
- *Stopwords*: Para generar el corpus en Mallet, se construyó una lista de *stopwords*, con el mismo criterio utilizado para los experimentos exploratorios (eliminar términos más frecuentes de la matriz de términos documentos, y aquellos que se repetían mucho en los tópicos sin agregarles sentido). El listado de stopwords se encuentra en el **Apéndice A**.

En las tablas 3.5 y 3.6 se detallan la cantidad de documentos sin balancear por época, mientras que en las tablas 3.7 y 3.8 se exponen la cantidad de documentos por época balanceados.

Cuadro 3.5: Corpus separado en décadas

| Ventana | Período | # textos |
|---------|-------------|----------|
| W1 | [1960-1969] | 754 |
| W2 | [1970-1979] | 1.338 |
| W3 | [1980-1989] | 2.652 |
| W4 | [1990-1999] | 7.092 |
| W5 | [2000-2009] | 14.454 |
| W6 | [2010-2014] | 4.679 |
| | | 30.969 |

Cuadro 3.6: Corpus separado en hitos históricos

| Ventana | Período | # textos |
|---------|-------------|----------|
| W1 | [1960-1974] | 1.549 |
| W2 | [1975-1982] | 1.063 |
| W3 | [1983-1990] | 2.444 |
| W4 | [1991-1997] | 4.684 |
| W5 | [1998-2003] | 7.557 |
| W6 | [2004-2014] | 13.672 |
| | | 30.969 |

Cuadro 3.7: Corpus balanceado separado en décadas

| Ventana | Período | # textos |
|---------|-------------|----------|
| W1 | [1960-1969] | 754 |
| W2 | [1970-1979] | 1.338 |
| W3 | [1980-1989] | 2.652 |
| W4 | [1990-1999] | 3.000 |
| W5 | [2000-2009] | 3.000 |
| | | 12.244 |

Cuadro 3.8: Corpus balanceado separado en hitos históricos

| Ventana | Período | # textos |
|---------|-------------|----------|
| W1 | [1960-1974] | 1.549 |
| W2 | [1975-1982] | 1.063 |
| W3 | [1983-1990] | 2.444 |
| W4 | [1991-1997] | 2.450 |
| W5 | [1998-2003] | 2.100 |
| W6 | [2004-2014] | 2.200 |
| | | 11.806 |

3.2.2.2. Word2Vec (W2V)

Se cargaron los datos de entrenamiento, utilizando Sklearn [Pedregosa et al., 2011] y el tokenizador de NLTK [Loper y Bird, 2002]. Se utilizó el corpus de **lemas sin balancear separado en décadas** (ver tabla 3.5).

Los parámetros utilizados fueron:

- Arquitectura: las opciones son *skip-gram* o *continuous bag of words*. *Skip-gram* es más lento, pero produce mejores resultados. Se eligió *skip-gram*, que es el algoritmo por defecto.
- Algoritmo de entrenamiento: las opciones son *softmax jerárquico* o *negative sampling*. Se eligió *softmax*.
- Submuestreo de palabras frecuentes: la documentación de google recomienda valores entre 0,00001 y 0,001. El valor elegido fue 0,001.
- Dimensionalidad del vector de palabras: Valores razonables son entre 10 y cientos, se eligió 300.
- Contexto: cuantas palabras del contexto debe tomar el algoritmo en cuenta. Se eligió un valor de 10.
- Procesos paralelos: depende de cada computadora, con 4 *workers*, el proceso corre en tiempo razonable.
- Cuenta mínima de palabras: 40 para limitar el tamaño del vocabulario a palabras significativas. Se ignora cualquier palabra que aparezca menos de 40 veces.

Se importó el algoritmo de gensim [Řehůřek y Sojka, 2010], y se entrenó el modelo con estos parámetros.

El número de filas en el modelo es el número de palabras en su vocabulario, y el número de columnas corresponde al tamaño del vector de *features*. Este modelo tiene 3.958 palabras x 300 *features*.

W2V crea grupos de palabras relacionadas semánticamente, por lo que se puede explotar la similitud entre palabras en un cluster. Primero se tienen que encontrar los centros de los clusters de palabras, lo que se puede realizar con *k-means*. Para decidir el número de clusters, se usó el criterio que agrupamientos con pocas palabras dan mejor resultado. Se dividió el número de palabras por 5 para obtener la cantidad de grupos.

El resultado fue un cluster o asignación de centroide para cada palabra, y se puede definir una función para convertir los documentos en *bags of centroids*. Es similar a *bag of words*, pero usa agrupamientos relacionados semánticamente en vez de palabras individuales.

Esto da como resultado un vector para cada documento, cada uno con un número de *features* igual al número de clusters. Se creó un *bag of centroids* para el conjunto de entrenamiento y *testing*. Se entrenó un modelo de *random forest* con 100 árboles.

3.2.2.3. VowpalWabbit

El proyecto Vowpal Wabbit (VW) es un sistema patrocinado por Microsoft Research y previamente por Yahoo! Research. El proyecto es sobre la creación de un algoritmo intrínsecamente rápido. Hay varios algoritmos disponibles con una línea base constituida por stochastic gradient descent [Langford et al., 2007b]. La única dependencia externa es la librería boost, que generalmente ya se encuentra instalada por defecto.

Las características del paquete son:

- Aprendizaje online y optimización por defecto.
- *Feature Hashing*: Permite aprender de representaciones sin procesar, reduciendo la necesidad de pre-procesamiento, acelerando la ejecución y en algunos casos mejorando la exactitud.
- Pila de reducción, permite que el algoritmo alcance problemas avanzados como clasificación de múltiples clases, sensitiva al costo.

La ventaja que tiene VW es que no necesita preprocesamiento de los datos. La estructura de los datos de entrada permite la combinación de variables numéricas y categóricas.

Para armar el conjunto de datos, se utilizaron como entrada los resultados del modelo de tópicos para $k = 20$, y el **corpus de lemas**. También se probó con el **corpus de sustantivos**, el cual arrojó en entrenamiento resultados muy pobres (menores al 30 %), por lo que se desestimó su uso para los experimentos finales.

El funcionamiento del modelo en entrenamiento fue mejor con los lemas solos, sin los porcentajes de participación de los tópicos. Se decidió dejar el modelo sólo con los lemas.

Inicialmente se tomaron los datos del corpus **sin balancear por décadas** (tabla 3.5), pero como en el entrenamiento no reconocía las tres primeras ventanas, se agruparon las décadas del 60 y del 70 en una sólo ventana, así como el período 2000-2014. Este modelo se definió con cuatro ventanas temporales, a diferencia del resto (ventana 1: 1960-1970, ventana 2: 1980, ventana 3: 1990, ventana 4: 2000-2014).

Se implementaron tres modelos, a los que luego se puso a votar en un ensemble :

- One against all (Oaa): El algoritmo reduce internamente el problema de k múltiples clases en k problemas binarios separados. Cada clasificación binaria aprende si un ejemplo es de esa clase o no. Se implementó un costo por error en cada clase.
- *Neural network reduction (Nn)*: Implementa una red neuronal *feedforward* con una única capa oculta.
- *Error correcting tournament (Ect)*: Representan las clases como un torneo de eliminación donde los pares de clases compiten entre ellas para ser la clase del ejemplo.

CAPÍTULO 4

Resultados

En el [Capítulo 3](#) se presentó el armado de un corpus de 30.969 letras del rock y del folklore argentinos escritas durante el período 1960-2014. Para analizar la existencia de una relación entre los períodos históricos y la letra de las canciones, se realizaron experimentos exploratorios y de clasificación. A continuación se analizan los resultados de estos de experimentos.

4.1. Experimentos Exploratorios

Los experimentos exploratorios consistieron en la comparación temporal de la evolución de tópicos entre el corpus dividido en ventanas de tiempo de 5 años versus el corpus sin dividir. El objetivo de comparar el corpus dividido en ventanas temporales versus el corpus entero fue determinar que técnica es más efectiva para detectar la existencia de tópicos emergentes, estables y en decadencia.

4.1.1. Ventanas de tiempo

En la [Subsubsección 3.2.1.1](#) se dividió el corpus en ventanas de 5 años. Sobre estas ventanas se experimentó con matrices de similitud y factorización matricial dinámica no negativa.

4.1.1.1. Matriz de similitud

Se muestran las matrices de similitud resultantes de la división del corpus en ventanas de tiempo, en aquellos casos donde aparecieron tópicos emergentes. Las dimensiones

de las ventanas son diferentes, debido a que para cada ventana, se eligió el número óptimo de tópicos (ver detalle en la [Subsubsección 3.2.1.1](#)).

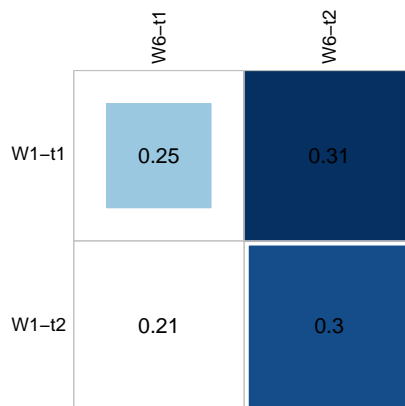


Figura 4.1: Matriz de similitud entre ventana 1 (1960-1964) y ventana 6 (1965-1969)

Las palabras propias del tópico 1 de la ventana 6 (1965-1969) no presentes en el tópico 2 de la ventana 1 (1960-1964) (divergencia = 0,21) están relacionadas con temas de María Elena Walsh.

Tópico estable: en esta ventana se muestra la poesía con humor, de inspiración surrealista [Brizuela, 2008]. Tópico levemente emergente (tópico 2 - ventana 6): Los discos con mayor participación de este tópico fueron Miguel Abuelo simple, Que pena me das simple (Manal) y No pibe Simple (Manal). Manal es uno de los grupos fundacionales del rock argentino.

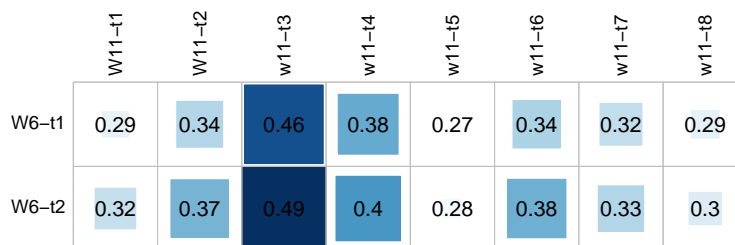


Figura 4.2: Matriz de similitud entre ventana 6 (1965-1969) y ventana 11 (1970-1974)

El tópico más estable es el tópico 5.

Tópico estable: En la figura 4.1 se muestran los discos con mayor porcentaje de participación de este tópico. Moris, Tanguito, Los Gatos, Almendra y Los Abuelos

de la Nada son las figuras descollantes para la historia del tránsito entre los 60 y 70 [Polimeni, 2002]. Polimeni llama a esta etapa la edad de la inocencia.

Tópico emergente (tópico 3 - ventana 11): Los discos con mayor proporción de este tópico son Candiles (Aquelarre) y Primavera para un valle de lágrimas (Roque Narvaja). Según las críticas de la época, Aquelarre se fue consolidando como una de las opciones más creativas y originales del rock argentino. Sus letras surrealistas no encajaban con los comunes denominadores de la época.

Cuadro 4.1: Discos con mayor proporción ventana 11 (1970-1974) del tópico 5

| # | disco | Tópico 5 |
|----|---|----------|
| 1 | hermano perro simple (Almendra) | 0.9886 |
| 2 | ciudad de guitarras callejeras (Moris) | 0.9863 |
| 3 | blues de dana (Arco Iris) | 0.9849 |
| 4 | david lebon (David Lebón) | 0.9844 |
| 5 | la historia de los chalchaleros disco 2 | 0.9825 |
| 6 | Tontos (Billy Bond) | 0.9602 |
| 7 | tiempo de resurrección (Arco Iris) | 0.7177 |
| 8 | Inéditos (Los Gatos) | 0.5378 |
| 9 | miguel abuelo simple (Miguel Abuelo) | 0.4955 |
| 10 | zitarrosa 74 edicion uruguaya (Zitarrosa) | 0.4943 |

| | w16-t1 | w16-t2 | w16-t3 | w16-t4 |
|--------|--------|--------|--------|--------|
| w11-t1 | 0.29 | 0.34 | 0.36 | 0.33 |
| w11-t2 | 0.36 | 0.43 | 0.44 | 0.38 |
| w11-t3 | 0.43 | 0.43 | 0.45 | 0.47 |
| w11-t4 | 0.41 | 0.45 | 0.47 | 0.4 |
| w11-t5 | 0.26 | 0.31 | 0.32 | 0.28 |
| w11-t6 | 0.35 | 0.39 | 0.4 | 0.37 |
| w11-t7 | 0.33 | 0.39 | 0.36 | 0.38 |
| w11-t8 | 0.34 | 0.37 | 0.36 | 0.37 |

Figura 4.3: Matriz de similitud entre ventana 11 (1970-1974) y ventana 16 (1975-1979)

El t3pico 1 es el m3s estable con respecto a la ventana 11 (1970-1974).

T3pico estable: en esta 3poca nace el rock progresivo neo sinf3nico, del cual Esp3ritu es el primer exponente. Pescado Rabioso y Ricardo Soule tambi3n forman parte de este subg3nero. Este surgimiento est3 directamente relacionado con los a3os de plomo. En la tabla 4.2 se enumeran los discos con mayor participaci3n de este t3pico. T3pico levemente emergente (t3pico 3): Los discos m3s representativos son: Polifemo (Polifemo), Guitarra Negra (Alfredo Zitarrosa) y Desde Tacuaremb3 (Alfredo Zitarrosa).

Cuadro 4.2: Discos con mayor proporci3n ventana 16 (1975-1979) del t3pico 1 “haber tener ver si querer cantar poder”

| # | disco | T3pico 1 |
|---|----------------------------------|----------|
| 1 | los chalchalers con alain debray | 0,9760 |
| 2 | vuelta a casa (Ricardo Soule) | 0,9534 |
| 3 | libre y natural (Esp3ritu) | 0,9439 |
| 4 | lo mejor de pescado rabioso | 0.9344 |
| 5 | Mundo (Ra3l Porchetto) | 0,6796 |

Los t3picos de la ventana 16 son distintos entre s3, a comparaci3n de las representaciones de otras ventanas. Esto se puede observar en la figura 4.4 donde se representan los t3picos en sus dos primeros componentes principales.

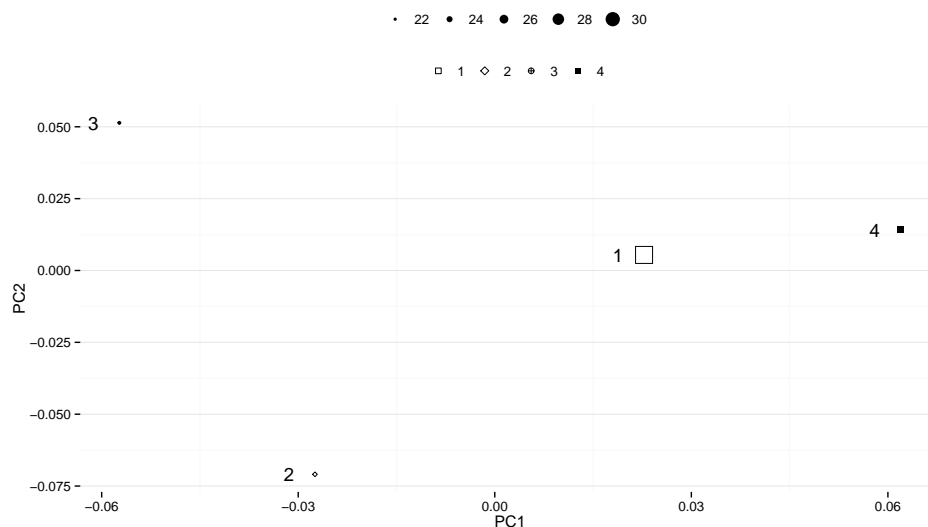


Figura 4.4: Representaci3n de t3picos en dos primeros componentes principales de ventana 16 (1975-1979)

Las canciones con mayor porcentaje de participación para el tópico 3 son folklóricas. Este tópico se presenta como el más divergente con respecto a la ventana 11, y es el que tiene mayor separación con respecto al resto de los tópicos en la ventana 16.

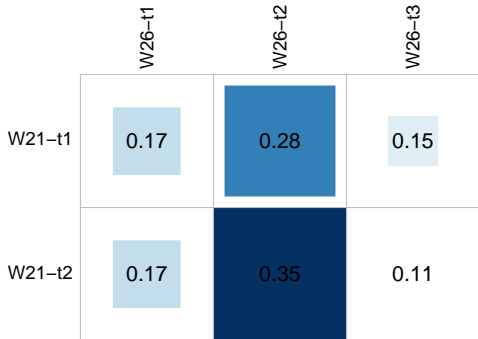


Figura 4.5: Matriz de similitud entre ventana 21 (1980-1984) y ventana 26 (1985-1989)

El tópico 3 es el que tiene menor divergencia de los tres tópicos de la ventana 26 con respecto a la ventana 21. El tópico 2 es levemente divergente en comparación con el resto (discos con mayor participación: Plan diabólico (Pappo), Fiebre (Sumo), El ritual de la banana (Los Pericos)).

A partir de la vuelta de la democracia en 1983 aparece una nueva camada de músicos. Los músicos que habían sido perseguidos y censurados durante la dictadura se orientaron a una nueva rama del rock. En los fines del rock sinfónico, el new-wave y el pop invadieron el rock nacional. La sociedad vivía con alegría el retorno de la democracia, creyendo que solucionaba los problemas del país. Soda Stereo aparece como el grupo mas icónico de este estilo. Lunardelli [Lunardelli, 2002] señala como no casual el surgimiento de esta era. Según su libro, un mensaje intencional está constituido por dos partes, la información propiamente dicha y un soporte visual. La autora afirma que se privilegiaron esos Raros peinados nuevos, para brindar al público Música Ligera. Charly García hace referencia al fenómeno en su disco Piano Bar, donde se habla de las canciones de amor, y los raros peinados nuevos, que ya no quiere criticar, sino ser enfermero. Los Soda en Canción Animal hablan de aquel amor de música ligera, de lo que ya nada les quedaba.

Los discos con mayor participación de este tópico se detallan en la tabla 4.3. Esta época refleja un estilo musical independiente de los debates ideológicos.

Cuadro 4.3: Ventana 26 discos con mayor proporción tópico 3

| | Disco | Tópico 3 |
|----|--------------------------------------|----------|
| 1 | ruido blanco (Soda Stereo) | 0,9950 |
| 2 | desnudita es mejor (Divina Gloria) | 0,9922 |
| 3 | maxi anfitreu (Pericos) | 0,9916 |
| 4 | Vii (Riff) | 0,9906 |
| 5 | riff n roll (Riff) | 0,9885 |
| 6 | Mami (Juan C. Baglietto) | 0,9873 |
| 7 | en el opera (Los abuelos de la Nada) | 0,9870 |
| 8 | travesia del alma (Julia Zenko) | 0,9858 |
| 9 | Demo (Le fou) | 0,9854 |
| 10 | barrios bajos (Raúl Porchetto) | 0,9844 |

| | W31-t1 | W31-t2 | W31-t3 | W31-t4 | W31-t5 |
|--------|--------|--------|--------|--------|--------|
| W26-t1 | 0.22 | 0.22 | 1.18 | 0.3 | 0.48 |
| W26-t2 | 0.29 | 0.32 | 0.69 | 0.38 | 0.47 |
| W26-t3 | 0.12 | 0.18 | 1.65 | 0.21 | 0.45 |

Figura 4.6: Matriz de similitud entre ventana 26 (1985-1989) y ventana 31 (1990-1994)

El tópico 1 es el que tiene menor divergencia de los cinco tópicos de la ventana 31 con respecto a la ventana 26. Ninguno de los tópicos de la ventana 26 se muestra en decadencia. Tienen todos alta divergencia con el tópico 3 de la ventana 31. Los temas con mayor proporción de este tópico pertenecen a Ignacio Copani, Vilma Palma e Vampiros y Los Calzones.

Este tópico marca el inicio del estilo tropical y de fusión latina en el rock argentino. Durante los 90 las medidas de apertura y desregulación produjeron un cambio social. El desempleo produjo una gran desigualdad social. Las clases bajas escuchaban cumbia, y esto se trasladó a la música de clase media y alta. "Lo que sí noto es que los países del mundo que están bien, sin problemas, escuchan blues y rock and roll. Los países que caen en desgracia escuchan cumbia. Hay que fijarse en eso..." (Pappo, en una entrevista para el diario Los Andes de Mendoza).

| | W41-t1 | W41-t2 |
|--------|--------|--------|
| W36-t1 | 0.18 | 1.4 |
| W36-t2 | 0.16 | 1.47 |
| W36-t3 | 0.29 | 1.31 |
| W36-t4 | 0.29 | 0.97 |
| W36-t5 | 0.05 | 1.51 |
| W36-t6 | 0.64 | 0.23 |

Figura 4.7: Matriz de estabilidad entre ventana 36 (1995-1999) y ventana 41 (2000-2004)

El t3pico 1 es el que tiene menos divergencia con respecto a la ventana anterior. Los temas m3s representativos de este t3pico pertenecen a Actitud Maria Marta, Sindicato Argentino del Hip Hop y a Le3n Gieco. Hacia fines de los 90 resulta evidente un crecimiento del rock suburbano, debido a la brecha creciente entre ricos y pobres. Despu3s de 1990 surge la cultura del Rap o Hip Hop en la Argentina, mucho despu3s que en otros pa3ses, debido a la no entrada de materiales importados. La segunda ola del rap surgi3 en 1998. Actitud Mar3a Marta desde sus inicios estuvo vinculado a reclamos sociales y de organismos de derechos humanos.

El t3pico 2 se presenta como un t3pico emergente. Corresponde a discos con temas en ingl3s (Highway on fire simple - Rata Blanca, por ejemplo).






En las 3ltimas ventanas no se ve el surgimiento de ning3n t3pico en especial (entre 2005 y 2014), lo cual coincide con la crisis post-Cromagnon.

4.1.1.2. Dynamic NMF

Luego que se crearon seis ventanas de t3picos temporales (teniendo en cuenta hitos hist3ricos), se combinaron los resultados de las ventanas para generar t3picos din3micos que abarcaron varias ventanas de tiempo (ver composici3n de ventanas en la tabla 3.6).

A los resultados producidos para cada ventana se les asign3 un color, donde cada color representa una tem3tica global (tabla 4.4).

Cuadro 4.4: Grupos DNMF

| Grupo | Color |
|----------------------|--|
| Naturaleza |  |
| Folklore/Campo |  |
| Sentimientos/Familia |  |
| Inglés |  |
| Rock |  |

Se reproducen los resultados para cada ventana:

Cuadro 4.5: Ventana 1 (1960-1974) Tópicos NMF

| Rank | w1_01 | w1_02 | w1_03 | w1_04 | w1_05 | w1_06 | w1_07 |
|------|---------|---------|-----------|----------|-----------|---------|-----------|
| 1 | río | amor | vida | noche | pa | niño | carnaval |
| 2 | sol | dolor | alma | luna | pago | navidad | copla |
| 3 | tierra | mujer | pena | guitarra | vino | calle | chacarero |
| 4 | viento | calle | día | zamba | cosa | mundo | vidala |
| 5 | flor | sol | ojo | copla | amigo | cuna | flor |
| 6 | agua | voz | tiempo | canto | don | madre | diablo |
| 7 | sangre | día | camino | vino | campo | pan | bombo |
| 8 | canción | corazón | corazón | monte | mama | hombre | albahaca |
| 9 | sombra | flor | recuerdo | estrella | chacarero | lugar | pena |
| 10 | mar | boca | esperanza | camino | rancho | luna | vino |

Cuadro 4.6: Ventana 2 (1975-1982) Tópicos NMF

| Rank | w2_01 | w2_02 | w2_03 | w2_04 | w2_05 | w2_06 | w2_07 |
|------|----------|--------|-----------|----------|----------|----------|--------|
| 1 | amor | nene | vida | noche | tiempo | vez | sol |
| 2 | corazón | dios | muerte | tierra | cosa | cosa | luz |
| 3 | dolor | casa | día | guitarra | piedra | hombre | mar |
| 4 | flor | calor | madre | pa | hora | amigo | ojo |
| 5 | tristeza | madre | ser | pena | hombre | día | niño |
| 6 | cielo | abismo | esperanza | canto | verdad | mundo | cuerpo |
| 7 | sueño | gana | mentira | copla | mano | casa | viento |
| 8 | niño | sombra | camino | corazón | amigo | historia | mañana |
| 9 | alma | noche | beso | alma | recuerdo | lugar | sueño |
| 10 | adiós | día | tierra | zamba | dios | gente | color |

Cuadro 4.7: Ventana 3 (1983-1990) Tópicos NMF

| Rank | w3_01 | w3_02 | w3_03 | w3_04 | w3_05 | w3_06 | w3_07 |
|------|---------|---------|---------|----------|-------|----------|---------|
| 1 | corazón | amor | vez | tiempo | you | vida | día |
| 2 | sol | corazón | cosa | lugar | and | mundo | noche |
| 3 | noche | cuerpo | lugar | libertad | the | gente | mañana |
| 4 | luz | canción | verdad | voz | my | cosa | casa |
| 5 | canción | dolor | final | momento | it | nene | sol |
| 6 | alma | mundo | razón | ojo | of | ciudad | suerte |
| 7 | cielo | ilusión | gana | viento | dont | libertad | boca |
| 8 | luna | sol | momento | amigo | im | calle | beso |
| 9 | flor | calor | mundo | camino | your | dios | ventana |
| 10 | ojo | palabra | soledad | calle | all | hombre | alegría |

Cuadro 4.8: Ventana 4 (1991-1997) Tópicos NMF

| Rank | w4_01 | w4_02 | w4_03 | w4_04 | w4_05 | w4_06 | w4_07 |
|------|---------|---------|--------|-------|--------|--------|--------|
| 1 | amor | corazón | vez | you | vida | tiempo | nene |
| 2 | dolor | canción | tiempo | and | cosa | noche | noche |
| 3 | beso | dolor | vuelta | the | día | sol | rock |
| 4 | piel | alma | verdad | my | mundo | día | día |
| 5 | pasión | voz | lugar | of | amigo | cielo | vuelta |
| 6 | adiós | ilusión | final | it | verdad | mar | bar |
| 7 | flor | razón | cosa | your | gente | luz | blues |
| 8 | sol | soledad | lado | that | mujer | lugar | amigo |
| 9 | razón | viento | suerte | dont | alma | sueño | mamá |
| 10 | canción | sueño | perdón | im | sos | ojo | perro |

Aparece por primera vez un tópico que relaciona el rock con bar, noche, blues, amigo. A mediados de los 90 el nuevo rock argentino crecía de manera lenta. En el año 94 se publica el disco Valentín Alsina, y se considera como el hito fundacional del rock barrial. La temática del rock barrial gira en torno a las peleas con la policía, el alcohol, y las drogas.

Cuadro 4.9: Ventana 5 (1998-2003) Tópicos NMF

| Rank | w5_01 | w5_02 | w5_03 | w5_04 | w5_05 | w5_06 | w5_07 |
|------|---------|---------|---------|-------|----------|----------|---------|
| 1 | amor | sol | vez | you | noche | vida | tiempo |
| 2 | corazón | corazón | cosa | and | nene | día | lugar |
| 3 | dolor | sueño | ve | the | rock | gente | cosa |
| 4 | beso | cielo | lugar | my | luna | cosa | momento |
| 5 | canción | luz | pie | your | día | mundo | verdad |
| 6 | pasión | alma | mas | it | estrella | verdad | final |
| 7 | flor | ojo | final | of | amigo | mentira | nene |
| 8 | mujer | mar | lado | that | chico | amigo | mas |
| 9 | adiós | voz | palabra | be | mañana | historia | vuelta |
| 10 | calor | dolor | piel | dont | calle | salida | camino |

Cuadro 4.10: Ventana 6 (2004-2014) Tópicos NMF

| Rank | w6_01 | w6_02 | w6_03 | w6_04 | w6_05 | w6_06 | w6_07 |
|------|--------|---------|----------|-------|----------|-------|---------|
| 1 | vida | amor | vez | you | tiempo | sol | corazón |
| 2 | día | pasión | final | and | lugar | noche | razón |
| 3 | cosa | error | pie | the | camino | luz | canción |
| 4 | dolor | piel | ve | my | miedo | cielo | dolor |
| 5 | noche | calor | verdad | your | verdad | mar | ilusión |
| 6 | dios | beso | historia | of | recuerdo | voz | alma |
| 7 | lado | dolor | lugar | it | momento | ojo | beso |
| 8 | mundo | razón | dolor | that | cosa | sueño | pena |
| 9 | camino | nene | mundo | all | silencio | lugar | ritmo |
| 10 | amigo | ilusión | cosa | be | sueño | mundo | pasión |

En esta ventana desaparece el rock barrial, y siguen los temas que se mantuvieron en todas las ventanas.

El resultado de la combinación de los resultados de las ventanas para generar tópicos dinámicos que abarcan varias ventanas de tiempo es el de la tabla 4.11.

Cuadro 4.11: Tópicos Dinámicos para $k=7$

| Rank | TD01 | TD02 | TD03 | TD04 | TD05 | TD06 | TD07 |
|------|---------|---------|----------|--------|------|----------|--------|
| 1 | corazón | amor | vez | vida | you | tiempo | nene |
| 2 | sol | corazón | cosa | día | and | lugar | noche |
| 3 | noche | dolor | lugar | cosa | the | cosa | casa |
| 4 | luz | beso | verdad | mundo | my | verdad | rock |
| 5 | cielo | canción | final | gente | it | momento | cosa |
| 6 | sueño | flor | amigo | muerte | of | camino | calle |
| 7 | viento | mujer | mundo | amigo | your | recuerdo | ciudad |
| 8 | ojo | ilusión | historia | camino | that | hora | dios |
| 9 | tierra | pasión | gana | alma | all | miedo | chico |
| 10 | luna | calor | casa | verdad | dont | viento | mujer |

Los tópicos que permanecen en el tiempo son los relacionados a los sentimientos, el rock identificado con la noche, la ciudad, la calle y los términos extranjeros.

El resultado del cálculo de la medida de coherencia para comparar diferentes modelos de tópicos y elegir un modelo con una cantidad de tópicos adecuado es el que se reproduce en la tabla 4.12.

Cuadro 4.12: Medidas de coherencia para el modelado dinámico (para k entre 4 y 10)

| k | coherence |
|----|-----------|
| 4 | 0,4210 |
| 5 | 0,3197 |
| 6 | 0,3582 |
| 7 | 0,3570 |
| 8 | 0,3394 |
| 9 | 0,3250 |
| 10 | 0,3240 |

El mejor valor de coherencia fue para $k = 6$. En la tabla 4.13 permanecen los tópicos relacionados con los sentimientos, términos en inglés, rock y naturaleza.

Cuadro 4.13: Tópicos dinámicos para $k=6$

| Rank | TD01 | TD02 | TD03 | TD04 | TD05 | TD06 |
|------|---------|---------|------|--------|--------|---------|
| 1 | sol | amor | you | vez | nene | tiempo |
| 2 | noche | corazón | and | cosa | noche | vida |
| 3 | corazón | dolor | the | lugar | casa | cosa |
| 4 | luz | vida | my | final | ciudad | día |
| 5 | cielo | beso | it | verdad | juego | lugar |
| 6 | ojo | flor | of | pie | mujer | verdad |
| 7 | sueño | canción | your | ve | cosa | mundo |
| 8 | alma | ilusión | that | vuelta | mamá | camino |
| 9 | luna | pasión | all | lado | calle | amigo |
| 10 | viento | mujer | dont | mas | chico | momento |

Si bien la medida de coherencia es levemente mejor para $k = 6$, de la observación de los resultados, surge $k = 7$ como una mejor representación de los tópicos que permanecen en el tiempo. El tópico 4 para $k = 6$ no agrega valor, mientras que los tópicos para $k = 7$ tienen más sentido.

En ambos casos, se observa un tópico de naturaleza, uno de rock y el tópico en inglés. La única diferencia son los tópicos relacionados con sentimiento y familia.

4.1.2. Corpus sin separar ventanas

4.1.2.1. Elección número de tópicos

En la [Subsubsección 3.2.1.2](#) teniendo en cuenta las medidas de Mallet (coherencia, distancia de corpus, número de palabras, por ejemplo), se decidió que el número de tópicos que mejor representaba el corpus era entre 10 y 50 tópicos. Para elegir el número óptimo de tópicos se graficaron los tópicos en ejes temporales para $k = 10$, $k = 20$ y $k = 50$. En base a los resultados obtenidos, se decidió que la mejor representación temporal estaba dada por $k = 20$. Para $k = 10$ no se observaron cambios de la participación de los tópicos en el tiempo, y para $k = 50$ se observaron pocos tópicos que representen un cambio temporal con un porcentaje de participación relevante.

Para relacionar los tópicos para $k = 20$, se agruparon jerárquicamente (figura [4.8](#)). Los tópicos que presentan menor distancia entre ellos, tienen parecido comportamiento temporal (figuras [4.9](#) y [4.10](#)).

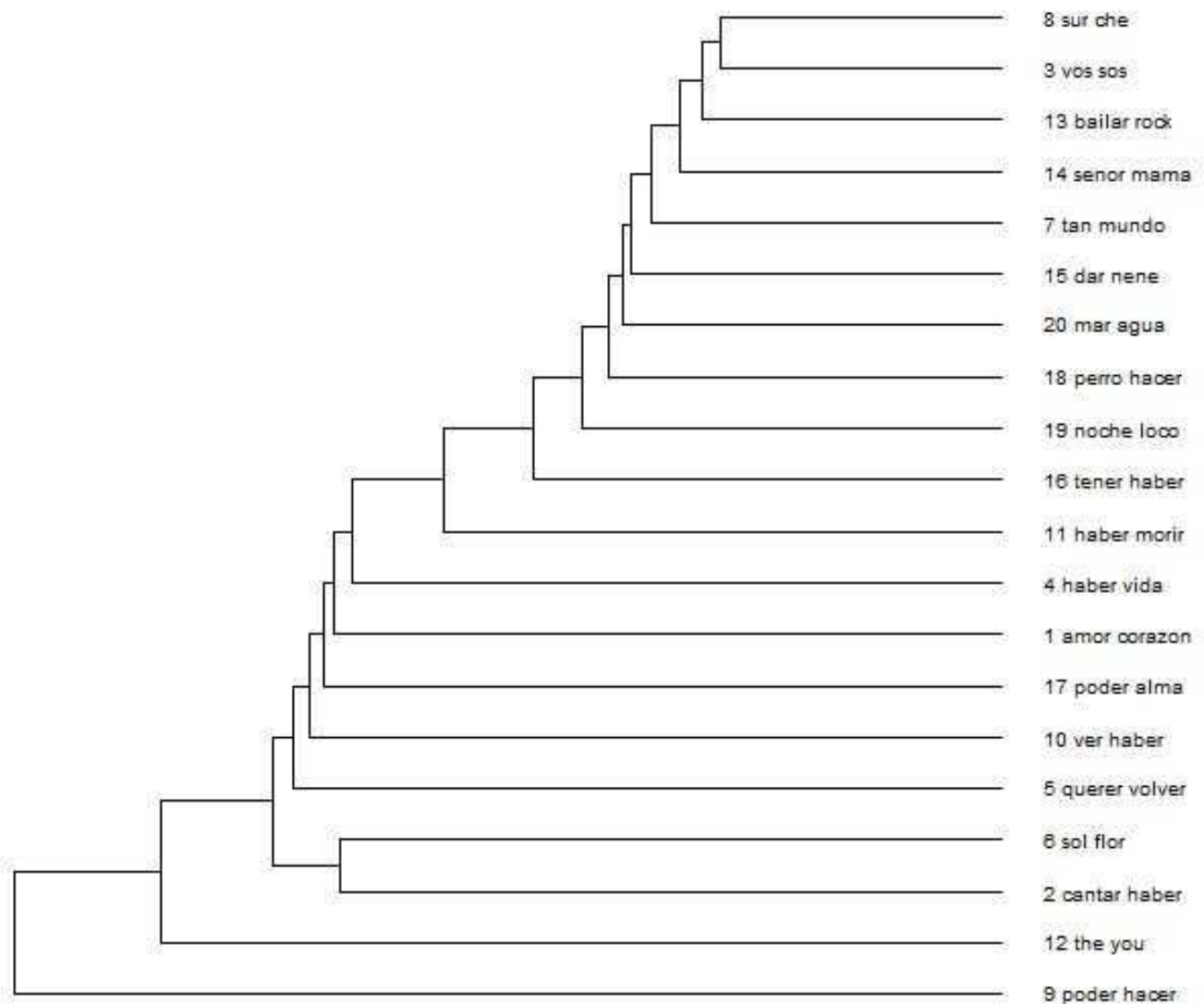


Figura 4.8: Agrupamiento jerárquico de tópicos para $k=20$

Los tópicos 3 (vos sos) y 8 (sur che) comparten la caída en los años 80, y el crecimiento en los 90. Se analizaron otros tópicos que tenían distancia mínima, como el 6 (sol flor) y el 2 (cantar haber), y también presentaron un comportamiento temporal similar (alta participación hasta los 80, y luego decadencia). Los tópicos 6 y 2 tienen mucha distancia con los mencionados antes (3 y 8), y su comportamiento temporal es inverso al que describen las figuras 4.9 y 4.10 para los tópicos 3 y 8.

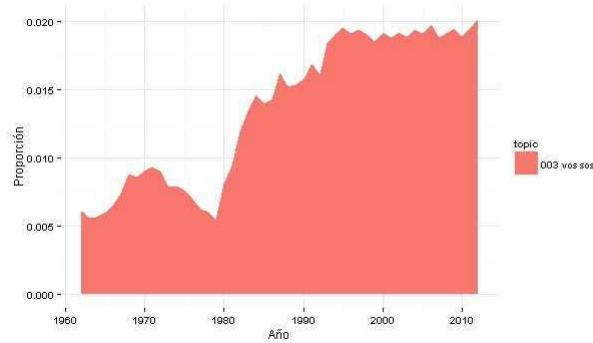


Figura 4.9: Tópico 3 vos sos

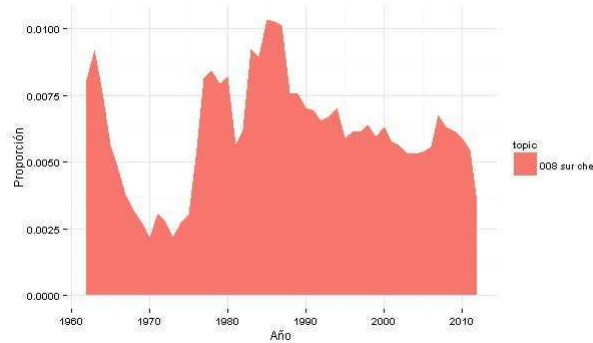


Figura 4.10: Tópico 8 sur che

4.1.2.2. Uso de las palabras a través del tiempo

Se dividió el tiempo en cuartiles, en función de la fecha, y se contaron las frecuencias en cada intervalo temporal. Luego se calculó la media de la medida Dunning log-likelihood. Esto permite ver los tópicos no como una única lista ordenada, sino como cuatro listas con un orden distinto dependiendo del período. Se seleccionaron los términos más relevantes por tópicos, y se graficaron en cada época, mostrando la diferencia de posición en el tópico según el período. Se comparó el resultado entre un tópico elegido al azar (figura 4.11) versus un mal tópico (figura 4.12). Se define un tópico como malo cuando su correlación entre las palabras y los años es muy baja.

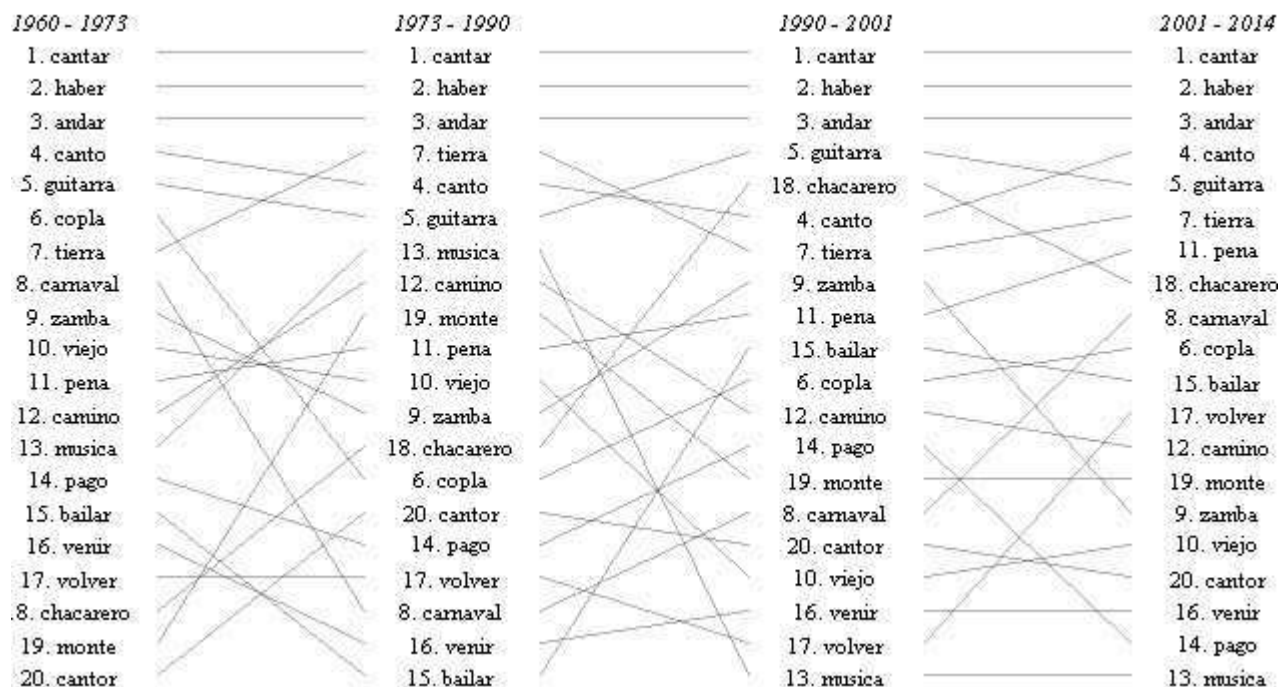


Figura 4.11: Evolución de palabras para un tópico elegido al azar (Tópico cantar haber)

En el primer cuartil (1960-1983), las 5 primeras palabras son cantar, haber, andar, canto. En el segundo cuartil (1973-1990) se mantienen las tres primeras, y se reemplaza canto por tierra, y guitarra por canto. Una palabra que ascendió de los últimos puestos fue monte, del puesto 19 a 9. En el tercer cuartil (1990-2001), suben guitarra, chacarero y bailar, mientras que descienden tierra, camino, monte. En el último cuartil, se mantienen las tres primeras palabras, baja chacarero y zamba, y sube carnaval. En este tópico hay palabras que pertenecen al mismo dominio como música, movimiento o viaje (andar, volver, venir), campo (pago, tierra, monte). Es decir la temática sigue siendo constante. Si bien cambia el orden de las palabras, siguen teniendo importancia palabras del mismo dominio.

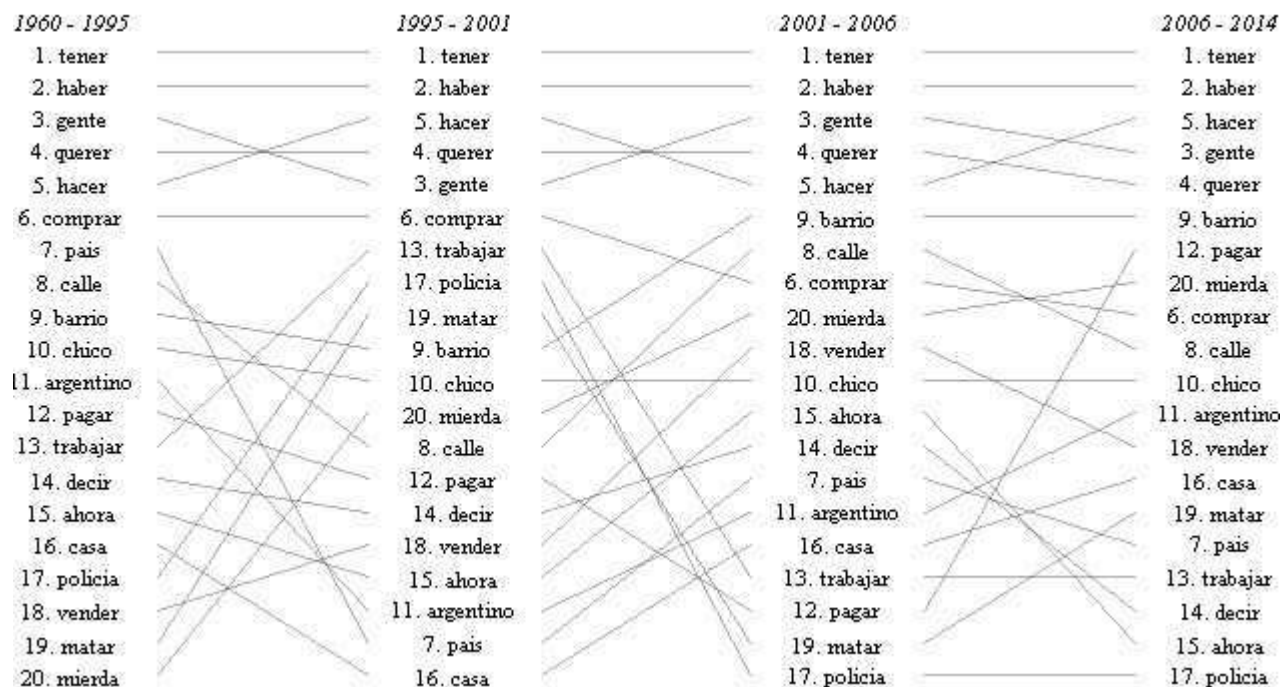


Figura 4.12: Tópico tener haber (peor correlación entre palabras y años)

Es interesante las palabras que ascienden en el período 1995-2001. El rock barrial surgió a comienzos del 90, también conocido como rock chabón. En la década del 80 el mundo del rock había sufrido cambios importantes, mientras en los 70 el movimiento estaba comprometido con la situación del país, en los 80 se fragmentó y despolitizó.

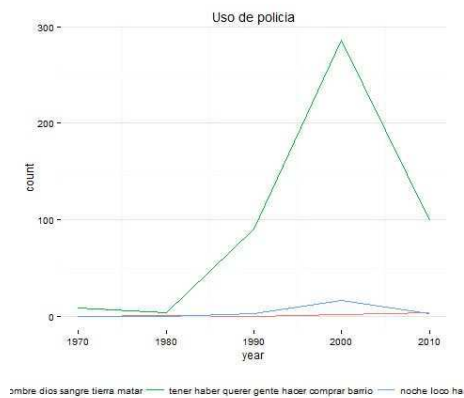
El rock barrial surgió como alternativa para jóvenes de clases populares y sectores medios empobrecidos. Sus letras hablaban con nostalgia del mundo de la infancia, se afirmaban en la ética de la fidelidad al barrio, lamentaban el fin del mundo del trabajo, y exponían críticamente el presente de pobreza y corrupción [Adamovsky, 2012]. Esto se refleja en la figura 4.12, con el ascenso de palabras como trabajar, policia, matar.

En el [Apéndice B](#) se reproducen los gráficos pertenecientes a otros tópicos.

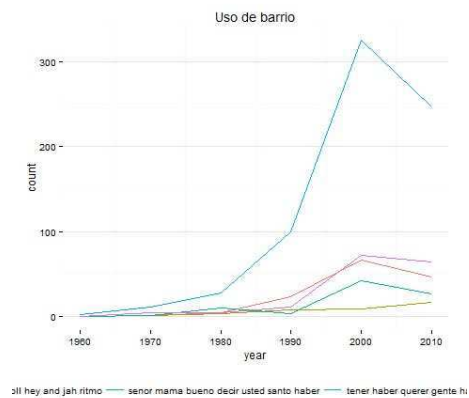
Se eligieron tres términos típicos del rock barrial para ver su evolución en el tiempo (figura 4.13).

4.1.2.3. Hot y Cold Topics

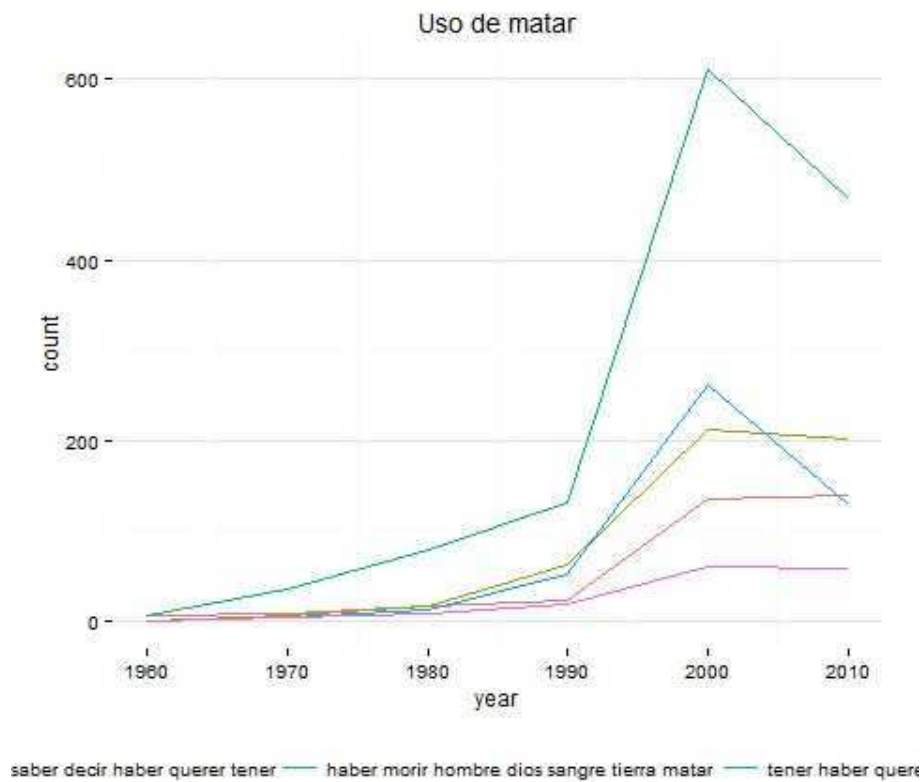
En la figura 4.14 se muestran los tópicos que crecen a medida que pasa el tiempo, por ejemplo, el tópico 1 (amor corazón). En la figura 4.14 se muestran las canciones con mayor proporción y en la figura 4.15 el gráfico de la evolución temporal, donde se ve la tendencia creciente, una caída en los 80, durante la dictadura, y luego un leve



(a) Policía



(b) Barrio



(c) Matar

Figura 4.13: Evolución temporal de palabras

amesetamiento en el 2000.

```
$top.words
[1] "amor corazon haber querer amar solo olvidar saber vida llorar tener decir morir dolor poder
hacer dejar quedar vivir nunca "
```

```
$alpha
[1] 0.21594
```

```
$top.articles
```

| | id | titulo | artista | fecha | disco | topic1 |
|-------|-------|----------------------|-----------------|-------|------------------------------|-----------|
| 4134 | 13718 | imillitay | los tekis | 2010 | mixtura | 0.9405650 |
| 26310 | 5804 | cariñito | los tekis | 2007 | mamapacha | 0.9250753 |
| 20855 | 28768 | deseo | juan terrenal | 2004 | nuestra forma | 0.9066009 |
| 3779 | 13399 | perfidia | la portuaria | 2002 | hasta despertar | 0.8976063 |
| 14833 | 23347 | bajo la rambla | andres calamaro | 2001 | duetos | 0.8929147 |
| 14640 | 23173 | jamás | los chachaleros | 1967 | los chachaleros por el mundo | 0.8789131 |
| 162 | 10142 | la resentida | mercedes sosa | 1983 | como un pajarito libre | 0.8677591 |
| 5945 | 15348 | no quisiera perderte | luciano pereyra | 2002 | soy tuyo | 0.8644656 |
| 723 | 10648 | lo olvidaras | los gatos | 1967 | los gatos | 0.8513091 |

Figura 4.14: Tópico 1 amor corazón - Hot Topic

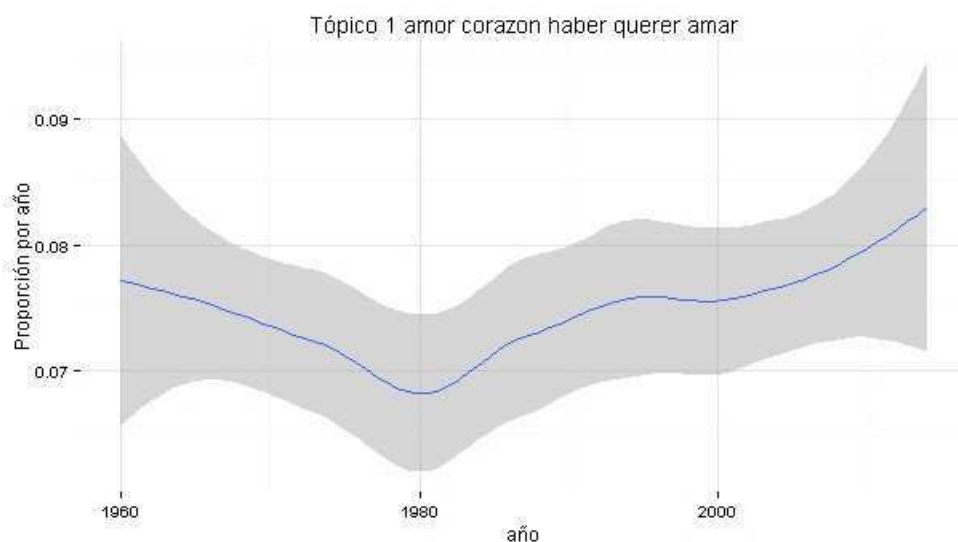


Figura 4.15: Evolución temporal tópico 1 amor corazón

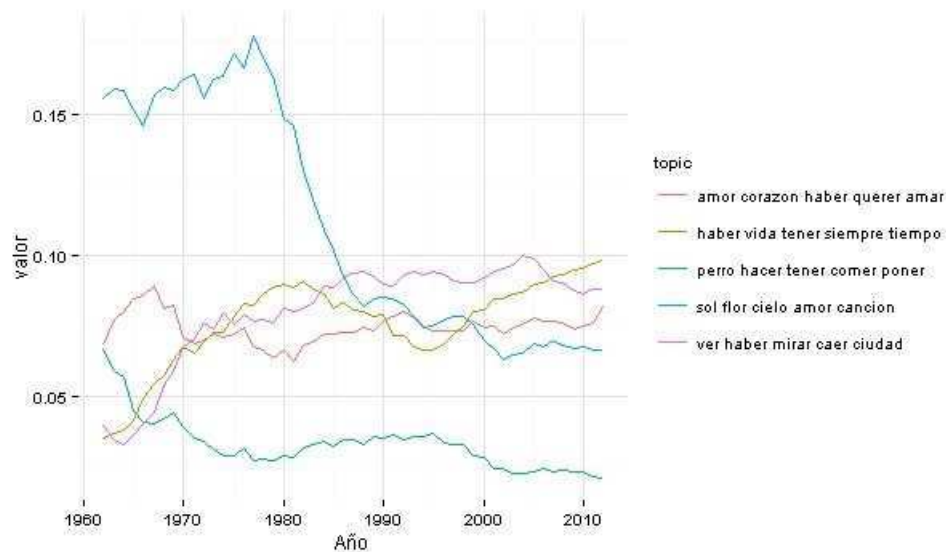


Figura 4.16: 5 tópicos más correlacionados negativamente (Cold Topics) con la fecha - p value = 0,05 para $k = 20$

Se reproducen los principales temas que tienen tendencia a la baja, como el tópico 19 noche loco (figura 4.17)

```
$top.words
[1] "noche loco hacer gustar tomar decir amigo chico tener mujer bailar bueno bar salir bien vino
fiesta pasar venir dormir "
```

```
$alpha
[1] 0.1061
```

```
$top.articles
```

| id | titulo | artista | fecha | disco |
|-------|------------------------------|---------------------|-------|----------------------------|
| 419 | 10374 todos van al news cafe | sui generis | 2000 | sinfonia para adolescentes |
| 19340 | 27403 baile frotado | babasonicos | 2012 | carolo |
| 21046 | 2894 bolero | sueter | 1995 | sueter 5 |
| 23637 | 3399 traka traka | el otro yo | 1996 | los hijos de alien |
| 18248 | 26420 fondo blando | rockeros y borregos | 2007 | rockeros y borregos |
| 24168 | 3877 estos pies peligrosos | gorriones | 1993 | peligrosos gorrones |
| 12880 | 2159 un vinito mas | bulldog | 1998 | el angel de la muerte |

```

      topic19
419  0.9424387
19340 0.8972423
21046 0.8336700
23637 0.8259218
18248 0.8233984
24168 0.8130956
12880 0.8112128

```

Figura 4.17: Tópico 19 noche loco - Cold Topic

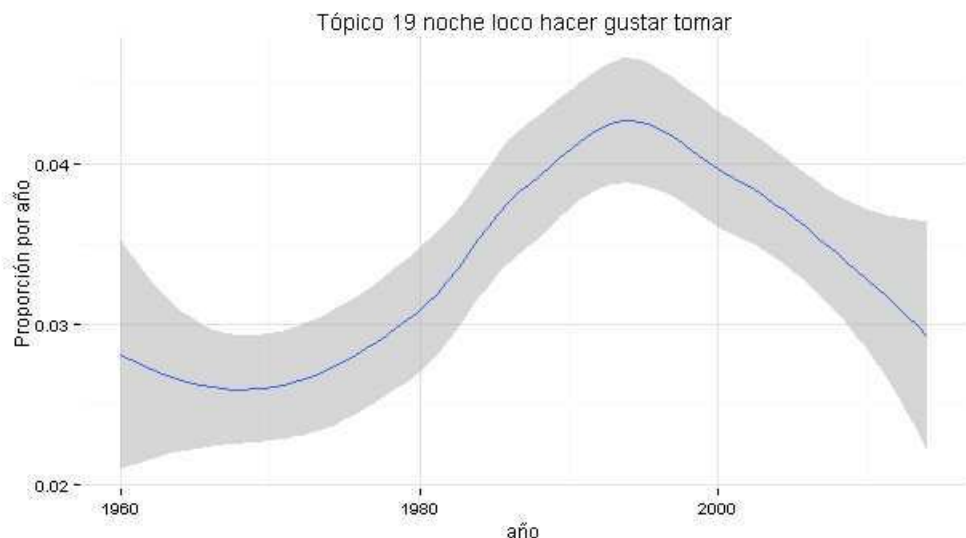


Figura 4.18: Evolución temporal tópico 19 noche loco

4.2. Clasificación

4.2.1. Métricas de evaluación

Con los **corpus de lemas y sustantivos** se ejecutaron experimentos de clasificación con distintos algoritmos y parámetros.

En Mallet se entrenaron los **corpus de lemas y sustantivos** con los algoritmos Naive Bayes y Máxima entropía, con el corpus balanceado y sin balancear, y dividido por distintas ventanas temporales (ver detalle en la [Subsubsección 3.2.2.1](#)).

Para Word2vec cada documento se representó como un vector, cada uno con un número de **features** igual al número de grupos generado. Se creó un **bag of centroids** para los conjuntos de entrenamiento y testing, y se entrenó un modelo **random forest** con 100 árboles (ver detalle en la [Subsubsección 3.2.2.2](#)).

Para VowpalWabbit se implementaron tres modelos: **One against all**, **Neural network reduction** y **Error correcting tournament**. Se pusieron a votar estos modelos con un ensemble (ver detalle en la [Subsubsección 3.2.2.3](#)).

Se exponen las medidas de accuracy (exactitud) y f1 para los conjuntos de entrenamiento (tabla [4.14](#)) y testing (tabla [4.15](#)).

Sea **accuracy** la medida que expresa la cantidad de casos correctamente clasificados sobre el total, y **F1** (F-Score o medida-F) la medida de precisión que tiene un test, ponderando la precisión y la exhaustividad, la medida f1 se calculó con la fórmula **Macro-averaged F-Measure** detallada en [[Özgür et al., 2005](#)].

Cuadro 4.14: Accuracy y F1 para modelos de aprendizaje para el conjunto de entrenamiento. NB: Naive Bayes, MaxEnt: Máxima entropía, CV: Validación cruzada, TR/TEST: dividido en training y testing, Hist: Histórico, Vw: VowpalWabbit, nn: Neural network reduction, etc: Error correcting tournament, oaa: One against all

| Modelo | Accuracy | F1 |
|--|----------|--------|
| NB-CV-Lemas-Décadas-Sin Balancear | 0,6869 | 0,5914 |
| MaxEnt-CV-Lemas-Décadas-Sin Balancear | 0,9717 | 0,9744 |
| NB-TR/TEST-Lemas-Décadas-Sin Balancear | 0,6834 | 0,5806 |
| MaxEnt-TR/TEST-Lemas-Décadas-Sin Balancear | 0,9917 | 0,9924 |
| NB-CV-Lemas-Décadas-Balanceado | 0,7729 | 0,7533 |
| MaxEnt-CV-Lemas-Décadas-Balanceado | 0,9974 | 0,9974 |
| NB-TR/TEST-Lemas-Décadas-Balanceado | 0,7729 | 0,7533 |
| MaxEnt-TR/TEST-Lemas-Décadas-Balanceado | 0,9878 | 0,9892 |
| NB-CV-Lemas-Hist-Sin Balancear | 0,6372 | 0,5112 |
| NB-CV-Sustantivos-Hist-Sin Balancear | 0,6132 | 0,4867 |
| MaxEnt-CV-Lemas-Hist-Sin Balancear | 0,9309 | 0,9286 |
| MaxEnt-CV-Sustantivos-Hist-Sin Balancear | 0,7795 | 0,7682 |
| NB-CV-Lemas-Hist-Balanceado | 0,6671 | 0,6446 |
| MaxEnt-CV-Lemas-Hist-Balanceado | 0,9791 | 0,9799 |
| Word2Vec-Lemas-Décadas-Sin Balancear | 0,9980 | 0,9954 |
| Vw-nn-Lemas-4w-Sin Balancear | 0,6271 | 0,2917 |
| Vw-etc-Lemas-4w-Sin Balancear | 0,8886 | 0,8649 |
| Vw-oaa-Lemas-4w-Sin Balancear | 0,7524 | 0,6455 |
| Ensemble vw-nn , vw-etc, vw-oaa | 0,7560 | 0,6442 |

Cuadro 4.15: Accuracy y F1 para modelos de aprendizaje para el conjunto de testing. NB: Naive Bayes, MaxEnt: Máxima entropía, CV: Validación cruzada, TR/TEST: dividido en training y testing, Hist: Histórico, Vw: VowpalWabbit, nn: Neural network reduction, etc: Error correcting tournament, oaa: One against all

| Modelo | Accuracy | F1 |
|--|----------|--------|
| NB-CV-Lemas-Décadas-Sin Balancear | 0.4642 | 0.2489 |
| MaxEnt-CV-Lemas-Décadas-Sin Balancear | 0.4156 | 0.2859 |
| NB-TR/TEST-Lemas-Décadas-Sin Balancear | 0.4674 | 0.2339 |
| MaxEnt-TR/TEST-Lemas-Décadas-Sin Balancear | 0.4065 | 0.2714 |
| NB-CV-Lemas-Décadas-Balanceado | 0.3367 | 0.2821 |
| MaxEnt-CV-Lemas-Décadas-Balanceado | 0.3285 | 0.3097 |
| NB-TR/TEST-Lemas-Décadas-Balanceado | 0.3367 | 0.2821 |
| MaxEnt-TR/TEST-Lemas-Décadas-Balanceado | 0.4006 | 0.2617 |
| NB-CV-Lemas-Hist-Sin Balancear | 0.4443 | 0.2594 |
| NB-CV-Sustantivos-Hist-Sin Balancear | 0.4327 | 0.2547 |
| MaxEnt-CV-Lemas-Hist-Sin Balancear | 0.4201 | 0.2865 |
| MaxEnt-CV-Sustantivos-Hist-Sin Balancear | 0.4013 | 0.2566 |
| NB-CV-Lemas-Hist-Balanceado | 0.3311 | 0.2922 |
| MaxEnt-CV-Lemas-Hist-Balanceado | 0.3098 | 0.3045 |
| Word2Vec-Lemas-Décadas-Sin Balancear | 0.4738 | 0.1613 |
| Vw-nn-Lemas-4w-Sin Balancear | 0.6188 | 0.2845 |
| Vw-etc-Lemas-4w-Sin Balancear | 0.5637 | 0.3930 |
| Vw-oaa-Lemas-4w-Sin Balancear | 0.6086 | 0.3954 |
| Ensemble vw-nn , vw-etc, vw-oaa | 0.5970 | 0.3693 |

En el trabajo anterior mencionado en la [Sección 1.3](#) la medida de accuracy no superó el 30 %.

En la mayoría de los casos los modelos sobreajustaron en training, y bajaron notablemente su performance en testing.

El modelo que mejor clasificó el conjunto de testing en términos de accuracy fue VowpalWabbit-Redes neuronales, con lemas, sin balancear, y cuatro clases por décadas (se agruparon las décadas del 60 y 70, así como el período 2000-2014, ver detalle en la [Subsubsección 3.2.2.3](#)). El resultado general fue de 61,88 %, pero esto se debe a las clases desbalanceadas. A continuación se muestra la respectiva matriz de confusión (tabla [4.16](#)).

Cuadro 4.16: VowpalWabbit-Redes Neuronales: Matriz de confusión para conjunto de testing

| Actual | Predicho | | | |
|--------|----------|----|---|------|
| | 1 | 2 | 3 | 4 |
| 1 | 276 | 2 | 0 | 489 |
| 2 | 146 | 15 | 4 | 865 |
| 3 | 183 | 22 | 8 | 2143 |
| 4 | 293 | 26 | 8 | 6489 |

A continuación se muestra la matriz de confusión en testing para el mejor modelo de Mallet (Naive Bayes dividido en training y testing, sin balancear, por décadas), con accuracy del 46 %. Si se incluyen las clases con un margen de error de 1 clase (por ejemplo, el caso corresponde a la ventana 2 y se clasificó como ventana 1 ó 3), el porcentaje de casos correctamente clasificados asciende al 86 % (ver tabla 4.17, las clases sombreadas en celeste son las que tienen una diferencia de uno con la clase real).

Cuadro 4.17: Naive Bayes-dividido en entrenamiento y testing-sin balancear-Décadas: Matriz de confusión para conjunto de testing

| Actual | Predicho | | | | | |
|--------|----------|----|----|-----|------|-----|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 19 | 55 | 12 | 68 | 67 | 2 |
| 2 | 8 | 80 | 20 | 76 | 213 | 4 |
| 3 | 11 | 49 | 42 | 142 | 529 | 20 |
| 4 | 13 | 41 | 41 | 397 | 1551 | 63 |
| 5 | 9 | 39 | 41 | 384 | 3753 | 111 |
| 6 | 6 | 5 | 9 | 109 | 1250 | 52 |

Los casos mal clasificados con un desvío de clase entre 4 y 5 corresponden a letras de álbumes que son recopilaciones de temas escritos anteriormente, o bien tópicos folklóricos (lo mismo que sucedió con los algoritmos Naive Bayes y Máxima Entropía).

Se seleccionaron aleatoriamente los resultados de testing de Mallet, para analizar los errores de clasificación. El modelo elegido fue **el experimento no balanceado, con períodos históricos, validación cruzada y máxima entropía**. Se separaron los casos en donde el algoritmo predijo que la letra pertenecía a la ventana 1, y pertenecía a las ventanas 4, 5 y 6 (ver matriz de confusión en tabla 4.18, las clasificaciones que se analizaron están sombreadas en celeste).

Cuadro 4.18: Máxima entropía-Validación cruzada-Lemas-Históricos: Matriz de confusión en conjunto de testing

| Actual | Predicho | | | | | |
|--------|----------|-----|-----|------|------|------|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 124 | 113 | 78 | 186 | 230 | 23 |
| 2 | 77 | 256 | 146 | 315 | 465 | 79 |
| 3 | 73 | 147 | 459 | 672 | 1100 | 201 |
| 4 | 96 | 194 | 438 | 2195 | 3611 | 558 |
| 5 | 82 | 228 | 612 | 2702 | 9112 | 1718 |
| 6 | 24 | 79 | 190 | 747 | 2774 | 865 |

Se reproducen dos casos a modo de ejemplo. En el primer caso, se trata de un disco lanzado en el año 2014, pero que es una versión de un tema de la primera ventana. En el segundo caso, el formulario de consulta en la página de SADAIC trajo un tema homónimo, pero que no corresponde a la letra de la base. El tema figura en la base con fecha 2012, pero en realidad fue editado en el año 1964.

```

      titulo      artista
7651 de tiempo adentro cristobal repetto

letra
7651 \n\t\t\tDE TIEMPO ADENTRO\n\t\t\tQuiero una copla que ruede\ncuando ya no ruede yo,\nsemilla
hermana del trigo\ndel tabaco y del arroz \nPanaderito de cardo\nque ande como sin razón \nsin qu
e ni el viento se acuerde\nde qué tallo era la flor. \nQuiero que ruede mi copla \ncomo la tierra
y el sol\nlejano ya de la mano\nque acaso los redondeó\nRecuerdo que ande penando\ncomo un olvido
de amor<U+0085>\nabrojo que nadie sepa \nni donde se le prendió.\nPolvo se hará mi guitarra,\nni
memoria<U+0085> cerrazón,\nni nombre, puede que muera,\nni copla<U+0085> puede que no. \t\t
      fecha      disco
7651 2014 tiempo y silencio

lemas
7651 <U+FEFF> de tiempo adentro querer uno copla que rodar cuando ya no rodar yo semilla hermano
de el trigo de el tabaco y de el arroz panadero de cardo que andar como sin razón sin que ni el v
iento se acordar de qué tallo ser el flor querer que rodar mi copla como el tierra y el sol lejan
o ya de el mano que acaso lo redondear recuerdo que andar penar como uno olvido de amor < U+0085
> abrojo que nadie saber ni donde se le prender polvo se hacer mi guitarra mi memoria < U+0085 >
cerrazón mi nombre poder que morir mi copla < U+0085 > poder que no character(0)

SUST
7651 tiempo, copla, semilla, hermano, trigo, tabaco, arroz, panadero, cardo, razón, viento, tallo
, flor, copla, tierra, sol, mano, recuerdo, olvido, amor, abrojo, polvo, guitarra, memoria, cerra
zón, nombre, copla
      origin length  id name cat
7651 r_6836 524 7651 7651 W6

```

Figura 4.19: Tema 7651 - De tiempo adentro

El autor es el poeta uruguayo Osiris Iris Castillo, que editó sus discos entre los años 1962 y 1974. El algoritmo clasifica esta letra como del período 1 (1960-1974), con un 75 % de probabilidad.

El tema fue escrito por Jaime Dávalos, editado en el año 1964 por Los Chalchaleros,

con el disco del mismo nombre. El algoritmo predice que fue escrito en el período 1 con un 64 % de probabilidad, lo cual es correcto.

Los casos donde $ClaseReal - ClasePredicha > 3$ se dividen en : errores de imputación de fecha en la base de origen, y letras que emulan un estilo campestre/folklórico tradicional.

En Word2Vec si se incluyen las clases con un margen de error de 1 clase el porcentaje de casos correctamente clasificados asciende al 84 % (ver tabla 4.19, las clases sombreadas en celeste son las que tienen una diferencia de uno con la clase real).

Cuadro 4.19: Word2Vec-Matriz de Confusión para conjunto de testing

| Actual | Predicho | | | | | |
|--------|----------|---|----|-----|------|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 3 | 1 | 2 | 43 | 176 | 0 |
| 2 | 3 | 7 | 6 | 44 | 273 | 8 |
| 3 | 3 | 4 | 27 | 61 | 606 | 7 |
| 4 | 11 | 8 | 14 | 245 | 1754 | 6 |
| 5 | 11 | 5 | 9 | 174 | 3387 | 4 |
| 6 | 1 | 3 | 3 | 34 | 806 | 5 |

5.1. Conclusión

En el presente trabajo se generó un corpus histórico de 30.969 letras disponibles en la web del rock y del folklore argentinos. Este corpus cuenta con información adicional como disco, intérprete, nombre del tema y fecha. Se utilizaron técnicas de minería de datos para la limpieza del corpus, eliminando letras repetidas a través de k-means, agrupamiento jerárquico y cálculo de distancias. Se implementaron estrategias exploratorias y de clasificación para encontrar patrones que permitan identificar cambios en la temática de las canciones del rock y del folklore nacional a través del tiempo.

Las técnicas exploratorias, ya sea a través de la creación de ventanas, o a través de la generación de tópicos para el conjunto del corpus sin separar ventanas, permitieron definir tópicos emergentes, y tópicos en decadencia, así como el auge o decadencia del uso de ciertas palabras, relacionado con hechos históricos. Los tópicos detectados relacionados con hechos históricos más estables en el tiempo fueron el rock neo sinfónico (Espíritu, Pescado Rabioso, Ricardo Soulé), la edad de la inocencia (Los Gatos, Almendra, Abuelos de la Nada), la era new-wave y pop (Soda Stereo, Viudas e Hijas), la fusión tropical (Vilma Palma, Calzones), el rock barrial (2 minutos), y la era post Cromagnon sin temas emergentes. Cada uno de estos estilos estuvo relacionado con hechos políticos y económicos (dictadura, democracia, crisis económica). La relación más clara entre la historia y la cultura se refleja en este trabajo en el surgimiento del rock barrial, producto de la crisis económica de los 90, que apareció en todos los experimentos exploratorios. Experimentando con el modelado dinámico (DNMF), se encontraron tópicos con cierta estabilidad en el tiempo, que era una de las hipótesis a demostrar.

Las tareas de clasificación si bien no tuvieron un buen desempeño en general, en el modelo VowpalWabbit-Redes Neuronales se alcanzó una medida de accuracy en el conjunto de testeo del 62 %. Del análisis del conjunto de testeo en Mallet, separando los casos clasificados con 5 períodos de diferencia, se detectaron documentos con el dato de origen incorrecto, ya sea porque son versiones de temas de otros períodos, o porque son temas homónimos, escritos en otro período.

En resumen, las técnicas exploratorias no solamente permitieron identificar tópicos, sino también identificar patrones temporales. Las técnicas de clasificación permitieron ejercer una función de curación para validar la fecha de la base. Luego de aplicar estas técnicas, se puede decir que existe una relación entre la historia y los cambios en los tópicos de las letras musicales en el corpus analizado.

5.2. Trabajos a futuro

Sería interesante obtener temas pertenecientes a periodos anteriores a la era digital, a través de fuentes escritas, para poder procesar esos textos y seguir ampliando la muestra. Asimismo, se podría crear un modelo de limpieza de fechas aplicando los resultados aprendidos a través de la clasificación, y volver a generar las pruebas de clasificación, con la aplicación de otros algoritmos (xgboost, gbm, por ejemplo), para analizar la importancia de cada palabra en el modelo.

Otro análisis posible es representar las relaciones entre tópicos, autores y palabras como redes, y poder estructurar la historia del rock y del folklore argentinos en distintas topologías. Este análisis se puede ampliar a cuentos, novelas, textos en periódicos, para tener una base más representativa de la cultura nacional.

APÉNDICE A

Lista de Stopwords

de a un al le esta también desde les ante algunos otra quienes estar mis ellas mía
tuyas nuestra vuestras estamos estéis estaréis estarían estuve estuviera estuvieses esta-
da ha hayamos habremos habríais habrían hubieron hubiese habido es seamos seremos
seríais eran fueron fuese sido tienen tendré tendría tenías tuvo tuviéramos tuvieseis
tenidas la los para lo ya entre me todo ni ellos qué él nada estas tú nosotras míos
suyo nuestros esos estáis estén estarán estaba estuviste estuvieras estuviésemos estados
hemos hayáis habréis habrían hube hubiera hubieses habida somos seáis seréis serían
fui fuera fueses tengo tenga tendrás tendrías teníamos tuvimos tuvierais tuviesen tened
que del con como o cuando hasta nos contra e unos tanto muchos algunas te vosotros
mías suya nuestras esas están estaré estaría estabas estuvo estuviéramos estuvieseis es-
tadas habéis hayan habrán había hubiste hubieras hubiésemos habidos sois sean serán
era fuiste fueras fuésemos tienes tengas tendrá tendríamos teníais tuvisteis tuvieran
teniendo el se no más este muy hay durante otros esto yo esa cual algo ti vosotras
tuyo suyos vuestro estoy esté estarás estarías estábamos estuvimos estuvierais estuvie-
sen estad han habré habría habías hubo hubiéramos hubieseis habidas son seré sería
eras fue fuéramos fueseis tiene tengamos tendremos tendríais tenían tuvieron tuviese
tenido en las una pero sí sin donde todos ese mí otro estos poco nosotros tu os tuya
suyas vuestra estás estés estará estaríamos estabais estuvisteis estuvieran estando he
haya habrás habrías habíamos hubimos hubierais hubiesen soy sea serás serías éramos
fuimos fuerais fuesen tenemos tengáis tendréis tendrían tuve tuviera tuvieses tenida y
por su sus porque sobre quien uno eso antes otras mucho ella mi tus mío tuyos nuestro
vuestros está estemos estaremos estaríais estaban estuvieron estuviese estado has hayas
habrá habríamos habíais hubisteis hubieran habiendo eres seas será seríamos erais fuis-
teis fueran siendo tenéis tengan tendrán tenía tuviste tuvieras tuviésemos tenidos x aca

acá adonde adónde ahí alavez algo alguien algun alguno alla allá allí ante aquel aquí así atrás atrás atravésde aun aunque como cómo con cual cuál desde despues después despuésde donde dónde encima ese estan estar este fué fui mas más pero por uno ser estar ir

APÉNDICE B

Gráficos de uso de las palabras a través del tiempo

A continuación se reproduce la evolución de las palabras de tres tópicos a través del tiempo (ver detalle en la [Subsubsección 4.1.2.2](#)).

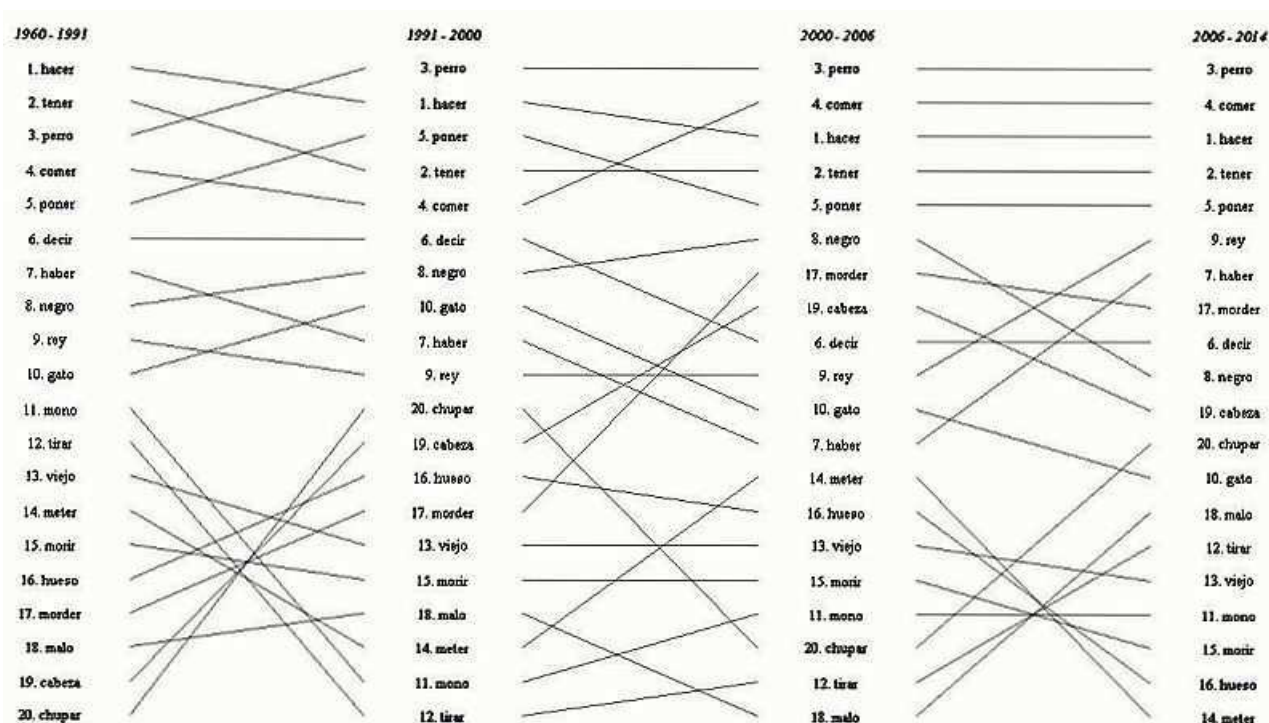


Figura B.1: Tópico Perro hacer

Los términos que presentaron un comportamiento distinto a través de los cuartiles fueron: chupar, cabeza, hueso, meter. Es notorio el aumento de posición del verbo chupar desde el primer cuartil al segundo, la baja en el tercero, y la suba en el último.

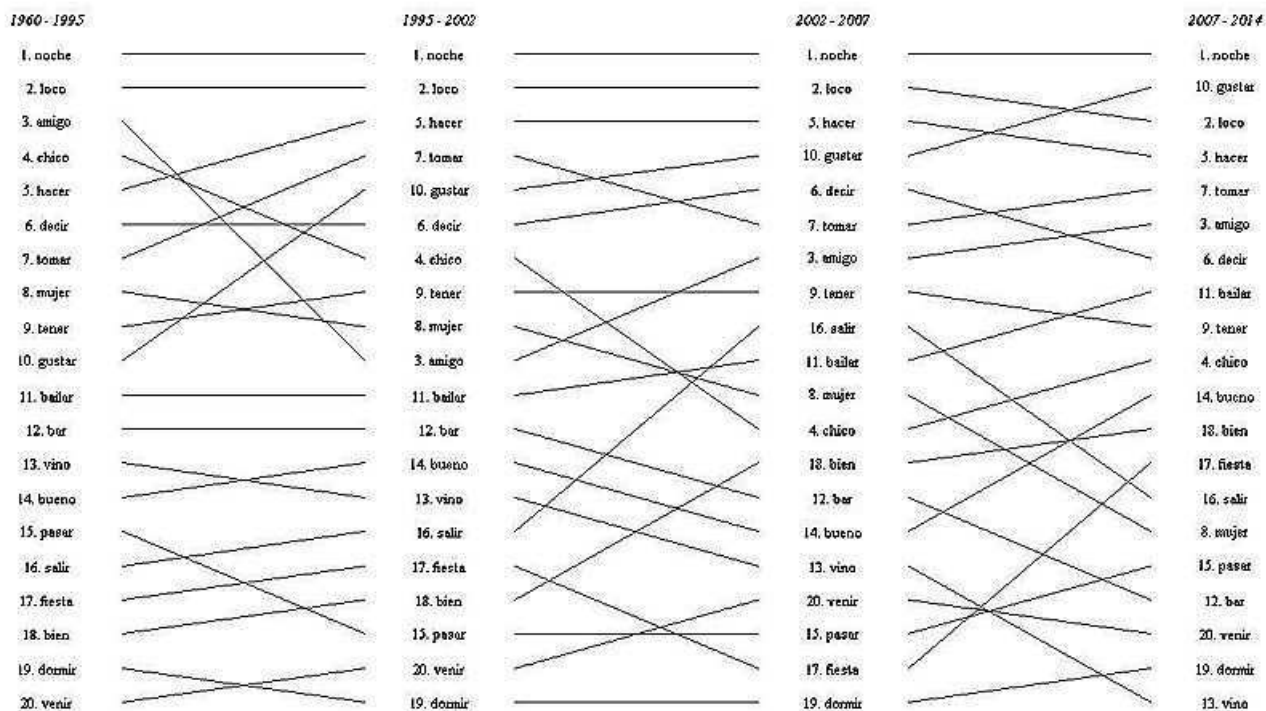


Figura B.2: Tópico Noche loco

En este tópico el término mujer desciende a través de los cuartiles. Otro término que muestra el mismo comportamiento es vino. Bar también desciende en su posicionamiento. Otro término que podría estar relacionado con bar y vino, que es tomar, mostró un comportamiento dispar, con subas y bajas. A lo largo del tiempo noche se mantuvo estable.

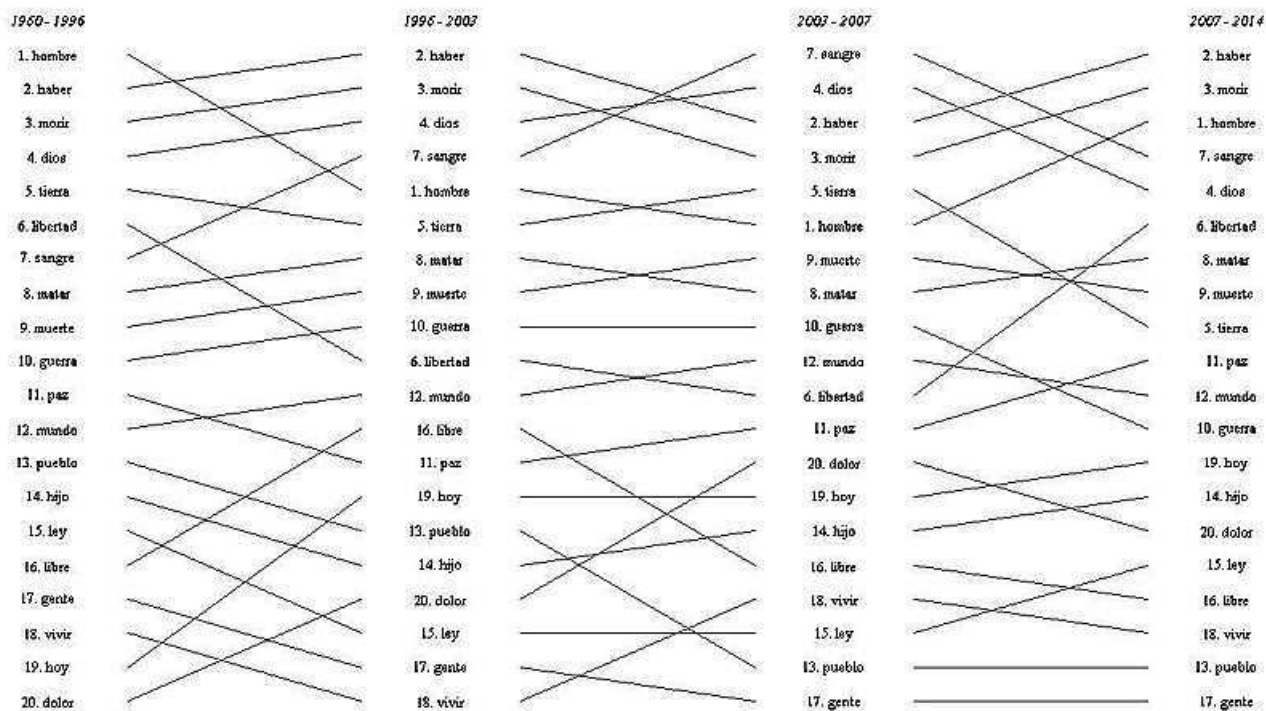


Figura B.3: Tópico Haber morir

Finalmente, se observan términos relacionados como muerte, sangre, matar, que conservan los primeros puestos a través del tiempo. Libertad sólo sube en el último cuartil. Gente, pueblo, ley y vivir, conservan los últimos puestos en distinto orden en los distintos cuartiles.

Índice de figuras

| | |
|--|----|
| 3.1. Flujo de trabajo para el armado del corpus | 21 |
| 3.2. Matriz de letras | 22 |
| 3.3. Metadata de letras del rock nacional | 23 |
| 3.4. Resultado de consulta de fecha de registro en SADAIC de letras de folklore | 23 |
| 3.5. Agrupamiento jerárquico de documentos de un cluster de k-means | 25 |
| 3.6. Letras repetidas, eliminadas luego del agrupamiento k-means | 25 |
| 3.7. Esquema de experimentos. DTM: Modelado dinámico de tópicos, ME: Máxima entropía, NB: Naive Bayes, w2v: Word2vec, vw: VowpalWabbit | 26 |
| 3.8. Histograma de documentos por fecha | 27 |
| 3.9. Histograma de cantidad de documentos por ventanas $w1$ (1960-1964) a $w4$ (1963-1967) | 28 |
| 3.10. KL Simétrica para $w1$ (1960-1964) | 29 |
| 3.11. Histograma de participación de cada tópico de la ventana 1 (1960-1964) | 30 |
| 3.12. Matriz de similitud $w1$ (1960-1964) versus $w2$ (1961-1965) | 31 |
| 3.13. Representación de tópicos en dos primeros componentes principales de la ventana 2 ($w2$: 1961-1965) | 32 |
| 3.14. Cantidad de tokens (en miles) según número de tópicos | 36 |
| 3.15. Largo de palabras según número de tópicos | 37 |
| 3.16. Coherencia según número de tópicos | 38 |
| 3.17. Distancia según número de tópicos | 39 |
| 3.18. Diferencia token/documento según número de tópicos | 39 |
| 3.19. Documentos de rango 1 según número de tópicos | 40 |
| 3.20. Entropía de documentos según número de tópicos | 40 |
| 3.21. Tópicos a través del tiempo para $k = 10$ | 41 |
| 3.22. Tópicos a través del tiempo para $k = 20$ | 42 |
| 3.23. Tópicos a través del tiempo para $k = 50$ | 43 |

| | |
|---|----|
| 3.24. Evolución temporal de tópicos para $k=10$ | 44 |
| 3.25. Evolución temporal de tópicos para $k=20$ | 45 |
| 3.26. Evolución temporal de tópicos para $k=50$ | 45 |
| 4.1. Matriz de similitud entre ventana 1 (1960-1964) y ventana 6 (1965-1969) | 57 |
| 4.2. Matriz de similitud entre ventana 6 (1965-1969) y ventana 11 (1970-1974) | 57 |
| 4.3. Matriz de similitud entre ventana 11 (1970-1974) y ventana 16 (1975-1979) | 58 |
| 4.4. Representación de tópicos en dos primeros componentes principales de ventana 16 (1975-1979) | 59 |
| 4.5. Matriz de similitud entre ventana 21 (1980-1984) y ventana 26 (1985-1989) | 60 |
| 4.6. Matriz de similitud entre ventana 26 (1985-1989) y ventana 31 (1990-1994) | 61 |
| 4.7. Matriz de estabilidad entre ventana 36 (1995-1999) y ventana 41 (2000-2004) | 62 |
| 4.8. Agrupamiento jerárquico de tópicos para $k=20$ | 68 |
| 4.9. Tópico 3 vos sos | 69 |
| 4.10. Tópico 8 sur che | 69 |
| 4.11. Evolución de palabras para un tópico elegido al azar (Tópico cantar haber) | 70 |
| 4.12. Tópico tener haber (peor correlación entre palabras y años) | 71 |
| 4.13. Evolución temporal de palabras | 72 |
| 4.14. Tópico 1 amor corazón - Hot Topic | 73 |
| 4.15. Evolución temporal tópico 1 amor corazón | 73 |
| 4.16. 5 tópicos más correlacionados negativamente (Cold Topics) con la fecha - p value = 0,05 para $k = 20$ | 74 |
| 4.17. Tópico 19 noche loco - Cold Topic | 74 |
| 4.18. Evolución temporal tópico 19 noche loco | 75 |
| 4.19. Tema 7651 - De tiempo adentro | 79 |
| B.1. Tópico Perro hacer | 85 |
| B.2. Tópico Noche loco | 86 |
| B.3. Tópico Haber morir | 87 |

Índice de cuadros

| | |
|--|----|
| 1.1. Distribución temporal de temas | 4 |
| 3.1. Parámetros para Latent Dirichlet Allocations | 34 |
| 3.2. Tabla de contingencia | 46 |
| 3.3. DTM Ejemplo 1 | 50 |
| 3.4. DTM Ejemplo 2 | 50 |
| 3.5. Corpus separado en décadas | 52 |
| 3.6. Corpus separado en hitos históricos | 52 |
| 3.7. Corpus balanceado separado en décadas | 52 |
| 3.8. Corpus balanceado separado en hitos históricos | 53 |
| 4.1. Discos con mayor proporción ventana 11 (1970-1974) del tópico 5 | 58 |
| 4.2. Discos con mayor proporción ventana 16 (1975-1979) del tópico 1 “haber tener ver si querer cantar poder” | 59 |
| 4.3. Ventana 26 discos con mayor proporción tópico 3 | 61 |
| 4.4. Grupos DNMF | 63 |
| 4.5. Ventana 1 (1960-1974) Tópicos NMF | 63 |
| 4.6. Ventana 2 (1975-1982) Tópicos NMF | 63 |
| 4.7. Ventana 3 (1983-1990) Tópicos NMF | 64 |
| 4.8. Ventana 4 (1991-1997) Tópicos NMF | 64 |
| 4.9. Ventana 5 (1998-2003) Tópicos NMF | 65 |
| 4.10. Ventana 6 (2004-2014) Tópicos NMF | 65 |
| 4.11. Tópicos Dinámicos para k=7 | 66 |
| 4.12. Medidas de coherencia para el modelado dinámico (para k entre 4 y 10) | 66 |
| 4.13. Tópicos dinámicos para k=6 | 67 |

| | |
|---|----|
| 4.14. Accuracy y F1 para modelos de aprendizaje para el conjunto de entrenamiento. NB: Naive Bayes, MaxEnt: Máxima entropía, CV: Validación cruzada, TR/TEST: dividido en training y testing, Hist: Histórico, Vw: VowpalWabbit, nn: Neural network reduction, etc: Error correcting tournament, oaa: One against all | 76 |
| 4.15. Accuracy y F1 para modelos de aprendizaje para el conjunto de testing. NB: Naive Bayes, MaxEnt: Máxima entropía, CV: Validación cruzada, TR/TEST: dividido en training y testing, Hist: Histórico, Vw: VowpalWabbit, nn: Neural network reduction, etc: Error correcting tournament, oaa: One against all | 77 |
| 4.16. VowpalWabbit-Redes Neuronales: Matriz de confusión para conjunto de testing | 78 |
| 4.17. Naive Bayes-dividido en entrenamiento y testing-sin balancear-Décadas: Matriz de confusión para conjunto de testing | 78 |
| 4.18. Máxima entropía-Validación cruzada-Lemas-Históricos: Matriz de confusión en conjunto de testing | 79 |
| 4.19. Word2Vec-Matriz de Confusión para conjunto de testing | 80 |

Bibliografía

- [Adamovsky, 2012] Adamovsky, E. (2012). *Historia de las clases populares en la Argentina: Desde 1880 hasta 2003*. Sudamericana.
- [Ahmed y Xing, 2012] Ahmed, A. y Xing, E. P. (2012). Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. *arXiv preprint arXiv:1203.3463*.
- [Arun et al., 2010] Arun, R., Suresh, V., Madhavan, C. V., y Murthy, M. N. (2010). *On finding the natural number of topics with latent dirichlet allocation: Some observations*. Springer.
- [Astor, 2008] Astor, M. (2008). Los ojos de sojo : El conflicto entre nacionalismo y modernidad en los festivales de música de caracas (1954-1966). Tesis de maestría, Universidad Central de Venezuela.
- [Berebeglia, 2007] Berebeglia, C. (2007). *Propuestas para una antropología argentina (Vol. 7)*. Editorial Biblos.
- [Blei, 2012] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.
- [Blei y Lafferty, 2006] Blei, D. M. y Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.
- [Blei et al., 2003] Blei, D. M., Ng, A. Y., y Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

- [Boyd-Graber et al., 2014] Boyd-Graber, J., Mimno, D., Newman, D., Airoldi, E. M., Blei, D., Erosheva, E. A., y Fienberg, S. E. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of Mixed Membership Models and Their Applications*.
- [Brizuela, 2008] Brizuela, L. (2008). La trama secreta de una revolución poética que cautivó a chicos y grandes. <http://www.lanacion.com.ar/1039056-la-trama-secreta-de-una-revolucion-poetica-que-cautivo-a-chicos-y-grandes>. Accedido 13-Enero-2016.
- [Bruni et al., 2014] Bruni, E., Tran, N.-K., y Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res. (JAIR)*, 49:1–47.
- [Dowle et al., 2014] Dowle, M., Short, T., Lianoglou, S., Srinivasan, A with contributions of Saporta, R., y Antonyan, E. (2014). data.table: Extension of data.frame. R package version 1.9.4.
- [Dunning, 1993] Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.
- [Eldén, 2007] Eldén, L. (2007). *Matrix methods in data mining and pattern recognition*, volume 4. SIAM.
- [Gaujoux y Seoighe, 2010] Gaujoux, R. y Seoighe, C. (2010). A flexible r package for nonnegative matrix factorization. *BMC bioinformatics*, 11(1):1.
- [Grainger, 2014] Grainger, C. (2014). Finding the natural number of topics for latent dirichlet allocation. <http://blog.cigrainger.com/2014/07/lda-number.html/>. Accedido 13-Enero-2016.
- [Greene y Cross, 2015] Greene, D. y Cross, J. P. (2015). Unveiling the political agenda of the european parliament plenary: A topical analysis. *arXiv preprint arXiv:1505.07302*.
- [Griffiths y Steyvers, 2004] Griffiths, T. L. y Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.
- [Handey y Potey, 2015] Handey, S. y Potey, M. (2015). Temporal text summarization of tv serial excerpts using lingo clustering and lucene summarizer. *SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE)*, 2:274–280.
- [Herrero y Bote, 1998] Herrero, F. d. M. A. V. y Bote, S. V. G. (1998). La aplicación de redes neuronales artificiales (rna): a la recuperación de la información. *Anuario SOCADI de Documentación e Información*, (2):147–164.

- [Heuer, 2015] Heuer, H. (2015). Semantic and stylistic text analysis and text summary evaluation. Tesis de maestría, Alto University.
- [Kleinberg, 2003] Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.
- [Koren et al., 2009] Koren, Y., Bell, R., y Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, (8):30–37.
- [Lang, 2013a] Lang, D. T. (2013a). Rcurl: General network (http/ftp/...) client interface for r. R package version 1.95-4.1.
- [Lang, 2013b] Lang, D. T. (2013b). Xml: Tools for parsing and generating xml within r and s-plus. R package version 3.98-1.1.
- [Langford et al., 2007a] Langford, J., Li, L., y Strehl, A. (2007a). Vowpal Wabbit. <http://hunch.net/~{vw}>. Accesado 13-Enero-2016.
- [Langford et al., 2007b] Langford, J., Li, L., y Strehl, A. (2007b). Vowpal wabbit online learning project. Technical report, <http://hunch.net>. Accesado 13-Enero-2016.
- [Loper y Bird, 2002] Loper, E. y Bird, S. (2002). Nltk: The natural language toolkit. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, pages 63–70.
- [Lunardelli, 2002] Lunardelli, L. (2002). *Alternatividad, divino tesoro. El rock argentino en los 90*. Editorial Biblos.
- [Ma, 2003] Ma, Junshui y Perkins, S. (2003). Online novelty detection on temporal sequences. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–618.
- [Marwick, 2013] Marwick, B. (2013). Simple text mining and document clustering of jstor journal articles. <https://github.com/benmarwick/JSTORr>.
- [McCallum, 2002] McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu>. Accesado 13-Enero-2016.
- [Mei y Zhai, 2005] Mei, Q. y Zhai, C. (2005). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 198–207.
- [Meyer et al., 2008] Meyer, D., Hornik, K., y Feinerer, I. (2008). Text mining infrastructure in r. *Journal of statistical software*, 25(5):1–54.

- [Michalke, 2015] Michalke, M. (2015). korpus: An r package for text analysis. (Version 0.05-6).
- [Michel et al., 2011] Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.
- [Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., y Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- [Mimno et al., 2011] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., y McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272.
- [Mølgaard et al., 2009] Mølgaard, L. L., Larsen, J., y Goutte, C. (2009). Temporal analysis of text data using latent variable models. *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing, 2009*.
- [Morinaga y Yamanishi, 2004] Morinaga, S. y Yamanishi, K. (2004). Tracking dynamics of topic trends using a finite mixture model. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 811–816.
- [O’Callaghan et al., 2015] O’Callaghan, D., Greene, D., Carthy, J., y Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657.
- [Özgür et al., 2005] Özgür, A., Özgür, L., y Güngör, T. (2005). Text categorization with class-based and corpus-based keyword selection. *International Symposium on Computer and Information Sciences*, pages 606–615.
- [Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P. andY Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., y Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- [Perkio et al., 2004] Perkio, J., Buntine, W., y Perttu, S. (2004). Exploring independent trends in a topic-based search engine. *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 664–668.
- [Polimeni, 2002] Polimeni, C. (2002). *Bailando sobre los escombros: historia crítica del rock latinoamericano (Vol. 3)*. Editorial Biblos.

- [R Core Team, 2013] R Core Team (2013). R: A language and environment for statistical computing.
- [Řehůřek y Sojka, 2010] Řehůřek, R. y Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. <http://is.muni.cz/publication/884893/en>.
- [Roy et al., 2002] Roy, S., Gevry, D., y Pottenger, W. M. (2002). Methodologies for trend detection in textual data mining. *Proceedings of the Textmine*, 2:1–12.
- [Sahlgren, 2005] Sahlgren, M. (2005). An introduction to random indexing. *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE*, 5.
- [Sanderson, 2010] Sanderson, M. (2010). Christopher d. manning, prabhakar raghavan, hinrich schütze, introduction to information retrieval, cambridge university press. 2008. isbn-13 978-0-521-86571-5, xxi+ 482 pages. *Natural Language Engineering*, 16(01):100–103.
- [Schilling, 2006] Schilling, C. (2006). Nos vamos poniendo simples. http://archivo.lavoz.com.ar/2006/0827/Espectaculos/nota438234_1.asp/. Accesado 13-Enero-2016.
- [Schmidt, 2012] Schmidt, B. M. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, 2(1):49–65.
- [Sievert y Shirley, 2014] Sievert, C. y Shirley, K. E. (2014). Ldavis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70.
- [Steyvers y Griffiths, 2007] Steyvers, M. y Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.
- [Suchanek y Preda, 2014] Suchanek, F. M. y Preda, N. (2014). Semantic culturomics. *Proceedings of the VLDB Endowment*, 7(12):1215–1218.
- [Tahmasebi et al., 2015] Tahmasebi, N., Borin, L., Capannini, G., Dubhashi, D., Exner, P., Forsberg, Markus y Gossen, G. y. J. F. D. y. J. R., Kågebäck, M., et al. (2015). Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries*, 15(2-4):169–187.
- [Tong y Koller, 2002] Tong, S. y Koller, D. (2002). Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66.

- [Wang y McCallum, 2006] Wang, X. y McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433.
- [Wickham, 2011] Wickham, H. (2011). The split-apply-combine strategy for data analysis.
- [Wickham, 2015] Wickham, H. (2015). stringr: Simple, consistent wrappers for common string operations. R package version 1.0.0.
- [Xu y Gong, 2003] Xu, Wei y Liu, X. y Gong, Y. (2003). Document clustering based on non-negative matrix factorization. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273.
- [Yegnanarayana, 2009] Yegnanarayana, B. (2009). *Artificial neural networks*. PHI Learning Pvt. Ltd.
- [Yuan et al., 2014] Yuan, Y., He, L., Peng, L., y Huang, Z. (2014). A new study based on word2vec and cluster for document categorization. *Journal of Computational Information Systems*, 10(21):9301–9308.