UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales

Técnicas matriciales para el análisis y clasificación de discursos presidenciales.

Tesis presentada para optar al título de Licenciado en Matemáticas de la Universidad de Buenos Aires

Autor: Ian Evangelos Bounos
Director: Dr. Juan Pablo Pinasco

Buenos Aires, 2023

Resumen

En este trabajo se muestra cómo pueden utilizarse métodos de reducción de dimensionalidad basados en matrices para el análisis y clasificación de autores de discursos presidenciales. La representación de textos como matrices de frecuencias, en la cual cada columna es una palabra del vocabulario, suele presentar el desafío de la alta dimensionalidad, por lo cual es preciso utilizar técnicas para reducir dicha dimensión. En este estudio, se emplean 1073 discursos de los presidentes Alberto Fernández, Cristina Fernández de Kirchner y Mauricio Macri, obtenidos mediante técnicas de *scraping* de páginas oficiales. Se utilizan dos métodos matriciales de reducción de dimensionalidad: el Análisis de Componentes Principales (PCA) y la Factorización No Negativa de Matrices (NMF). Esto nos permitirá, en primer lugar, asignar tópicos o temáticas a los presidentes con conjuntos términos que identifiquen sus discursos. En segunda instancia, las matrices resultantes de los métodos serán utilizadas para entrenar un modelo de clasificación con K vecinos más cercanos, cuyos resultados serán analizados y comparados en el trabajo.

Abstract

In this work, we demonstrate how matrix-based dimensionality reduction methods can be employed for the analysis and classification of presidential speech authors. Representing texts as frequency matrices, where each column represents a word from the vocabulary, often poses the challenge of high dimensionality. Hence, it is necessary to employ techniques to reduce this dimension. In this study, we use 1073 speeches from Presidents Alberto Fernández, Cristina Fernández de Kirchner, and Mauricio Macri, obtained through web scraping from official websites. We utilize two matrix-based dimensionality reduction methods: Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF). Firstly, this allows us to assign topics or themes to the presidents using sets of terms that identify their speeches. Secondly, the resulting matrices from these methods are used to train a K-Nearest Neighbors classification model, the results of which will be analyzed and compared in the study.

Agradecimientos

Quiero agradecer especialmente a mis padres por su amor y apoyo incondicional en todas mis decisiones.

A mi familia y amigos por acompañarme durante todo este trayecto.

A Juan Pablo, mi director, por su tiempo, su paciencia, por las charlas y los consejos.

Al jurado, por tomarse el trabajo de evaluar esta tesis.

También vienen a mi mente Marta, Susana, Mario, Santiago, Diego y aquellas personas que, quizás sin darse cuenta, me hayan encendido la llama de la curiosidad durante la infancia y adolescencia.

A la Universidad de Buenos Aires y todos los que la hacen posible.

Índice general

1. Introducción **7**

2. Técnicas matriciales en NLP **9**
2.1. Representaciones matriciales del texto 9
2.1.1. Bolsa de Palabras y Matriz de Frecuencias 9
2.1.2. Matriz TF-IDF 16
2.2. Reducción de Dimensión 17
2.2.1. Lematización y Stemming 17
2.2.2. Analisis de Componentes Principales (PCA) 18

3. Factorización No Negativa de Matrices **25**
3.1. Introducción 25
3.1.1. Historia y comparación con otros métodos 27
3.2. NMF: definiciones formales y principales resultados 29
3.2.1. Algoritmos de Actualización Multiplicativa 31
Algoritmo de Lee y Seung con Norma de Frobenius 31
Algoritmo de Lee y Seung con Divergencia KL . 34
3.2.2. La función de costo 35
3.3. Algoritmo de Mínimos cuadrados alternados 38
3.4. Interpretación como modelado de tópicos 39

4. Aplicaciones prácticas **41**
4.1. Datos 41
4.1.1. Tratamiento de los Datos 41
4.1.2. Análisis exploratorio 42
4.2. Análisis por año 46
4.2.1. Matriz A de frecuencias relativas con $k = 3$ 48
4.2.2. Matriz TF-IDF con $k = 3$ 50
4.2.3. Matriz TF-IDF con $k = 15$ 53
4.2.4. Matriz TF-IDF restringida a discursos durante roles presidenciales con $k = 3$ 55

4.3. Análisis por discurso 58
4.3.1. Utilización de NMF para clasificación de discursos . . . 60
4.3.2. Utilización de PCA para la clasificación de discursos . 63
4.3.3. Comparación 66

5. Conclusiones **69**

Capítulo 1

Introducción

El Procesamiento del Lenguaje Natural (NLP por sus siglas en inglés) es un campo interdisciplinario que combina áreas como la inteligencia artificial, las ciencias de la computación, la lingüística computacional y teórica, entre otras, con el objetivo de permitir que las computadoras comprendan, interpreten y generen lenguaje humano [5]. Esto implica desarrollar algoritmos y técnicas para procesar y analizar el texto y el lenguaje en diferentes formas, como el reconocimiento del habla, la comprensión del lenguaje, la generación automática de texto y la traducción. Recientemente, en esta línea existen múltiples avances de modelos de aprendizaje profundo y *Large Language Models* (LLM), que están siendo investigados de forma sistemática con sólidos resultados. En estos modelos y técnicas, si bien existen numerosas investigaciones respecto a desarrollar su explicabilidad, en muchas ocasiones los mismos adolecen de una falta de capacidad de rendir cuentas de qué es lo que efectivamente están realizando. Con este contexto en mente, nuestro trabajo buscará la aplicación de métodos matriciales más simples para el análisis de tópicos y la clasificación de discursos de presidentes y expresidentes. En particular, el objetivo será aplicar las técnicas de Factorización no negativa de matrices (NMF) y Análisis de componentes principales (PCA) a discursos de los presidentes Cristina Fernández de Kirchner, Mauricio Macri y Alberto Fernández, obtenidos por medio de *scraping* de páginas oficiales.

El primer paso será definir las bases de la representación de textos de forma vectorial con matrices de frecuencias. Esto no estará exento de problemas y costos en términos de pérdida de información, que serán discutidos a lo largo del trabajo. Por ejemplo, ignoraremos el orden de las palabras en el texto, con lo cual nuestro enfoque será puramente semántico. Otro aspecto que será central en nuestro trabajo es el hecho de que esta representación matricial será de dimensiones relativamente altas que requerirán técnicas de reducción de la dimensionalidad.

Las técnicas de PCA y NMF nos permitirán tratar estas matrices de una forma comprimida, perdiendo la menor información posible, para utilizar métodos clásicos de clasificación, como K vecinos más cercanos, evitando la conocida *maldición de las altas dimensiones*. En otras palabras, proyectaremos nuestra matriz, que será de dimensión 1073×9272, en un espacio de 1073×5 y obtendremos una tasa de aciertos en la clasificación de discursos de 92,6 % para la matriz de NMF y de 98 % para PCA.

Por otra parte, se mostrará cómo estas técnicas nos permitirán identificar ciertos términos con cada presidente. Esto será un análisis más semántico y cualitativo del mismo, que puede servir de herramientas para investigadores de las ciencias sociales. En particular, NMF se mostrará especialmente efectivo en esta tarea, puesto que suele utilizarse para el análisis de tópicos, el cual segmenta textos en clústers según el uso de palabras en simultáneo, que *a posteriori* pueden ser identificados con ciertas temáticas. Este tipo de análisis tiene numerosos antecedentes, un ejemplo puede verse en [14].

El trabajo estará estructurado de la siguiente forma:

- En el capítulo 2 se presentan las técnicas de representación matricial de textos que serán utilizadas a lo largo del trabajo. Se introducen, por ejemplo, las matrices de frecuencias y sus ponderaciones TF-IDF para lidiar con la presencia de términos muy frecuentes, que aparecen en la mayoría de los textos y no aportan demasiada información a la hora de distinguirlos. Se presenta la técnica de PCA. En todos los casos habrá ejemplificaciones prácticas con las obras de Borges y Arlt.

- En el capítulo 3 se expone la técnica de NMF de forma conceptual y heurística basada en el paper originario de Lee y Seung [8], que permite una interpretación comparada de NMF y PCA. Se presentan también resultados teóricos de los algoritmos más utilizados para la descomposición de las matrices de forma no negativa, basados en otro trabajo de Lee y Seung [9].

- En el capítulo 4 se aplican los resultados de los capítulos anteriores. En primer lugar, se hace un análisis exploratorio de los discursos disponibles similar al realizado con las obras de Borges y Arlt. Luego, se aplicará NMF a los discursos acumulados por año y se verá como este agrupamiento estará fuertemente vinculado a los períodos presidenciales. Finalmente, se utiliza NMF y PCA para la clasificación y análisis de la totalidad de los discursos. Se presentarán y compararán resultados.

Capítulo 2

Técnicas matriciales en NLP

En este capítulo se realizará una breve exposición de técnicas matriciales aplicadas al procesamiento del lenguaje natural. La misma no pretende ser exhaustiva y se limitará a presentar las que serán utilizadas en la sección práctica. Esto abarcará cómo representar textos por medio de matrices, qué problemas implican estas representaciones y qué transformaciones se pueden realizar para enfrentar dichos problemas.

2.1. Representaciones matriciales del texto

2.1.1. Bolsa de Palabras y Matriz de Frecuencias

Una bolsa de palabras (BOW, por sus siglas en inglés) es una representación comúnmente utilizada en NLP, donde se construye un vocabulario a partir de un corpus de texto y se registra la frecuencia de aparición de cada palabra de dicho corpus. En esta representación, todas las cuestiones sintácticas, como el orden y la estructura gramatical de las palabras, se ignoran, y solo se considera la presencia o ausencia de cada palabra en el texto, lo cual es un enfoque más "semántico". Esto, si bien es una limitación ostensible, se ha mostrado eficiente para ciertas técnicas de NLP.

Para ilustrar el concepto, consideremos la siguiente oración:

> **Oración 1**:"El perro corre en el parque y el gato duerme en el parque"

Si nosotros quisiéramos representar esa oración como un vector, podríamos hacerlo con el Cuadro 2.1.

Este proceso de dividir un texto en palabras o, más en general, en unidades más pequeñas o **"tokens"** se denomina **tokenización**. Un token puede ser

Cuadro 2.1: Bolsa de palabras

Palabra	**Frecuencia**
el	4
perro	1
corre	1
en	2
parque	2
y	1
gato	1

una palabra, un carácter, una subpalabra o cualquier otra unidad definida, lo cual en muchos contextos puede resultar muy ventajoso[10]. Sin embargo, en este trabajo nos limitaremos a tokens compuestos de palabras.

Las **stopwords** (palabras vacías) son palabras comunes que suelen ser filtradas o eliminadas durante el procesamiento de lenguaje natural, ya que generalmente no aportan información relevante para el análisis de texto y aparecen en muchas ocasiones. Estas palabras incluyen artículos, conjunciones, preposiciones y otros términos muy frecuentes en un idioma determinado. Cabe aclarar que las stopwords pueden tener un valor sintáctico y de estilo muy relevante. Por ejemplo, en [13] encontramos un estudio de las obras de Shakespeare que muestra agrupamientos de las obras según su género literario que **solo fue realizado teniendo en cuenta la proporción de stopwords en el texto.** Esto puede observarse en la Figura 2.1. Sin embargo, en nuestro contexto de análisis de frecuencia de palabras, pueden distorsionar resultados y por eso las excluiremos del mismo. La lista de stopwords en español puede obtenerse del paquete *tm*[1]. Además, en el Cuadro 2.2 se presenta la Oración 1 en frecuencia absoluta removiendo stopwords.

Cuadro 2.2: Bolsa de palabras (sin stopwords)

Palabra	**Frecuencia**
parque	2
perro	1
corre	1
gato	1

Un tercer procesamiento posible para tener comparabilidad entre distintas oraciones es considerar la frecuencia relativa en lugar de la absoluta. Esto

[1]https://cran.r-project.org/web/packages/tm/index.html

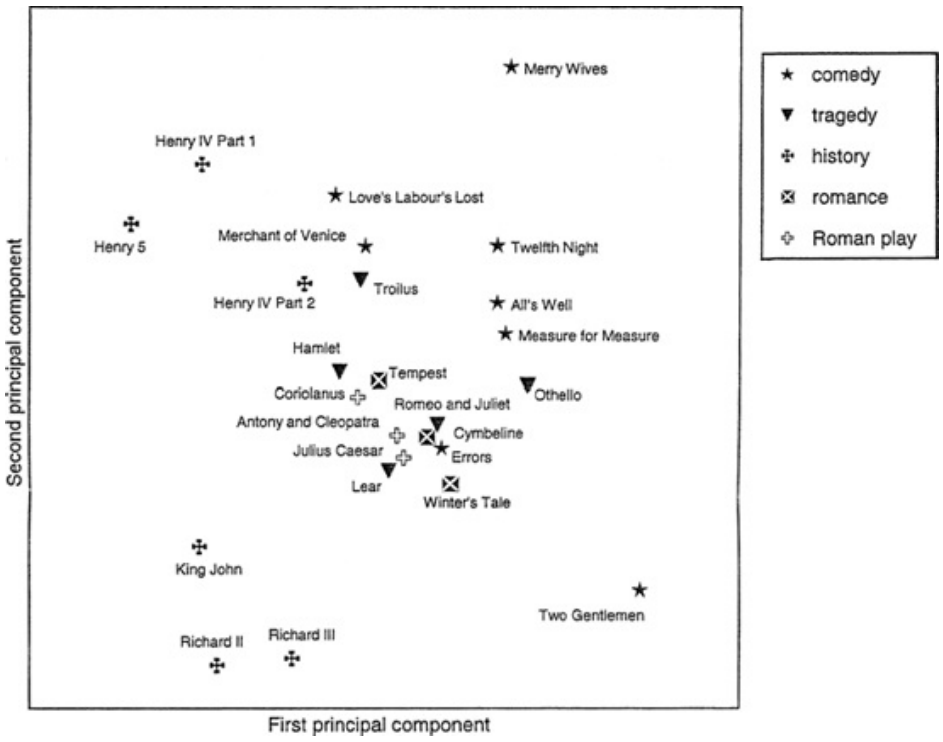


Figura 2.1: Agrupamiento de obras de Shakespeare utilizando pura y exclusivamente stopwords[13].

puede verse en el Cuadro 2.3.

Cuadro 2.3: Bolsa de palabras frecuencia relativa (sin stopwords)

Palabra	**Frecuencia Relativa**
parque	0.4
perro	0.2
corre	0.2
gato	0.2

Una vez analizadas las frecuencias absolutas y relativas de un texto y, si se considera preciso, haber eliminado las stopwords, resulta de interés obtener herramientas que permitan comparar entre textos de un mismo idioma. En este caso, tendremos un **corpus**, que será un conjunto de textos, los cuales serán denominados **documentos**. De este modo, podremos armar una matriz en la cual las columnas serán el **vocabulario** del corpus, es decir, el conjunto de palabras que aparecen en él y las filas serán la bolsa de palabras de cada texto. Como en los casos anteriores, estas frecuencias podrían ser relativas o absolutas según resulta de interés. En términos formales, una matriz de frecuencia relativa será

$$M_{ij} = \frac{TF_{ij}}{\sum_{k \in V} TF_{ik}}$$

Donde TF_{ij} es la cantidad de veces que aparece la j-ésima palabra en el documento i y V es el conjunto de índices de palabras del vocabulario.

Este primer enfoque, a pesar de su sencillez, nos permite analizar textos con un enfoque "frecuentista". Para ver un ejemplo de qué se puede hacer con estas herramientas mostraremos ejemplos extraídos del texto de Silge y Robinson [15], realizados con datasets de textos en español de Borges y Arlt[2].

El primer nivel de análisis es el de términos más frecuentes de la Figura 2.2. Aquí pueden observarse dos cuestiones: En primer lugar, ninguno de los términos tiene contenido semántico relevante, lo cual es un fuerte argumento para remover stopwords. Por otra parte, vemos un decrecimiento exponencial en la cantidad de apariciones de palabras. Esto está relacionado con la Ley de Zipf que afirma que la frecuencia de aparición de una palabra en un texto es inversamente proporcional a su posición en un ranking de frecuencias. Esto parece sugerir que incluso quitando stopwords debería mantenerse un decaimiento similar. Esto lo podemos ver gráficamente en las Figuras 2.3 y 2.4.

[2]En https://github.com/karen-pal/borges/tree/master/datasets pueden encontrarse los datasets

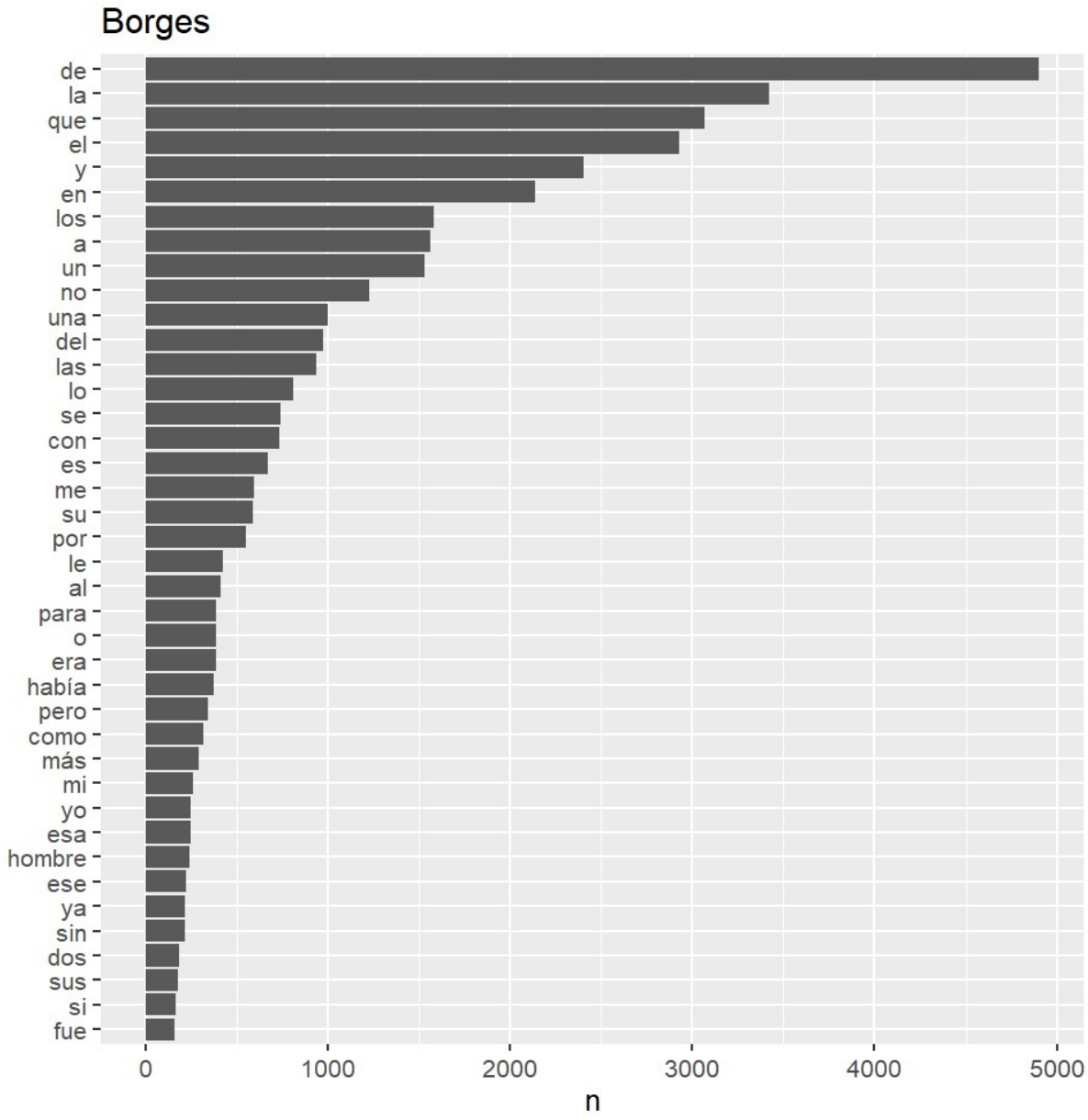


Figura 2.2: Términos más frecuentes - Borges

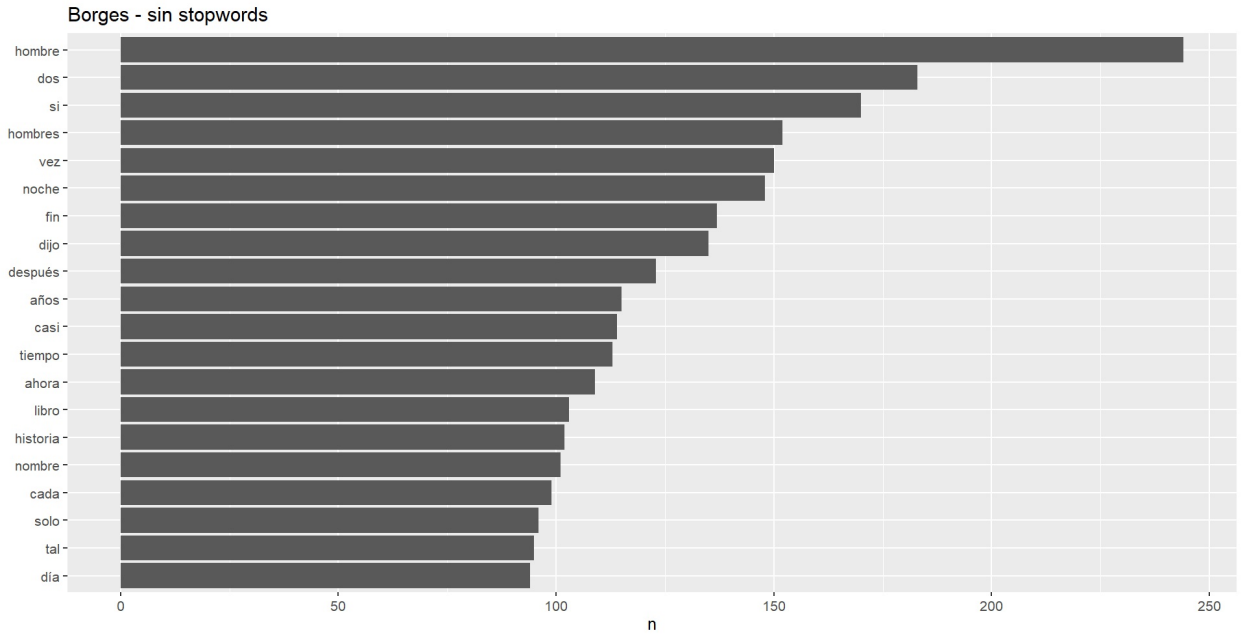


Figura 2.3: Términos más frecuentes sin stopwords - Borges

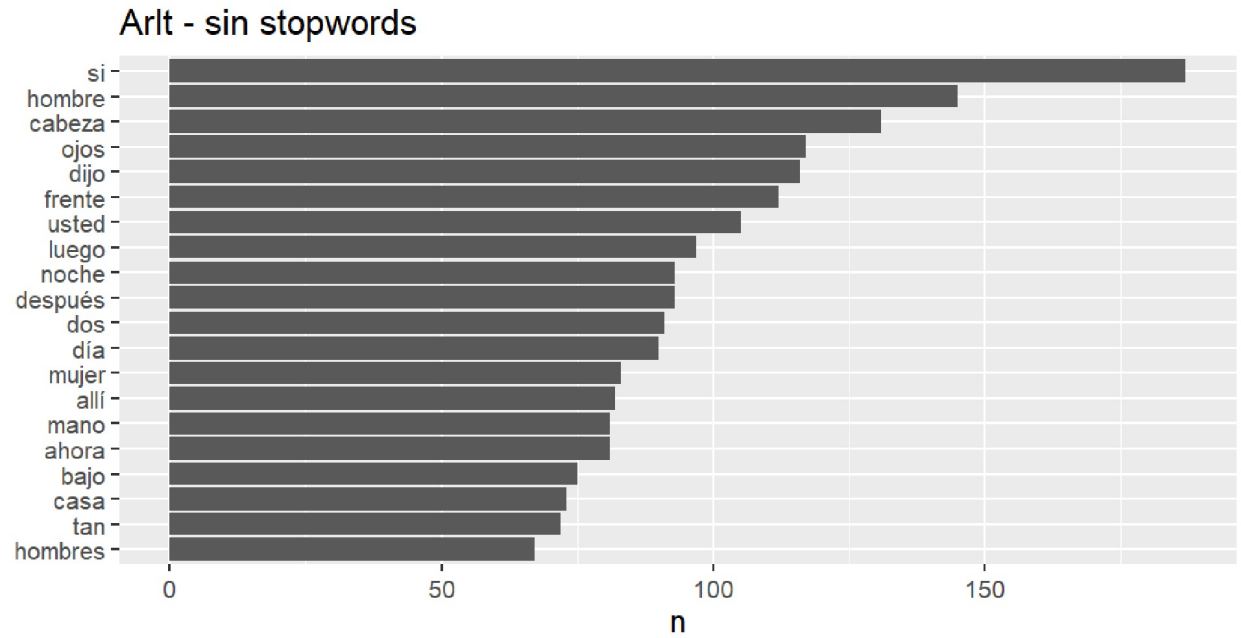


Figura 2.4: Términos más frecuentes sin stopwords - Arlt

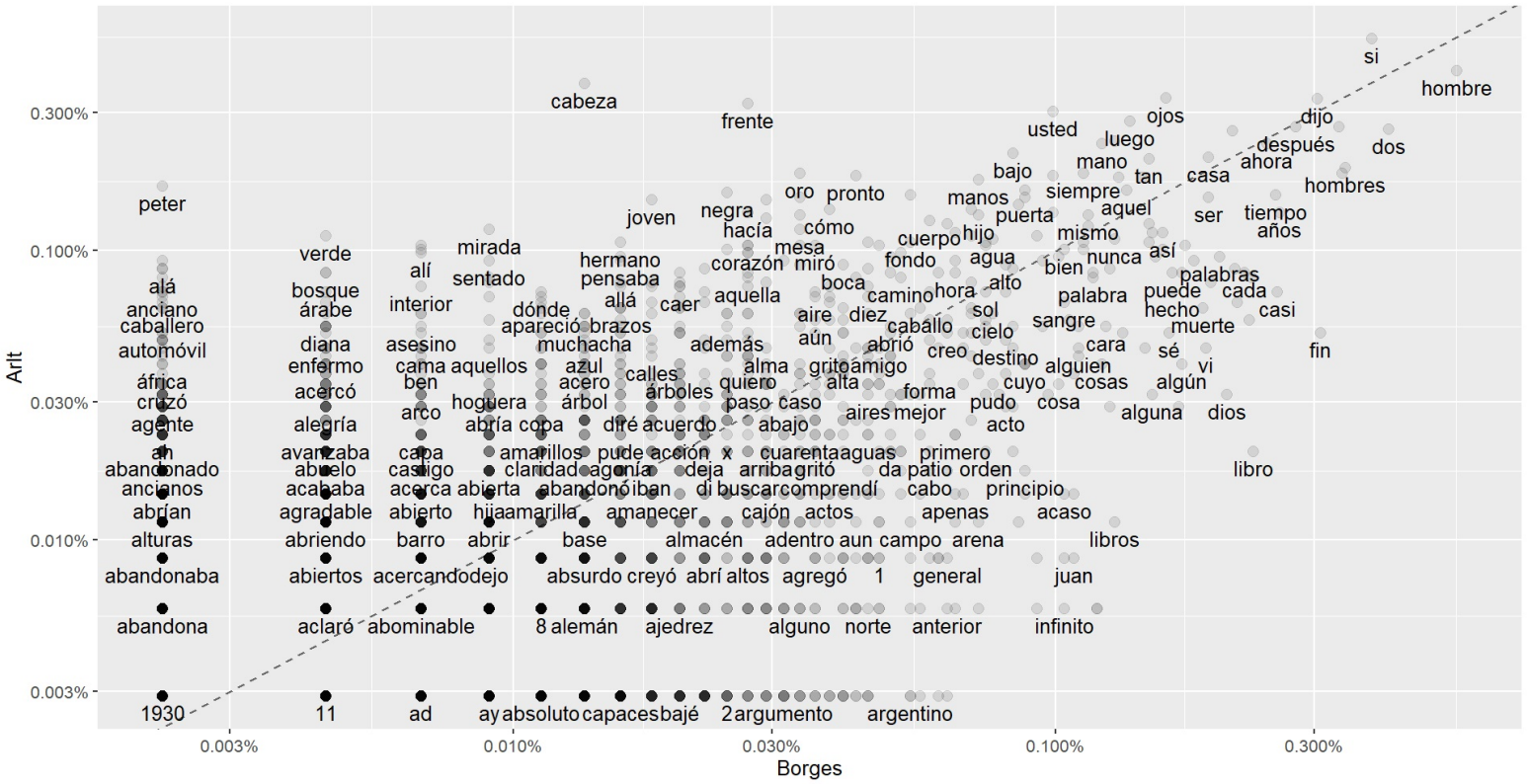


Figura 2.5: Gráfico de dispersión de proporción de palabras según autor. Escala logarítmica.

Otro enfoque de esta primera aproximación es comparar la proporción de veces que aparece cada palabra en cada autor y hacer un gráfico de dispersión, el cual tendrá una escala logarítmica por las consideraciones sobre la Ley de Zipf. En algún sentido, esto y la correlación resultante nos dan una idea de la similitud semántica entre dos autores. Podemos verlo en la Figura 2.5

2.1.2. Matriz TF-IDF

Las observaciones realizadas en la subsección anterior nos dió fuertes motivos para remover stopwords. Sin embargo, la Ley de Zipf y las Figuras 2.3 y 2.4 nos sugieren que hay términos que tendrán mucho peso en los análisis y será necesario realizar alguna transformación como la logarítmica de la Figura 2.5. Esto nos presenta un enfoque, que bien puede ser considerado alternativo o complementario: la matriz TF-IDF. La misma intenta medir la importancia de una palabra en un documento. Para ello, se basa en un equilibrio de dos ideas: La primera, captada en la matriz de frecuencias, es que **las palabras con mayor presencia en un texto son, en iguales condiciones, más relevantes que aquellas con menor presencia.**

La segunda idea es que **las palabras que aparecen en más documentos son menos relevantes que aquellas que aparecen en menos documentos, pues lo identifican particularmente**. Es decir, si sabemos que un texto contiene la palabra "algebraico", es probable que hable de matemáticas, incluso si la misma tiene una baja frecuencia relativa en dicho texto. Para captar este concepto, definimos la frecuencia inversa del documento (IDF) del i-ésimo término del documento j:

$$IDF_{ij} = \ln(\frac{J}{J_i})$$

Donde J es la cantidad de documentos totales y J_i es la cantidad de documentos que poseen el i-ésimo término. Observemos que lo que hace IDF es penalizar los términos que aparecen en muchos documentos (y que, por lo tanto, en algún sentido, aportan menos información). Es más: aquellos términos que aparecen en todos los documentos tienen un IDF igual a 0. Esto, si bien en algún sentido es deseable, también tiene la desventaja de que si una palabra aparece en todos los textos, pero con proporciones muy distintas, IDF decide ignorar su efecto.

Con estas consideraciones en mente, definimos la matriz TF-IDF, que notaremos con A:

$$A_{ij} = TF_{ij}IDF_{ij}$$

Heurísticamente, esta matriz mide la relevancia de una palabra en un texto haciendo un *trade off* entre la presencia de una palabra en el documento y la inversa el conjunto de documentos.

Finalmente, observemos que TF es la frecuencia absoluta. Por lo tanto, puede ser recomendable normalizar la matriz por fila para tener comparabilidad entre documentos de distintos tamaños.

2.2. Reducción de Dimensión

Las matrices descritas en la sección anterior tendrán necesariamente dos características que dificultarán su uso, interpretabilidad y cómputo:

- En primer lugar, dado que un texto suele tener muchas palabras distintas, tendrán un gran número de columnas (y dependiendo de la cantidad de documentos, filas).
- Por otra parte, dado que en todo texto hay muchas palabras que no se utilizan, la matriz tendrá muchos ceros (es decir, será "rala").

Esto requiere algunas consideraciones relativas a reducir la dimensión de la matriz que serán abordadas en esta sección:

2.2.1. Lematización y Stemming

Si se observa la Figura 2.3 puede observarse que aparece en primer lugar la palabra "hombre" y en cuarto lugar, "hombres". Dependiendo del tipo de análisis que se esté realizando, puede considerarse que esencialmente se trata de la misma palabra y agruparlas puede ser una buena estrategia para reducir dimensiones y ruido. El mismo fenómeno ocurre, entre muchos otros, con:

- Diferentes tiempos verbales: "comer", "comieron".
- Relación adjetivo/sustantivos: "serio" y "seriedad".
- Adverbios: "Rápido" y "Rapidamente"

La **lematización** es el proceso de reducir una palabra a su lema o forma base, que es una forma canónica o diccionario. El lema representa la raíz léxica de una palabra y puede ser un sustantivo, verbo, adjetivo, adverbio, etc. Por ejemplo, el lema del verbo "corriendo" es "correr", mientras que el lema del sustantivo "ratones" es "ratón". La lematización suele tener en cuenta la morfología y la estructura gramatical de las palabras para determinar su forma base correcta.

Por otro lado, el **stemming** es el proceso de reducir una palabra a su raíz o stem mediante la eliminación de sufijos y prefijos comunes. A diferencia de la lematización, el stemming no considera la estructura gramatical o el contexto de las palabras. En lugar de ello, se basa en reglas heurísticas simples para cortar o truncar los extremos de las palabras. Por ejemplo, el stemming

podría convertir las palabras "corriendo", "corre", "corrió" y "correrá" a la raíz "corr". El stemming es más rápido que la lematización, pero puede generar resultados menos precisos y palabras truncadas que pueden no ser reconocibles o legibles.

2.2.2. Analisis de Componentes Principales (PCA)

Los métodos de Lematización y Stemming pueden ser muy útiles para reducir considerablemente las dimensiones con la ventaja de no perder demasiada información semántica del texto. Sin embargo, lo que ocurrirá es que las matrices de frecuencias, si bien tendrán menos columnas, seguirán siendo muy grandes. Por lo tanto, a efectos de tener dimensiones más pequeñas y realizar clustering o clasificación con alguna métrica adecuada, evitando la *maldición de la dimensión*, pueden utilizarse otras técnicas de reducción de la dimensión. En este caso, veremos brevemente el Ánalisis de Componentes Principales (PCA, por sus siglas en inglés). Esta es una técnica no supervisada cuyo objetivo es transformar un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas, denominadas componentes principales, que capturan la mayor variabilidad presente en los datos originales. La bibliografía utilizada para esta sección será el famoso texto de James y Garret[6].

En este caso, supongamos que tenemos n observaciones de p variables $X_1, \cdots, X_p$, que serán las columnas de la matriz de frecuencias o TF-IDF, es decir, las palabras que pueden estar agrupadas con algún método de la subsección anterior. La idea es encontrar pesos ϕ_{i1} de modo que la variable Z_1 (a la que llamaremos *primera componente principal*) resuma información de los X_i. Z_1 será definida como:

$$Z_1 = \phi_{11}X_1 + \cdots + \phi_{p1}X_p$$

Además, pediremos que los ϕ_{i1} estén normalizados para cada j, es decir $\sum_{i=1}^{p} \phi_{i1}^2 = 1$. La cuestión posterior es decidir qué criterio utilizar para elegir los ϕ_{i1}. Heurísticamente que la idea es buscar "maximizar la información" que se preserva, en Z_1 se buscará una maximización de la varianza. Esto se puede fundamentar intuitivamente porque datos con más variabilidad permiten extraer mayor información, por ejemplo, a través de clústers. Por lo tanto, el problema resultante es maximizar respecto a las ϕ_{i1}:

$$\sum_{j=1}^{n}(\phi_{11}x_{1j} + \cdots + \phi_{p1}x_{pj})^2$$

donde los x_{ij} son la realización de la j-ésima observación de X_i (las cuales se asumirán estandarizadas). Además, todo este problema de optimización tiene la mencionada restricción de normalización de las ϕ_{i1}. El problema puede ser calculado con técnicas matriciales.

Análogamente calcular la segunda componente principal, buscaremos una combinación lineal:

$$Z_2 = \phi_{12}X_1 + \cdots + \phi_{p2}X_p$$

Nuevamente, pediremos que esté normalizada y que maximice la varianza de Z_2 **dentro de los vectores** ϕ_2 **no correlacionados con** ϕ_1. Esto implica que serán vectores ortonormales.

$$Z_j = \phi_{1j}X_1 + \cdots + \phi_{pj}X_p$$

Uno podría extender esto a k componentes principales con $k \leq p$. Para ello utilizamos notación matricial y el problema se reduce a: dado $X \in \mathbb{R}^{n\times p}$ encontrar una matriz ortonormal $W \in \mathbb{R}^{p\times k}$ (cuyas filas serán las ϕ_j) de forma que cada columna Z_j de la matriz $Z = X^tW$ tenga máxima varianza dentro de todas las combinaciones lineales de vectores ortonormales al conjunto de ϕ_r con $r < j$. Una representación visual del funcionamiento de PCA en tres dimensiones puede verse en la Figura 2.6.

Típicamente la matriz Z es llamada matriz de *Scores*. En lo práctico, esto nos permite graficar las primeras componentes para tener una visualización global en dos dimensiones que capte gran parte de la varianza de todas las variables.

A modo ilustrativo, retomemos los textos de Borges y Arlt con sus matrices de frecuencias. En este caso, nos hemos quedado solo con los términos que aparecen más de 5 veces en total y realizamos PCA, graficamos las dos componentes y los resultados pueden verse en la Figura 2.7. Pueden observarse dos cosas:

1. Tenemos una clara diferenciación entre grupos que un simple algoritmo de clasificación supervisada logrará diferenciar

2. La componente que separa a los dos autores no es la primera, si no la segunda. Esto puede apreciarse aún más en el histograma de la Figura 2.8.

Esta segunda observación resulta fundamental para entender una particularidad del método y que ha de ser tenida en cuenta: este separa en componentes que maximizan la varianza, con lo cual la varianza de la primera componente ha de ser necesariamente mayor que la de la segunda, **pero no**

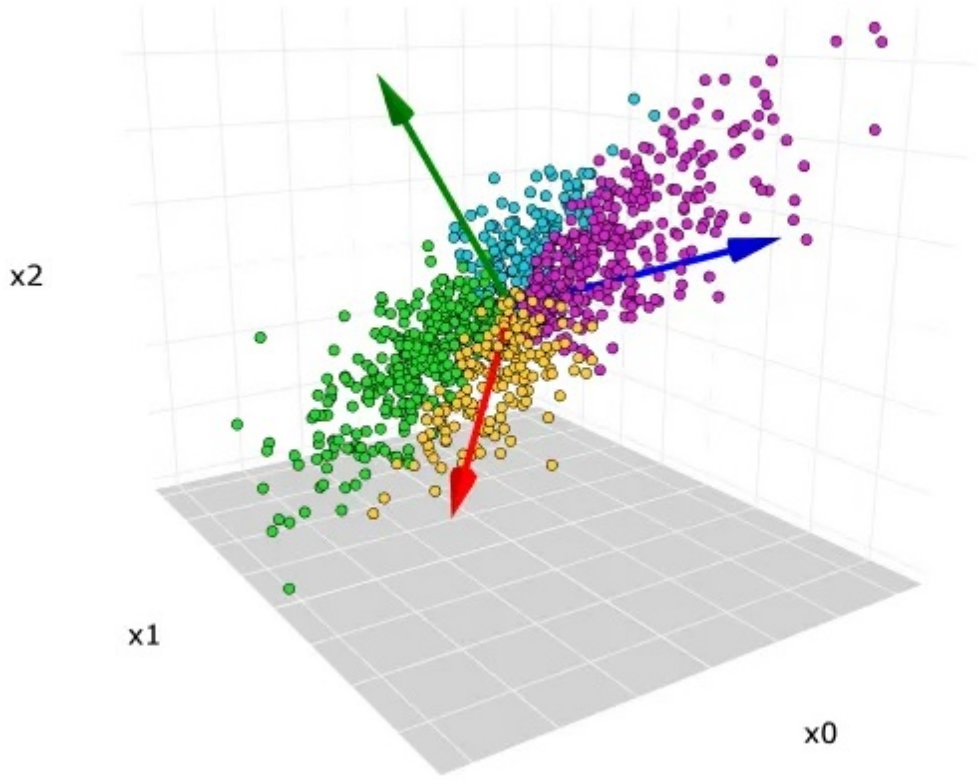


Figura 2.6: Gráfico ilustrativo del funcionamiento de PCA. Las flechas rojas, azules y verdes son las direcciones de la primera, segunda y tercera componentes principales, respectivamente. Elaborada por Casey Cheng en [3].

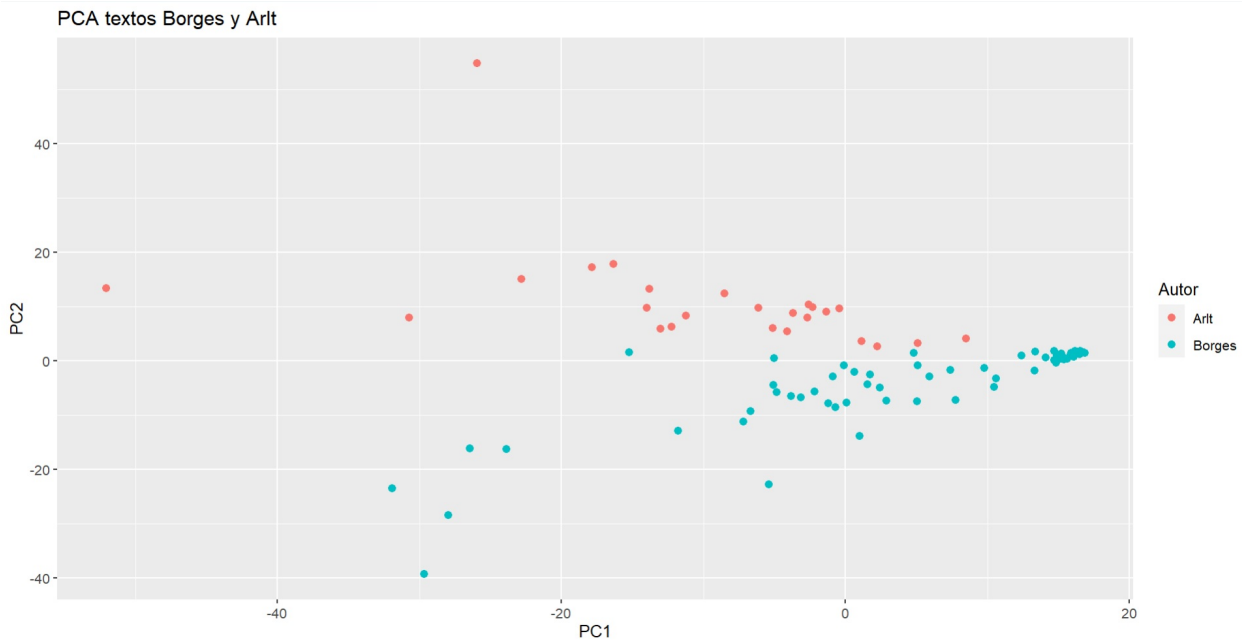


Figura 2.7: Primeras Componentes PCA de textos de Borges y Arlt

tiene por qué traducirse en una mejor separación de los dos autores, pues al ser no supervisado, en ningún momento el método los tiene en cuenta. Esto resulta de particular interés y es compartido con la técnica central de este trabajo (Factorización No Negativa de Matrices), que es el hecho de usar técnicas matriciales, computacionalmente no demasiado pesadas, no supervisadas y con una interpretación clara.

Por último, para mostrar la interpretabilidad de los resultados, podemos observar en el Cuadro 2.4 en el cual se ven las palabras que mayor peso ϕ_i tienen en valor absoluto ordenado de forma decreciente para cada una de las dos primeras componentes. Vemos que en la PC1 se encuentran palabras de uso más frecuente en ambos autores (como "si", "hombre", "usted"), lo cual puede comprobarse en las Figuras 2.3 y 2.4. Estas palabras inducen variabilidad, por ejemplo, uno podría hipotetizar que la palabra "usted" aparecerá o bien mucho o bien muy poco según el registro del texto, pero no logran diferenciar entre autores. De hecho, puede verse que el conjunto de textos de Borges acumulados en la parte derecha del eje de abscisas corresponde a poesías. Con lo cual, la primera componente podría resultar útil para diferenciar poesías de cuentos. Por otra parte, al ver la PC2, vemos cómo aparecen nombres propios que permiten incluso diferenciar textos por nombres propios cómo Anderson o Peter. Analizando el histograma, podemos inferir también

qué términos con valores más negativos en la PC2 serán más propios de Borges; mientras que los más positivos, de Arlt. Por ejemplo, "Tlön" tiene un peso de $-0{,}13$ y proviene del texto de Borges "Tlön, Uqbar, Orbis Tertius".

Cuadro 2.4: Palabras PCA

PC1-Palabra	**Peso**	PC2-Palabra	**Peso**
Si	0.30	Anderson	0.30
Hombre	-0.21	Peter	0.29
Usted	-0.16	Cabeza	0.14
Dos	-0.15	Tlön	-0.13
Dijo	-0.15	Vi	-0.12
Después	-0.14	Libro	-0.11
Hombres	-0.13	Frente	-0.10
Noche	-0.13	Mundo	-0.10

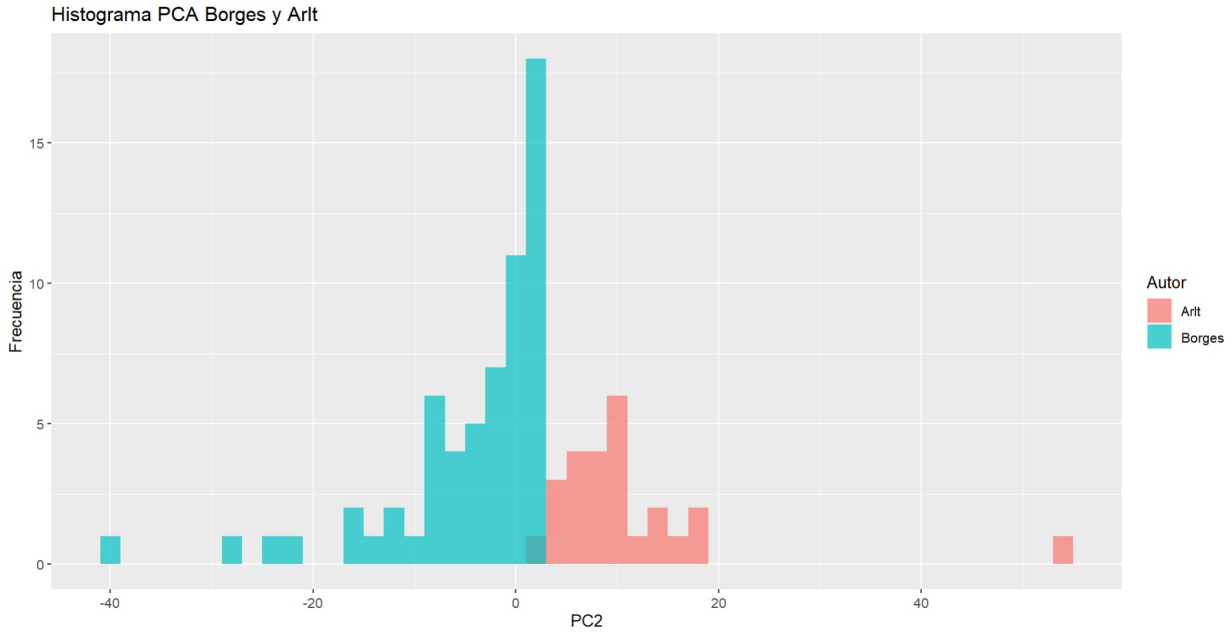


Figura 2.8: Histograma PC2 Borges y Arlt

Capítulo 3

Factorización No Negativa de Matrices

3.1. Introducción

Supongamos que poseemos una matriz de términos y documentos y queremos extraer las palabras o términos más importantes que los componen. Cómo mencionamos en la sección anterior, podemos representar cada documento como un vector de términos y cada término como un vector de frecuencia en todos los documentos. Por otro lado, se pueden hacer matrices de pesos de términos como las TF-IDF. Esta matriz probablemente se vea como la observada en la figura 3.1, es decir, sin patrones ni estructuras visibles, similar a un "ruido blanco".

Nosotros queremos generar un agrupamiento basado en el supuesto de que **documentos similares utilizan términos similares de forma simultánea**. Una estrategia posible es reordenar los documentos y términos de forma que se formen bloques que permitan diferenciar grupos de documentos como en la Figura 3.2

El problema es conseguir un posible criterio de agrupamiento. Una posibilidad para lograrlo es a partir de la matriz de términos y documentos no negativos V de tamaño $n \times m$, donde n es el número de términos y m es el número de documentos y aproximarla como producto de dos matrices no negativas $W \in \mathbb{R}^{n \times k}$ y $H \in \mathbb{R}^{k \times m}$. Típicamente, se tendrá $k \ll \text{mín}(n, m)$ y la elección de k será una cuestión a determinar en función del problema, al igual que en la mayoría de los métodos de aprendizaje no supervisado. Es decir, en algún sentido, lo que buscamos es que:

$$V \approx WH.$$

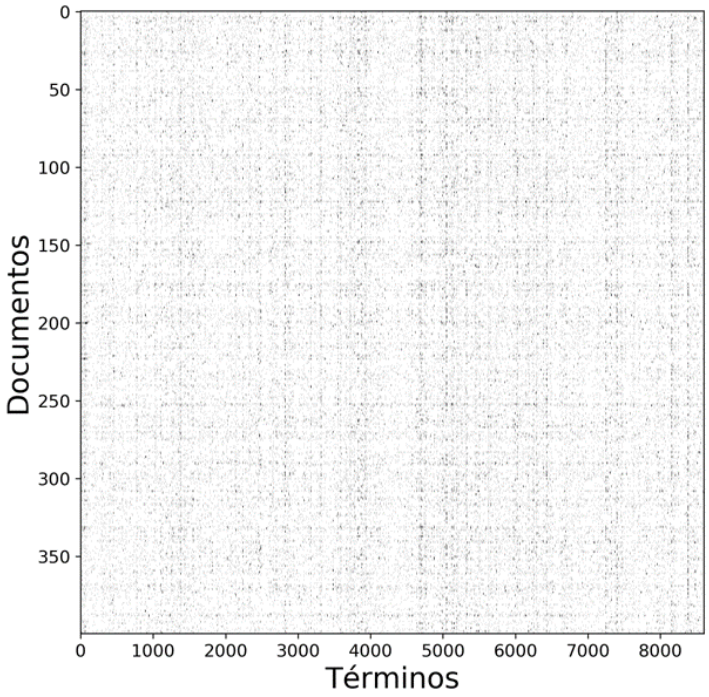


Figura 3.1: Matriz de frecuencia de términos y documentos. Extraída de [12].

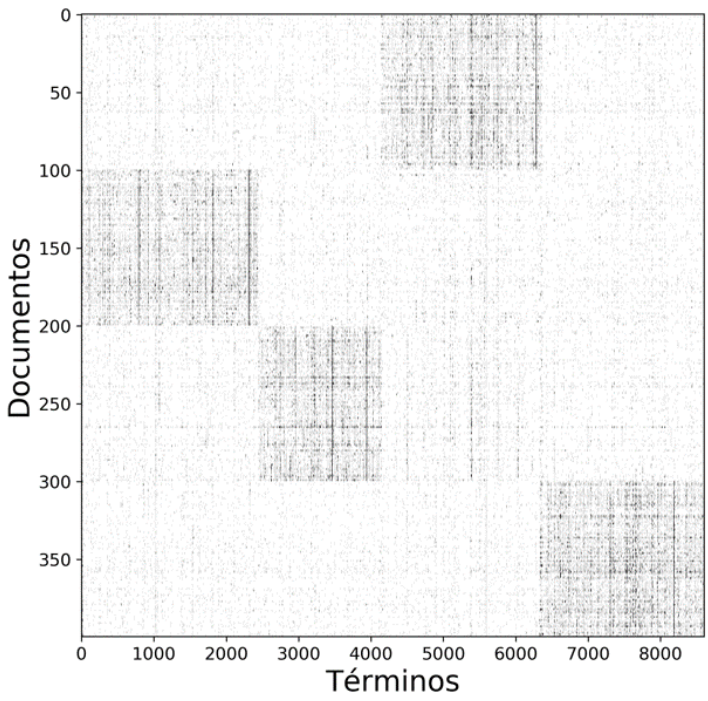


Figura 3.2: Matriz reordenada de frecuencia de términos y documentos. Extraída de [12].

A lo largo del trabajo, el producto WH será denominado **factorización como producto de matrices no negativas** (NMF, por sus siglas en inglés), aunque cabe aclarar que no es una factorización estricta, sino una aproximación, pues WH no necesariamente es igual a V. Por ello hay algunos autores [4] que proponen llamarla **aproximación por producto de matrices no negativas** (NNMA por sus siglas en inglés). Una representación ilustrativa de qué es lo que hace esta factorización puede verse en la Figura 3.3.

3.1.1. Historia y comparación con otros métodos

Para comprender mejor el método, revisaremos su historia. Su primera aparición fue en 1994 en el artículo de Paatero y Tapper *"Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values"* [11]. Allí bautizó al método con el desacertado nombre

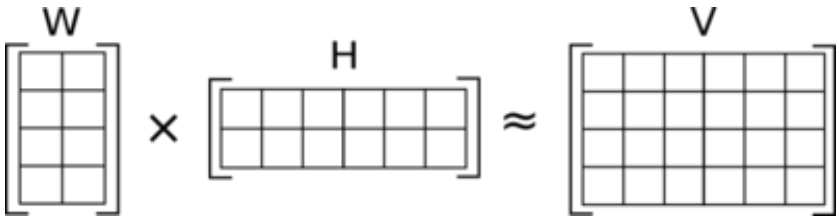


Figura 3.3: Descomposición NMF.

de Factorización Positiva de Matrices. Sin embargo, el método se popularizó en 1999 con un artículo de la revista Nature de Lee y Seung [8]. En el mismo se propone su uso para la compresión de imágenes y se sugiere su utilización en NLP, además, compara NMF con *Vector Quantization* (VQ) y PCA, el método visto en el capítulo anterior. Lee y Seung argumentan que la restricción de no negatividad genera resultados que son conceptualmente diferentes. Esto puede apreciarse en la Figura 3.4. En el trabajo, estos tres métodos se aplicaron a una base de datos de 2.429 imágenes faciales, donde cada imagen tenía una resolución de 19×19 píxeles y se representaban mediante la matriz V de tamaño $n \times m$. Aunque todos ellos buscan factorizaciones aproximadas de la forma $V \sim WH$, utilizan diferentes restricciones en las matrices W y H. En los resultados, se muestra que cada método ha aprendido un conjunto de 49 imágenes base que representan distintas características faciales. Estas imágenes base se combinan linealmente para aproximadamente representar una instancia particular de un rostro. Además, los valores positivos se ilustran con píxeles negros y los valores negativos con píxeles rojos.

El hecho de que los pesos de NMF sean no negativos genera que cada una de las imágenes base sea una parte de la cara que se van combinando para formar un rostro. Este fenómeno puede observarse en la primera fila de la Figura 3.4. En el caso de VQ, que no explicaremos formalmente, la restricción es que los pesos tengan solo un elemento con valor 1 y todos los demás, 0. Por lo tanto, las imágenes base funcionan en algún sentido como "rostros arquetípicos" y a cada persona se le asigna una de las 49 caras que más se le parecen en algún sentido. Finalmente, PCA no tiene restricciones más que la normalización de los pesos, por lo tanto, pueden tomar valores negativos, lo cual implica una difícil interpretación de las imágenes base, pues cada una contiene varios elementos que no son claramente distinguibles como si lo son en NMF y VQ. Como el único método que separa el rostro en partes es NMF, los autores dicen que es una "representación basada en partes", mientras que los otros métodos son denominados "representaciones holísticas".

En resumen, de forma informal y simplificada, podemos afirmar que, a efectos de reconstruir un rostro:

- NMF lo hace asignando cada una de las partes que más "le corresponden". Es decir, generará el rostro a partir de colocar la nariz, la boca, los ojos y otras partes del cuerpo que más le corresponde dentro de la base predeterminada (o combinaciones lineales de los mismos).

- VQ seleccionará el más similar de una base de rostros predeterminada.

- PCA lo reconstruirá como combinación lineal de píxeles positivos y negativos de difícil interpretación conceptual.

3.2. NMF: definiciones formales y principales resultados

En la subsección anterior se ha mencionado que, *en algún sentido*, necesitamos realizar una aproximación

$$V \approx WH.$$

Tendremos la restricción de que todas las matrices involucradas sean no negativas. Para determinar en qué sentido se da la aproximación, un primer acercamiento sería buscar minimizar el cuadrado de la norma de Frobenius:

$$f(W, H) = ||V - WH||_F^2,$$

donde

$$||A||_F = \sqrt{\mathrm{tr}(A^T A)}.$$

Puede probarse que

$$||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^2},$$

lo cual hace el problema equivalente a minimizar el error cuadrático medio. Sin embargo, esta optimización implica problemas numéricos de dos tipos [8]:

1. **Existencia de Mínimo Global**: La función objetivo $f(W, H)$ no es convexa al mismo tiempo en W y H. Con lo cual, a priori, no tenemos garantizada la existencia de mínimo global.

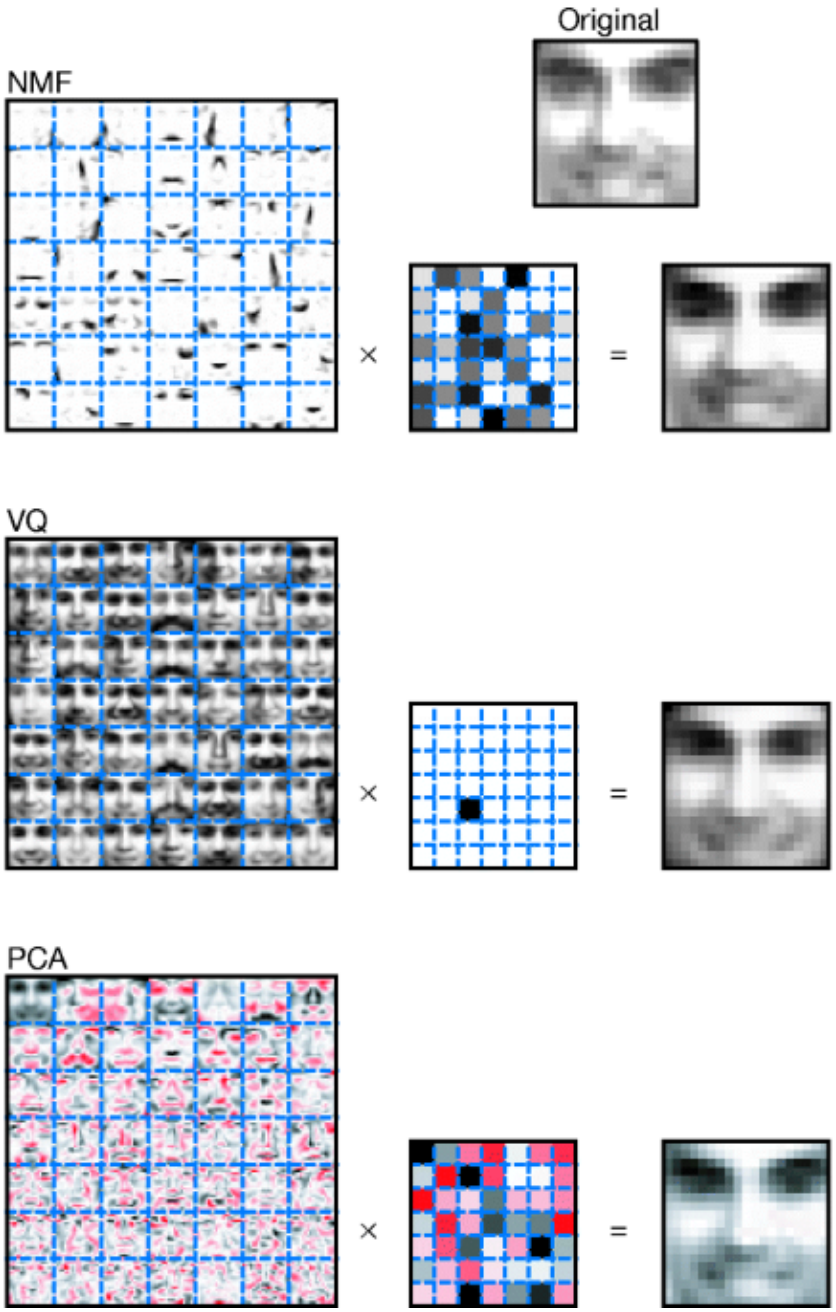


Figura 3.4: Representaciones faciales con NMF, VQ y PCA. Extraído de [8].

2. **No unicidad de mínimos:** si encontramos un mínimo W_0H_0, nada garantiza que este sea global. De hecho, dado este producto, si tomamos cualquier matriz diagonal positiva $D \in \mathbb{R}^{k\times k}$, claramente $(W_0D)(D^{-1}H_0)$ también lo será. Lo cual muestra la imposibilidad de unicidad.

3. **Convergencia:** Cuestiones relacionadas a la convergencia de los algoritmos utilizados.

Por otra parte, el problema puede ser pensado de una forma más general: podemos considerar que, dada V, buscamos minimizar una función de costo $f(W, H)$ que no necesariamente sea la dada por la norma de Frobenius, con la restricción de la no negatividad de las matrices involucradas.

3.2.1. Algoritmos de Actualización Multiplicativa

Algoritmo de Lee y Seung con Norma de Frobenius Lee y Seung [9] [1]consideraron el problema de NMF explicado en la subsección anterior para el caso en que $f(W, H) = ||V - WH||_F^2$.

Problema 1: Dada V matriz no negativa, minimizar $||V - WH||_F^2$ con respecto a W y H sujeto a la restricción de que estas sean no negativas.

Para ello propusieron el siguiente algoritmo:

Algoritmo de Lee y Seung para NMF con Norma de Frobenius

1. **Inicialización:** Seleccionar aleatoriamente dos matrices no negativas W y H de dimensiones $n \times k$ y $k \times m$, respectivamente, donde m y n son las dimensiones de la matriz original V y r es la dimensión de la matriz de factorización.

2. **Actualización de W:** Calcular la matriz H^TV y H^THW. Actualizar W como $W \leftarrow W \odot \frac{(VH^T)}{(H^THW)}$, donde $\odot$ denota la multiplicación elemento por elemento.

3. **Actualización de H:** Calcular la matriz VW^T y WW^TH. Actualizar H como $H \leftarrow H \odot \frac{(W^TV)}{(WW^TH)}$.

4. **Repetir:** Repetir los pasos 2 y 3 hasta que se alcance un criterio de parada. Los criterios de parada comunes incluyen un número máximo de iteraciones o una pequeña mejora en la función de costo.

[1]Todas las demostraciones de esta subsección están basadas en este artículo

La fundamentación de este algoritmo se basa en el siguiente teorema [9]:

***Teorema* 1.** *$||V - WH||_F^2$ es no decreciente respecto de W y de H bajo las reglas de actualización*

$$W \leftarrow W \odot \frac{(VH^T)}{(H^THW)}$$
$$H \leftarrow H \odot \frac{(W^TV)}{(WW^TH)}$$

Además, f es invariante bajo estas actualizaciones si y solo si W y H están en un punto estacionario de la distancia.

Para demostrar el Teorema hemos de establecer algunas definiciones y lemas:

Definición **1** (Función Auxiliar)**.** *$G(h, h')$ es una función auxiliar para la función $F(h)$ si cumple simultáneamente:*

$$G(h, h) = F(h)$$
$$G(h, h') \geq F(h) \quad \forall h'.$$

***Lema* 1.** *Si G es una función auxiliar para F, F es no decreciente para la regla de actualización*

$$h^{t+1} = argmin_h G(h, h^t). \tag{3.1}$$

Demostración. Usando la definición de función auxiliar, obtenemos que $F(h^{t+1}) \leq G(h^{t+1}, h^t)$. Por otro lado, por el hecho de que h^{t+1} proviene de minimizar $G(h, h^t)$ concluimos que

$$F(h^{t+1}) \leq G(h^{t+1}, h^t) \leq G(h^t, h^t) = F(h^t).$$

□

En este punto resulta importante notar que, por definición, $F(h^{t+1}) = F(h^t)$ solo si h^t es un mínimo local de $G(h, h^t)$. Además, bajo hipótesis de suavidad de G en un entorno de h^t, esto implica que $\nabla F(h^t) = 0$ (pues $F(h) = G(h, h)$).

Por otra parte, esto nos permite armar una sucesión decreciente:

$$F(h^0) \geq F(h^1) \geq \cdots \geq F(h^t) \geq \ldots$$

Puede verse en la Figura 3.5 cómo minimizar $G(h, h^t) \geq F(h)$ garantiza que $F(h^{t+1}) \leq F(h^t)$ para $h^{t+1} = argmin_h G(h, h^t)$

El próximo lema nos muestra una elección adecuada de función auxiliar para nuestra función objetivo.

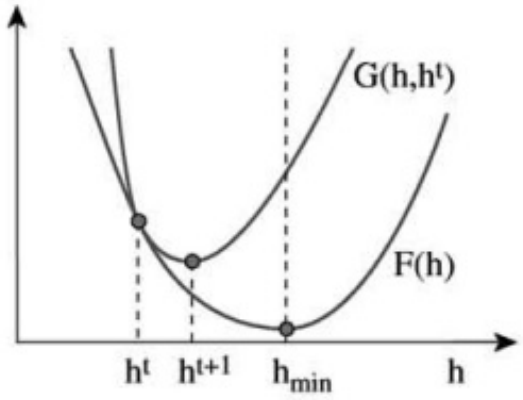


Figura 3.5: La minimización de la función $G(h, h^t) \geq F(h^t)$ garantiza que $F(h^{t+1}) \leq F(h^t)$ para $h^{t+1} = argmin_h G(h, h^t)$. Imagen extraída de [9]

***Lema* 2.** *Sea $K(h^t)$ la matriz diagonal*

$$K_{ab}(h^t) = \delta_{ab}(W^T W h^t)_a / h^t_a,$$

Supongamos que h^t es vector no negativo. Luego,

$$G(h, h^t) = F(h^t) + (h - h^t)\nabla F(h^t) + \frac{1}{2}(h - h^t)^T K(h^t)(h - h^t) \qquad (3.2)$$

es una función auxiliar para

$$F(h) := \frac{1}{2}\sum_i (v_i - \sum_a W_{ia} h_a)^2$$

donde v_i es un vector de n filas.

Demostración. Es inmediato que $G(h, h) = F(h)$ con lo cual hemos de demostrar que $G(h, h^t) \geq F(h)$. Para ello observamos que:

$$F(h) = F(h^t) + (h - h^t)\nabla F(h^t) + \frac{1}{2}(h - h^t)^T W^T W (h - h^t)$$

y si lo comparamos con la definición dada por la Ecuación 3.2, observamos que $G(h, h^t) \geq F(h)$ resulta equivalente a probar que:

$$0 \leq (h - h^t)^T (K(h^t) - W^T W)(h - h^t)$$

lo cual resulta equivalente a probar que $K(h^t) - W^T W$ es semi definida positiva, que a su vez se reduce a probarlo para M definida como

$$M_{ab}(h^t) := h_a^t (K(h^t) - W^T W)_{ab} h_b^t$$

En efecto, veamos que es semi definida positiva:

$$\nu^T M \nu = \sum_{ab} \nu_a M_{ab} \nu_b = \sum_{ab} h_a^t (W^T W)_{ab} h_b^t \nu_a^2 - \nu_a h_a^t (W^T W)_{ab} h_b^t \nu_b$$

$$= \sum_{ab} (W^T W)_{ab} h_a^t h_b^t [\frac{1}{2}\nu_a^2 + \frac{1}{2}\nu_b^2 - \nu_a \nu_b] = \frac{1}{2} \sum_{ab} h_a^t (W^T W)_{ab} h_b^t (\nu_a - \nu_b)^2 \geq 0$$

Esta última desigualdad vale porque todos los términos involucrados en la última suma son no negativos.

□

Ahora estamos en condiciones de probar el Teorema 1.

Demostración. Reemplazamos el $G(h, h^t)$ de la Ecuación 3.2 en 3.1, obtenemos la regla de actualización:

$$h^{t+1} = h^t - K(h^t)^{-1} \nabla F(h^t)$$

Como G es una función auxiliar para F, esta es no creciente bajo esta regla de actualización por el Lema 1. Si reescribimos los componentes de esta G de forma explícita, obtenemos:

$$h_a^{t+1} = h_a^t \frac{(W^T V)_a}{(W^T W h^t)_a}$$

Observemos que la regla preserva la no negatividad de h^t porque todos los términos involucrados son no negativos.

La demostración para la regla de actualización de W es completamente análoga.

□

Algoritmo de Lee y Seung con Divergencia KL El segundo problema que se plantean en su paper de 2001 Lee y Seung [9] es el siguiente:

Problema 2: Dada V matriz no negativa, minimizar $f(W, H) := \sum_{ij} V_{ij} \log \frac{V_{ij}}{WH_{ij}} - V_{ij} - WH_{ij}$ con respecto a W y H sujeto a la restricción de que estas sean no negativas.

La función de costo f por no ser simétrica no es una distancia y por eso se le llama "divergencia de Kullback-Leibler". A su vez, el algoritmo que proponen para resolver el problema es el siguiente:

Algoritmo de Lee y Seung para NMF con Divergencia de KL

1. **Inicialización:** Seleccionar aleatoriamente dos matrices no negativas W y H de dimensiones $n \times k$ y $k \times m$, respectivamente, donde m y n son las dimensiones de la matriz original V y r es la dimensión de la matriz de factorización.
2. **Actualización de W:** Calcular la matriz $H^T V$ y $H^T HW$. Actualizar W como $W \leftarrow W \odot \frac{(VH^T)}{(H^T HW)}$, donde $\odot$ denota la multiplicación elemento por elemento.
3. **Actualización de H:** Calcular la matriz VW^T y $WW^T H$. Actualizar H como $H \leftarrow H \odot \frac{(W^T V)}{(WW^T H)}$.
4. **Repetir:** Repetir los pasos 2 y 3 hasta que se alcance un criterio de parada. Los criterios de parada comunes incluyen un número máximo de iteraciones o una pequeña mejora en la función de costo.

La fundamentación del algoritmo surge de un teorema análogo al del Problema 1. No incluiremos su demostración por su similitud con el Teorema 1. Sin embargo, puede encontrarse en el texto de Lee y Seung [9].

***Teorema* 2.** *La divergencia de Kullback-Leibler* $f(W,H) := \sum_{ij} V_{ij} \log \frac{V_{ij}}{WH_{ij}} - V_{ij} - WH_{ij}$ *es no decreciente bajo las reglas de actualización*

$$W \leftarrow W \odot \frac{(VH^T)}{(H^T HW)}$$

$$H \leftarrow H \odot \frac{(W^T V)}{(WW^T H)}.$$

Además, f *es invariante bajo estas actualizaciones si y solo si* W *y* H *están en un punto estacionario de la distancia.*

3.2.2. La función de costo

El Problema 2 de la subsección anterior plantea el interrogante de cuál es el rol de la función de costo. La misma, típicamente tiende a seleccionar qué tipo de errores queremos penalizar en nuestras estimaciones/aproximaciones.

Si se considera el caso del Problema 1, la norma de Frobenius, si bien no carece de ventajas, tiende a darle un peso muy alto a la presencia de valores individuales particularmente altos. Es decir, la "distancia" entre dos matrices puede ser arbitrariamente grande cambiando un solo valor de una de ellas. Que esto sea una propiedad deseable o no, dependerá de la naturaleza misma del problema que se esté considerando. Por ejemplo, en algunos casos uno puede privilegiar el hecho de que haya un dato mal cargado no afecte a la totalidad de la estimación/aproximación. Concretamente, para la Norma de Frobenius, las matrices $\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$ y $\begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ tienen la misma distancia respecto a la identidad.

En conclusión, en el contexto de NMF modificar la función de costo nos puede permitir definir en qué sentido queremos que nuestra factorización se parezca a la matriz original. Por ejemplo, en el caso del procesamiento de imágenes, en la Figura 3.6 puede verse cómo ciertas modificaciones a la función de costo puede ser de mayor utilidad que la de Frobenius para limpiar ruidos "discretos"[2].

Por último, Dhillon y Sra en *"Generalized Nonnegative Matrix Approximations with Bregman Divergences"* [4] prueban versiones más generales de los teoremas y algoritmos de la sección anterior para funciones de costo que puedan ser definidas en términos de **Divergencias de Bregman**.

Definición **2.** *Dados S un conjunto cerrado y convexo y $\phi : S \to \mathbb{R}$ una función estrictamente convexa y diferenciable, su correspondiente* ***Divergencia de Bregman*** *$D_\phi : S \times S^o \to \mathbb{R}_+$ se define como:*

$$D_\phi(x, y) = \phi(x) - \phi(y) + \nabla\phi(y) \cdot (x - y)$$

Observemos que la divergencia de Bregman satisface las siguientes propiedades:

- No negatividad: $D_\phi(x, y) \geq 0$, con igualdad si y solo si $x = y$.
- Convexidad: $D_\phi(\lambda x_1 + (1 - \lambda)x_2, y) \leq \lambda D_\phi(x_1, y) + (1 - \lambda)D_\phi(x_2, y)$, para todo $\lambda \in [0, 1]$.
- Monotonía: $D_\phi(x, y) \geq D_\phi(x, z) + D_\phi(z, y)$.

En el trabajo, los autores se restringen a divergencias separables.

Figura 3.6: Reducción del ruido en la imagen. Primera fila: imágenes contaminadas. Segunda fila: Reducción de ruido con NMF con norma de Frobenius. Tercera fila: Reducción de ruido con NMF con función de costo con regularización L_1. Imagen obtenida de [2].

Definición **3.** *Dada una Divergencia de Bregman D_ϕ, decimos que esta es separable si la función ϕ se puede descomponer en una suma de funciones más simples, es decir:*

$$\phi(x, y) = \sum_{i=1}^{n} \phi_i(x_i)$$

Observemos que en este caso, la divergencia también se puede separar:

$$D_\phi(x, y) = \sum_{i=1}^{n} D_{\phi_i}(x_i, y_i)$$

Formalmente, dada la matriz V, el problema de optimización que se plantea en el trabajo es:

$$\min_{W,H \geq 0} D_\phi(WH, V) + \alpha(W) + \beta(H)$$

donde α y β sirven como penalizaciones para inducir regularizaciones como las de la Figura 3.6. En el trabajo, Dhillon y Sra [4] muestran teoremas y algoritmos completamente análogos a los de Lee y Seung [9], que son efectivamente una generalización de los resultados porque la norma de Frobenius corresponde a $\phi(x) = x^2$ y $\alpha = \beta = 0$, mientras que la de Kullback-Leidler corresponde a $\phi(x) = x \log(x)$ y $\alpha = \beta = 0$. Esto permite potencialmente extensiones de resultados obtenidos.

3.3. Algoritmo de Mínimos cuadrados alternados

Una alternativa más veloz a los algoritmos de actualización multiplicativa es conocido como mínimos cuadrados alternados (ALS, por sus siglas en inglés). Como su nombre lo indica, el método consiste en estimar alternadamente las matrices W y H con mínimos cuadrados de forma iterada. Es decir, resuelven el problema de $||V - WH||^2$ en dos etapas separadas. Como esto no garantiza la no negatividad de la matriz, esto se realiza de forma forzada, asignando 0 a todos los valores negativos. Como es de esperar, esto si bien es veloz, será problemático para garantizar la convergencia.

Algoritmo ALS con no negatividad forzada para la descomposición NMF

1. **Inicialización:**W será definida inicialmente como una matriz no negativa aleatoria o con algún criterio definido.
2. **Iteración:** Calcular H como aquel que minimiza $||V - WH||^2$ por mínimos cuadrados. Si algún valor es negativo se modifica a 0.

   Una vez calculada H, W será aquel que minimiza $||V - WH||^2$ por mínimos cuadrados. Si algún valor resulta negativo se modifica a 0.
3. **Criterio de cierre** Repetir el paso anterior hasta que se alcance el número máximo de iteraciones o W y H se mantengan inalterados bajo una tolerancia a establecer.

A lo largo del trabajo, con el uso del paquete *RcppML*[2] hemos utilizado este método, pues es el que permitió un equilibrio adecuado entre tiempo de ejecución y resultados adecuados. Sin embargo, cabe aclarar que existen numerosos debates respecto a su optimalidad. Por ejemplo, una discusión al respecto puede encontrarse en [7].

3.4. Interpretación como modelado de tópicos

Como se ha mencionado anteriormente, en el contexto del análisis de textos NMF puede ser interpretado como un modelado de tópicos. Para visualizar esto, observemos la Figura 3.7. En esta versión simplificada para efectos ilustrativos, tenemos un corpus de 4 documentos, que en total presentan 6 palabras y queremos separarlo en dos tópicos.

- La matriz W contendrá información de **qué peso de cada tópico tiene cada documento** . Por ejemplo, W_{11} y W_{12} tiene los pesos que tienen en el primer documento los tópicos 1 y 2, respectivamente.
- La matriz H codifica qué participación tienen las palabras del vocabulario en cada tópico. De esta forma, H_{11} y H_{21} representan el peso de la palabra 1 en el tópico 1 y 2 respectivamente.

Teniendo esto en cuenta, podríamos "recuperar" la matriz V a partir de W y H. Por ejemplo, para V_{11}, que representa el valor de la palabra 1 en

[2]https://cran.r-project.org/web/packages/RcppML/index.html

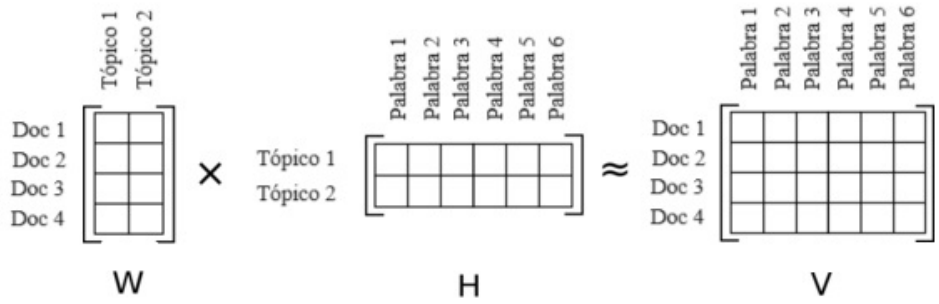


Figura 3.7: Descomposición NMF con detalles de filas y columnas.

el documento 1, lo estimaríamos con $W_{11}H_{11} + W_{12}H_{21}$, que no es más que considerar la proporción de tópicos 1 y 2 que compone el documento junto a la participación de la palabra 1 en cada tópico.

Aquí es menester hacer dos aclaraciones. En primer lugar, esta forma de "recuperar" la matriz V y de segmentar el documento 1 en dos tópicos tiene como subyacente la idea de que cuando se habla de determinadas temáticas, ciertas palabras se combinan juntas. Además, el etiquetado de estos tópicos se realiza *a posteriori* y de manera manual, analizando qué palabras tienen mayor peso en cada temática, lo cual estará condensado en la matriz H.

Por otra parte, hemos de hablar qué implica "recuperar" la matriz V. En el caso de las imágenes de los rostros se obtiene la imagen original con cierta pérdida de resolución. En ese sentido, W y H son una versión comprimida de la imágen original. En cuanto a los textos, solo se estaría recuperando cierta proporción de palabras, pero la información del orden de las palabras está completamente perdido en el momento que decidimos hacer un análisis puramente semántico de frecuencias. Con lo cual, el sentido de la "recuperación" y "compresión" del corpus original tiene un sentido muy limitado y el aspecto que realmente se capitaliza de utilizar este método está relacionado con su poder de segmentar los textos en tópicos.

Capítulo 4

Aplicaciones prácticas

4.1. Datos

Se dispone de discursos de presidentes y expresidentes de la República Argentina obtenidos por medio de técnicas de *scraping* realizadas en el lenguaje *Python* con la librería *BeautifulSoup*[1]. El resto del tratamiento de los datos fue realizado con el lenguaje R. Los discursos obtenidos son:

- 483 discursos de la vicepresidenta en curso **Cristina Fernandez de Kirchner** (CFK) en el lapso de tiempo desde 2008 hasta 2022, en el cual ocupó diversos roles: presidenta, vicepresidenta y senadora. De los mismos, 420 corresponden a su rol como presidenta y 63 a discursos fuera de su mandato. Los textos fueron obtenidos de su página personal oficial[2].
- 242 discursos del expresidente **Mauricio Macri** (MM) entre 2015 y 2019, es decir, en su período de gobierno. Los mismos fueron obtenidos de la página oficial de la Casa Rosada[3].
- 348 discursos del presidente en curso **Alberto Fernández** (AF) durante su período de gobierno hasta 2022. Los mismos también fueron obtenidos a través de la página de la Casa Rosada.

4.1.1. Tratamiento de los Datos

Como se comentó en el primer capítulo, es necesario hacer un tratamiento de datos. En primer lugar, se realizaron procedimientos típicos de limpieza

[1] https://pypi.org/project/beautifulsoup4
[2] https://www.cfkargentina.com/
[3] https://www.casarosada.gob.ar/informacion/discursos

de texto para la obtención de la matriz de frecuencias:

- Se detectaron y descartaron discursos en inglés.
- Se removieron signos de puntuación, números y caracteres especiales.
- Se descartaron *stopwords*.
- Se eliminaron los nombres propios y apellidos de los presidentes involucrados en el análisis.

Finalmente, se realizó un procedimiento de **lematización**, con una particularidad. En lugar de agrupar un conjunto de palabras asignándole una palabra genérica, se le asignará aquella que tenga mayor representatividad en el grupo. Por ejemplo, si tenemos las palabras "presidente", "presidenta" y "presidencia", en una primera instancia serán agrupadas bajo el término en infinitivo "presidir". Nosotros, en un segundo paso, **cambiaremos la etiqueta en infinitivo por aquella que tenga mayor cantidad de apariciones dentro del conjunto**. Esto se hace porque consideramos que en muchas ocasiones el infinitivo quita el espíritu de lo que la palabra original refiere, lo cual dificulta la interpretabilidad de lo obtenido.

4.1.2. Análisis exploratorio

Consideramos que es una buena práctica realizar un adecuado análisis exploratorio previo a la implementación de técnicas estadísticas. Por lo tanto, haremos un análisis similar al realizado a modo ilustrativo con los textos de Borges y Arlt.

En las Figuras 4.1, 4.2 y 4.3 pueden verse las palabras más frecuentes en términos absolutos tras haber removido *stopwords*. A primera vista, podemos observar que no hay marcadas diferencias entre los tres presidentes, con lo cual parece que será necesario utilizar una matriz TF-IDF.

Por otra parte, si hacemos un gráfico de dispersión como el de la Figura 2.5 en la 4.4 podemos ver cómo se arma un grupo de palabras con una frecuencia alta en Alberto Fernández y relativamente baja en Cristina Kirchner. Es comprensible, pues estas palabras son ligadas a la pandemia de COVID-19, como "aplausos", "enorme", "contagios" y "virus". En la Figura 4.5 puede verse que este efecto desaparece cuando nos restringimos a discursos durante la presidencia de Alberto Fernández, como es de esperar.

Otro aspecto a analizar es la longitud y diversidad de vocabulario en los discursos.

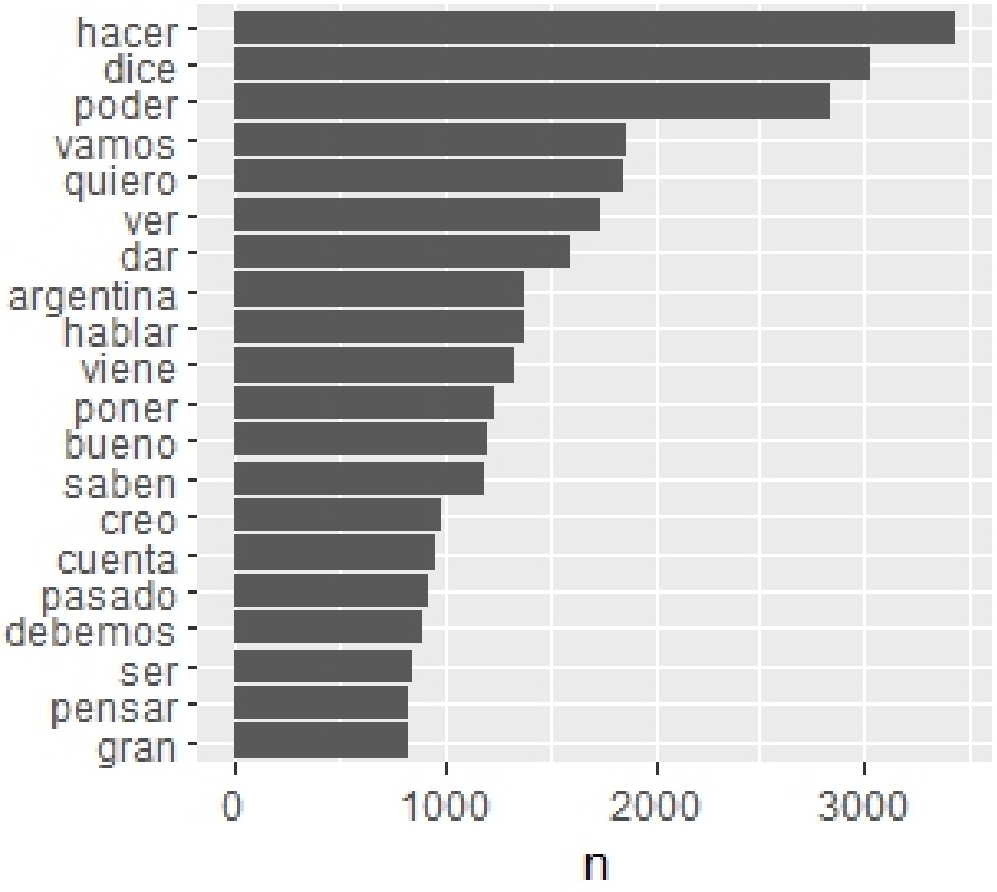


Figura 4.1: 20 palabras más frecuentes en los discursos de Cristina Kirchner.

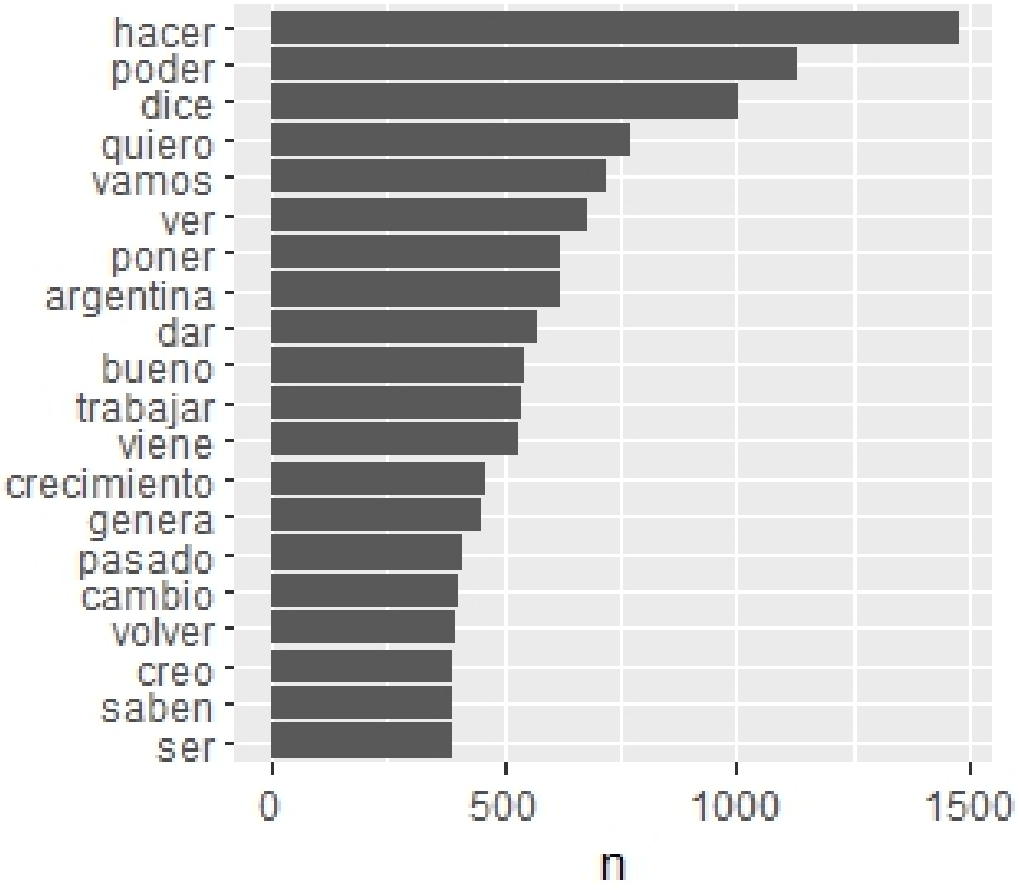


Figura 4.2: 20 palabras más frecuentes en los discursos de Mauricio Macri.

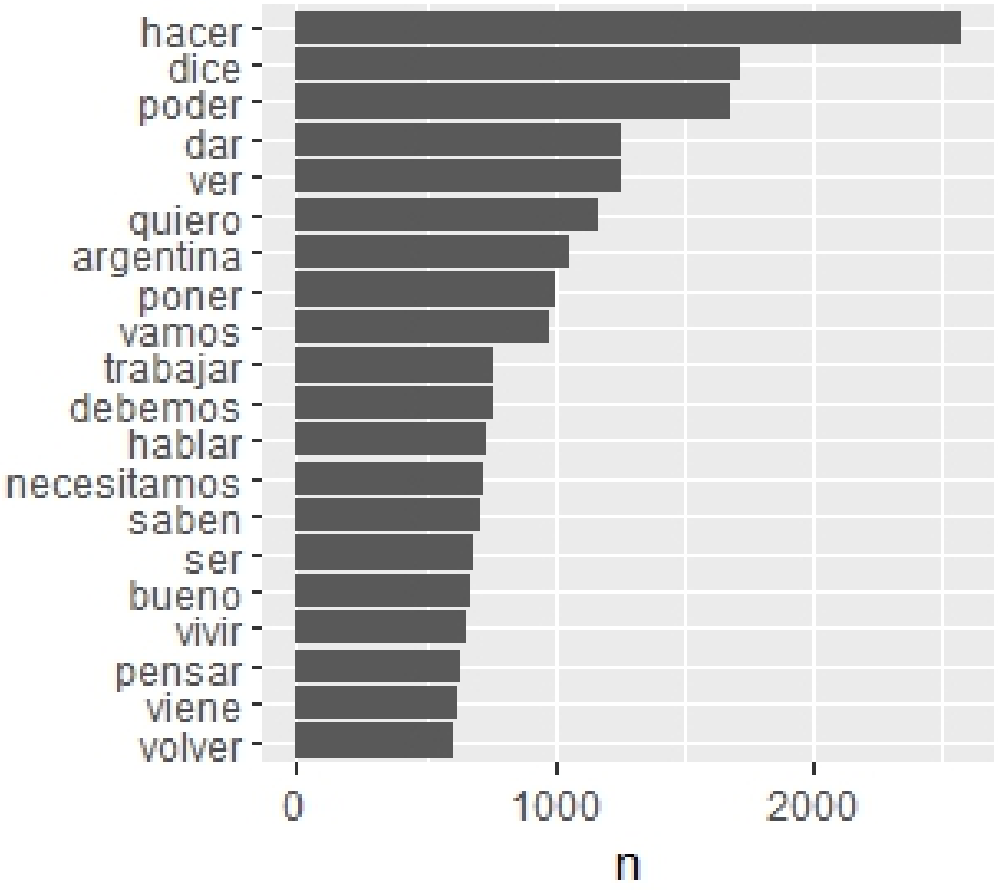


Figura 4.3: 20 palabras más frecuentes en los discursos de Alberto Fernandez.

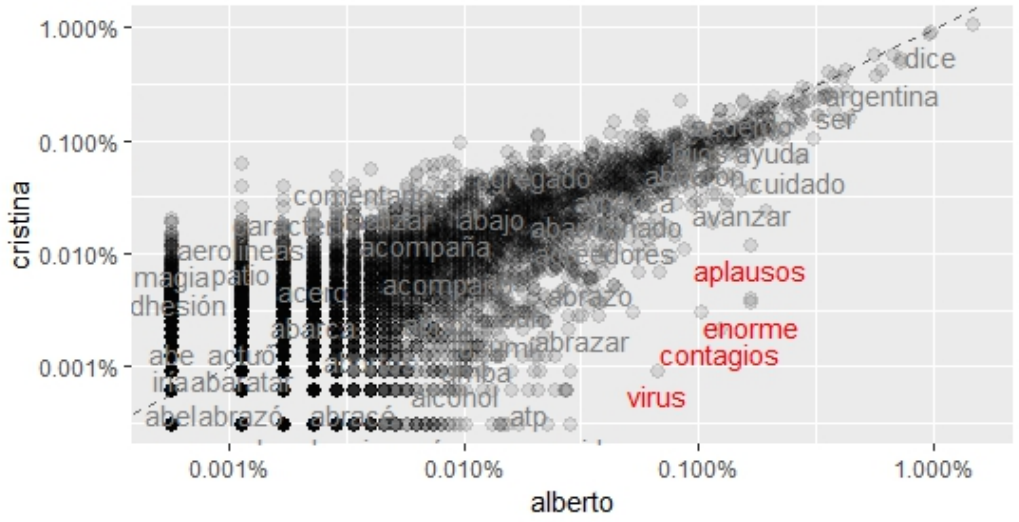


Figura 4.4: Gráfico de dispersión de proporción de palabras según si es de Cristina Kirchner o Alberto Fernández. En rojo, un grupo de palabras asociadas a la pandemia. Escala logarítmica.

Presidente	Palabras por discurso	Palabras distintas por discurso
Cristina F. Kirchner	1357	682
Mauricio Macri	711	430
Alberto Fernández	891	500

Cuadro 4.1: Cuadro de palabras promedio por discurso de cada presidente.

4.2. Análisis por año

El primer análisis que realizaremos será por año. Es decir, generamos una matriz de frecuencias relativas de palabras A en la cual cada fila será el año y cada columna el término lematizado, excluyendo *stopwords*. A_{ij} indicará el porcentaje de veces que aparece el término j en el año i. Simultáneamente, se considerará la matriz TF-IDF restringida a que todas las filas sumen 1. Esta matriz será llamada B. **En ambos casos, se eliminarán las columnas propias de palabras que solo aparecen en un texto.**

Como disponemos de discursos de 2008 a 2022, ambas matrices tendrán 15 filas. Además, debido al vocabulario obtenido, encontramos 9272 columnas.

Comenzaremos haciendo una descomposición NMF de la matriz A, luego de la matriz B y compararemos resultados.

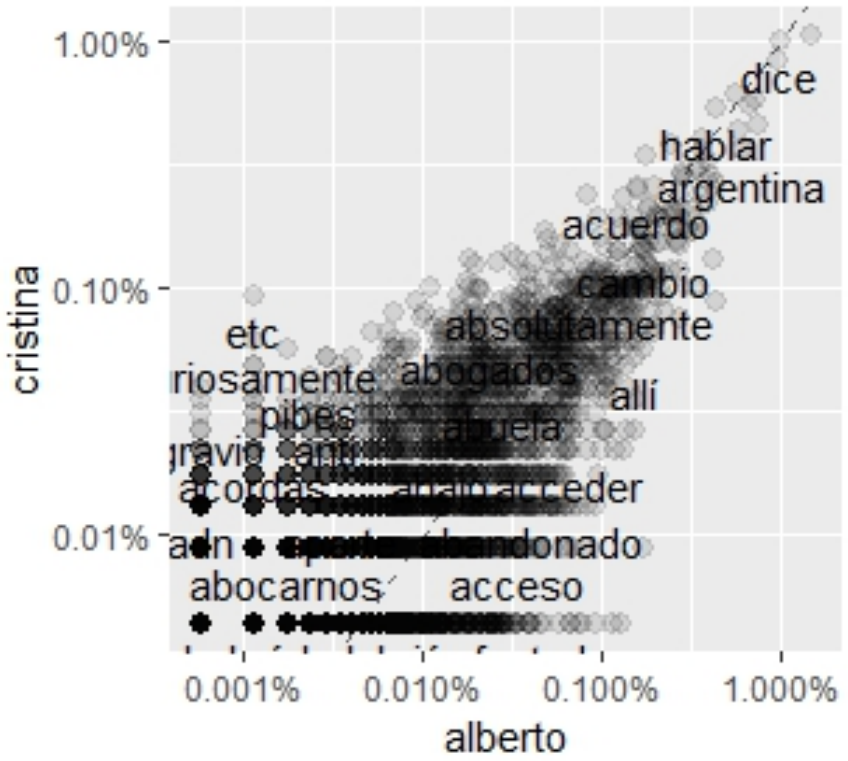


Figura 4.5: Gráfico de dispersión de proporción de palabras según si es de Cristina Kirchner o Alberto Fernández. Se excluyen discursos anteriores a 2020. Escala logarítmica.

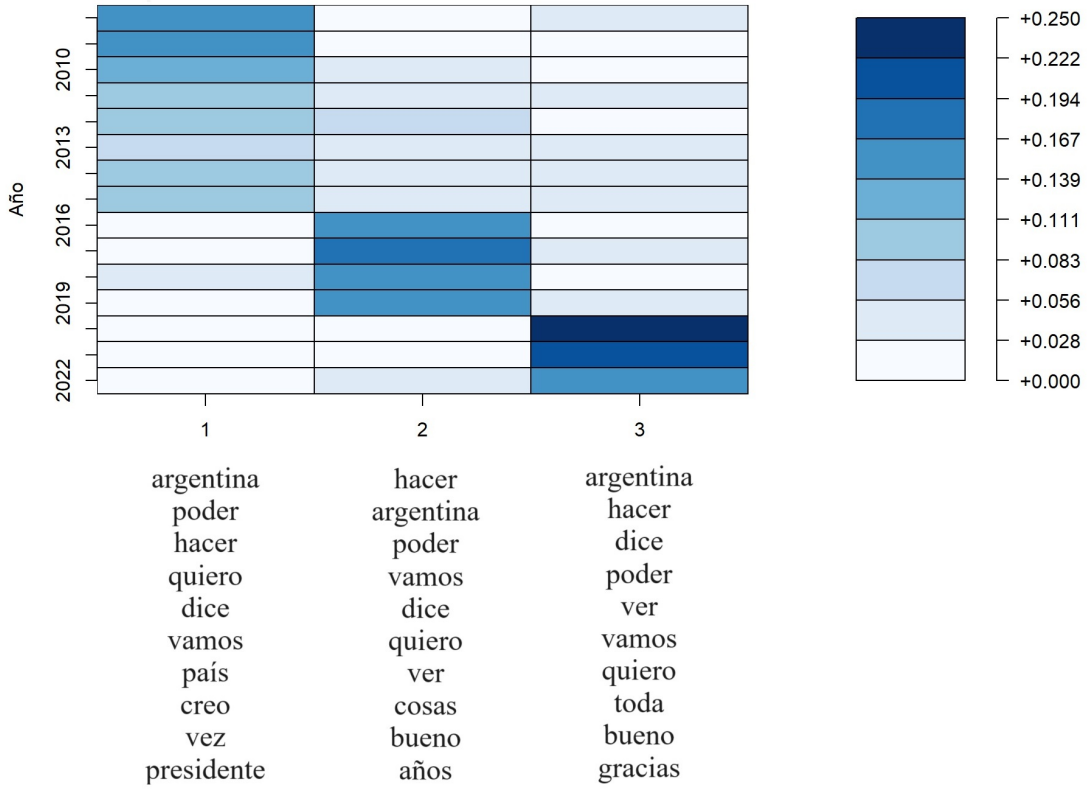


Figura 4.6: Mapa de calor de Matriz W_A con $k = 3$. Se muestran las 10 palabras de mayor peso en cada tópico. Observar que el primer tópico corresponde a las presidencias de Cristina Fernández de Kirchner, el segundo y tercero a la de Mauricio Macri y Alberto Fernández, respectivamente.

4.2.1. Matriz A de frecuencias relativas con $k = 3$

Aproximaremos a la matriz $A \in \mathbb{R}^{15\times 9292}$ como producto de dos matrices no negativas $W_A \in \mathbb{R}^{15\times k}$ y $H_A \in \mathbb{R}^{k\times 9292}$.

$$A \approx W_A H_A$$

Para la estimación de W_A y H_A utilizaremos la función *nmf* del paquete *RcppML*[4], que utiliza el algoritmo ANLS. Nosotros seleccionaremos $k = 3$ porque sabemos *a priori* que tenemos tres autores y veremos qué es lo que ocurre.

[4]https://cran.r-project.org/web/packages/RcppML

Observemos en la Figura 4.6 que los resultados obtenidos son consistentes con los períodos presidenciales, pues los años correspondientes a la presidencia de CFK (en el dataset, serían los discursos de 2008-2015), tendrán mayor peso de la primera columna. Mientras que aquellos correspondientes a la presidencia de MM tendrán un peso de la segunda y los de AF, de la tercera. Podemos decir que, *en algún sentido*, la factorización logra armar *clústers* coherentes con los períodos presidenciales. Un segundo aspecto a considerar es las palabras asociadas de mayor peso asociadas a cada uno de esos *clústers*. Esta información está contenida en la matriz H_A y serían, para cada fila, las palabras con mayor valor. En el Figura 4.6 se ve el ranking de palabras por fila. Puede observarse que en los tres casos se trata de palabras muy frecuentemente utilizadas por los tres presidentes que no logran identificarlos, como se vio en el análisis exploratorio. Esto dificulta la interpretación tradicional de NMF como descomposición de tópicos y **se debe principalmente a que no hay un peso IDF en la matriz que les reste importancia a dichos términos**. Finalmente, cabe aclarar que se han probado otros valores de k pero se consideró que no aportaban demasiado al análisis en este caso.

Del análisis de esta subsección podemos concluir que el algoritmo de NMF sobre una matriz de frecuencias relativas sin pesos TF-IDF se muestra eficiente para separar los períodos presidenciales de forma muy marcada. Sin embargo, de forma simultánea carece de cierta interpretabilidad en los tópicos.

4.2.2. Matriz TF-IDF con $k = 3$

Repetimos el análisis realizado en la subsección anterior para la matriz B normalizada por filas. En este caso, la factorizamos como $B \approx W_B H_B$.

En la Figura 4.7 parece diferenciarse claramente el período de AF de los demás períodos en la columna 2. Sin embargo, parece difícil discernir entre la presidencia de CFK y la de MM. Sin embargo, si graficamos $log(W_b)$, en la Figura 4.8 puede apreciarse que la columna 3 parece captar más el primer período de la presidencia de CFK y abarca temáticas más vinculadas con las relaciones internacionales e incluso una clara referencia al conflicto con el campo de 2008. En contraste, la primera columna incluye la presidencia de MM y la de CFK a partir de 2010, pero resulta difícil establecer una temática común a las palabras propias del tópico.

Por otro lado, puede llamar la atención la presencia de palabras como "Hilton". La explicación del fenómeno es que el peso TF-IDF tiende a ponderar en valores elevados a palabras demasiado infrecuentes. Si bien acarrea muchos beneficios que ya hemos mencionado, también puede traer consigo ciertas distorsiones indeseables.

En lo que respecta a las principales palabras por fila de H_B, los resultados son mucho más interpretables en lo que respecta al significado. En la Figura 4.7 puede verse que la fila 2 tiene palabras fuertemente ligadas a la pandemia, temática que claramente dominó al período presidencial de AF. Por otro lado, no aparecen términos comunes a todos los presidentes como "argentina", "hacer" o "poder", lo cual sí ocurría en el caso de la matriz A. Esto parece sugerir que el hacer pesos TF-IDF genera una interpretación más ligada a la tradicional, es decir, de cada columna asociada a un tópico. Esta última interpretación sugiere incorporar un mayor número de tópicos al análisis.

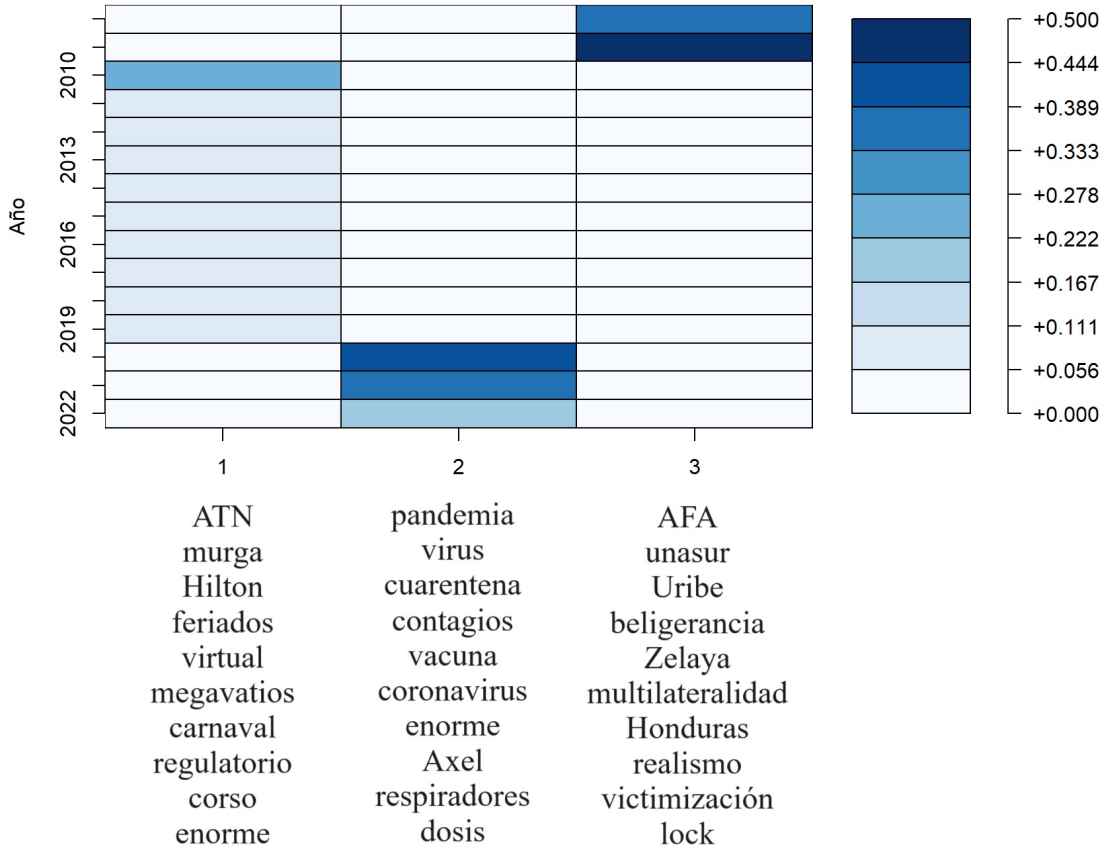


Figura 4.7: Mapa de calor de Matriz W_B, que proviene de factorizar matriz TF-IDF B con $k = 3$. Se muestran las 10 palabras de mayor peso en cada tópico.

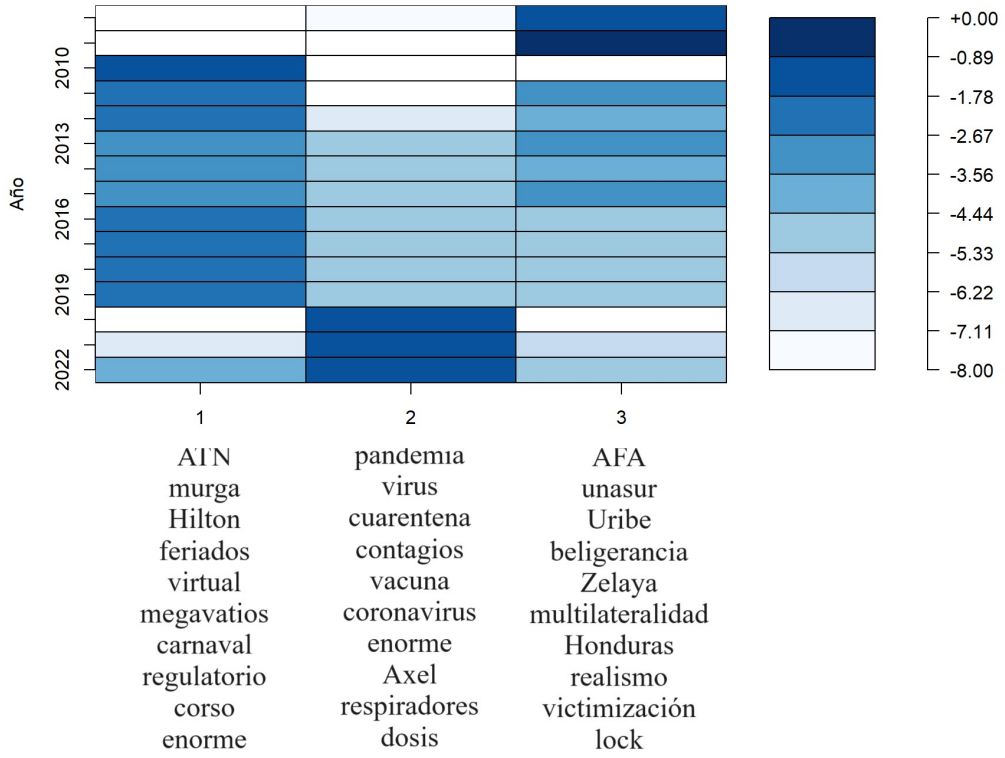


Figura 4.8: Mapa de calor del **logaritmo** de la Matriz W_B, que proviene de factorizar matriz TF-IDF B con $k = 3$. Se muestran las 10 palabras de mayor peso en cada tópico.

4.2.3. Matriz TF-IDF con $k = 15$

A efectos de tener una mayor cantidad de tópicos, resulta de particular interés el caso en el que $k = 15$, pues, con la excepción de 2013/2014 y 2020/2021, asigna un tópico a cada año. Esto puede apreciarse en la Figura 4.9 y las palabras con mayor peso asignado a cada tópico en el Cuadro 4.2. Observemos cómo en términos generales las palabras tienen un correlato directo con los tópicos que marcaron la agenda política de cada año. Por mencionar algunos casos, 2008 y el conflicto del campo o 2013/2014 y el fallo de Griesa respecto a la deuda externa Argentina. En el caso de 2020 a 2022 vemos la evolución de los términos relacionados con la pandemia, mientras que en 2020 hay una fuerte mención de la palabra "cuarentena", en 2021 aparecen términos como "dosis" o "vacuna".

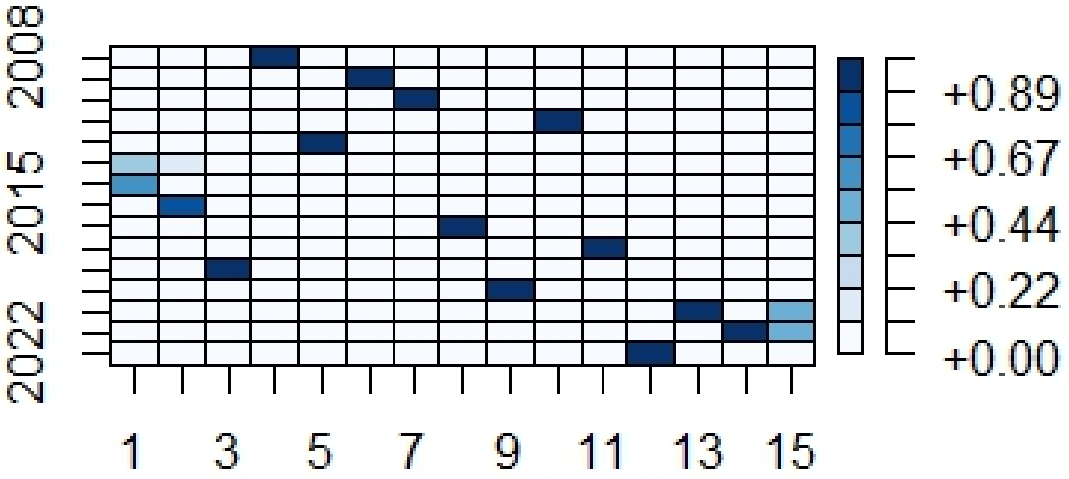


Figura 4.9: Mapa de calor de la Matriz W_B con $k = 15$

Por otra parte, en el Cuadro 4.2, podemos observar que 2018 contiene palabras como "tarifazo", propias de la crítica de CFK a la política de precios de la energía del gobierno de 2015-2019, en simultáneo con otras como

Año	P1	P2	P3	P4	P5
2008	realismo	lock	elusión	espejitos	admisión
2009	AFA	unasur	uribe	zelaya	beligerancia
2010	ATN	hilton	murga	feriados	virtual
2011	roy	fitz	pintores	palestina	veto
2012	divorcio	rattenbach	codificación	FONID	subtes
2013/2014	griesa	canje	fiduciario	Pro.Cre.Ar	reestructuración
2015	nisman	nuclear	megavatios	quintil	sable
2016	enorme	estudiantil	canadienses	herramientas	toti
2017	atajo	enorme	herramientas	desafuero	aplausos
2018	tarifazo	enorme	tormenta	pymes	rajoy
2019	libro	marcelo	eugenia	fmi	enorme
2020	cuarentena	virus	coronavirus	pandemia	contagios
2020/2021	pandemia	contagios	virus	respiradores	vacuna
2021	pandemia	dosis	vacuna	enorme	axel
2022	pandemia	aplausos	enorme	caribe	ercolini

Cuadro 4.2: Palabras con mayor peso en cada tópico asignado a cada año en la descomposición con 15 tópicos.

"tormenta", que se hallan principalmente en los discursos de MM en referencia a las recurrentes devaluaciones de los últimos años de su presidencia. Esta simultaneidad de discursos de un presidente (en este caso, Mauricio Macri) con los de Cristina Fernández sin un cargo presidencial genera en algún sentido dos grupos dentro de dicho período presidencial. Lo mismo ocurre con los discursos del mandato de Alberto Fernández. Por ello, en la próxima sección excluiremos los discursos de CFK fuera de su rol presidencial y compararemos resultados.

En conclusión, podemos ver que, mientras que la matriz A se muestra eficiente para diferenciar períodos presidenciales **de forma no supervisada**, la matriz TF-IDF nos aporta un agrupamiento más "semántico" que permite un análisis cualitativo de las palabras que identifican determinados textos.

4.2.4. Matriz TF-IDF restringida a discursos durante roles presidenciales con $k = 3$

Finalmente, se repetirá el análisis realizado en la subsección anterior, excluyendo los discursos de Cristina Fernández fuera de su rol presidencial, lo cual descarta 63 discursos. Solo se analizarán los resultados en el caso de la ponderación TF-IDF, puesto que en el caso de las frecuencias relativas los resultados son esencialmente indistinguibles.

En la Figura 4.10 podemos ver el mapa de calor para $k = 3$ y observar cómo se recupera la división de períodos presidenciales. En el caso del tópico correspondiente a la presidencia de CFK, se ve una preponderancia del primer período. Por otra parte, vemos que las palabras del Cuadro **??** vemos que efectivamente hay un fuerte vínculo entre los tópicos y las presidencias correspondientes.

Si repetimos el caso particular en el que la cantidad de tópicos es igual a la cantidad de años disponibles, podemos asignar un tópico a cada año. En el Cuadro 4.3 vemos las primeras 5 palabras en términos de pesos de la matriz H para cada año. Como fue comentado, las mayores diferencias se encuentran en el período de MM. En particular, en el año 2018 desapareció la palabra “tarifazo” del primer lugar, que fue ocupado por “tormenta”.

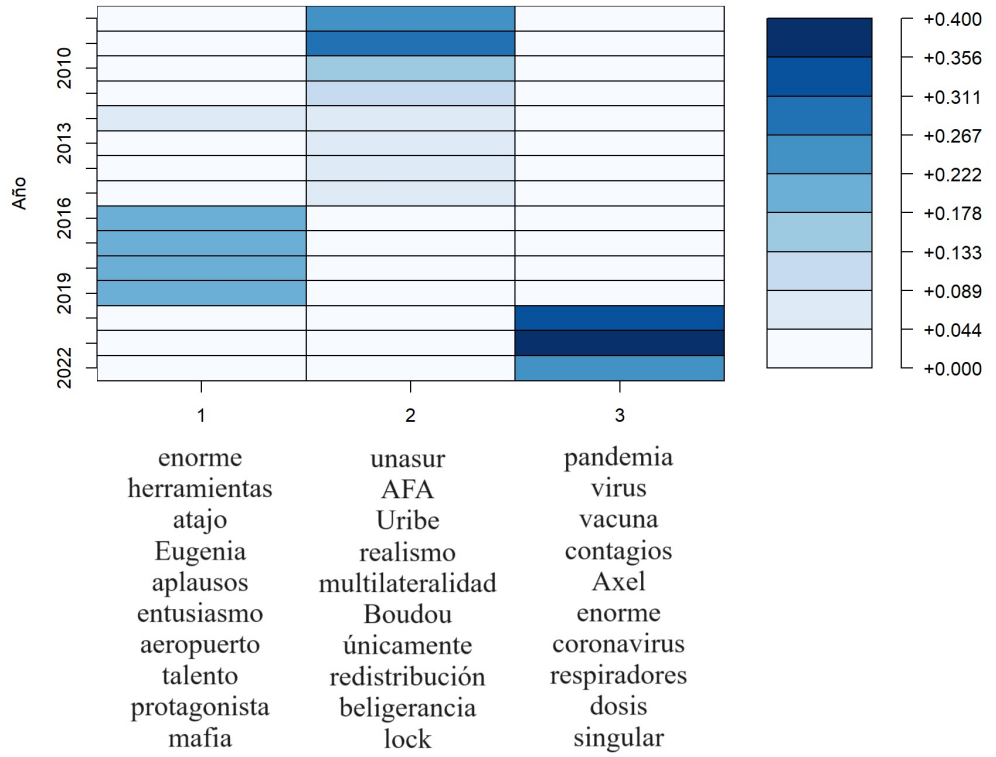


Figura 4.10: Mapa de calor de la Matriz W_B **excluyendo discursos de CFK fuera de su rol presidencial** con $k = 3$.

Año	P1	P2	P3	P4	P5
2008	realismo	lock	elusión	out	boudou
2009	unasur	AFA	uribe	beligerancia	dispositivos
2010	ATN	murga	hilton	feriados	desendeudamiento
2011	roy	fitz	pintores	formidable	veto
2012	divorcio	FONID	rattenbach	subtes	vélez
2013	cautelares	virgen	santacruceños	turbio	afortunadamente
2014	griesa	buitre	fíjense	canje	reestructuración
2015	nisman	nuclear	compatriotas	rusa	megavatios
2016	enorme	canadienses	herramientas	trudeau	toti
2017	atajo	enorme	herramientas	mafia	aplausos
2018	tormenta	enorme	davos	rajoy	herramientas
2019	enorme	eugenia	mechita	atajo	aplausos
2020	pandemia	virus	contagios	coronavirus	axel
2021	pandemia	vacuna	virus	dosis	contagios
2022	pandemia	aplausos	enorme	caribe	singular

Cuadro 4.3: Palabras con mayor peso en cada tópico asignado a cada año en la descomposición con 15 tópicos, excluyendo discursos de CFK fuera de su rol presidencial.

4.3. Análisis por discurso

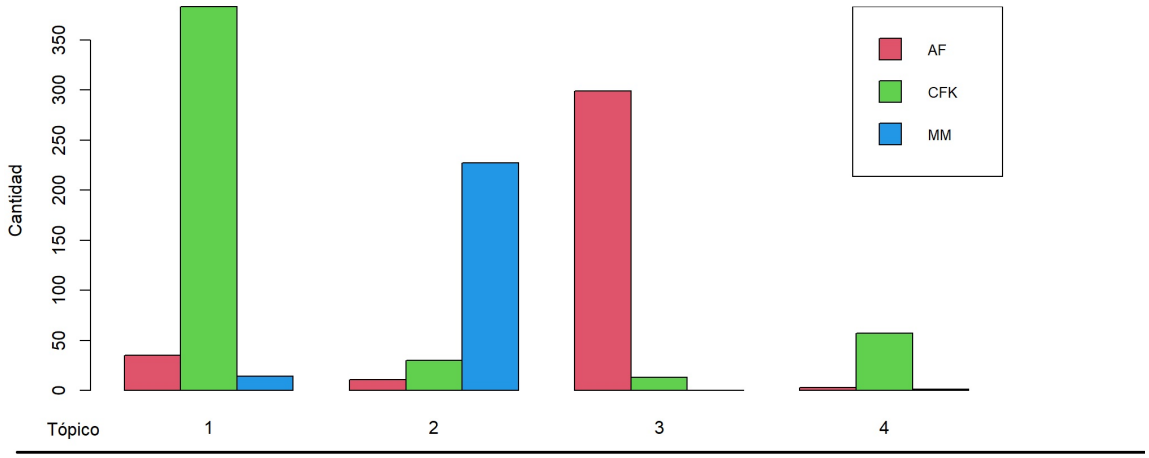
En este caso, haremos un análisis por discurso, es decir, cada fila de nuestra matriz, que llamaremos V, representará un discurso. En esta sección, trabajaremos con la matriz TF-IDF normalizada por filas. Factorizamos $V \approx W_V H_V$ con la función *nmf* utilizada anteriormente. La matriz V tendrá 1073 filas y 9272 columnas. Esta cantidad de filas implicará que no podrán realizarse visualizaciones con mapas de calor con la misma facilidad que en la sección precedente. Por lo tanto, para visualizar los resultados, a cada discurso le asignamos un tópico, que es aquel que tiene mayor valor en la fila correspondiente de la matriz W_V. Formalmente, el tópico del discurso i es el j que maximiza W_{Vij}. Manualmente, hemos determinado un $k = 4$. Es decir, cada discurso tendrá asignado una temática, lo que nos permite ver relaciones entre cada presidente y el tópico que tiene asignado. En resumen:

1. Se realiza factorización NMF con 4 tópicos.
2. A cada discurso se le asigna el tópico de mayor presencia en el mismo.
3. Se observa la distribución de presidentes por cada tópico en la Figura 4.11

En la Figura vemos una separación muy marcada por cada presidente, en la cual los tópicos 1 y 4 tienen una mayor presencia de CFK; el 2, de MM y el 3,de AF. Por otra parte, puede advertirse que los tópicos 1 a 3 pueden relacionarse en su contenido semántico con los presidentes que enunciaron cada discurso.

Cristina Fernández de Kirchner: Para el caso de CFK, el tópico 1 abarca temáticas relacionadas con los ideales vinculados con la democracia. Por ello encontramos términos como “república”, “derecho” “democracia”, “votar”, “pueblo”, “patria”. Además, podemos encontrar referencias a las fuerzas armadas y la deuda, que son temas con cierta recurrencia en sus discursos.

Por otra parte, en el tópico 4 encontramos palabras que tienen menor relación entre sí. Sin embargo, al hacer una búsqueda en los textos puede verse que en muchos casos están vinculadas a expresiones típicas de la vicepresidenta, por ejemplo, la palabra “selectividad” se vincula con la “selectividad de la información”, que menciona en numerosos discursos refiriéndose a los medios de comunicación; o “nariz” con la expresión “llevar de la nariz”, que curiosamente repite con cierta frecuencia y es un efecto potenciado por el peso TF-IDF, ya que los otros presidentes no la utilizan nunca.



república	energía	vacuna	selectividad
defensa	obras	pandemia	catastróficas
derecho	equipo	contagios	nariz
armadas	eugenia	salud	internacional
democracia	maría	axel	organizado
votar	conectar	virus	acontecimiento
deuda	camino	aplausos	curso
pueblo	ruta	cuidado	planeta
países	futuro	hospital	formación
patria	aplausos	universidad	adquirido

Figura 4.11: Tópico asignado a cada discurso según presidente para $k = 4$. Debajo, las diez palabras de mayor peso en cada tópico.

Mauricio Macri: Aquí puede argumentarse la presencia de referencias a cuestiones más vinculadas a la construcción y la obra pública. En esta línea están los términos "energía", "obras", "camino", "ruta". Además, hay dos términos que refieren directamente a la en aquel entonces gobernadora de Buenos Aires, María Eugenia Vidal.

Alberto Fernández: A esta altura, no ha de sorprendernos que los términos del tópico correspondiente a Alberto Fernández sean directamente vinculados con la pandemia. Por otra parte, resulta llamativa la presencia simultánea del gobernador de la provincia de Buenos Aires, Axel Kicillof.

Finalmente, cabe mencionar que este análisis cualitativo de las palabras

en los discursos no pretende ser exhaustivo, pero consideramos que tiene la utilidad de ilustrar cómo el análisis de tópicos puede servir como herramienta para profesionales de las ciencias sociales.

4.3.1. Utilización de NMF para clasificación de discursos

Un nuevo enfoque que pretendemos aplicar en esta subsección es utilizar la matriz W_V como input para un algoritmo de clasificación. Es decir, de cada discurso disponemos de:

- El nombre del presidente que lo enunció.
- El vector de la fila correspondiente al discurso en la matriz W_V que nos indica el peso de cada tópico para dicho discurso.

Es importante destacar que **en ningún momento del proceso de cálculo de la matriz W_V se tuvo en cuenta qué presidente pronunció el discurso, por lo tanto, si este se puede predecir a partir de la matriz, esto será debido puramente al contenido del texto.**

Se ha decidido utilizar el algoritmo de clasificación conocido como K vecinos más cercanos (KNN, por sus siglas en inglés). El algoritmo KNN determina la clase o el valor objetivo de una instancia de datos de entrenamiento al encontrar las K instancias de datos más cercanas en el espacio de entrenamiento. La "cercanía" será calculada utilizando la distancia euclidiana. Una vez que se encuentran las K instancias más cercanas, el algoritmo asigna la etiqueta más común. Un desarrollo teórico del mismo puede encontrarse en [6].Para implementar el algoritmo utilizaremos la función *knn* del paquete *class*[5].

Se ha decidido un valor de $K = 10$ del algoritmo KNN por medio de prueba y error. Para evitar el riesgo de sobreajuste se realizó el método conocido como *cross-validation*. El mismo consiste en dividir los datos en 5 subconjuntos aleatorios. Luego, en cada uno de los subconjuntos se predice qué presidente expresó cada discurso **con la información de los demás subconjuntos**. Esto nos permite obtener una matriz de confusión que puede observarse en el Cuadro 4.4. La misma implica una tasa de aciertos del 92,6 %, la cual teniendo en cuenta la distribución de autoría de los discursos es considerablemente elevada.

Hasta aquí, habíamos fijado la cantidad de tópicos en 4, sin embargo, lo que podemos hacer es repetir este mismo proceso para distintos valores de

[5]https://cran.r-project.org/web/packages/class

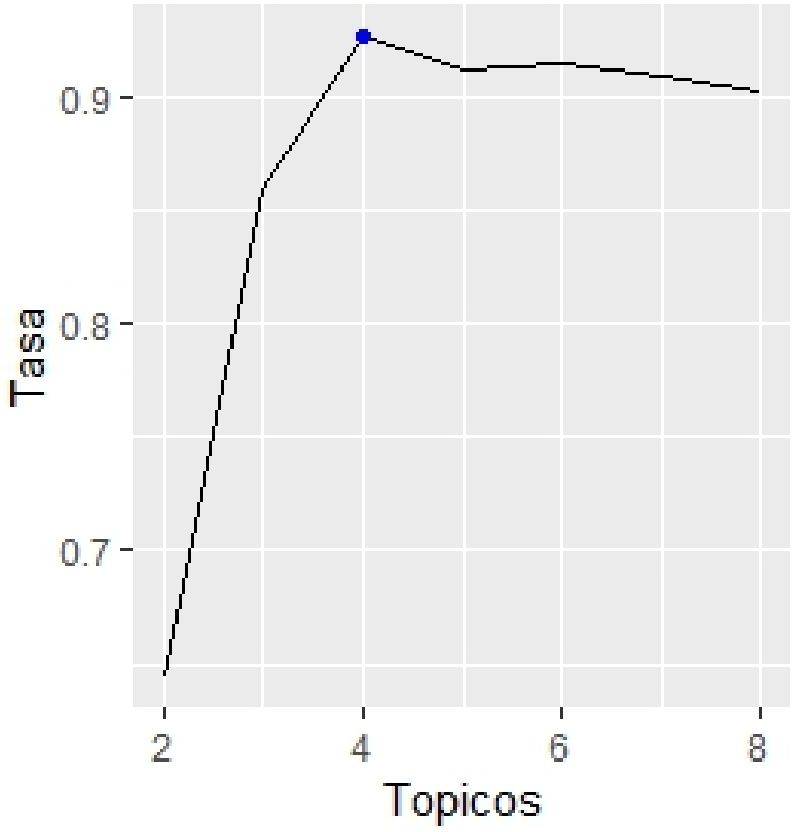


Figura 4.12: Tasa de aciertos con cross-validation de 5 pliegues en función de cantidad de tópicos. El punto azul muestra el máximo que se da con 4 tópicos.

k y calcular la tasa de aciertos del algoritmo (cabe recordar en este punto que la k minúscula es la cantidad de tópicos y K mayúscula, la cantidad de vecinos). Esto lo graficamos en la Figura 4.12, en la cual puede verse que el máximo se alcanza efectivamente en el cuarto tópico. Con lo cual, con este criterio, la selección de 4 tópicos parece adecuada. Además, se puede observar que, en los primeros pasos, el aporte marginal que hace cada tópico es considerable puesto incorporar dos tópicos se obtiene una tasa del 64,9 % y con tres, 85,9 %. Tras haber considerado 4 tópicos, la tasa se estaciona alrededor del 92 %.

Otra observación pertinente es que hay una muy adecuada diferenciación entre los discursos de Alberto Fernández y Mauricio Macri, mientras que los errores suelen provenir de textos de Cristina Fernández. Una primera explicación que se tuvo está basada en las consideraciones previas acerca de

	CFK	MM	AF
Predicho CFK	457	21	25
Predicho MM	9	219	5
Predicho AF	17	2	318

Cuadro 4.4: Tabla de Confusión del Algoritmo KNN sobre la matriz W de NMF realizada con 4 tópicos y evaluada por medio de cross-validation con 5 pliegues.

que CFK es la única de la que disponemos de discursos en todo el período. Es decir, la ausencia de discursos contemporáneos de AF y MM debería contribuir a facilitar la tarea de diferenciarlos. Sin embargo, se repitió el proceso excluyendo los discursos de CFK fuera de su rol presidencial y no se encontraron diferencias significativas.

4.3.2. Utilización de PCA para la clasificación de discursos

Como método alternativo a NMF dentro del conjunto de técnicas matriciales, repetiremos el análisis de componentes principales realizado en el capítulo 2 en las obras de Borges y Arlt. El algoritmo de PCA fue utilizado con la función interna de R *prcomp* sobre la matriz estandarizada V, la cual estaba ponderada con TF-IDF, excluyendo aquellos términos que aparecen únicamente en un discurso. Cuando decimos que la matriz está escalada nos referimos a que restamos la media y dividimos por el desvío estándar de cada columna, puesto que PCA es muy sensible a cambios de escala. Si hacemos el gráfico de dispersión en la Figura 4.13 de la primera componente contra la segunda, observamos que los discursos de los distintos presidentes tienden a agruparse. Sin embargo, como se observó en el análisis de Borges y Arlt, esta separación no tiene por qué ser más marcada en las primeras componentes, por lo tanto, en la Figura 4.14 vemos que la primera contra la **tercera** componente generan una separación aún más clara. Por lo tanto, *a priori* podemos pensar que tomar como input de un modelo de clasificación a la matriz reducida por PCA debería producir buenos resultados.

Podemos analizar en el histograma de la Figura 4.15 cómo la primera componente logra separar los discursos de CFK (asociado a valores positivos) de los de AF y MM (asociado a valores negativos). Nuevamente aparece identificada con CFK la palabra "República", además de ciertos términos relacionados con la justicia, como "juez", "embargo" y "causa". En el caso de la parte negativa de la componente es difícil encontrar un hilo semántico que las unifique. En resumen, podemos decir que la primera componente brinda información relevante a la hora de determinar si un discurso es de CFK.

Por otra parte, 4.16 parece ser relevante para diferenciar discursos de AF y MM. Nuevamente, aparece en primer lugar la palabra "pandemia" identificada con AF. Sin embargo, las 9 palabras restantes no guardan relación con el COVID. Por otra parte, las palabras correspondientes a MM, son más bien ligadas a la economía en muchos sentidos ("inversión", "ciento", "pesos", "empresas", etc.).

Para determinar cuántas componentes involucrar al análisis, repetimos el mismo proceso de *cross-validation* con 5 pliegues de la subsección anterior, medimos la tasa de aciertos para las primeras m componentes con m entre 1 y 15 y graficamos el resultado. El mismo puede verse en la Figura 4.17. Podemos observar cómo en la incorporación de las primeras cinco componentes hay una ganancia considerable y luego la misma se estaciona en 98 %.

El resultado medido en tasa de aciertos es ostensiblemente mejor tomando

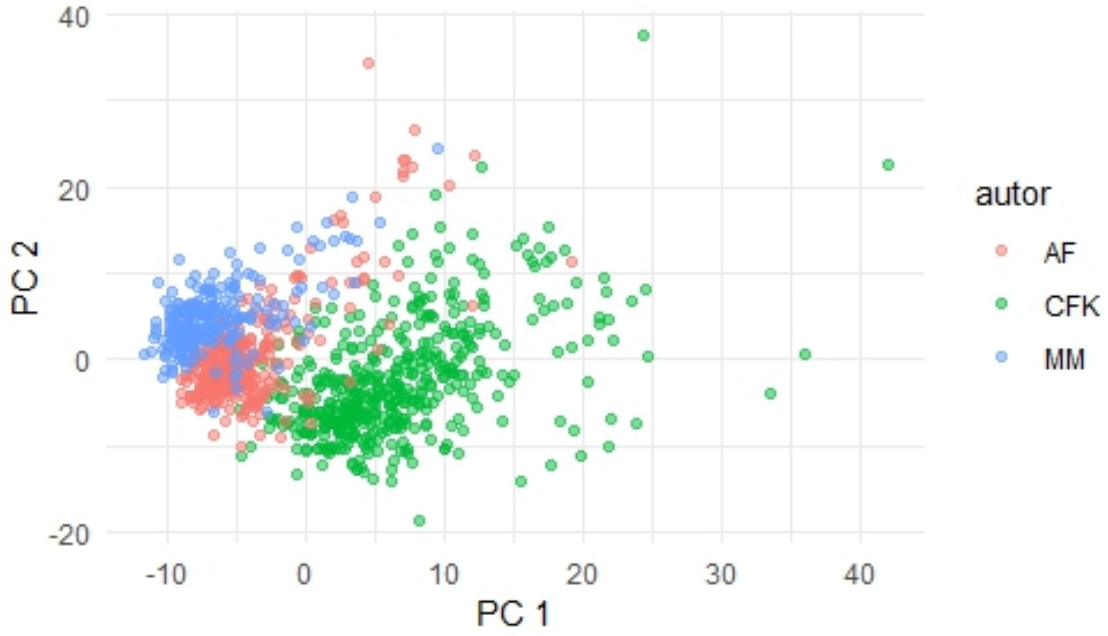


Figura 4.13: Gráfico de dispersión de primera componente contra segunda componente de PCA de la matriz V.

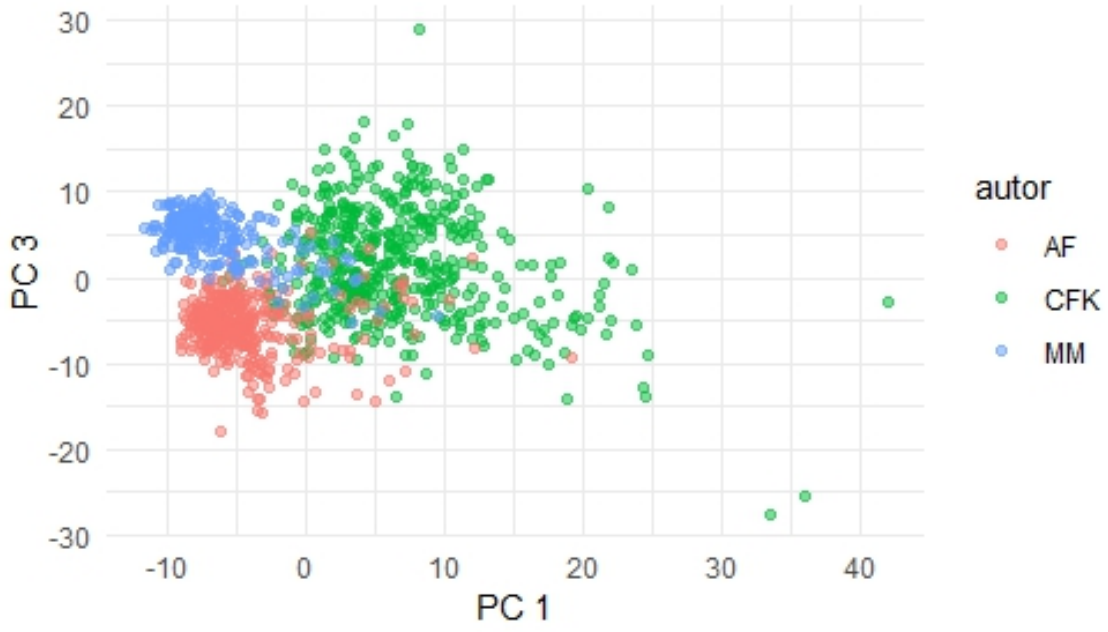


Figura 4.14: Gráfico de dispersión de segunda componente contra tercera componente de PCA de la matriz V.

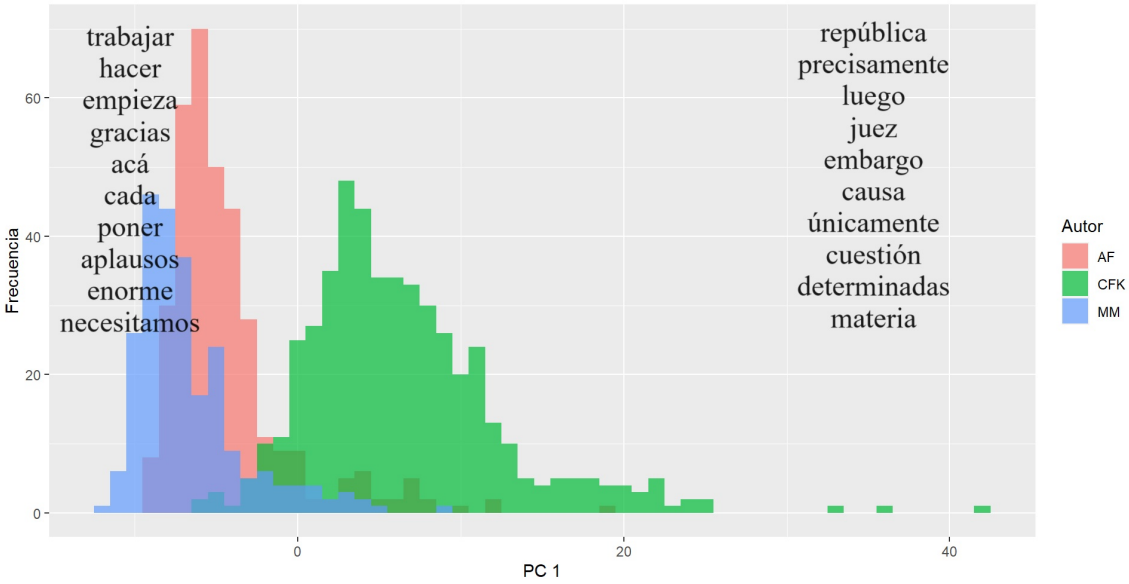


Figura 4.15: Histograma de primera componente por cada presidente. Encontramos a la derecha y a la izquierda las palabras que posee mayor ponderación en valor absoluto para generar valores respectivamente positivos y negativos en la componente.

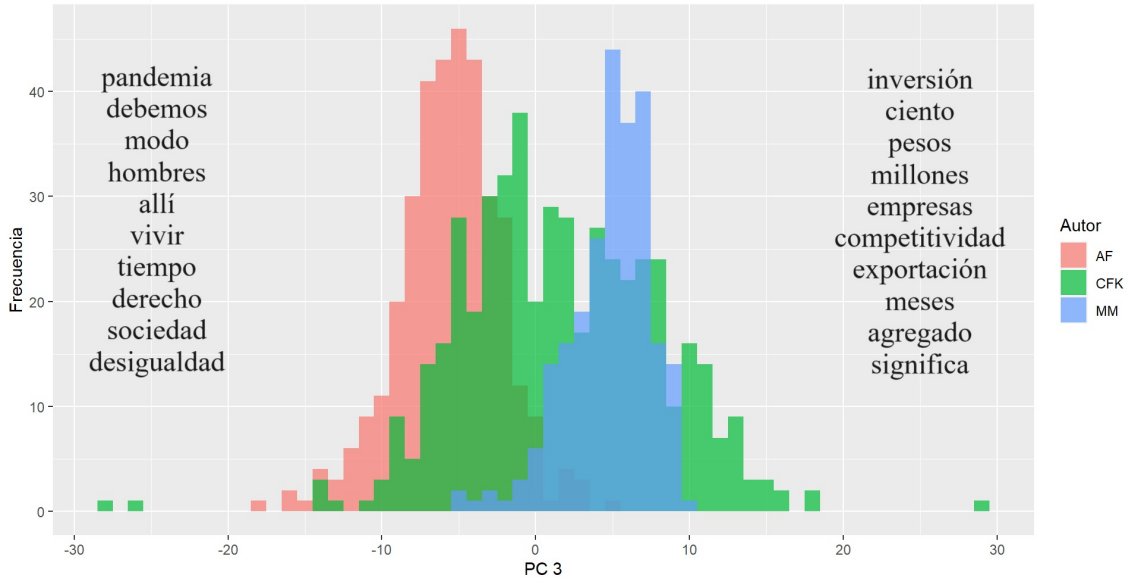


Figura 4.16: Histograma de tercera componente por cada presidente. Encontramos a la derecha y a la izquierda las palabras que poseen mayor ponderación en valor absoluto para generar valores respectivamente positivos y negativos en la componente.

como input el método de PCA. Si observamos los detalles en la matriz de confusión del Cuadro 4.5, podemos ver en detalle cómo su principal fuente de error proviene de confundir discursos de AF con los de CFK.

	CFK	MM	AF
Predicho CFK	474	2	6
Predicho MM	2	238	2
Predicho AF	7	2	340

Cuadro 4.5: Tabla de Confusión con cross-validation con 5 pliegues del método KNN que toma como input las 5 primeras componentes de PCA.

4.3.3. Comparación

Si hacemos una comparación de ambos métodos, podemos tener varias cuestiones en consideración:

- Lo primero que surge es el mejor desempeño predictivo de la técnica de análisis de componentes principales. Mientras una compresión del

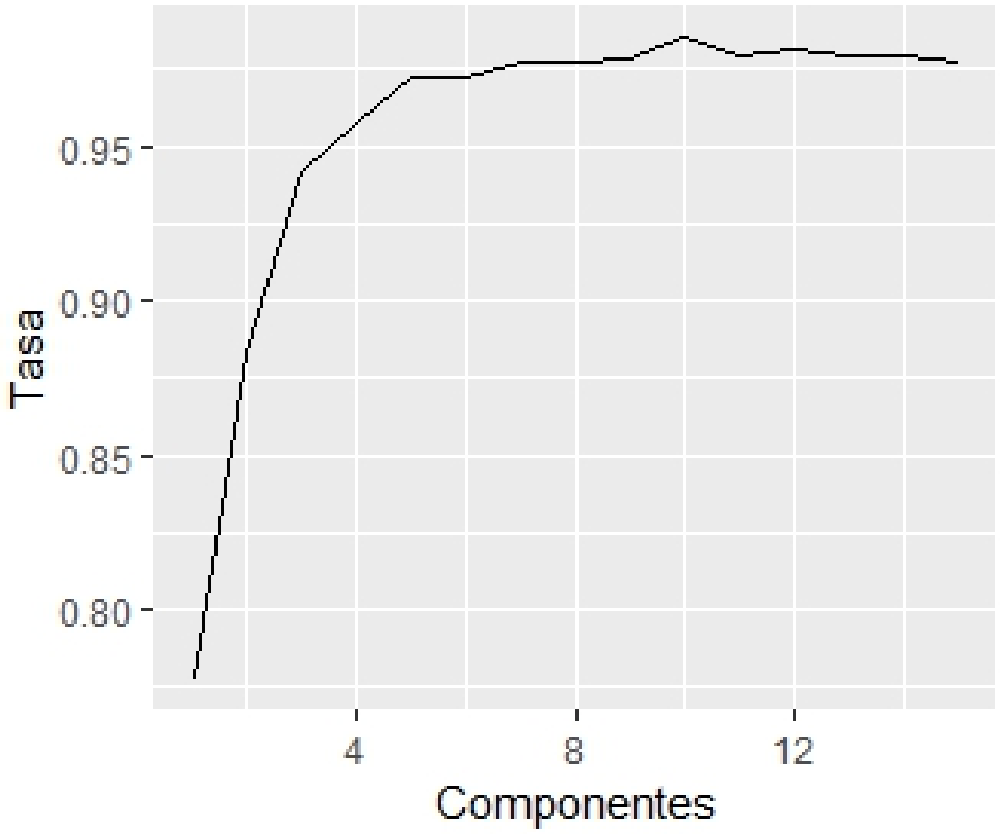


Figura 4.17: Tasa de aciertos por cross-validation con 5 pliegues según cantidad de componentes consideradas en el análisis.

dataset con este método logró una tasa de aciertos del 98 %, la técnica de NMF obtuvo 92,6 %. En este sentido hay una clara dominancia de PCA.

- En cuanto a la visualización, PCA se muestra muy efectivo a la hora de realizar gráficos de dispersión como la Figura 4.13. Por otra parte, existe la forma de analizar qué componentes separan a los discursos de cada presidente a través de histogramas como los de la Figura 4.15.

- NMF se muestra mucho más eficiente a la hora de asignar palabras a los tópicos de forma coherente. En este sentido, la Figura 4.11 permite ver que los términos correspondientes a cada presidente tiene una interpretación mucho más clara que los de las Figuras 4.15 y 4.16, correspondientes a PCA. Por ello, consideramos que, a pesar de tener un peor desempeño predictivo en términos relativos, NMF es una herramienta adecuada para un análisis semántico de segmentación de temáticas.

En resumen, ambos métodos presentan resultados positivos en términos de métricas de clasificación e interpretabilidad semántica. Sin embargo, PCA resulta superior en el primer caso y NMF, en el segundo.

Capítulo 5

Conclusiones

A lo largo del trabajo hemos logrado introducir e implementar los métodos básicos de representación del contenido semántico de los textos a través de matrices de frecuencias. Además, hemos presentado una discusión acerca de los problemas que surgen de la Ley de Zipf y la existencia de palabras con una elevada presencia en los textos y que, por lo tanto, aportan poca información y mucha variabilidad. Esta dificultad nos ha llevado a considerar la eliminación de *stopwords* e introducir la matriz TF-IDF.

En segundo lugar, hemos presentado resultados teóricos acerca de la factorización no negativa de matrices y breves comentarios acerca de su interpretación como segmentación en tópicos del texto. Además, hemos presentado un análisis comparativo con PCA y VQ de carácter heurístico, lo cual nos permitió comprender mejor la naturaleza conceptual que subyace al método, tal como lo presentaron Lee y Seung en su paper original [8].

En la primera de las aplicaciones prácticas, que consistió en un análisis de los discursos de Cristina Fernández de Kirchner, Mauricio Macri y Alberto Fernández de forma anual, se ha observado cómo el método de NMF, a pesar de su simpleza, permite separar de forma no supervisada los períodos presidenciales. Incluso, puede apreciarse una interpretación semántica tanto de cada presidente, como de cada año en la que se disciernen los términos de mayor relevancia y pueden ser de utilidad como herramienta para los investigadores de las ciencias sociales en diversos contextos.

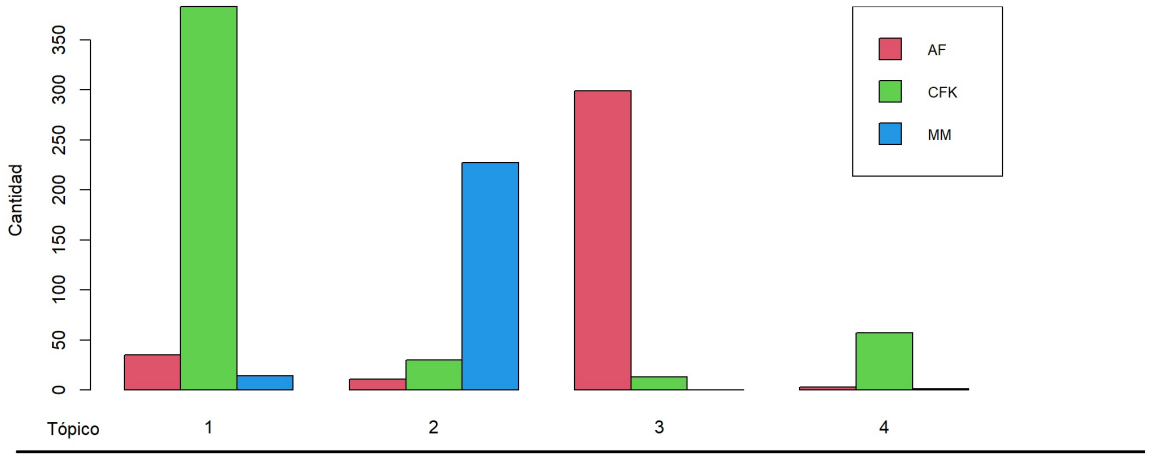
En segundo lugar, el Análisis de NMF realizado discurso a discurso permitió comprender mejor el contenido semántico presente en cada presidente. Esto lo podemos apreciar si volvemos a tener en cuenta la Figura 5.1. Además, utilizar KNN sobre la matriz comprimida W proveniente de la factorización de 4 tópicos logra una tasa de aciertos de 92,6 %, la cual es considerable, nuevamente teniendo en cuenta la simpleza del mecanismo. Además, hemos de resaltar el hecho de que la tasa de aciertos está calculada usando *cross*

validation, lo cual excluye la posibilidad del sobreajuste.

En lo que respecta al uso de PCA, se han obtenido métricas de clasificación muy superiores al de NMF. Estas fueron del 98 % considerando exclusivamente las cinco primeras componentes principales. Además, los histogramas de las componentes nos permitieron visualizar cómo se da el proceso de separación entre discursos. Sin embargo, el modelo de PCA no se mostró tan eficiente como NMF a la hora de segmentar semánticamente los discursos, pues las palabras que mayor ponderación presentan en cada componente carecen de un hilo conductor.

En lo que respecta a posteriores investigaciones es posible pensar en modificaciones a los métodos que presenten cierta plasticidad, ya que hay términos que identifican muy fuertemente a ciertos presidentes y si encontramos un texto que los menciona en repetidas ocasiones será clasificado como perteneciente a dicho presidente. De este modo, un discurso de Mauricio Macri que hable de vacunas será inevitablemente clasificado como discurso de Alberto Fernández.

Finalmente, es lícito mencionar que si bien actualmente existen métodos más complejos y sofisticados para la clasificación de autores con mejor poder predictivo, cabe resaltar que los resultados tanto de PCA como de NMF resultan relevantes por su equilibrio entre desempeño en métricas de clasificación, simpleza e interpretabilidad.



república	energía	vacuna	selectividad
defensa	obras	pandemia	catastróficas
derecho	equipo	contagios	nariz
armadas	eugenia	salud	internacional
democracia	maría	axel	organizado
votar	conectar	virus	acontecimiento
deuda	camino	aplausos	curso
pueblo	ruta	cuidado	planeta
países	futuro	hospital	formación
patria	aplausos	universidad	adquirido

Figura 5.1: Tópico asignado a cada discurso según presidente para $k = 4$. Debajo, las diez palabras de mayor peso en cada tópico.

Bibliografía

[1] Berry, Browne, et. al. *Algorithms and Applications for Approximate Nonnegative Matrix Factorization.* Computational Statistics & Data Analysis. 2007.

[2] Bin Shen, Luo Si, et. al. *Robust Nonnegative Matrix Factorization via L1 Norm Regularization.* 2014 IEEE International Conference on Image Processing. 2015.

[3] Cheng, C. *Principal Component Analysis (PCA) Explained Visually with Zero Math* Towards Data Science. 2022.

[4] Dhillon, I. & Sra, S. *Generalized Nonnegative Matrix Approximations with Bregman Divergences.* The Univ. of Texas at Austin. 2005.

[5] Eisenstein, J. *Natural Language Processing.* Notas de clase Massachusetts Institute of Technology. 2018.

[6] James, Gareth, et al. *An introduction to Statistical Learning.* Vol. 112. New York: springer, 2013.

[7] Kim, J. He, Y. & Park, H. *Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework.* Springer. 2013.

[8] Lee, D. & Seung, H. *Learning the parts of objects by non-negative matrix factorization.* Nature. 1999.

[9] Lee, D. & Seung, H. *Algorithms for Non-Negative Matrix Factorization.* Advances in Neural Information Processing Systems. 2001.

[10] Mielke, S. et al. *Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP* Arxiv. arXiv:2112.10508. 2021.

[11] Paatero, P. & Tapper, U. *Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values.* Environmetrics. 1994.

[12] Pinto, S. *Análisis de la influencia de los medios de comunicación en la formación de opinión : de modelos basados en agentes a análisis de datos.* Tesis Doctoral. Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales, 2020.

[13] Schreibman, S.; Siemens, R. & Unsworth J. *A Companion to Digital Humanities* Oxford: Blackwell. 2004.

[14] Semerano, N. *Analysis of Presidential Speeches throughout American History.* Towards Data Science. 2020.

[15] Silge, J. & Robinson, D. *Text Mining with R: A Tidy Approach.* O'Reilly. 2017.