**Informe minería de texto**

1. Preprocesamiento y limpieza de datos

El análisis de texto partió de la recolección de 115 documentos, entre discursos y documentos de política públicas, de los últimos tres presidentes colombianos y argentinos. El preprocesamiento de los textos se hizo en RStudio, en el que los documentos se cargaron en pdf y se normalizaron, esto es, se pasaron a minúscula, se eliminaron acentos, números y direcciones web. Además, se hizo eliminación de stopwords y se hizo la selección de algunos términos adicionales que, por su alta frecuencia de repetición, se consideran como stopwords en este contexto (por ejemplo, términos como país, gobierno, nacional, ley, Argentina y Colombia). Para el caso de los discursos de la expresidenta Cristina Fernandez de Kirchner, se utilizó webscraping para descargar todos los discursos de la presidenta entre los años 2007 y 2015. La base de datos resultante se organizó siguiendo un esquema de ordenación de bases de datos (formato tidy) en el que los discursos se incluyeron dentro de una base de datos .csv, en la que se incluyó información adicional al discurso como el año y la fecha exacta en la que se hizo el discurso. Los discursos y documentos de los demás presidentes se construyeron siguiendo la misma lógica, de tal forma que al final se obtiene un data frame, con las columnas correspondientes a discurso / documento, autor, año y fecha de realización. Tanto el scraping como el preprocesamiento, procesamiento y posterior análisis y visualización de datos se hizo usando el lenguaje de programación R.

1. Frecuencia de términos

El primer análisis correspondiente a la frecuencia de términos para cada uno de los presidentes. De manera más ilustrativa, se grafican dichas frecuencias en una nube de palabras (ver Gráfico 1) para el caso de los presidentes Petro y Milei. Nótese que los términos más grandes son los términos de mayor aparición en los discursos de ambos presidentes.

*Gráfico 1. Nube de palabras para los discursos de Gustavo Petro y Javier Milei*

Texto

El contenido generado por IA puede ser incorrecto.Texto

El contenido generado por IA puede ser incorrecto.

Ahora bien, se puede hacer este mismo análisis para todos los presidentes del periodo analizado. Así, el gráfico 2 permite ver las tablas de frecuencias para los primeros 10 términos. En dicho gráfico se puede ver que para cada uno de los presidentes se pueden encontrar términos particulares (por ejemplo, ser, para los expresidentes argentinos) así como términos generales para todos ellos (desarrollo, economía, información, servicios) y algunos específicos para los presidentes colombianos (por ejemplo, social, justicia, para Petro, gestión para el caso de Santos).

*Gráfico 2. Tabla de frecuencia de términos, por presidente.*

Imagen que contiene Gráfico

El contenido generado por IA puede ser incorrecto.

1. Análisis tf – idf

Los términos por presidente nos permiten ver aquellos que tienen una mayor aparición. Sin embargo, puede ser más interesante revisar cuáles son aquellos términos que son propios de cada presidente y que no se mapean exclusivamente a partir de la frecuencia de aparición. Para ello, se hará uso de la frecuencia inversa de documento (idf) y de su métrica complementaria tf-idf, que es una métrica usada en marketing digital y recuperación de información, la cual disminuye la ponderación dada a las palabras más frecuentes e incrementa el peso de aquellos términos que no necesariamente son los más usados en un conjunto de documentos, como en este caso. Así, por ejemplo, una palabra como “de” tendrá un tf (frecuencia de término) muy alta no solo por cada documento sino por todos ellos, pero por ello mismo tiene un idf muy bajo, pues resulta ser un término que no es particular de algún documento del conjunto de datos. Matemáticamente, el idf de un término se define como el logaritmo natural del número de documentos que componen un corpus sobre el número de documentos que contienen ese término en ese corpus.

Así, términos muy frecuentes tendrán valores idf muy bajos, mientras que términos menos frecuentes en el conjunto total de documento, pero muy usados en algunos de ellos tendrán un idf muy alto, lo que indica que son particulares para cada documento.

Esto se puede ver en la Gráfica 3, en donde vemos que hay términos que no están dentro de los más frecuentes por cada presidente (su tf no es muy alto, como reindustrialización en el caso de Petro) aunque su idf es más alto, pues es *relativamente más importante* para ese presidente en relación con los demás términos dentro del conjunto de documentos. Así, la relación tf-idf es el producto de ambas mediciones y busca medir cuán importante es una palabra para un documento o autor (Silge y Robinson, 2025), como en este caso.

Escala de tiempo, Gráfico de barras

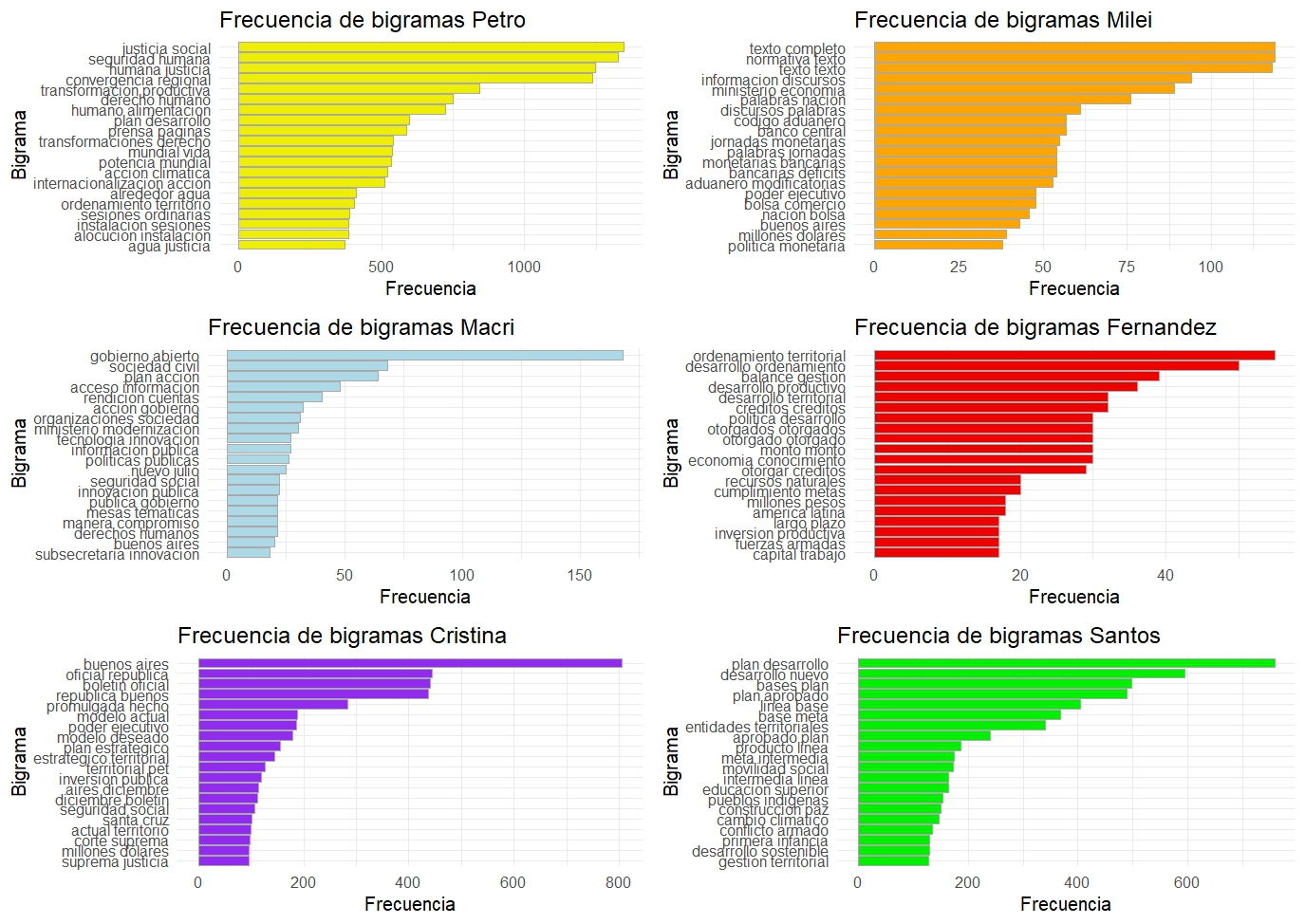
El contenido generado por IA puede ser incorrecto.*Gráfico 3. TF\_IDF, para los términos consultados de los presidentes de Argentina y Colombia.*

Nótese que aquí ya podemos encontrar diferencias significativas entre los diferentes presidentes, con términos en el caso de Milei como *mercadería, desregulación, aduanero, hiperinflación, fiscales, monetarias*, a diferencia de Petro, en el que los términos son *reindustrialización internacionalización, convergencia, pdet* (un conjunto de municipios priorizados en los Acuerdos de Paz de la Habana). Si a este análisis le agregamos los términos más frecuentes por presidente, obtendremos una idea clara de lo que habla cada presidente. Así, en Petro tenemos *social, justicia, transformación*, mientras que en Milei tenemos *modificatorias, economía, hiperinflación*.

1. Bigramas

Ahora bien, el análisis de bigramas nos permite ver cuáles de esos términos aparecen juntos en dicho análisis. Para ello, hacemos el procesamiento de la información dividiendo el discurso en pares de términos consecutivos. Una vez realizado, buscamos los bigramas más frecuentes y, como en el caso anterior, revisaremos el tf-dif de estos bigramas. El gráfico 4 permite ver la frecuencia de bigramas para los presidentes seleccionados:

*Gráfico 4. Bigramas más frecuentes de los presidentes de Argentina y Colombia*



* 1. Métrica tf-idf

1. Espectro político
2. Topic modeling
3. Conclusiones
4. Referencias

En este caso, se puede ver que hay pares de términos que son distintos de cada uno de los presidentes seleccionados. Así, por ejemplo, *justicia social, seguridad humana, transformación productiva, derecho humano, acción climática*, son bigramas frecuentes en Petro, pero no así en Milei o Macri. En este último, por ejemplo, *gobierno abierto, rendición cuentas, innovación tecnología, seguridad social, innovación pública* son algunos de los bigramas más frecuentes, dando así una idea sobre la importancia de la innovación en la rendición de cuentas a los ciudadanos. Por parte de Milei tenemos marcadores claramente asignables a su pensamiento libertario: *código aduanero, banco central, jornadas monetarias, política monetaria*, dando una relevancia a lo económico como eje central de su gobierno.

1. TF-IDF para bigramas

Al igual que en el caso de los términos individuales, tenemos el caso de la medición del tf-idf para los bigramas. En ese caso, vamos a encontrarnos con términos que son exclusivos de cada conjunto de documentos. En algunos casos, resultan relevantes para entender las posiciones de cada presidente, pero en otros casos, debido a los datos recogidos, los bigramas resultan estar más relacionados a la fuente de la cual se retomaron.

*Gráfico 5. Bigramas con el tf-idf más alto, por presidente seleccionado*

Escala de tiempo, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Este análisis combinado nos da una idea de las temáticas sobre las que habla y da prioridad cada uno de los presidentes. También nos muestra, de una manera muy clara, que los términos y lo bigramas no se intersectan, es decir, no hay bigramas que se encuentren en diferentes presidentes, lo cual es un indicador de la diferencia temática entre cada presidente. Esto también se puede ver en el grafo de bigramas (ver gráfico 6).

*Gráfico 6. Grafo con los bigramas por cada presidente.*

*Diagrama

El contenido generado por IA puede ser incorrecto.*

1. Enfoque político

A partir del análisis de los bigramas se puede hacer una relación más cercana entre aquellos conceptos que se consideran definitorios de una ideología específica y las correspondientes menciones a dichos bigramas. Así por ejemplo, los siguientes términos se asocian tanto a izquierda como a derecha, siguiendo la idea planteada por XXXX (XXX, 2019) en la que las ideas definitarias de cada uno de los espectros políticos se pueden ligar a algunos conceptos centrales, a partir del concepto de la noción lingüística de marco cognitivo, heredada de la lingüística cognitiva de George Lakoff (Lakoff y Johnson, 2003) y de la semántica de marcos de Charles Filmore (Filmore, 1982), y que en algunos otros estudiosos del lenguaje se entiendo como campos semánticos (cf. Coseriu) según la cual estos marcos actúan como un conjunto semánticamente relacionado de conceptos interconectados entre sí por un concepto central o prototipo. De esta manera, conceptos como izquierda y derecha, desde el punto de vista cognitivo, se puede asociar a algunos de los conceptos que se mencionan en la imagen siguiente, y que surgen de las definiciones de los principales autores de referencia en el aspecto económico y político.

Tabla

El contenido generado por IA puede ser incorrecto.

Ahora, es de interés revisar si efectivamente los conceptos antes señalados se pueden mapear dentro de los discursos de los presidentes de Colombia y Argentina. Se espera que exista un conjunto de estos términos que se pueda mapear en el análisis de bigramas que se ha hecho hasta ahora. Esto se puede ver en la siguiente tabla:

*Tabla 1. Filiación en el espectro político de los presidentes de Argentina y Colombia*

|  |  |  |
| --- | --- | --- |
| Presidente | Puntuación Izquierda | Puntuación Derecha |
| Cristina Fernández de Kirchner | 0,063 | 0,000 |
| Iván Duque | 0,125 | 0,077 |
| Alberto Fernández | 0,063 | 0,000 |
| Mauricio Macri | 0,063 | 0,000 |
| Javier Milei | 0,063 | 0,077 |
| Gustavo Petro | 0,188 | 0,000 |
| Juan Manuel Santos | 0,125 | 0,077 |

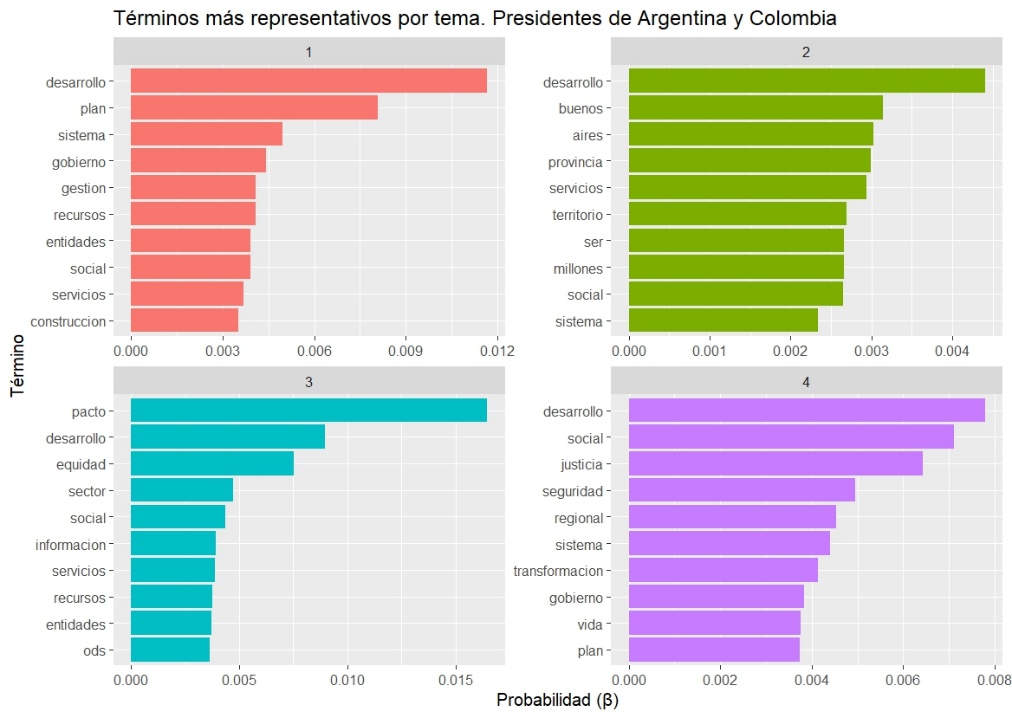
Las puntuaciones que aparecen en esta tabla surgen de revisar las coincidencias entre los bigramas seleccionados y el listado preliminar de términos previamente mencionados. Así, el puntaje se calcula tomando el número de coincidencias entre el corpus seleccionado y los bigramas sobre la cantidad total de términos seleccionados como representativos del marco semántico. Así, por ejemplo, Gustavo Petro tiene un puntaje de 0,18 para el marco semántico asociado a la izquierda pero 0 para el marco asociada a la derecha, de allí que su discurso sea considerado más hacia la izquierda que hacia la derecha.

Es importante señalar que el match no se hizo sobre la totalidad de bigramas, sino sobre los 500 más frecuentes, cuyo valor oscila entre 4 y 5 apariciones del bigrama, una frecuencia considerada mínima para poder ver que dichos términos tienen un verdadero peso en la función. Para el caso de Milei, las puntuaciones tanto para la izquierda como para la derecha son muy cercanas: 0,063 para izquierda y 0,077 para derecha. Esta ligera diferencia puede deberse a que varios de los términos asociados a la izquierda aparecen en Milei en sentido de crítica y no tanto como parte del efecto programático de sus ideas.

1. Topic modeling.

Como ejercicio final para la parte de minería, se llevó a cabo un ejercicio de topic modeling, a partir de la aplicación de un modelo de semántica latente, el de Alocación de Dirichlet (LDA), que es un modelo generativo en minería de texto en el que los términos de los discursos se tratan como se agrupan en clusters, que no son evidentes en la tokenización de los términos, lo que presupone que cada documento es una sumatoria de categorías. En este análisis de la información, los términos y sus posibles temas se disponen en una matriz termino – documento, donde se estudian las probabilidades de que un término pertenezca a un clúster específico. En el caso de los discursos de los presidentes, se hizo la revisión tanto para 2 como para 4 posibles temas, lo que se puede ver reflejado en los siguientes gráficos:

*Gráfico 7. Modelamiento de tópicos usando LDA, para k=4*



Nótese que al identificar las palabras en cada clúster encontramos términos similares (desarrollo) con algunas variaciones que no son claramente asignables a un tema específico: por ejemplo, el grupo 4 incluye el tema de justicia, seguridad, vida y regional, que podría asociarse a problemas de seguridad a nivel territorial, mientras que el tema 2 parece referirse específicamente a problemáticas de ordenamiento territorial y social en Argentina, dados los términos más frecuentes en este clúster.

Es interesante revisarse, cuando el número de clústers es igual a 2 (k = 2), los términos pueden asociarse a las problemáticas de cada país, esto es, cada cluster representa los problemas propios de cada país: así, en el caso del primer cluster se habla de *pacto, plan, desarrollo, transformación, seguridad, salud, proyectos, educación, infraestructura, recursos*, temas que son recurrentes en los planes de desarrollo de los presidentes colombianos. En el caso del otro cluster, tenemos menciones a términos como *ser, millones, ministerio, social, sistema, territorio, provincia, república, poder y servicios*, que en el caso de la Argentina, puede estar más asociados a los enfoques sectoriales que presentan los presidentes de esa nación y la necesidad de generar mayor empoderamiento en las provincias, en comparación con la provincia central en términos políticos y económicos, que es la provincia de Buenos Aires.

*Gráfico 8. Modelamiento de tópicos usando LDA, para k=2*

