Análisis de noticias de la crisis en Venezuela

Tahís Ahtty*

*Escuela Superior Politécnica del Litoral

Campus Gustavo Galindo Km 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador
{msegarra, aahtty, pmulloa}@fiec.espol.edu.ec

Abstract*—RESUMEN AL FINAL AQUÍ***.**

***Index Terms*—Twitter, tweets, toxic tweets, toxic users, LDA, Data Mining, ROC Curves**

I. Dataset

El dataset que se usa en esta investigación consta de 312 artículos o noticias que se obtuvieron a través de un Scraper en la sección de diario ecuatoriano *'El Universo'* denominada***"CRISIS EN VENEZUELA"*** ; en esta sección diario *El Universo* dedica el acceso directo a cualquier artículo relacionado con Venezuela y que tenga alguna relación con la opinión pública sobre la crisis venezolana. Estas noticias son correspondientes al año 2018 y van desde febrero a diciembre de ese año. Se extrajo la fecha, la url, el título y el cuerpo de la noticia. Posteriormente se delegó a un experto la clasificación de cada artículo por las categorías; resultando ser estas: Represion, Migracion, Economia, Politica, Hambre, Salud y Sociedad.

A. Categorías

- **Represion:** Artículos relacionados al sometimiento de cualquier ídole que pudiere expresar la noticia, generalmente violentando derechos humanos o civiles
- **Migracion:** Artículos relacionados a la actividad migratoria de los habitantes venezolanos hacia otros países (generalmente Ecuador) debido a la crisis que mantiene el país
- **Economia:** Artículos relacionados a decisiones por parte de los organismos competentes o del Presidente Nicolás Maduro que afectan a la economía del país
- **Politica:** Artículos relacionados a la actividad nacional o internacional por parte de los mandatarios correspondientes repsecto a la crisis de Venezuela
- **Hambre:** Artículos relacionados a hechos que afectan la adquisición de artículos de primera necesidad (esencialmente para alimentarse) en Venezuela
- **Salud:** Artículos relacionados a propagación de epidemias o enfermedades producto de las malas condiciones en que deja la crisis de Venezuela
- **Sociedad:** Artículos relacionados a la crisis de Venezuela que involucran personajes de la farándula internacional o aquellos que tratan de algún encuentro entre el primer mandatario venezolano con autoridades internacionales

B. Descripción de las columnas

Las columnas resultantes del dataset contienen la siguiente información:

- ***fecha:*** time_stamp en formato ISO de la fehca y hora del artículo en línea.
- ***url:*** dirección del artículo en la web
- ***titulo:*** Título de la noticia
- ***texto:*** Cuerpo de la noticia
- ***categoria:*** clasificación otorgada por un experto a cada artículo

II. Métodos

Para el presente proyecto se ha usado Metodología aplicada para el análisis exploratorio de datos en data science. Como herramientas tecnológicas para el análisis, se ha usado las librerías de pandas para análisis de dataframes, seaborns para el ploting multivariado y matplotlib para el plotting

A. Análisis Exploratorio

Al inicio se hicieron análisis básicos explorando el dataset de artículos en el tiempo y conforme a la categoría. Se observó si existía data faltante y se hizo feature engineering para agregar nuevas columnas que ayuden al posterior analisis de los atículos a través de la metodología de Análisis de Texto

1) ***Análisis de categoría vs tiempo****:* Primero se quitó el componente de la hora al dataset y se agregó en la columna fecha_string con la finalidad de separar el mes en la nueva columna mes. Se hizo una tabla cruzada entre categoria y mes para establecer la cantidad de artículos por mes diferenciando la categoría y finalmete se ploté un gráfico de barras múltiple en representación del análisis anterior

2) ***Análisis en el tiempo****:* Con la ayuda de pandas DateTimeIndex se hizo análisis de la cantidad de artículos por día de la semana, y por mes. Posteriormente se enfatizó en los picos resultantes del gráfico de líneas usado. Previo a todo lo anterior se hizo que la fecha sea el índice de cada artículo para poder ordenar con facilidad la data además de hacer uso de DateTimeIndex.

Para el análisis de texto siguiente se tokenizó, stemizó, lematizó y limpió la data considerando las stopwords propias del dataset más aquellas que son provistas por el módulo de nltk. Se consideró como corpus a todas las rows de texto del dataset y como documentos a cada texto (cuerpo de la noticia). Se hizo una variación del hiperparámetro número de tópicos

B. Nube de palabras

Con la importación de WordCloud se graficó las palabras más comunes a través de una nube de palabras tanto a todo el dataset sin considerar la categoría, como nubes por categoría

C. Análisis de sentimiento

Usando la libería nltk y su sentiment analyzer se estableció la polaridad del texto de cada artpiculo agregando una nueva fila de sentimiento, posteriormente se estableció etiquetas de positivo, negativo y neutro para hacer un gráfico de porcentajes. Todo este análisis a todo el dataset en general

D. Análisis de tópicos

Para el análisis de tópicos se hizo análisis con variación en el hiperparámetro de número de tópicos; considerando 3 y 5 tópicos, un análisis a todo el data set como corpus y análisis de corpus por categoría

1) ***Cluster map****:* Se usó el método cluster map de seaborn y se trata del ploteo de la correlación por densidad entre las palabas más relevantes del corpus yendo desde correlación negativa en rojo a positiva en azul y su intensidad dependía de la cifra resultante de la correlación. Se varío el número de tópicos probando con 3 y 5 tópicos tatno a todo el dataset como por categoría

2) ***LDA-TF IDF****:* Se obtuvo la tabla de palabras más significantes por tópico tanto para todos los artículos como por categoría y se plotéo un gráfico de distancias intertópicas junto con las palabras o términos más relevantes asociadas a cada tópico, en contraste con su frecuencia presente en el propio tópico y a todo el corpus. Además de variar el número de tópicos entre 3 y 5, se varío los valores de lambda entre 0.06 y 0.8

III. Resultados

A. Análisis de categoria

Como observación incial obtuvimos la cantidad de artículos por categoría, siendo política la catagoría que destaca en contraste con las demás, casi duplicando la cantidad de artículos a su segunda en cabecera

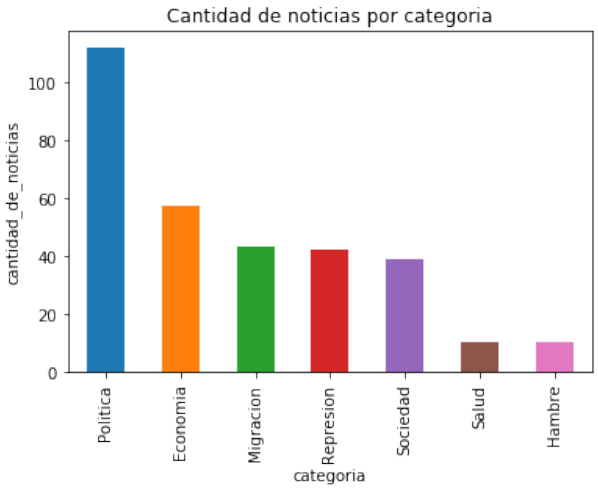

Figure 1. Cantidad de artículos por categoría

B. Análisis de Categoría vs Tiempo

mes categoria	02	04	05	06	07	08	09	10	11	12
Economia	2	4	7	3	6	19	8	3	3	2
Hambre	0	0	0	0	0	3	4	2	1	0
Migracion	0	0	0	0	0	39	2	0	2	0
Politica	1	13	28	17	7	18	17	5	4	2
Represion	1	2	3	7	4	8	8	3	5	1
Salud	1	0	0	1	0	1	1	2	3	1
Sociedad	0	16	2	5	6	8	0	2	0	0

Figure 2. Tabla de frecuencias de artículos por mes en cada categoría

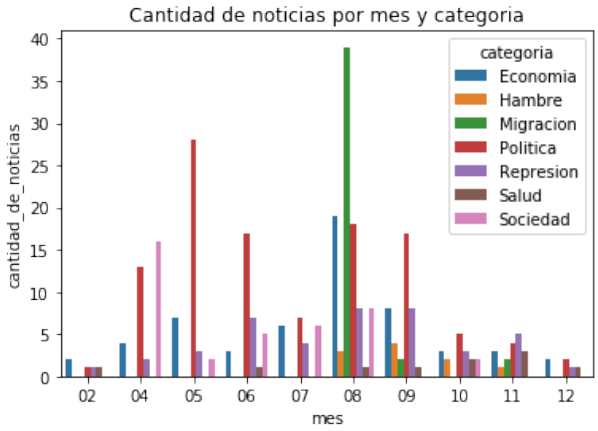

Figure 3. Gráfico de barras múltiple de frecuencias de artículos por mes en cada categoría

En la figura 3 se puede observar que los meses anteriores a agosto no tenía artpuculos relacionados a hambre o migración, además que la crisis migratoria tiene su cúspide en el mes de agosto.

C. Análisis en el tiempo

En la figura 4 se puede observar que los artículos relacionados con la crisis venezolana pertenecen al día lunes y en el transcurso de la semana hay un pico en el día viernes. En la figura 5 se observa al mes de agosto como el mes con mayor contenido relacionado a la crisis de Venezuela.

1) Análisis en el mes de agosto: En la figura 6 se puede observar que en los días finales del mes de agosto las noticias relacionadas a la crisis tomaron

Leyenda

dia_de_la_semana	Días
0	Lunes
1	Martes
2	Miércoles
3	Jueves
4	Viernes
5	Sábado
6	Domingo

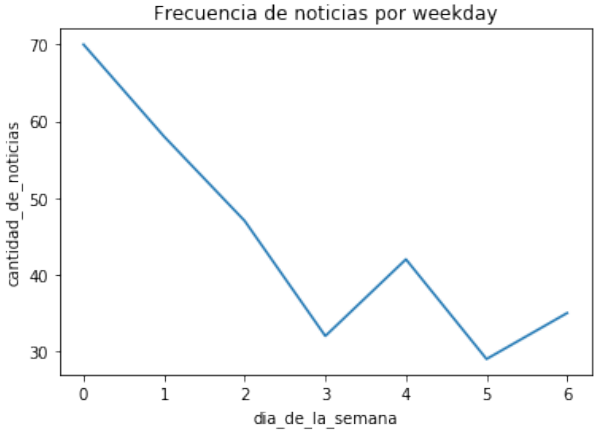

Figure 4. Cantidad de artículos por día de la semana

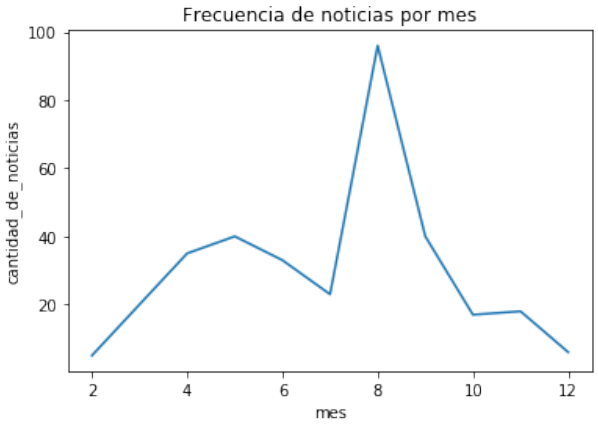

Figure 5. Cantidad de artículos por mes

gran papel en el diario en comparación con el resto de días

D. Análisis de sentimiento

En la figura 7 se observan los porcentajes de acuerdo al sentimiento de los artículos de acuerdo a las polaridades establecidas. Se puede observar que más del 70% de los artículos poseen un sentimiento negativo y menos del 20% negativo.

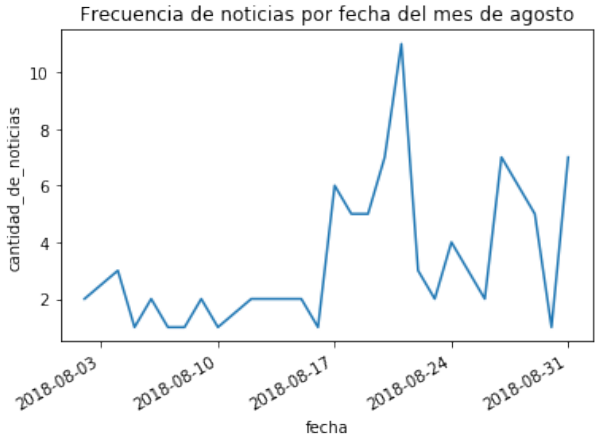

Figure 6. Cantidad de artículos en las fechas de agosto

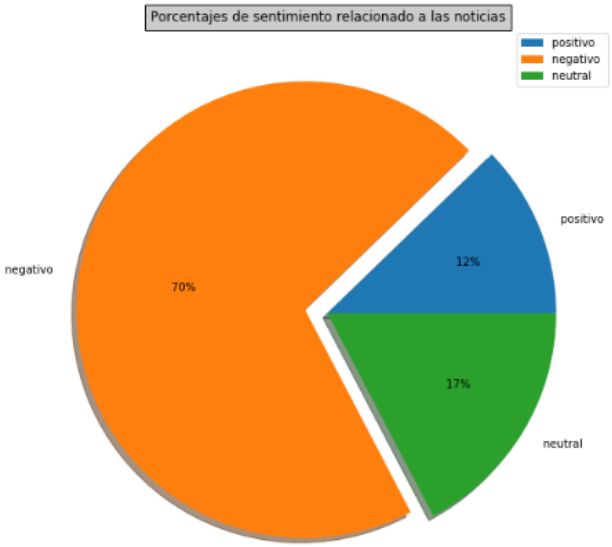

Figure 7. Porcentaje en sentimiento de todos los artículos del dataset

E. Word cloud

En la figura 8 se observa la nube de palabras para todo el dataset y en las figuras 9, 10, 11, 12 y 13 se observa las nubes de palabras por cada una de las categorías

F. Cluster map

1) Todo el dataset

: En la figura 14 se puede observar las correlaciones entre los términos más frecuentes considerando separación de tópicos de 3 y en la figura 15 se puede observar las correlaciones entre los términos más frecuentes considerando separación de tópicos de 5

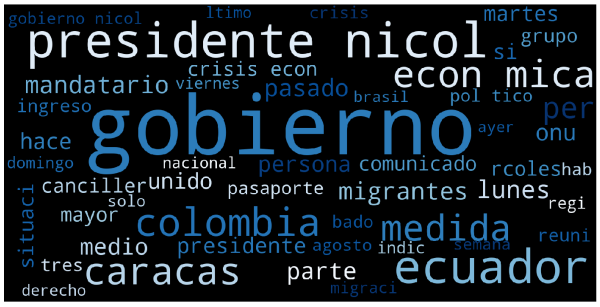

Figure 8. Nube de palábras con los términos más frecuentes de todo el dataset

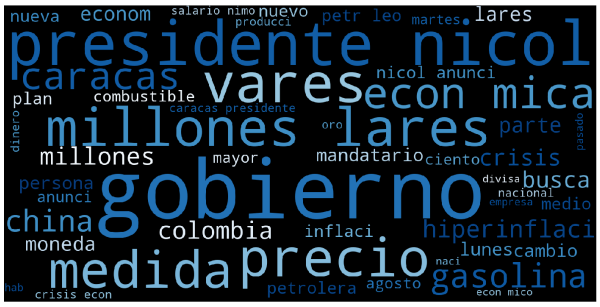

Figure 9. Nube de palábras con los términos más frecuentes en la categoría Economia

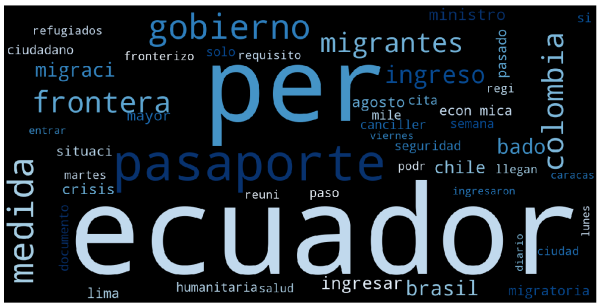

Figure 10. Nube de palábras con los términos más frecuentes en la categoría Migracion

2) Por categoría

: Para Economia, en la figura 16 se puede observar las correlaciones entre los términos más frecuentes considerando separación de tópicos de 3 y en la figura 17 se puede observar las correlaciones entre los términos más frecuentes considerando separación de tópicos de 5

Para Migracion, en la figura 18 se puede observar las correlaciones entre los términos más frecuentes considerando separación de tópicos de 3 y en la figura 19 se puede observar las correlaciones entre

Figure 11. Nube de palábras con los términos más frecuentes en la categoría Politica

Figure 12. Nube de palábras con los términos más frecuentes en la categoría Represion

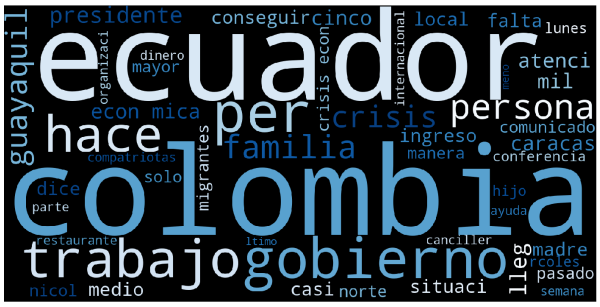

Figure 13. Nube de palábras con los términos más frecuentes en la categoría Sociedad

los términos más frecuentes considerando separación de tópicos de 5

G. LDA-TF IDF

1) Todo el dataset

: En la figura 20 se puede observar la tabla de correlaciones de los términos más frecuentes considerando separación de tópicos de 3 y en la figura 21 se puede observar la tabla de correlaciones de los términos más frecuentes considerando separación de tópicos de 5. Al considerar 3 tópicos se obtuvieron

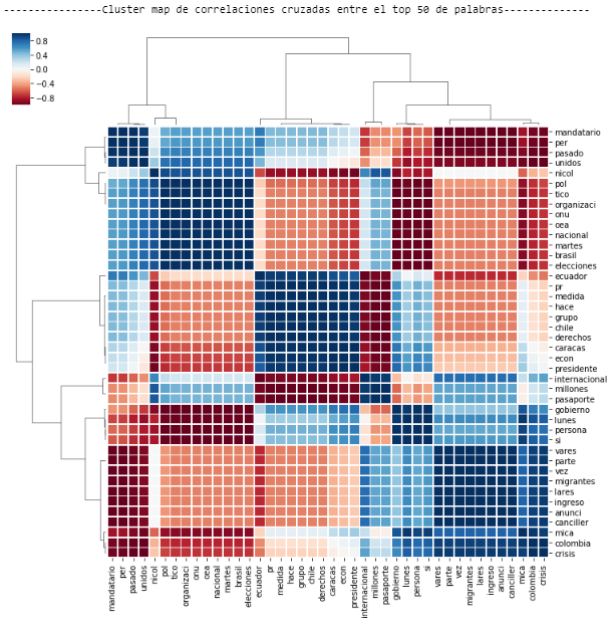

Figure 14. Cluster map de las palabras más frecuentes en todos los artículos

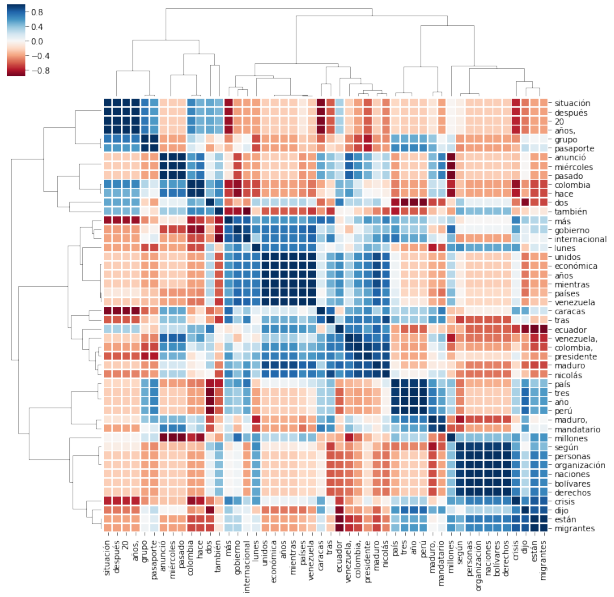

Figure 15. Cluster map de las palabras más frecuentes en todos los artículos

un total de 42 términos más frecuentes y al considerar 5, se obtuvo un total de 50

En las figuras 22, 23, 24 y 25 se puede observar capturas de los tópicos 1 y 5 del gráfico dinámico de distancias intertópicos y a su vez las palabras más frecuentes en cada tópico. Haciendo una variación el el parámetro lambda en 0.06 y 0.8 del algoritmo.

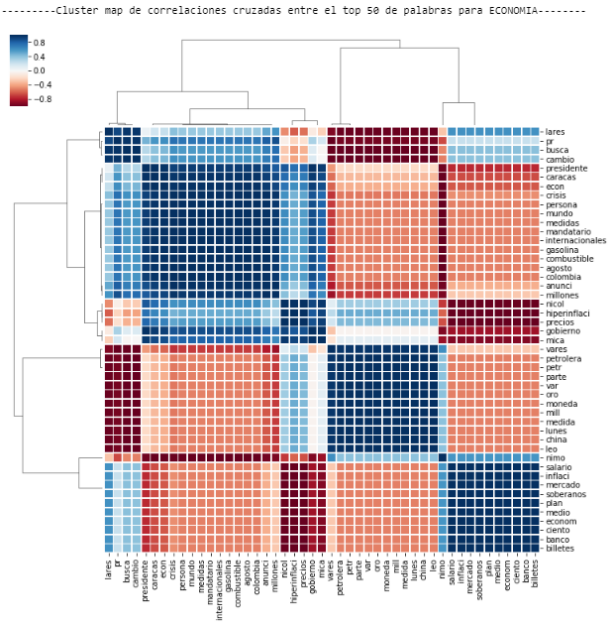

Figure 16. Cluster map de las palabras más frecuentes en la categoría Economia

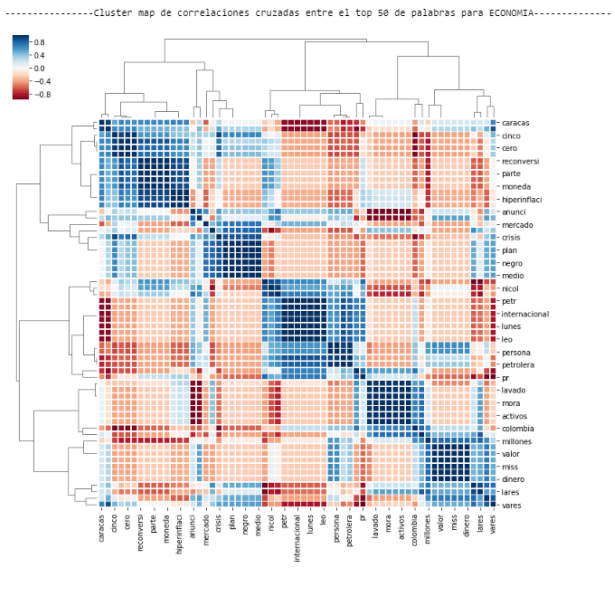

Figure 17. Cluster map de las palabras más frecuentes en la categoría Economia

Considerando todos los artículos com el corpus. En el lado de las frecuencias de las palabras, la parte azul de la barra representa a la presencia en todos los artpiculos y la roja a la presencia en el tópico correspondiente.

2) Por categoría

: Para la categoria Economia, las figuras 26,27,28 y 29 son capturas de los tópicos 1 y 3 del gráfico

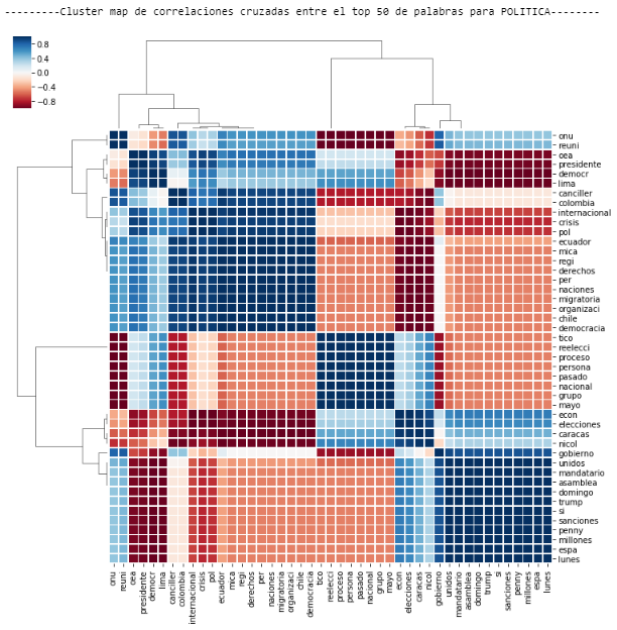

Figure 18. Cluster map de las palabras más frecuentes en la categoría Politica

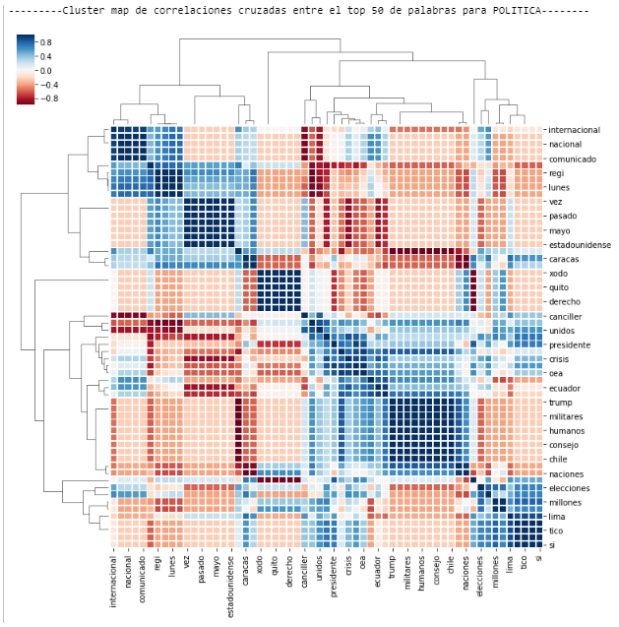

Figure 19. Cluster map de las palabras más frecuentes en la categoría Politica

	anunci	brasil	canciller	caracas	chile	colombia	crisis	derechos	econ	ecuador
0	0.000000	0.001914	0.000000	0.002782	0.000000	0.003201	0.004430	0.000000	0.002853	0.003888
1	0.000000	0.000000	0.000000	0.004225	0.001651	0.003536	0.005078	0.001652	0.004279	0.005075
2	0.002137	0.000000	0.002106	0.003025	0.000000	0.004140	0.006651	0.000000	0.003170	0.003298

Figure 20. Tabla parcial de los términos más frecuentes por tópicos en todos los artículos

	anunci	brasil	canciller	caracas	chile	cinco	colombia	crisis	derechos	econ
0	0.000000	0.000000	0.00000	0.002760	0.000000	0.000000	0.003321	0.005497	0.000000	0.003839
1	0.000000	0.000000	0.00228	0.003313	0.000000	0.000000	0.003764	0.004517	0.000000	0.002064
2	0.000000	0.002049	0.00000	0.004168	0.000000	0.000000	0.003163	0.006559	0.000000	0.004875
3	0.002632	0.000000	0.00000	0.002807	0.000000	0.000000	0.003260	0.004582	0.002167	0.003167
4	0.000000	0.000000	0.00000	0.003688	0.001951	0.001656	0.004732	0.005500	0.001615	0.003207

Figure 21. Tabla parcial de los términos más frecuentes por tópicos en todos los artículos

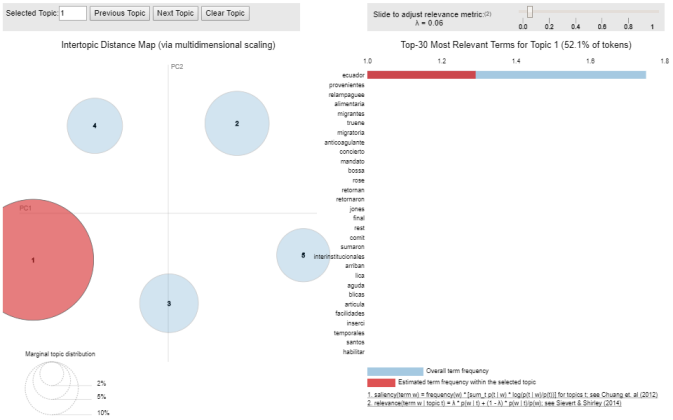

Figure 22.

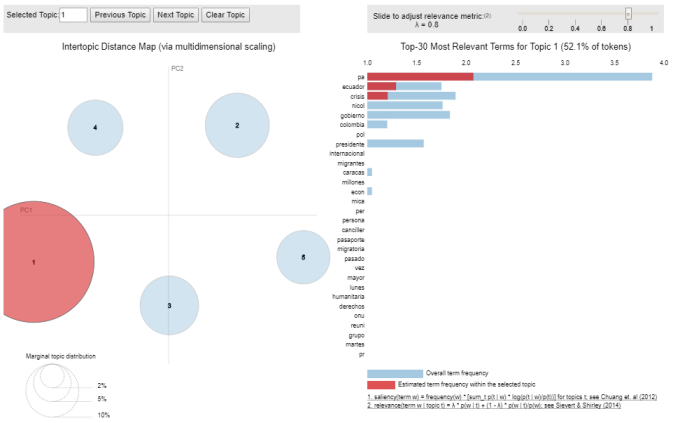

Figure 23.

dinámico de distancias intertópicos y a su vez las palabras más frecuentes en cada tópico, considerando un total de 3 tópicos. Haciendo una variación en el parámetro lambda en 0.06 (en 26 y 28) y 0.8 (en 27 y 29) del algoritmo.

Así mismo las figuras 30,31,32 y 33 son capturas de los tópicos 1 y 3 del gráfico dinámico de distancias intertópicos y a su vez las palabras más frecuentes en cada tópico, considerando un total de 3 tópicos.

Haciendo una variación en el parámetro lambda en 0.06 (en 30 y 32) y 0.8 (en 31 y 33) del algoritmo

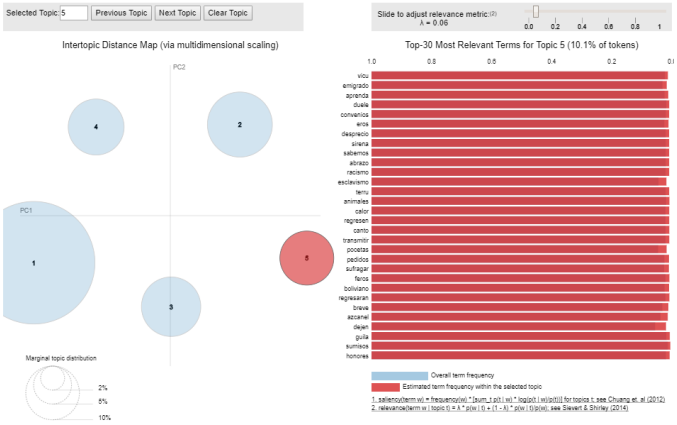

Figure 24.

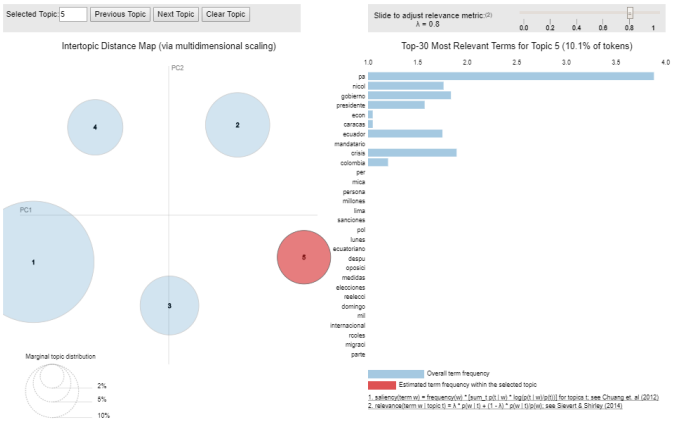

Figure 25.

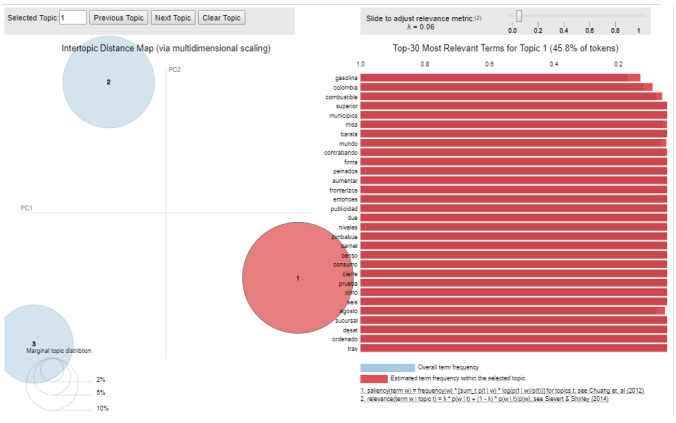

Figure 26.

Para la categoria Politica, las figuras 34,35,36 y 37 son capturas de los tópicos 1 y 3 del gráfico dinámico de distancias intertópicos y a su vez las palabras más frecuentes en cada tópico, considerando un total de 3 tópicos.

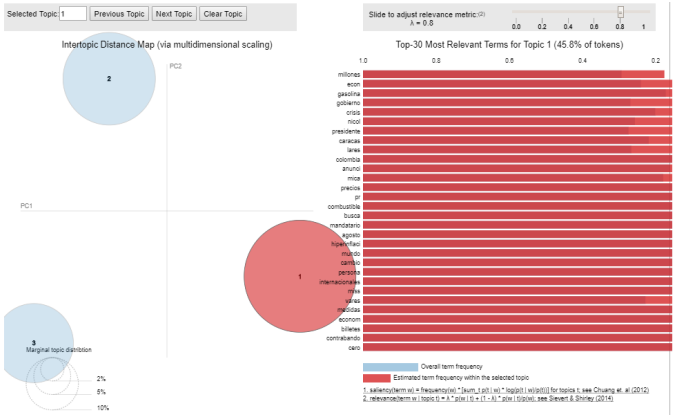

Figure 27.

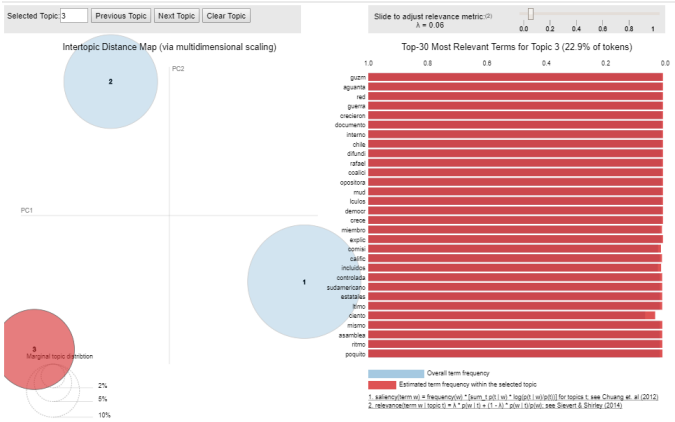

Figure 28.

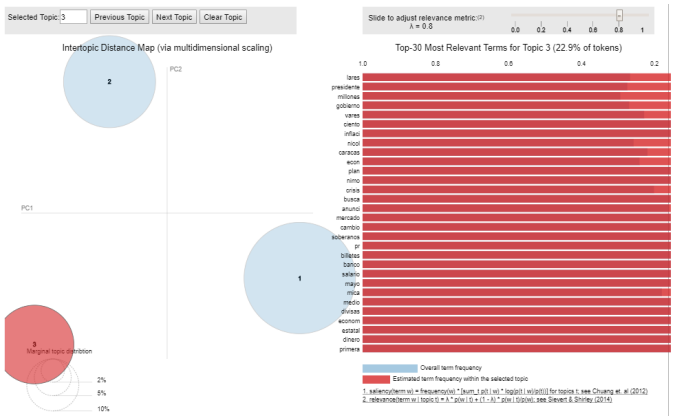

Figure 29.

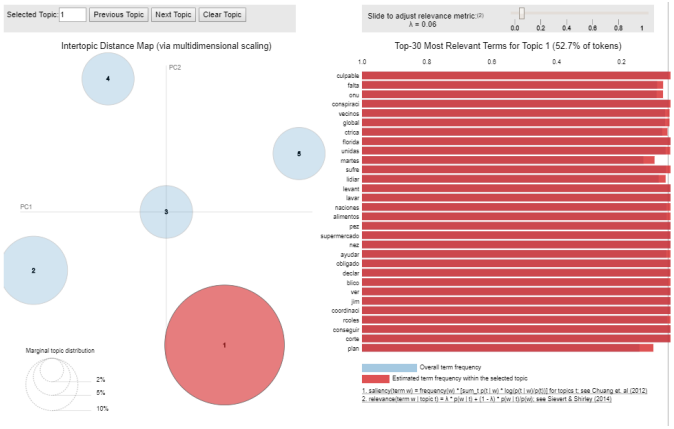

Figure 30.

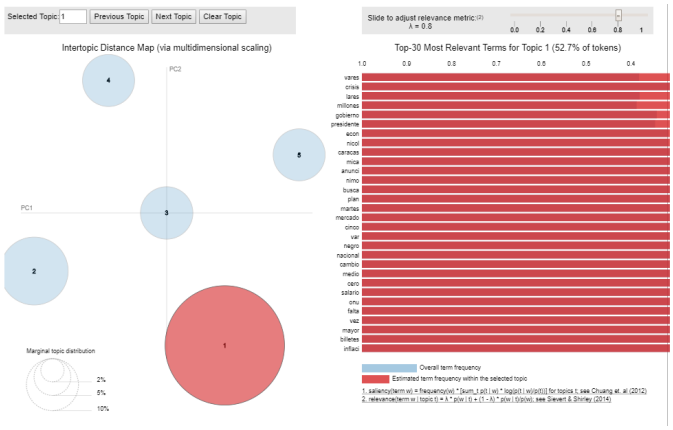

Figure 31.

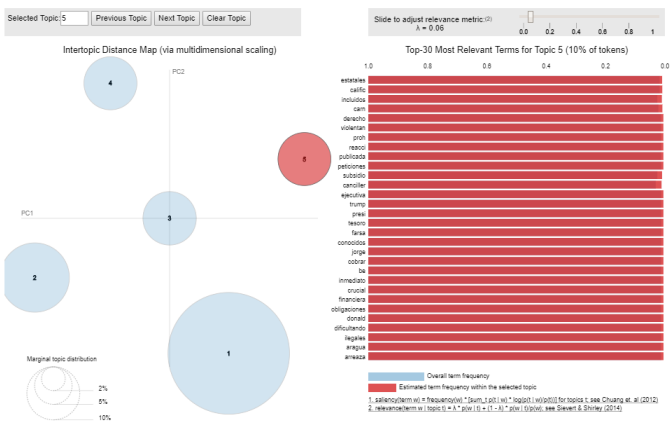

Figure 32.

Haciendo una variación en el parámetro lambda en 0.06 (en 34 y 36) y 0.8 (en 39 y 37) del algoritmo.

Así mismo las figuras 38,39,40 y 41 son capturas de los tópicos 1 y 3 del gráfico dinámico de distancias intertópicos y a su vez las palabras más frecuentes en cada tópico, considerando un total de 3 tópicos.

Haciendo una variación en el parámetro lambda

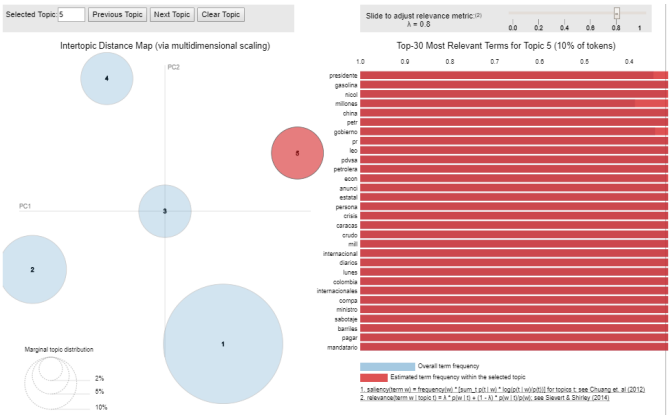

Figure 33.

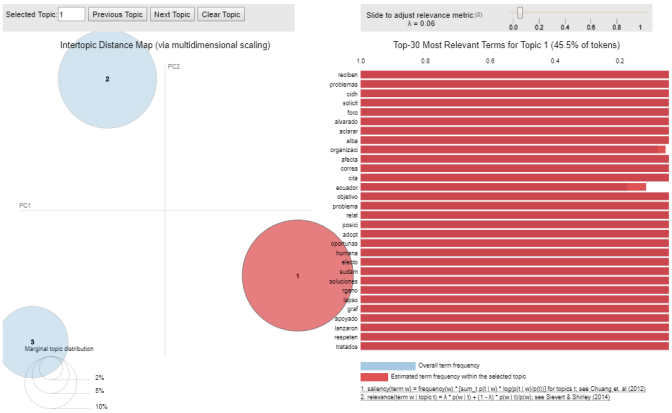

Figure 34.

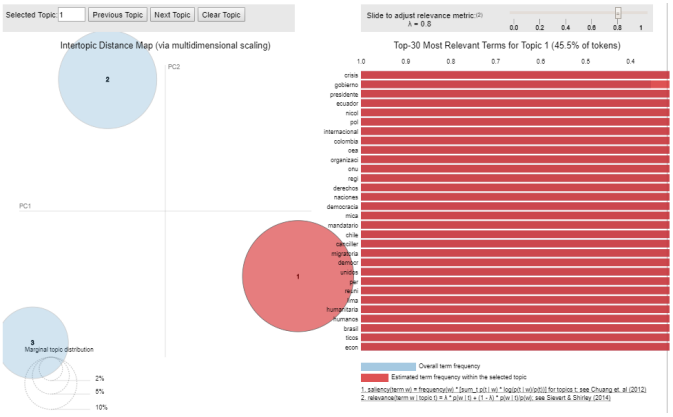

Figure 35.

en 0.06 (en 38 y 40) y 0.8 (en 39 y 41) del algoritmo

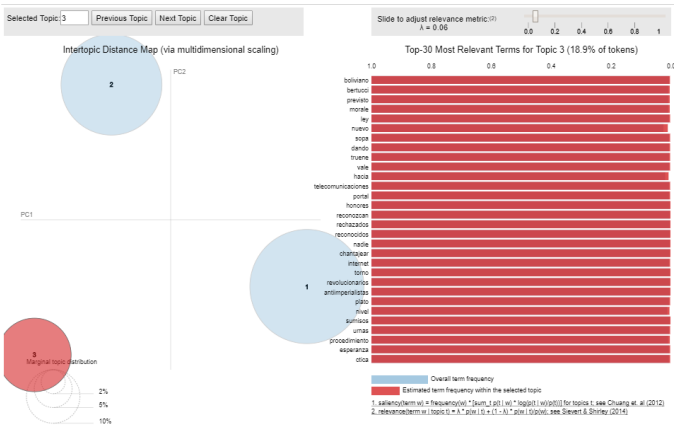

Figure 36.

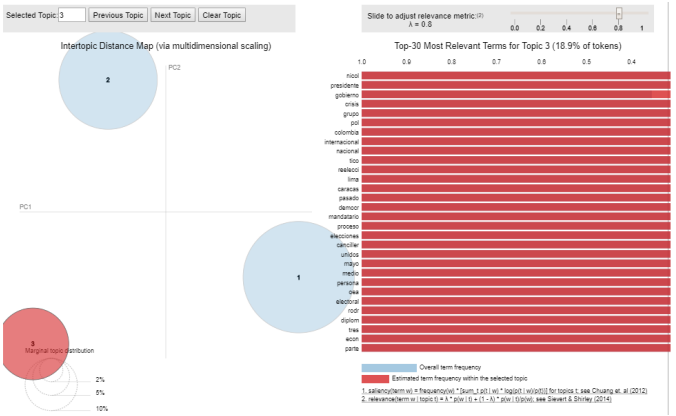

Figure 37.

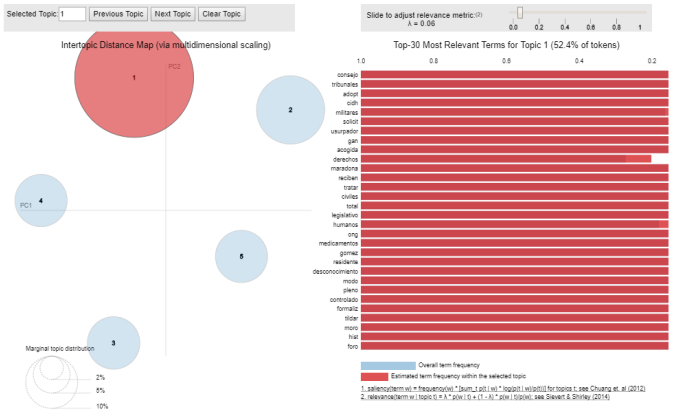

Figure 38.

IV. Discusión y Conclusiones

A. *Conclusiones generales*

- El mes de agosto fue cuando más movimiento migratorio se observó en los medios de prensa,

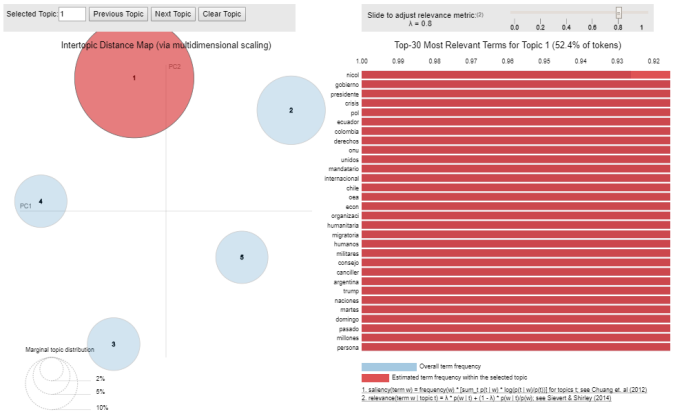

Figure 39.

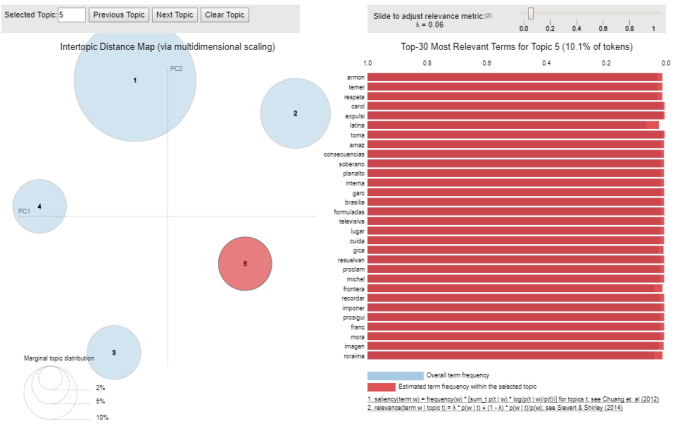

Figure 40.

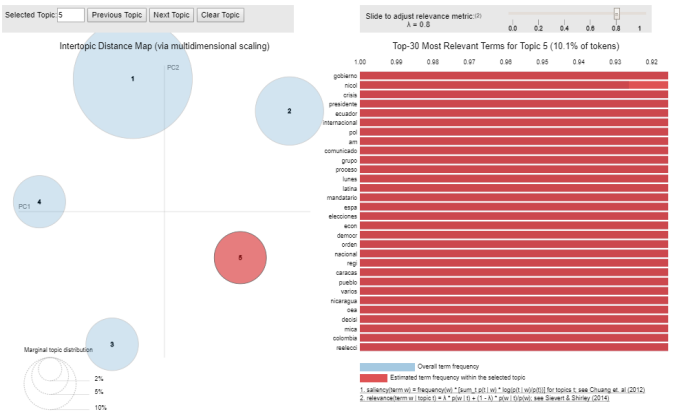

Figure 41.

eso explica que el mes de agosto tenga mayor cantidad de artículos de migración

- Las categorías de Salud y Hambre resultaron menos relevantes entre los artículos por lo que se decidió no usar para un análisis de comparación más profundo
- Los artículos de febrero a diciembre de 2018 de diario el Universo relacionados con la crisis en Venezuela tienen un sentimiento mayormente negativo.
- Dado que la mayoría de artículos se encuentran concentrados en Politica y en Economia, los análisis comparativos se dieron en esas 2 categorías con mayor profundidad.

B. Discusión sobre el análisis de tópicos

- La variación de la cantidad de tópicos en el análisis como hiperparámetro para los algoritmos, determinó en el clustermap que la correlación entre las palabras más frecuentes de los tópicos dependa estrechamente del número de tópicos. Se puede observar en la figura 14 que las palabras en general son en extremo positiva o negativamente correlacionadas; en cambio, en la figura 16 se observa una correlación débil entre las palabras frecuentes. Relación que se mantiene al hacer un análisis por categorías Economia y Politica (Si se observa la categoría Economia en las figuras 16 y 17)
- El análisis por separado de categorías permitió que las correlaciones entre las palabras más frecuentes del cluster map sea aún más estrecha, sin embargo no dejaba de observarse el mismo patrón de correlación
- La variación de la cantidad de tópicos en el análisis como hiperparámetro para los algoritmos, determinó en el LDA-TF IDF que el número de tópicos siendo menor establecía una mayor distinción entre la frecuencia en el tópico y la frecuencia en todo el corpus de documentos.
- El análisis por categoría permitió que haya una menor distancia intertópico cuando se trataba de una mayor cantidad de tópicos. Esto se puede observar en 22 y 30, correspondientes al análisis total y por Economia.
- La variación del parámetro lambda varió los términos frecuentes en cada categoría pero mantuvo el patrón de frecuencia asociado sea un mayo o menor lambda; la frecuencia en el tópico seleccionado y en el corpus era prácticamente la misma. No así cuando se

hacía un análisis sin distinción de categoría; la variación de lambda en un mismo tópico podía generar desde una distinción evidente en la frecuencia (figura 25 con lambda 0.8) y una distinción casi nula (figura 24 con lambda 0.06)

- Al mismo tiempo que la cantidad de términos considerablemente frecuentes disminuía con un menor lambda al tratarse de un análisis sin considerar categorías
- La cantidad de stopwords considerados para la limpieza del texto afectaba directamente la cantidad de términos frecuentes a los tópicos y sus correlaciones. La limpieza del texto era importante para no tener términos que representen ruido en la data.