

# Tarea Grande 2 - IIC1005

Diego Andai Castilla

15 de Mayo 2016

## 1. Introducción

## 2. Data

Para el siguiente estudio se utilizó un dataset de cuatro medios de noticias online chilenos. Los medios son Emol, La Tercera, El Mostrador y La Nación. De cada uno se extraen artículos (noticias), la cantidad de artículos por medio se muestra en el Cuadro 1.

| Medio        | Artículos |
|--------------|-----------|
| Emol         | 1000      |
| La Tercera   | 484       |
| El Mostrador | 631       |
| La Nación    | 832       |

Cuadro 1

De cada uno de estos artículos se extrajo el texto de la noticia. Estos textos se analizarán mediante un conteo de las palabras que contienen, lo que se explicará más adelante. Como antecedente, conviene observar las 30 palabras más comunes para cada medio (Figura 1).

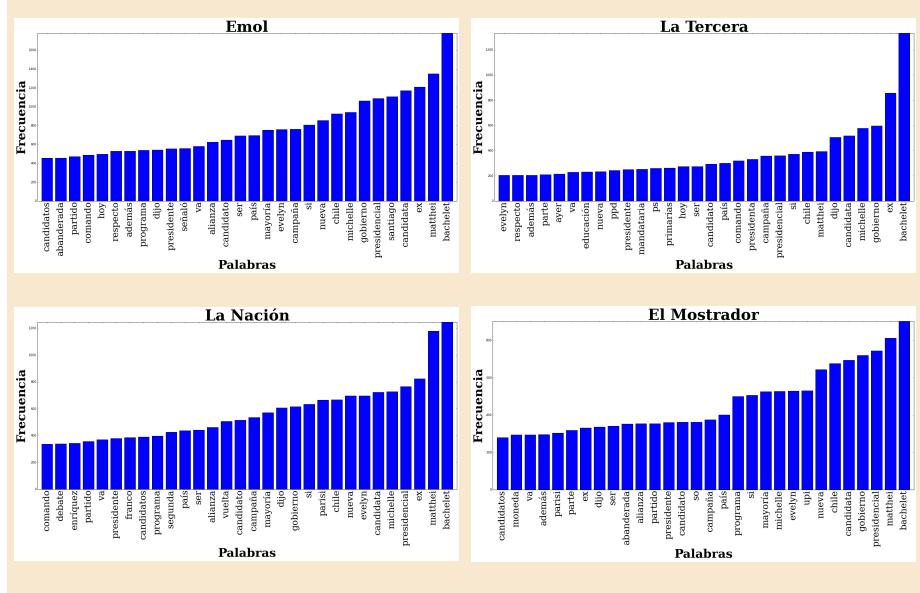
Las noticias refieren al proceso de elecciones presidenciales realizadas el año 2013 en Chile. Podemos extraer de las palabras más frecuentes ciertos conceptos temáticos preliminares que se discutirán más profundamente en la sección cuatro. Algunas cosas a rescatar:

- El término mas recurrente en los cuatro medios es “bachelet”, que alude a Michelle Bachelet, precisamente la presidenta electa en aquellas elecciones. El término “matthei” se encuentra en la segunda posición en tres de los medios, este término alude a Evelyn Matthei, candidata que justamente, salió segunda. Podría ser esto entonces la motivación para futuros estudios sobre la correlación entre atención de prensa y resultados electorales.

■ Pueden observarse ciertas tendencias de los medios según los gráficos. El medio La Tercera tiene entre sus términos más recurrentes “ppdz “ps”, dos partidos políticos de izquierda. Estos términos no se mencionan en Emol, en cambio en este se encuentra el término “alianza”, de derecha, que no está en La Tercera. En La Nación y El Mostrador aparece el término “parisi”, Franco Parisi era candidato independiente. Un Estudio con este tipo de información podría aplicarse, por ejemplo, para recomendar distintos diarios a las personas según sus preferencias.

■ Más de estas observaciones se discutirán en la sección cuatro.

Figura 1: 30 palabras más comunes por medio



## 2.1. Adaptación de los datos

Para poder trabajar con los datos, hay que estandarizarlos. Estos se representan mediante una matriz [Artículos x Palabras] (ver Cuadro 2).

|            | Palabra 1                          | ... | Palabra n                          |
|------------|------------------------------------|-----|------------------------------------|
| Artículo 1 | Frecuencia Palabra 1 en Artículo 1 | ... | ...                                |
| Artículo 2 | Frecuencia Palabra 1 en Artículo 2 | ... | ...                                |
| ...        | ...                                | ... | ...                                |
| Artículo n | ...                                | ... | Frecuencia Palabra n en Artículo n |

Cuadro 2  
(la matriz corresponde a la sección completamente encerrada)

Para obtener esta matriz, cada documento se representa como un “bag of words”, que consiste en una lista con cada palabra del artículo por separado. Luego se toman todos los documentos representados de esta forma y a cada palabra distinta se le asigna un índice, formando así un diccionario. Los documentos se pueden representar como una lista cuyos elementos son de la forma (índice de la palabra, frecuencia) para cada palabra que contiene el texto, esta representación se denomina corpus. Finalmente se arma la matriz requerida, cuyas filas corresponden al artículo y cuyas columnas corresponden al índice de la palabra, siendo las entradas de esta la frecuencia de la palabra en el documento.

## 2.2. TF y TF-IDF

La matriz descrita anteriormente se denomina TF, y presenta simplemente las ocurrencias de cada palabra por artículo. Un aspecto muy importante en la clasificación de documentos que se discutirá en la siguiente sección es la relevancia de las palabras y la representatividad de estas. Una palabra que aparece en gran cantidad de documentos es menos representativa que una que aparece en solo un tipo de artículos. Para corregir esto se ocupa una matriz llamada TF-IDF, que multiplica la frecuencia de la palabra por la frecuencia inversa de documento de dicha palabra, quitándole peso a las palabras comunes al momento de la predicción. En este estudio, se ocupó el modelo TFIDF de la librería scikit-learn, y se modificó el parámetro de normalización a ‘*norm=None*’ pues de este modo se obtenían mucho mejores resultados.

## 3. Clasificación con scikit-learn

Nuestro primer objetivo es armar modelos predictivos con el dataset. Estos modelos consisten en algoritmos computados que, después de ser entrenados con parte de los datos, intentan predecir de qué medio provienen los artículos de la otra parte. Se ocuparon tres modelos:

1. Árbol de Decisión
2. Regresión Logística
3. Naive Bayes (para este estudio, Multinomial)

### 3.1. Resultados según métricas

Para analizar el rendimiento de los modelos se ocupó el método de K-fold, con k=5. Este consiste en separar los datos en 5 partes, luego se entrena el modelo con 4 de las 5 partes y se testea con la otra. Se pueden obtener entonces 5 predicciones, las cuales se clasifican con las siguientes métricas:

1. Recall
2. F1-score

### 3. Precision

### 4. Accuracy

#### 3.1.1. TF

Primero se realizó el procedimiento anterior con la matriz TF explicada en la sección 2. Se tabulan en el Cuadro 3 el promedio de las métricas para cada modelo con su respectiva desviación estándar.

|         |           | Modelo utilizado |                     |               |
|---------|-----------|------------------|---------------------|---------------|
|         |           | Decision Tree    | Logistic Regression | Naive Bayes   |
| Métrica | Recall    | 0.7977±0,0151    | 0.8379±0,0209       | 0.5869±0,0280 |
|         | F1-score  | 0.7977±0,0153    | 0.8455±0,0194       | 0.6033±0,0260 |
|         | Precision | 0.7999±0,0165    | 0.8560±0,0169       | 0.6428±0,0246 |
|         | Accuracy  | 0.8028±0,0150    | 0.8449±0,0192       | 0.5921±0,0250 |

Cuadro 3

#### 3.1.2. TF-IDF

Luego se realizó el procedimiento anterior con la matriz TF-IDF explicada en la sección 2, los resultados se presentan en el Cuadro 4.

|         |           | Modelo utilizado |                     |               |
|---------|-----------|------------------|---------------------|---------------|
|         |           | Decision Tree    | Logistic Regression | Naive Bayes   |
| Métrica | Recall    | 0.7918±0,0129    | 0.8035±0,0218       | 0.5423±0,0395 |
|         | F1-score  | 0.7909±0,0113    | 0.8125±0,0210       | 0.5467±0,0353 |
|         | Precision | 0.7916±0,0109    | 0.8253±0,0196       | 0.5559±0,0327 |
|         | Accuracy  | 0.7974±0,0109    | 0.8086±0,0207       | 0.5270±0,0345 |

Cuadro 4

Algunas observaciones sobre los resultados que se obtuvieron:

- El modelo de regresión logística obtuvo mejores resultados en todas las métricas con ambos tipos de frecuencia.
- Las predicciones hechas a partir de la matriz TF obtuvieron mejores resultados, lo que quiere decir que para este dataset, quitarle relevancia a las palabras por ser comunes no afecta positivamente el rendimiento.

## 3.2. Resultados según matriz de confusión

Otra forma de medir los resultados es la llamada matriz de confusión, la forma de esta se muestra en el Cuadro 5. La notación debe interpretarse:

- Correctas(i,i): cantidad de veces en que el resultado esperado es i, y se predijo i.
- Incorrectas(i,j): cantidad de veces en que el resultado esperado es i, y se predijo j.

|            |   | Clase predicha   |                  |                  |
|------------|---|------------------|------------------|------------------|
|            |   | 1                | 2                | 3                |
| Clase real | 1 | Correctas(1,1)   | Incorrectas(1,2) | Incorrectas(1,3) |
|            | 2 | Incorrectas(2,1) | Correctas(2,2)   | Incorrectas(2,3) |
|            | 3 | Incorrectas(3,1) | Incorrectas(3,2) | Correctas(3,3)   |

Cuadro 5

Para este estudio, el orden de las clases es: Emol(1), La Nación(2), El Mos-trador(3), La Tercera(4). A continuación se presentan las matrices de confusión con cada modelo, para TF y TF-IDF, obtenida con la librería scikit-learn.

### 3.2.1. Árbol de decisión

Figura 2: Matrices de confusión modelo AD

```
([[259, 32, 2, 11],
 [ 25, 171, 10, 43],
 [ 4, 20, 159, 6],
 [ 7, 27, 7, 102]])
```

(a) TF

```
([[269 25 2 8]
 [ 23 174 13 39]
 [ 4 19 159 7]
 [ 4 23 6 110]])
```

(b) TF-IDF

### 3.2.2. Regresión Logística

Figura 3: Matrices de confusión modelo RL

```
([[273, 25, 0, 6],
 [ 29, 197, 10, 13],
 [ 12, 22, 153, 2],
 [ 5, 25, 1, 112]])
```

(a) TF

```
([[259 32 6 7]
 [ 33 189 18 9]
 [ 15 25 147 2]
 [ 10 22 3 108]])
```

(b) TF-IDF

### 3.2.3. Naive Bayes

Figura 4: Matrices de confusión modelo NB

```
[[229, 38, 31, 6],
 [108, 105, 33, 3],
 [ 58, 30, 101, 0],
 [ 43, 15, 1, 84]])
```

(a) TF

```
[[175 58 53 18]
 [ 77 118 45 9]
 [ 50 42 95 2]
 [ 29 15 5 94]])
```

(b) TF-IDF

Como se puede ver en las matrices de confusión y según las métricas de la sección 3.1, el mejor clasificador resultó ser el de regresión logística. El problema del modelo de Árbol de decisión con estos datos es que las palabras están muy relacionadas entre ellas, por lo que pueden ser multicolineales. Como todos los artículos hablan del mismo tema, es probable que algunas palabras y sus frecuencias respondan a más de una clase, es más, si dos variables son aproximadamente linealmente independientes (por ejemplo si ‘ps’ aparece un tercio de las veces que aparece ‘bachelet’ en gran cantidad de diarios), puede causar la confusión del modelo porque no aportaría diferenciación entre los medios.

El problema con el Naive Bayes es un poco más claro, la principal asunción de este modelo es que la probabilidad de las variables es independiente. Para este dataset esto tiene gran posibilidad de no ser correcto, muchas de las palabras están relacionadas y esta relación determina el medio. Esta dependencia de las variables no puede ser modelada.

El modelo de regresión logística probablemente respondió mejor a este dataset porque la *odds ratio* debe estar bien diferenciada, por ejemplo, debe ser mucho mas probable encontrarse con la palabra ‘matthei’ que con la palabra ‘parisi’ en Emol. Esta diferencia entre probabilidades permite trazar una linea de decisión para la predicción bien definida.

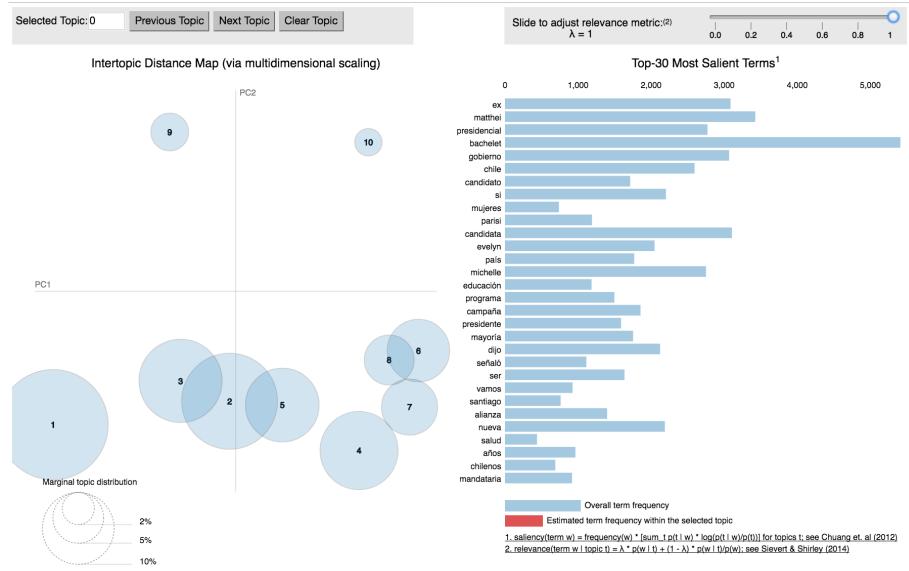
La clase más difícil de predecir es la de la nación, podemos calcular que su precisión es aproximadamente 0,67, mientras que los otros tres tienen una precisión mayor a 0,75 (Esto se calcula  $TP/(TP+FP)$ , desde la matriz de confusión, donde TP es verdadero positivo y FP es falso positivo). Una tendencia similar se da con las otras métricas, para las cuales La Nación en general tiene peores resultados.

## 4. Detección de Tópicos con pyLDAvis

Otra forma de analizar los documentos es con Latent Dirichlet Allocation (LDA), modelo que, en términos generales, encuentra tópicos dentro de un conjunto de textos. Los tópicos se definen como agrupaciones particulares de términos que tienden a repetirse. Para implementar esto se ocupa la librería pyLDA-

vis. Se debe crear un diccionario y un corpus. Luego se prepara el modelo que arroja las visualizaciones en las Figuras 5 y 6.

Figura 5: Visualización con 10 tópicos

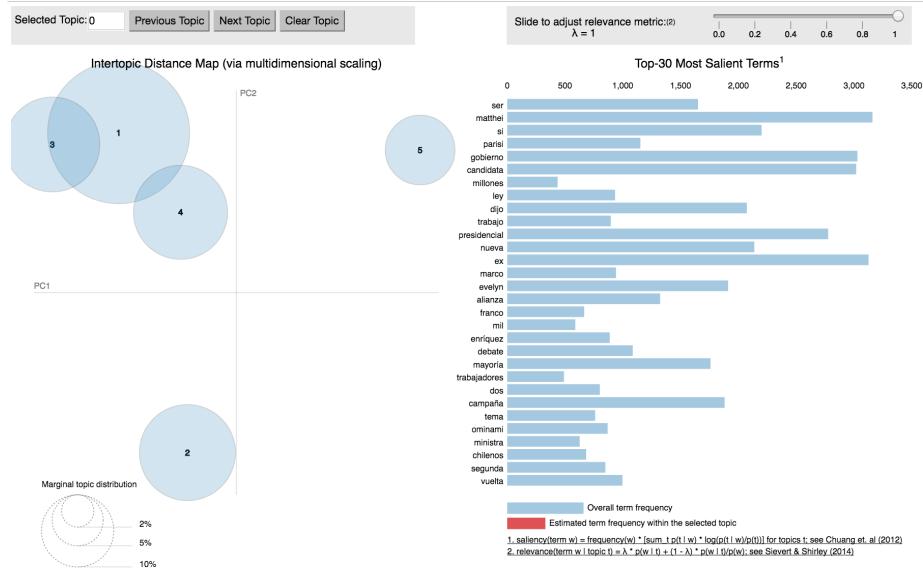


En esta primera figura se definen 10 tópicos distintos del dataset. Se pueden observar las palabras predominantes que encontramos en la sección 2. Luego se dividen en 5 tópicos. El detalle de los 10 y los 5 tópicos se muestra en la sección de anexos.

Nótese que las visualizaciones contienen un parámetro ajustable lambda, este tiene un objetivo parecido a la implementación de la matriz TF-IDF. En palabras generales este se usa para ajustar el peso de las palabras según su relevancia total en el corpus. Si una palabra ‘x’ aparece mucho en mi colección entera de artículos, esta va a tender a aparecer en muchos de mis tópicos. El parámetro lambda (mientras sea más pequeño) va a tender a quitarle relevancia a estas palabras ‘comunes’ arreglando esto y dando paso a palabras menos frecuentes pero más representativas del tópico. Si se cambia de, por ejemplo, 1 a 0.2, van a aparecer en la lista de cada tópico palabras que tienen menos frecuencia pero que lo diferencian más de los otros tópicos. Se ha estudiado y el valor óptimo es de aproximadamente 0.6 (Sievert and Shirley, 2014).

El mejor número de tópicos es 10. Al colocar lambda=0.6, se pueden ver tópicos relacionados con salud, reforma laboral, y distintos candidatos. En cambio con 5 tópicos tiende a juntarse todo y solo se puede distinguir la lucha Bachelet-Matthei con distintos tonos. Esto se denota en el área que abarcan los círculos que representan los tópicos.

Figura 6: Visualización con 5 tópicos



## 5. Bonus

### 5.1. Bonus 1

El código del bonus esta al final de la parte 1 (notebook) y en el repositorio (crawler). No alcancé a terminarlo pero estaba demasiado orgulloso de mi crawler como para dejarlo afuera.

### 5.2. Bonus 2

El código del bonus 2 está en el repositorio. En las siguientes figuras se puede apreciar 5 tópicos de cada dupla de medios. Se distingue un sesgo, por ejemplo, en que no se nombran a más candidatos que Bachelet y Matthei en Emol y La Tercera. Estas diferencias se evidencian en el detalle que está en el anexo. Palabras como ‘parisi’, ‘ominami’ y ‘claude’ denotan estas diferencias.

Figura 7: Visualización con 5 tópicos La Nación y El Mostrador

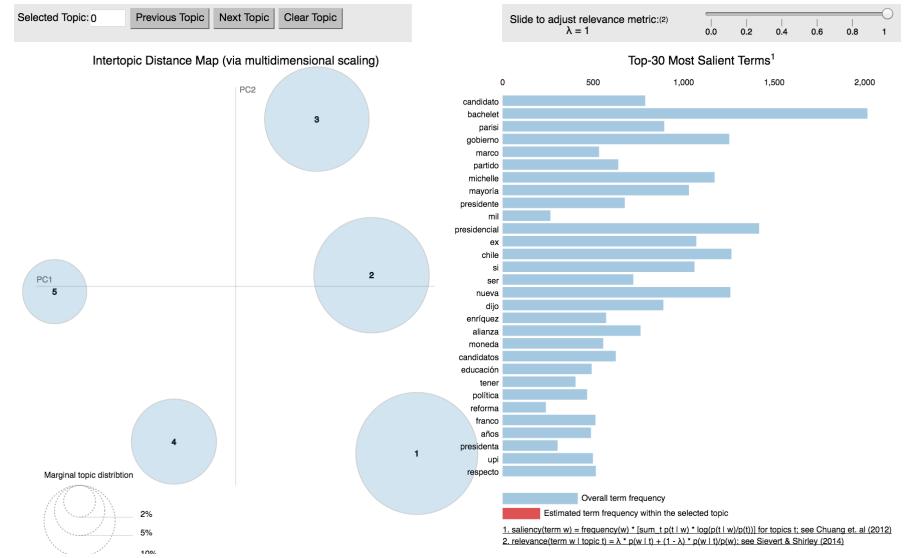
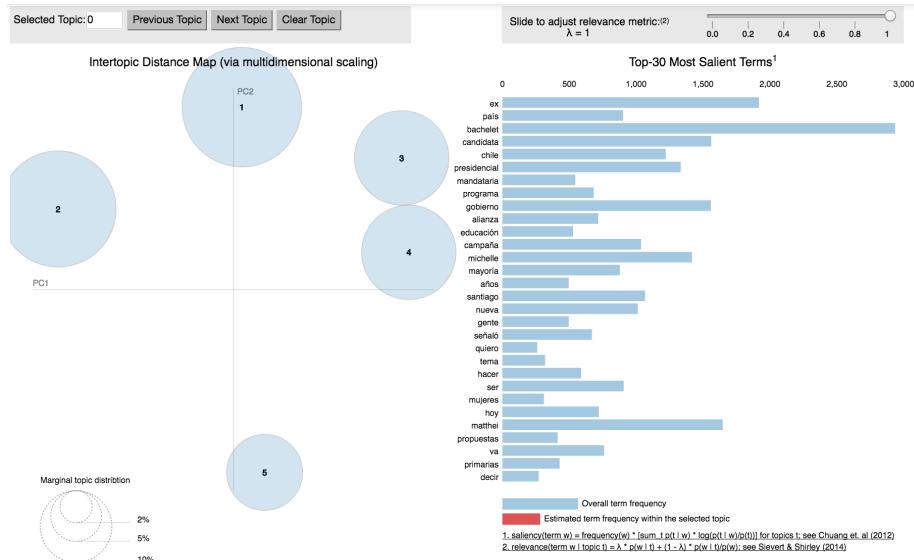


Figura 8: Visualización con 5 tópicos Emol y La Tercera

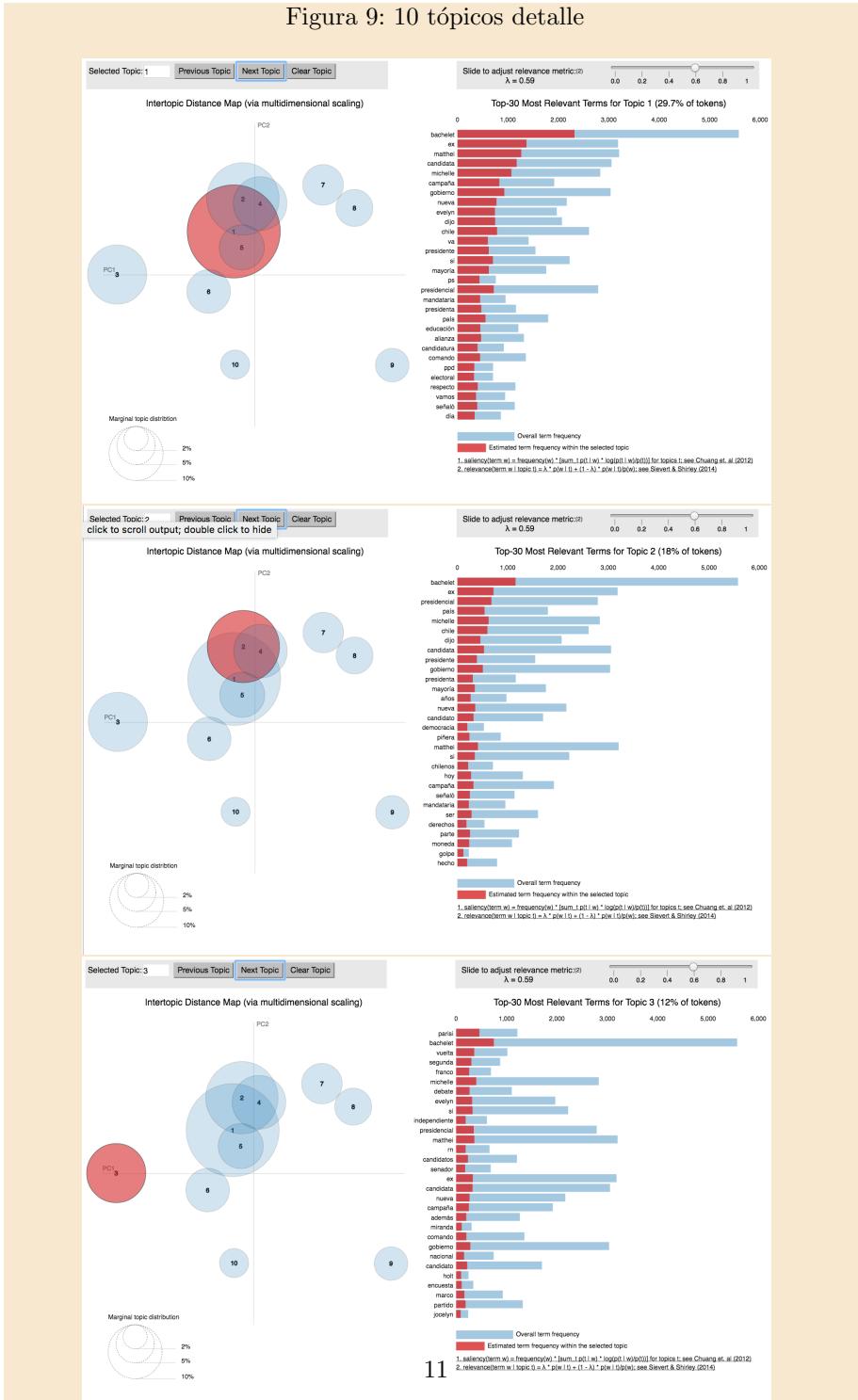


## 6. Bibliografía

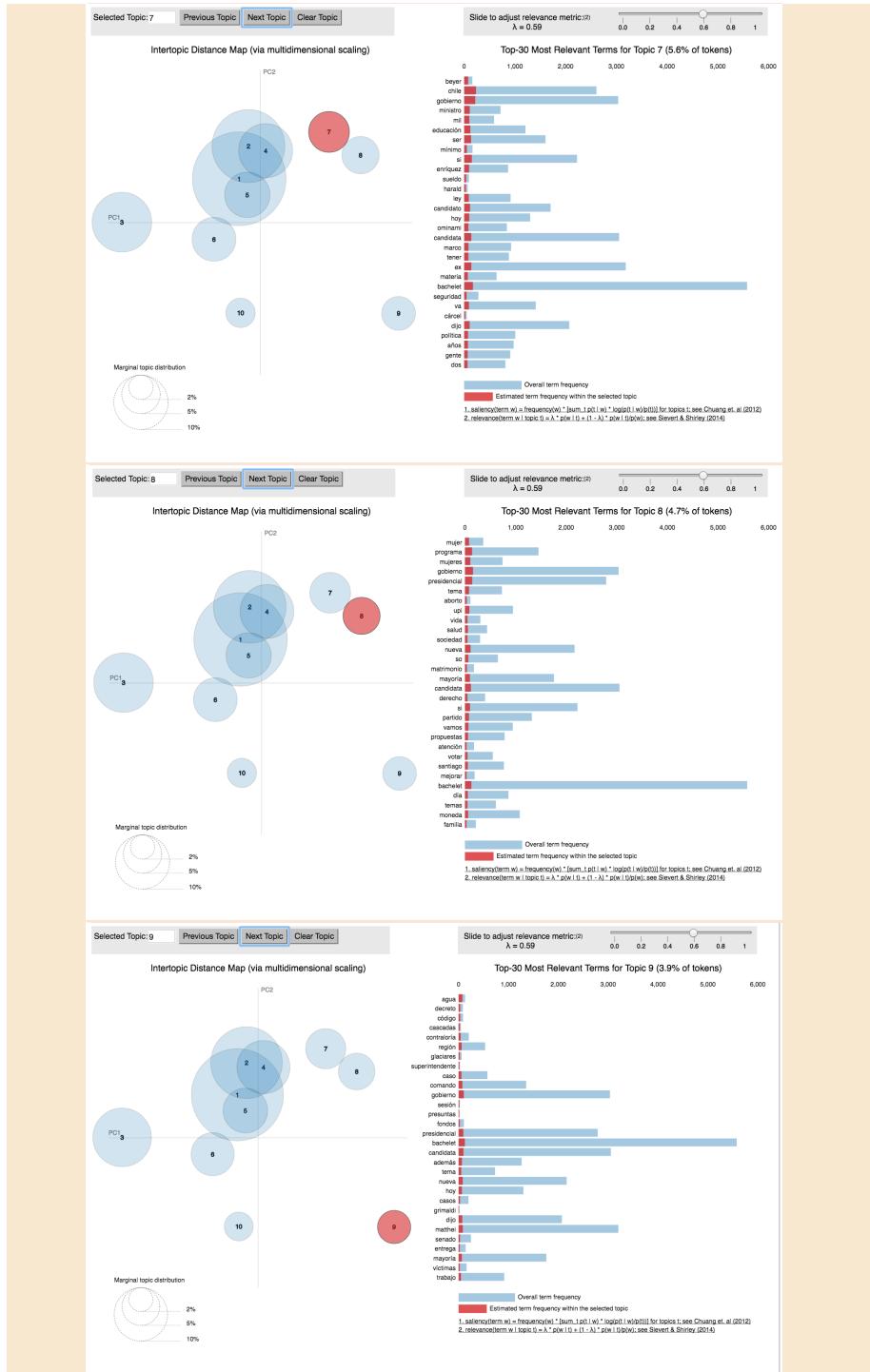
Carson Sievert, Kenneth E. Shirley. 2014. *LDAvis: A method for visualizing and interpreting topics*

## 7. Anexos

Figura 9: 10 tópicos detalle







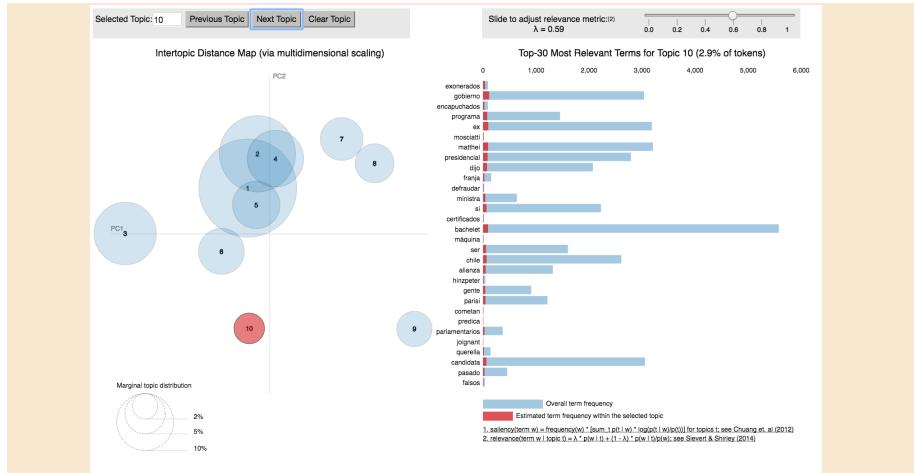
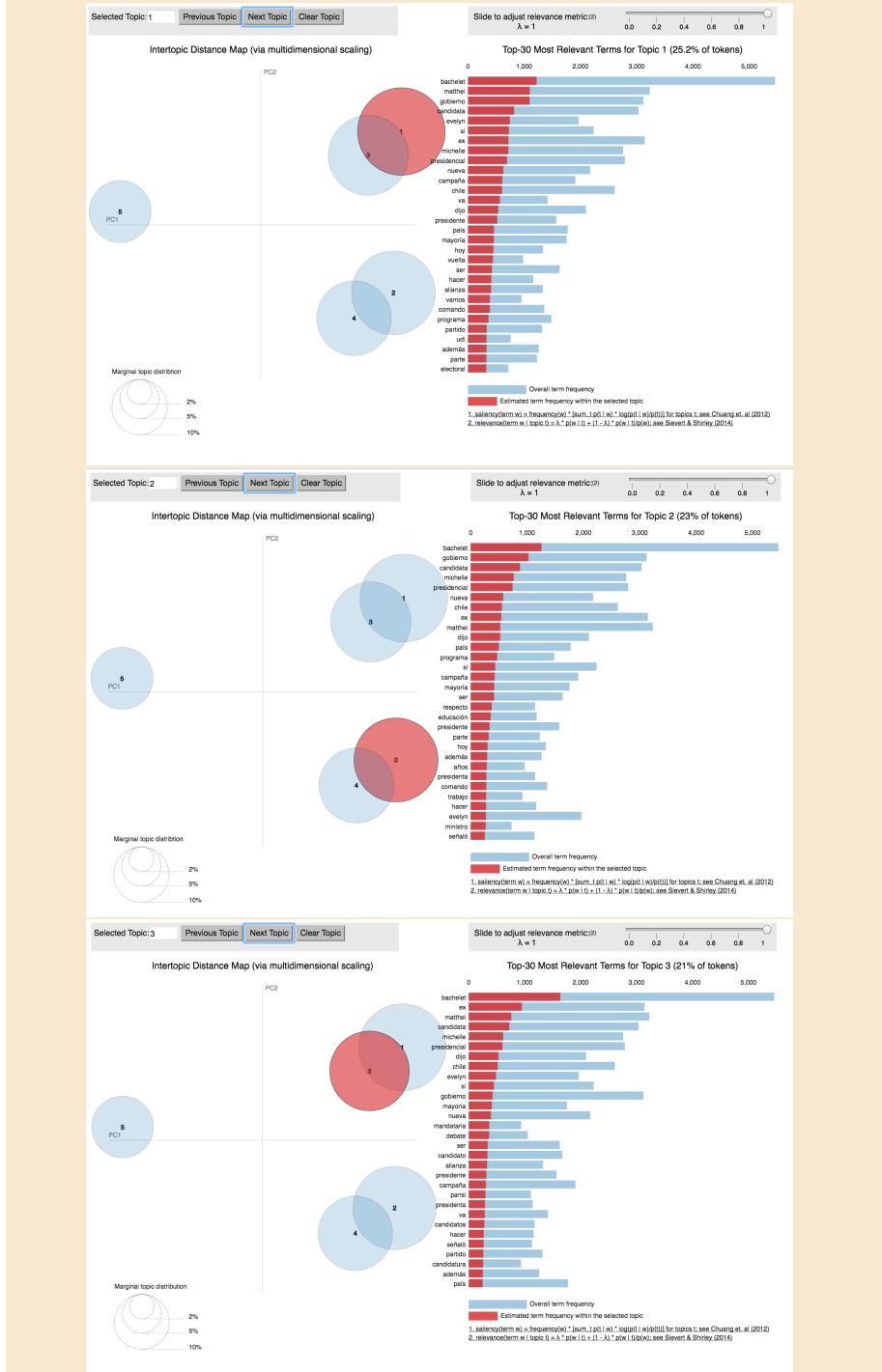


Figura 13: 5 tópicos detalle



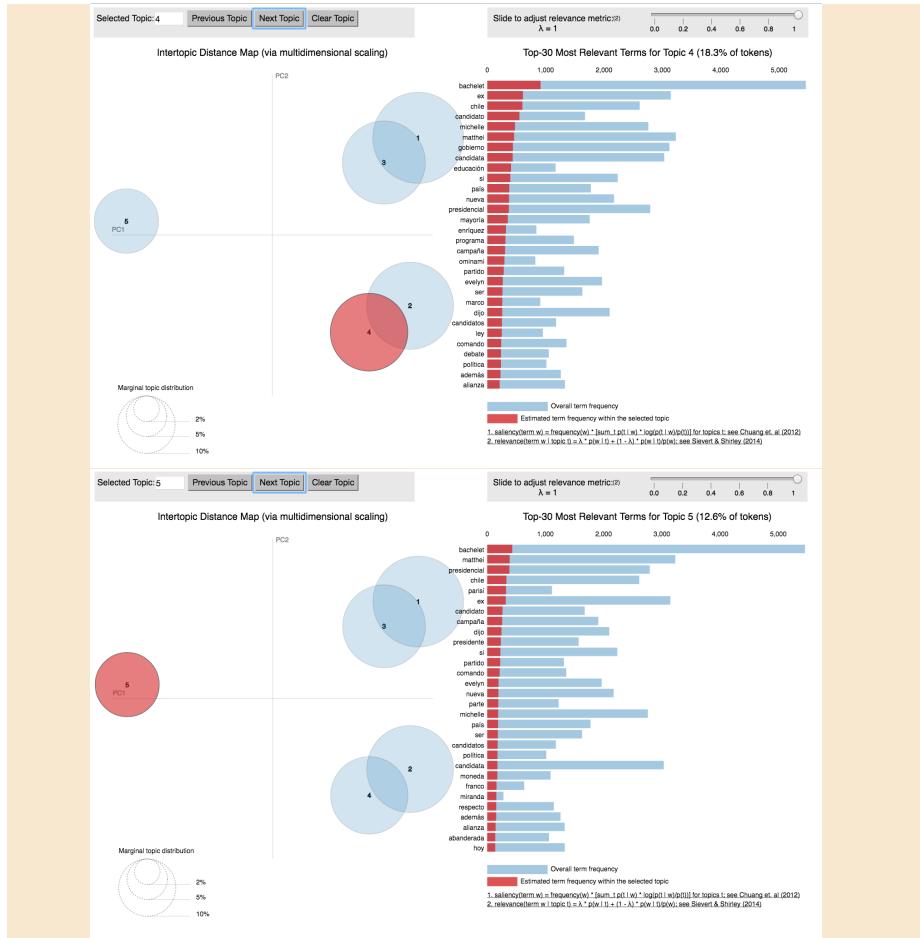
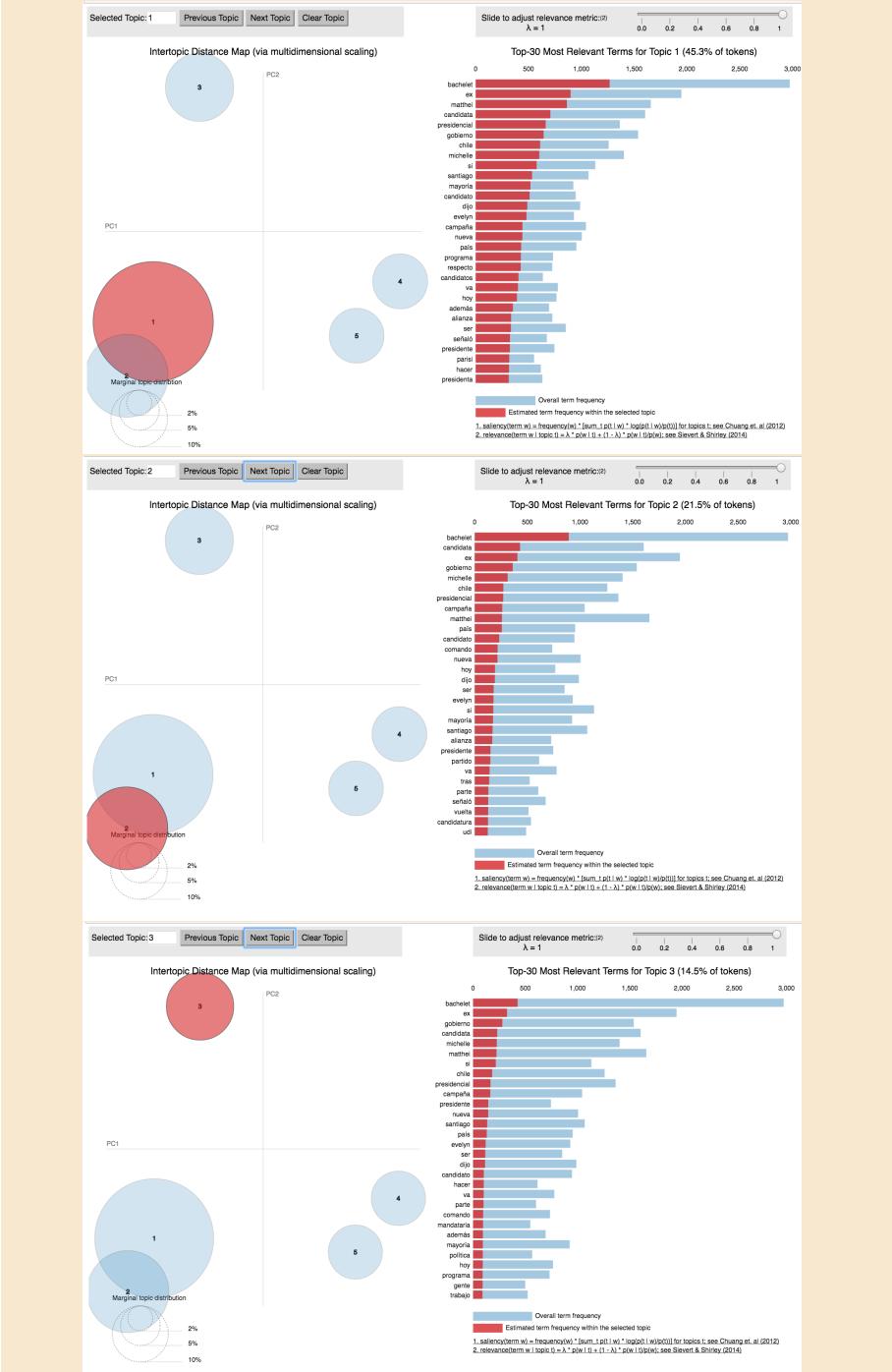


Figura 15: bonus 5 tópicos detalle Emol y La Tercera



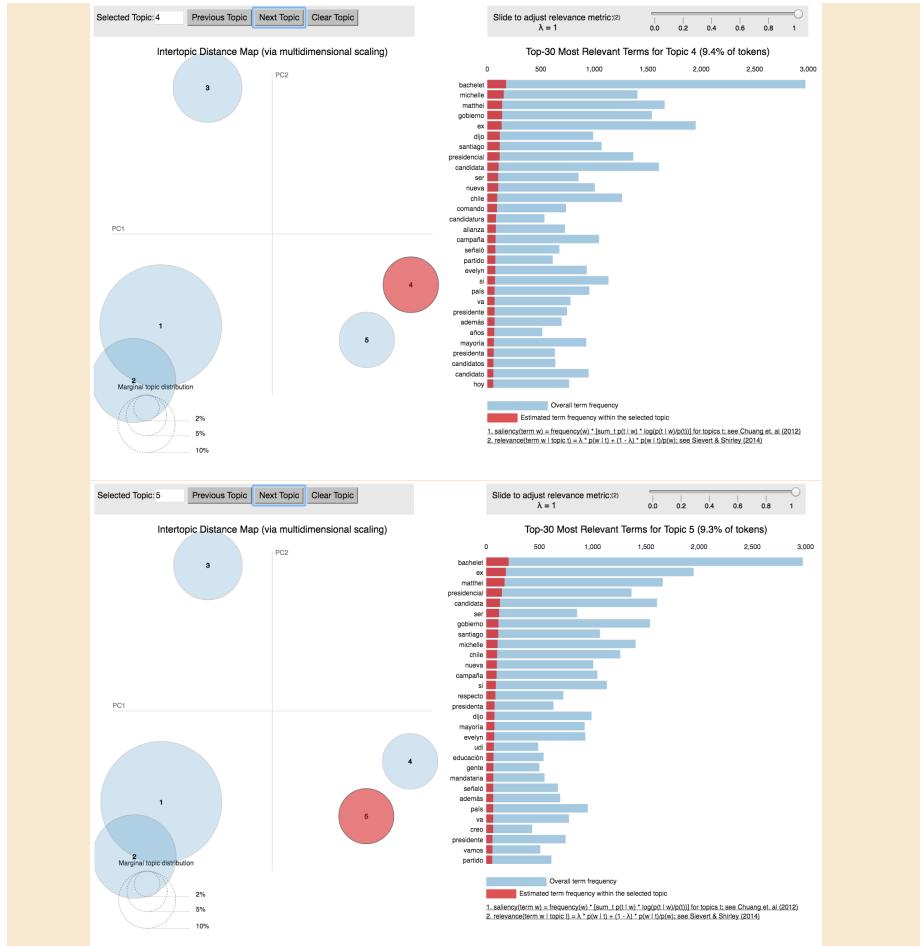


Figura 17: bonus 5 tópicos detalle La Nación y El Mostrador

