APLICACIÓN EN LOS MEDIOS DE PRENSA DE UN AGRUPAMIENTO K-MEANS (CLUSTERING K-MEANS)

Diana López*

Amaru Fernández**

RESUMEN

El objetivo de este trabajo es determinar la cercanía lingüística entre distintos medios de prensa nacionales e internacionales, aplicando un modelo no supervisado de *Agrupamiento K-Means* sobre una base documental conformada por los diferentes sitios web de medios de prensa digitales, que son: Agencia de Prensa Francesa (afp.com), Prensa Argentina (argenpress.info), El Mercurio On-Line (emol.com) y The Clinic On-Line (theclinic.cl). Estos medios se seleccionaron por su importancia y seriedad al abordar las distintas noticias, teniendo cada uno, una posición ideológica definida. Es por esto que se agruparon como medios de prensas internacionales, correspondientes a las *Agencia de Prensa Francesa* y *Prensa Argentina*, centrados en temáticas sobre acontecimientos mundiales, y por otro lado los nacionales, como El Mercurio On-Line y The Clinic On-Line, con temáticas mixtas, nacionales como internacionales, pero fuertemente marcados al área local.

Las temáticas de las noticias recogidas son de diferentes áreas, como política, deportes, economía, entre otras. Lo relevante de la selección de estos medios es que utilizan un lenguaje formal y, en su generalidad, evitan el uso de palabras coloquiales o de la jerga local, por lo que se espera una fuerte relación entre ellas, separándolas de una posible marcada ideología del periódico.

Palabras Clave: agrupamiento K-Means, medios de prensa

ABSTRACT

The objective of this work is to determine the linguistic proximity among different national and international media applying an unsupervised model of K-Means Clustering based on a corpus of digital media from different websites: Agencia de Prensa Francesa (afp.com), Prensa Argentina (argenpress.info), El Mercurio On-Line (emol.com) and The Clinic On-Line (theclinic.cl). These media were selected on the basis of their importance and accuracy to present the news considering the different ideological points of views. The international media - Agencia de Prensa Francesa and Prensa Argentina- were analyzed in terms of world affairs, whereas the national media - El Mercurio On-Line and The Clinic On-Line – were examined around both national and international affairs, but mainly local news.

The topics chosen cover different areas such as politics, sports, and economics, among others. The relevance of the selected media lies on the formal language and the absence of colloquial and local expressions; a strong relation is expected among them only being separated by their ideological points of views.

Keywords: K-Means clustering, media.

Códigos JEL C01, C5, C42

Fechas de Recepción 30 abril 2018

Fechas de Aceptación 30 mayo 2018

*Ingeniera Comercial, Mención Economía. Universidad de Chile, con especialización en Metódos Estadísticos y Big data, con diplomados en Pontificia Universidad Católica de Chile para las respectivas áreas, además de un Magíster en Finanzas (c) , en la Universidad de Chile, con 5 años de experiencia en recopilación de información, compilación de estadísticas, investigación y análisis de mercados internacionales, primero en la Facultad de Economía y Negocios de la Universidad de Chile y desde junio de 2012 en el Banco Central de Chile. Correo electrónico: dlopez@bcentral.cl

**Ingeniero informático de la Universidad de Playa Ancha, Diplomado en Big Data de la Universidad Católica y Arte Sonoro de la Universidad de Chile. Especializado en desarrollo informático, Machine Learning y construcción de espacios de características para Machine Learning.

1. INTRODUCCIÓN

El objetivo de este trabajo es determinar la cercanía lingüística entre distintos medios de prensa nacionales e internacionales, aplicando un modelo no supervisado de ***Agrupamiento K-Means*** sobre una base documental conformada por los diferentes sitios web de medios de prensa digitales, que son: Agencia de Prensa Francesa (afp.com), Prensa Argentina (argenpress.info), El Mercurio On-Line (emol.com) y The Clinic On-Line (theclinic.cl). Estos medios se seleccionaron por su importancia y seriedad al abordar las distintas noticias, teniendo cada uno una posición ideológica definida. Es por esto que se agruparon como medios de prensas internacionales, correspondientes a las **Agencia de Prensa Francesa** y **Prensa Argentina**, centrados en temáticas sobre acontecimientos mundiales, y por otro lado los nacionales, como El Mercurio On-Line y The Clinic On-Line, con temáticas mixtas, tanto nacionales como internacionales, pero fuertemente marcados al área local.

Las temáticas de las noticias recogidas son de diferentes áreas, como política, deportes, economía, entre otras. Lo relevante de la selección de estos medios es que utilizan un lenguaje formal y, en su generalidad, evitan el uso de palabras coloquiales o de la jerga local, por lo que se espera una fuerte relación entre ellas, separándolas de una posible marcada ideología del periódico.

2. ESTADO DEL ARTE

Antes de comenzar a desarrollar este ejercicio, se deben definir ciertos conceptos que explican de mejor manera lo relacionado con un proceso de agrupamiento *K-Means*. En este sentido, los conceptos de aprendizaje de máquinas, vectores de soporte, análisis no supervisado, entre otras definiciones, son necesarias incorporar para entender íntegramente el proceso y resultados obtenidos.

2.1. Tratamiento de bases documentales

Las bases de datos se clasifican en jerárquicas, de red, transaccionales, relacionales, multidimensionales, orientadas a objetos, deductivas y documentales. Es en esta última categoría donde se enmarca este ejercicio. Una ***Base de Datos Documental*** está constituida por un conjunto de programas que almacenan, recuperan y gestionan datos estructurados o extraídos de una variedad de archivos con distintas características.

Las Bases Documentales se construyen con el fin de tener una plataforma de información estructurada para el desarrollo de estudios o análisis de documentos, como una necesidad informática que permita el acceso directo a los datos en lenguaje natural[1]. El valor de las Bases Documentales es muy alto si están bien ordenadas, contienen la información correcta y están limpias de ruidos, por lo que la manera de construirlas y tratarlas incide en la calidad de los resultados que se obtengan en su procesamiento.

2.2. Procesamiento del lenguaje natural

El Procesamiento del Lenguaje Natural (de ahora en adelante NLP, por sus iniciales en inglés: *Natural Lenguage Processing*) es un concepto utilizado cuando se realizan distintas tareas sobre el lenguaje natural, que lo adecuan para su procesamiento en máquinas o herramientas computacionales, establecido según los objetivos de la información que se pretende extraer, las características de los dispositivos computacionales que se utilizarán y de acuerdo con las características de la información inicial. El NLP busca que los sistemas computacionales, máquinas y/o herramientas, sean capaces de "comprender" de algún modo la información tanto objetiva como subjetiva contenida en los textos.

1. Lenguaje natural: es la lengua o idioma hablado o escrito por humanos para propósitos generales de comunicación.

2.3. Tópicos de clasificación

Una de las aplicaciones de NLP son los tópicos de clasificación, que corresponden a la tarea de asignar etiquetas a los documentos, entendida como "categorización de texto", de esta forma se convirtierte en un punto de referencia dentro del desarrollo de problemas en las comunidades de Procesamiento del Lenguaje Natural - Máquinas de Aprendizaje. Lo anterior debido a la simplicidad y claridad de su formulación y a la relativa abundancia de textos *on line* clasificados de forma manual, susceptibles a la experimentación y a las muchas aplicaciones de esta naturaleza existentes en el mundo real.

2.4. Máquinas de aprendizaje

Existen dos definiciones para el concepto de Máquinas de Aprendizaje. La primera, más antigua y general, es la de Arthur Samuel que las describe como: "el área de estudio que da a los computadores la capacidad de aprender sin que se les hubiera explicitado la información"[2]. La segunda, con mayor aceptación en la actualidad, es la de Tom Michell: "Se dice que un programa de ordenador aprende experiencia de E respecto de alguna clase de tareas T y medida de rendimiento P, si su rendimiento en las tareas en T, medido por P, mejora con la experiencia E"[3].

Las Máquinas de Aprendizaje, también conocidas como ML por sus siglas en inglés, *Machine Learning*, han sido uno de los pilares de las tecnologías de información en las pasadas dos décadas, aunque muchas veces ello no fuera visible. No obstante, la creciente cantidad de información que se requiere procesar en la actualidad constituye una buena razón para considerar que su utilización en la actualidad es imperiosa.

Las Máquinas de Aprendizaje cumplen la función de obtener información valiosa y específica a través de su capacidad de procesamiento y análisis sobre cuantiosas cantidades de datos.

2.5. Análisis de texto

El término Análisis de Texto, también conocido como ***Minería de Datos***, fue formalizado en 2004, cuatro años después de que el profesor Ronen Feldman, reconocido especialista en esta área, modificara la descripción de Minería de Texto como un área en la que convergen muchas actividades, ya que para realizar un análisis de texto se deben tener bases documentales, NPL y ML. Lo anterior porque la información de calidad se obtiene a través del reconocimiento de patrones y tendencias, mediante el uso de herramientas que tengan la capacidad de realizar un aprendizaje estadístico de estos mismos, como lo hacen las Máquinas de Aprendizaje. A su vez, la Minería de Textos por lo general implica un proceso de estructuración del texto de entrada, pues un buen análisis incorpora algunas características lingüísticas y la eliminación de otras, lo que se puede obtener mediante la aplicación de tareas de NLP, con la posterior inserción en una base de datos o base documental, generando datos de salida que pueden manipularse para análisis y/o estudios.

El Análisis de Texto busca aprovechar la excesiva cantidad de datos digitales escritos para extraer de ellos el mayor número de información de calidad posible, dado su espacio de aplicación, abarca una gran diversidad de áreas del conocimiento.

2.6. Clasificación de noticias

Existen páginas que extraen de manera automática noticias de diarios de diferentes procedencias, relacionándolas a partir del contenido del texto, permitiendo tener un historial de las noticias y/o ver cómo es abordada la misma en diferentes medios.

2. Samuel, Arthur (1959). Some studies in machine learning using the game of Checkers. *IBM Journal of Research and Development*.

3. Mitchell, Tom (1997). *Machine Learning*. McGraw-Hill Science.

2.7. Construcción de bases documentales

La construcción de una Base Documental está conformada por dos partes. La primera es la extracción de información, a cargo de una Araña Web, también conocida como ***Web Crowler***, que se encarga de extraer la materia prima desde internet. La segunda parte es la regularización de los documentos, ordenando la información a una misma escala para adecuarla a la realización de las próximas tareas.

2.7.1. Araña web

Araña Web es un programa que inspecciona y/o extrae las páginas que se encuentran en internet siguiendo algún criterio, de manera tanto manual como automatizada; por ejemplo, mediante bots[4]. Estas arañas inician su procedimiento con la visita a una URL o a una lista de URLs, identificando los enlaces que se encuentran alojados allí, agregándolos, todos o una selección de ellos, a una lista de URLs que luego se revisa. En este punto la araña repite el proceso, que solo se detiene con los requerimientos del usuario. Las páginas así seleccionadas pueden ser descargadas o simplemente analizadas, dependiendo del objetivo para el cual fue realizado este proceso.

2.7.2. Regularización de documentos

En la actualidad un documento, una página web, etc., son documentos no estructurados debido a que no siguen reglas respecto de cómo se debe ordenar la información, por lo que siempre al iniciar algún trabajo de investigación, es necesario estructurar los datos extraídos para adecuarlos a la realización de un trabajo, creando con ello una base documental.

En esta investigación se verá que al realizar un análisis sobre noticias periodísticas, se encontrará con que cada medio de comunicación *on line* tiene su propia manera de estructurar la información; es decir, si los medios de trabajo van a ser emol.com y elmostrador.cl, se encontrará con que la ubicación donde Emol almacena la noticia está entre ... , diferenciándose de lo que hace El Mostrador, que las almacena entre <div class=”news-body gsf cf”>... . Entonces, cuando se logra extraer las informaciones de los dos periódicos, es importante que estas sean almacenadas con una misma estructura, sin perder los datos importantes, como la procedencia de estos documentos, construyendo así una ***Base Documental***.

2.7.3. Normalización

En el lenguaje cotidiano tenemos muchas formas de expresar lo mismo, tanto en la manera escrita como hablada. Lo anterior, cuya interpretación resulta natural para una persona, no lo es para una máquina, donde cada forma de expresión de una misma idea es analizada como una idea independiente de las otras. Esta característica de los idiomas ha llevado a los programadores a desarrollar la tarea de normalización, en la búsqueda de convertir a todos los elementos de un texto con idéntico significado a una misma forma de expresión.

Por ejemplo, en el idioma inglés se encuentran muchos modos de simplificar la escritura y el habla, como el caso de “*she isn’t*” y “*she’s not*” para sustituir a “*she is not*”, entonces, en la normalización se opta por usar una de estas tres formas, pues al final son todas las mismas, pudiéndose convertir a todas ellas a la forma “*she is not*”, y en español estos casos pueden ser encontrados en el lenguaje informal, ejemplo: “xq”, “tb”, etc.

Otro punto importante es en referencia al uso de letras mayúsculas, ya que en términos computacionales las letras mayúsculas tienen un valor diferente a las letras en minúsculas. La normalización, en este caso,

4. Bot (aféresis de robot) es un programa informático, cuyo objetivo es imitar el comportamiento de un humano.

requiere que el programador opte por la transformación de todos los textos a solo letras mayúsculas o a solo letras minúsculas, eliminando toda combinación. Por otro lado, para las máquinas computacionales es irrelevante considerar la variable tiempo en el lenguaje, sin importar si se trata de presente, pasado o futuro, por lo que en la normalización de todos los verbos se llevan a su infinitivo. Por ejemplo, si se tiene "ella caminará", "ella caminó" y "ella camina", todas estas frases se transforman en "ella caminar".

A su vez, es necesario referirse a los sustantivos plurales y singulares, llevándolos todos a su forma singular; es decir sin la letra *s* o *es* finales, que les diferencia en la mayoría de los casos.

2.8. Segmentación de sentencias

La función principal de la segmentación es la de identificar y separar los tokens[5] o componentes léxicos presentes en el texto, de manera que cada palabra individual y cada signo de puntuación constituya un token diferente. El módulo considera las abreviaturas, las siglas y los números con decimales, o las fechas en formato numérico, con el objetivo de evitar separar el punto, la coma o la barra, de los elementos anteriores y/o posteriores. Para ello se utiliza un diccionario de abreviaturas y otro de siglas, así como un conjunto de reglas para la detección de números con decimales y de fechas en formato numérico (dd/mm/aa).

2.9. N-Grama

La intención de poder capturar el contenido semántico[6] de una información escrita, se ha desarrollado con una herramienta capaz de extraerlo, el N-Grama. Consiste en agrupar las palabras en una cierta cantidad, donde su número corresponde a la variable "N", generando los Uni-Grama, Bi-Grama, Tri-Grama hasta llegar a N-Grama. Así, es posible capturar las frases coincidentes en un determinado texto o en comparación con otro. Por ejemplo la frase: "el mundo es agradable"
Y algunas de sus posibles separaciones en N-Grama serían:

Uni-Gramas serían: "el" - "mundo" - "es" - "agradable"
Bi-Gramas serían: "el mundo" - "mundo es" - "es agradable"

2.10. Jerarquización de recuperación de la información

El texto por sí solo carece de valor para las máquinas computacionales, por lo que la información escrita debe pasar del medio subjetivo a un medio cuantificable.

2.10.1. Diccionario

Al tratar con textos es importante fijar una base de trabajo, que en este caso corresponde a los N-Gramas ubicados en el "diccionario". Un diccionario consiste en ubicar dentro de una matriz a todos los N-Gramas existentes, extraídos de la revisión de todos los documentos. Esto permite conocer los datos con los cuales se está trabajando y permite saber, además, de manera intuitiva si la información que está expuesta es capaz de satisfacer lo que se busca o si la información fue tratada de manera coherente. Los N-Gramas repetidos solo se deben ubicar una vez en el diccionario.

5. Token: un elemento del vocabulario en el texto.

6. *Semántica* es la ciencia de los significados de los signos lingüísticos o de los enunciados orales o escritos, donde su objetivo es poder extraer el qué quiere decir.

Documento Uno					
la	ciudad	estaba	muy	muy	tranquila

Documento Dos			
la	mar	estaba	serena

Diccionario						
la	ciudad	mar	estaba	muy	tranquila	serena

2.11. Matrices

Este ejercicio requiere la construcción de matrices especiales que tengan características particulares, para que el tratado y la extracción de la información de bases documentales sean realizadas con éxito. Las matrices son descritas a continuación.

2.11.1. Matriz de incidencia del término en el documento

Matriz Incidencia del Término en el Documento, también conocida como ***La Matriz Booleana***, indica los términos utilizados y en cuales documentos. Se trata de una matriz $M_{(ixj)}$ donde cada columna representa un documento y cada fila representa un $(n-grama)_i$. En cada celda se coloca un 1 o un 0, de acuerdo con la existencia o no del $(n-grama)_i$ en el $(documento)_j$.

Matriz de Incidencia del Término en el Documento							
Diccionario	la	ciudad	mar	estaba	muy	tranquila	serena
Documento 1	1	1	0	1	1	1	0
Documento 2	1	0	1	1	0	0	1

2.11.2. Matriz de la frecuencia del término en el documento

La Matriz de la Frecuencia del Término en el Documento, como su nombre lo indica, muestra la frecuencia con que se utiliza un término en un documento. Se trata de una Matriz $M_{(ixj)}$, donde son los N-Gramas del diccionario j son los documentos. Dentro de cada celda se ubican las frecuencias de los N-Gramas en cada documento.

Matriz de la Frecuencia del Término en el Documento							
Diccionario	la	ciudad	mar	estaba	muy	tranquila	serena
Documento 1	1	1	0	1	2	1	0
Documento 2	1	0	1	1	0	0	1

2.11.3. Frecuencia de término - frecuencia inversa de documento

Frecuencia del Término - Frecuencia Inversa del Documento (ahora en adelante TF-IDF, del Inglés *Term Frequency - Inverse Document Frequency*) corresponde a una construcción matemática en el área de la estadística numérica que refleja la importancia de una palabra en un documento dentro de una base documental o corpus[7]. Se utiliza a menudo como un factor de peso en la recuperación de información dentro de la minería de texto. La matriz TF-IDF incrementa su valor en proporción al número de veces que la palabra aparece en el documento, pero se contrarresta con la frecuencia de la palabra en el corpus, pudiéndose detectar y controlar aquellas que son más comunes que otras, como los conectores.

Las variaciones de los esquemas de preponderancia de TF-IDF son a menudo usadas por máquinas de búsqueda como una herramienta central en el valor de los resultados y en el *ranking* de importancia que tiene el documento en la consulta.

7. Corpus: conjunto de datos, textos u otros materiales sobre determinada materia que pueden servir de base para una investigación o trabajo.

2.11.4. Cálculo de frecuencia del término

Frecuencia del término (ahora en adelante TF, del Inglés *Term Frequency*) se expresa como $tf(t,d)$, donde la frecuencia del término en el documento, tf, corresponde al número de veces con que el término aparece en el documento d . También, se puede indicar la frecuencia de t por $f(t,d)$, por lo que, en este caso $tf(t,d) = f(t,d)$.

Otras posibilidades incluidas:

• Frecuencia Booleana:

$$tf(t,d) = 1 \; si \; t \; existe \; en \; d \; y \; 0 \; si \; no \; existe$$

• Frecuencia de escala Logarítmica:

$$tf(t,d) = 1 + logf(t,d) \; (y \; 0 \; cuando \; f(t,d) = 0)$$

Frecuencia normalizada:

Esta frecuencia se usa para evitar un sesgo hacia los documentos más largos, correspondiendo a la frecuencia del término dividida por la frecuencia máxima del término con mayor frecuencia en el documento:

$$tf(t,d) = \frac{f(t,d)}{\max\{f(w,d): w \in d\}}$$

2.11.5. Cálculo de la frecuencia inversa de documento

La Frecuencia Inversa de Documento (ahora en adelante IDF, del Inglés *Inverse Document Frequency*) es una medida que indica si el término es común o poco común a través de todos los documentos. Se obtiene dividiendo el número total de documentos por el número de documentos que contienen el término, y luego a ese cociente se le aplica el logaritmo.

$$idf(t,d) = log\frac{|D|}{|\{d \in D : w \in d\}|}$$

2.11.6. Cálculo de frecuencia de término - frecuencia inversa de documento

El Cálculo de Frecuencia de Término - Frecuencia Inversa de Documento (ahora en adelante **TF-IDF**, del Inglés *Term Frequency - Inverse Document Frequency*) corresponde al resultado de la multiplicación entre la frecuencia del término y la frecuencia inversa del término. Este indica la relevancia que tiene la palabra en el conjunto de documentos (corpus) asociado a su documento, debido a que si la palabra es más común en los distintos documentos esta tendrá menor valor a si la palabra es menos común.

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

2.12. Análisis no supervisado

El Aprendizaje no Supervisado es un método que no requiere de información previa para la obtención de nuevo conocimiento, puesto que a partir de un conjunto de datos dados una máquina es capaz de generalizar una información a través de un modelo que los ajuste, relacionándolos entre sí y distinguiendo la distancia de los datos entre los posibles grupos conformables, para su clasificación final.

2.12.1. Agrupación (Clustering)

Una manera para desarrollar el Análisis no Supervisado es la separación y posterior agrupación de los elementos, donde en su mayoría son vectores, en conjuntos diferentes, caracterizados porque sus elementos comparten propiedades comunes. Estas propiedades comunes provienen de criterios pre-determinados, siendo los más generales: distancia, que es el espacio que separa a los elementos entre sí, y similitud, donde esta corresponde al grado de igualdad de los elementos, no pudiendo aplicarse los dos de manera simultánea, ya que los modelos de cálculo son compatibles entre sí.

Una determinada función de distancia entrega el valor de cercanía, definida como la geometría euclidiana. Por su parte, la medida más utilizada para medir la similitud entre los casos es las matriz de caso (nxn). Sin embargo, también existen muchos algoritmos que se basan en la maximización de una propiedad estadística llamada verosimilitud[8].

8. Verosimilitud es una función de los parámetros de un modelo estadístico que permite realizar inferencias acerca de su valor a partir de un conjunto de observaciones. (Evans y Rosenthal (2004). Probabilidad y Estadística. Capítulo 6, Inferencia basada en la verosimilitud. Barcelona, España: Editorial Reverté).

2.12.2. K-Means

K-Means es un método de agrupación que tiene como objetivo la separación n de elementos en k grupos o clústeres, como se representa en la Figura 1, donde k es la cantidad de clúster y $x^{(1)}, x^{(2)}, x^{(3)}, .., x^{(m)}$ son el conjunto de entrenamiento donde $x^{(i)} \in \mathbb{R}^n$ y siendo $x_0 = 1$ por convención.

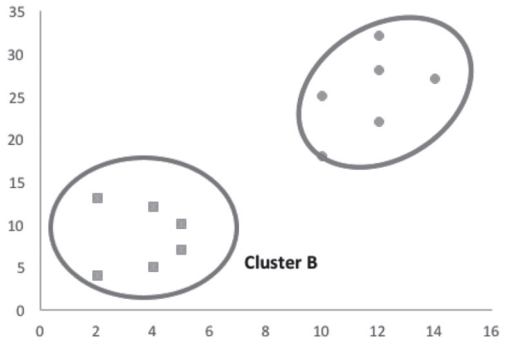


Figura 1. (*K-Means*): Clúster B, donde K=2.

A continuación, en la Figura 2, se representan los datos de entrenamiento no clasificados, mientras que en Figura 3, Figura 4 y Figura 5 se observa que la ubicación de los Centroide (centros creados por *K-Means* a los cuales se encargan de clasificar los elementos) va cambiando, y que con ello también cambia la pertenencia de los elementos asociados a cada cual, pues los elementos se ubican en el centroide que le es más cercano de entre los escogidos.

Figura 2. Datos no clasificados.

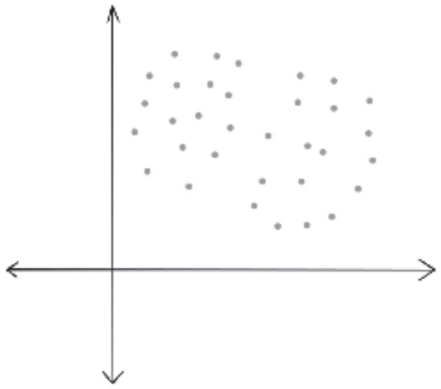


Figura 3. Datos clasificados A.

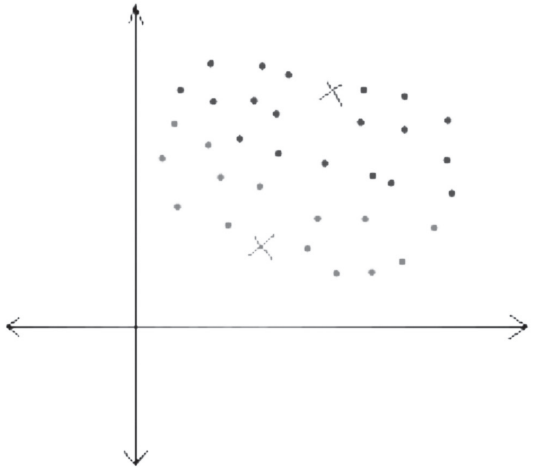


Figura 4. Datos clasificados B.

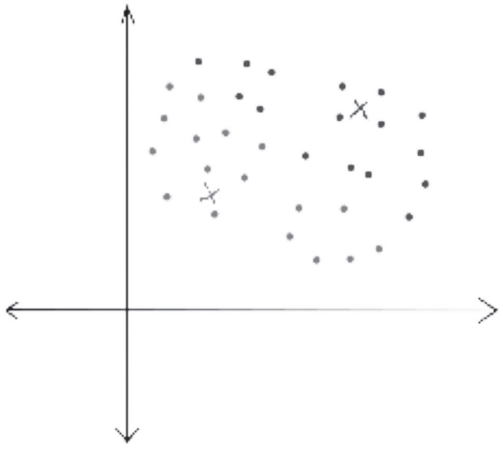
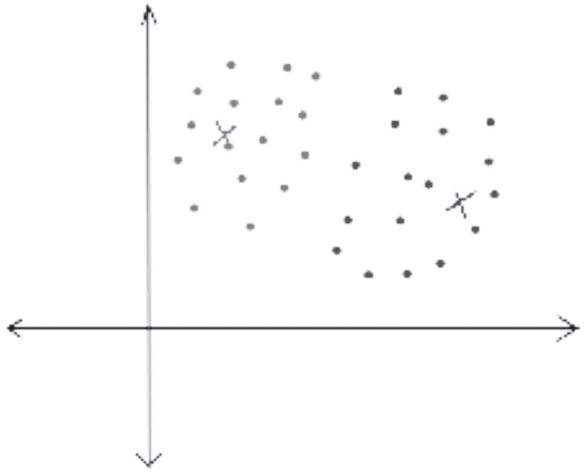


Figura 5. Datos clasificados C.



El problema para detectar la mejor ubicación de los centroides, es computacionalmente difícil, por esta razón se considera NP-Hard[9]. Sin embargo, hay algunas maneras bastante eficientes que se emplean en general y que convergen rápidamente a un óptimo local. Por ejemplo, se pueden ir ubicando los centroides de manera aleatoria en el espacio hasta llegar a la mejor solución, mediante iteraciones. Matemáticamente *K-Means* se expresa:

$x^{(i)}$ =corresponde a los datos que se van a clasificar.
$c^{(i)}$ =corresponde a la indexación de los clústeres (1, 2, 3,...,k) donde será asignado el dato $x^{(i)}$.
μ_k=corresponde al centroide del clúster (μ_k=$\mathbb{R}^n$) .
$\mu_c^{(i)}$ =corresponde al centroide del clúster donde el dato $x^{(i)}$ fue asignado.

Por ejemplo, si el dato $x^{(7)}$ es asignado al clúster 5 (μ_5), descrito como $c^{(7)}$ queda $\mu_c^{(7)} = \mu_5$.

Por otra parte, como *K-Means* corresponde a la teoría de Máquinas de Aprendizaje, su funcionamiento también es través de una función de costo que está determinada por la expresión:

$$Min\, J(C^{(1)}, \dots, C^{(m)}, \mu_1, .., \mu_k)$$

Esto quiere decir que se busca la minimización de la suma de las distancias entre los datos $x^{(i)}$ y el Centroide al cual fueron asignados y que se lo describe como $c^{(i)}$. Lo cual se expresa como la distancia que existe entre el dato $x^{(i)}$ y el Centroide μ_k , dando como resultado lo siguiente:

$$J(C^{(1)}, \dots, C^{(m)}, \mu_1, .., \mu_k) = \frac{1}{m}\sum_{i=1}^{m} \left\| x^{(i)} - \mu_c^{(i)} \right\|^2$$

En la Figura 6 se muestra la representación gráfica de lo anterior:

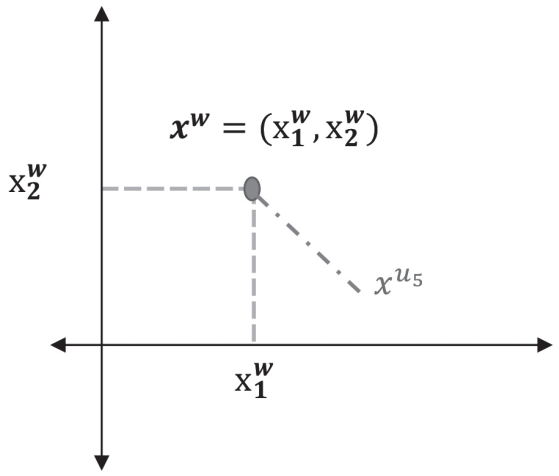


Figura 6. Distancias entre el punto $x^{(i)}$ y el centroide x^{u_5}

9. En términos básicos el NP-Hard corresponde a aquellos algoritmos que tienen un tiempo de ejecución polinomial y retornan un valor óptimo, pero no absoluto. Y se garantiza que el valor óptimo se encuentre bien definido. (Klein y Young, 1999).

3. METODOLOGÍA

La aplicación se basa en determinar una relación por medio del agrupamiento por *K-Means* sobre medios de prensa digitales, con el objetivo de determinar si las frases o el vocabulario utilizados por ellos se pueden asociar a líneas generales de pensamiento ideológico. Por ello, se establece que el área dentro de la cual se desarrolla este trabajo no excluye ningún texto de los medios internacionales, pero sí restringe los textos que se utilizarán en los medios nacionales; en este caso, solo al ámbito político. Así, lo primero es recopilar la información por analizar, extrayendo un número determinado de documentos en los medios internacionales, los cuales no están restringidos a ninguna temática en particular. Mientras que para el caso de los medios nacionales se selecciona un número indeterminado de documentos, debido a que se desconoce cuál de ellos se refiere a política y cuáles son de otra área de prensa, esto tiene como justificación alcanzar un número de documentos necesarios para comparar con el resto de los medios digitales.

Luego, una vez concluida esta fase de extracción de información, se sigue a una segunda etapa, donde estos documentos se procesan a través de un conjunto de actividades sucesivas, consistentes en limpiar y transformar expresiones lingüísticas y matemáticas, dejando la información estructurada para utilizar el agrupamiento por *K-Means*. De esta forma los datos estarán compuestos por un conjunto de vectores de N dimensiones, que permiten reconocer si los elementos, palabras o frases, se asocian o no a través de su vocabulario.

3.1. Desarrollo

Para realizar la extracción de datos, se desarrolla un *script* en MATLAB®[10], encargado de activar las funciones necesarias para llevar a cabo las distintas tareas mencionadas.

3.1.1 Recopilación de datos

La manera de obtener la información es muy diferente dependiendo del sitio web. Sin embargo, se utiliza como lenguaje principal MATLAB®, utilizando como complemento herramientas como wget[11] o *script*[12] en phpu otros lenguajes.

Las páginas emol.com y theclinic.cl entregan URLs, a través de sus sitios, que permiten acceder a los RSS3.0[13] que ellos proporcionan. Por su parte, los sitios afp.com y argenpress.info entregan, dentro de cada noticia, un vínculo a las entradas pasadas.

3.1.2. Araña RSS

La araña RSS funciona a través de un *daemon*[14] que se ejecuta cada cierto tiempo, descargando la información que se encuentra contenida en www.emol.com/rss.asp?canal=1 (emol.com) y www.theclinic.cl/feed/ (theclinic.cl). Los códigos de las arañas están en MATLAB®, siendo las funciones: **AranaTheCl** y **AranaEMOL**, que no necesitan ningún elemento extra

10. En la actualidad existen software que ya tienen incorporados este tipo de procesos, lo que hace mucho más rápido y simple la extracción de información desde los medios digitales. Uno de los más conocidos es R-Studio.

11. GNU Wget es una herramienta libre que permite la descarga de contenidos desde servidores web de una forma simple.

12. Son las siglas en inglés de "Hypertext Pre-Processor", que significa "Lenguaje de Programación Interpretado". Este lenguaje, permite la visualización de contenido dinámico en las páginas web

13. RSS corresponden a las siglas de Rich Site Summary, en referencia a un sistema que permite a los usuarios de Internet recibir las últimas informaciones de un sitio web determinado.

14. Daemon en Unix corresponde a las siglas de Disk And Execution Monitor, también conocido como "Servicio" en Windows o "Programa Residente" en MS-DOS, siendo un proceso informático que se ejecuta en segundo plano, realizando tareas de manera frecuente o iterativas.

para funcionar, sino que solo se ejecutan. Luego de descargar el contenido, se analiza mediante un código que busca la ubicación de los enlaces y, a medida que estos se encuentran, se extraen y almacenan en una variable temporal.

Posterior a esta etapa, el documento es revisado con el propósito de extraer sus enlaces, preparándose para el proceso de descarga de contenido. En ocasiones en el documento RSS hay texto que no está completo, en ese caso se tuvo que recurrir a la fuente original de la información.

La Araña RSS crea un archivo en el cual están todas las URLs establecidas para incorporar, donde se chequea si esta URL ya está contenida o es necesario incluirla hasta completar la capacidad establecida de descargas de vínculos, esto evita el duplicado de información en la recopilación de estos datos.

Figura 7. Código que invoca las arañas web de emol.com y theclinic.cl.

```
>>AranaEMOL; %Funcion para descargar los documentos de emol.com

>>AranaTheCl; %Funcion para descargar los documentos de %theclinic.cl
```

3.1.2. Araña por extracción de link

Dentro de los sitios afp.com y argenpress.info existe un *link* que permite revisar las noticias previas –con fechas y/u horas anteriores- colocando, además, al final de cada noticia un *link* que lleva a la entrada noticiosa previa o, también, si la noticia no es la última, a la entrada siguiente. Entonces, como en general el proceso de extracción de la información se inicia con las últimas noticias del día, se elige hacer el recorrido hacia la entrada anterior, por lo que el *script* necesita que se le ingrese el número de elementos por descargar. Como se comentó, en MATLAB® están las funciones **AranaAFP(x)** y **AranaAP(x)** (Figura 8), donde la x

corresponde al número de elementos que se desean descargar, y donde el funcionamiento de esta araña consiste en que primero se descarga la página inicial, donde se busca el *link* de la página anterior, el cual se almacena en una variable. Este *link* es comparado con una lista que contiene las URL ya descargadas y, si no se le encuentra allí, se descarga, repitiéndose el proceso de revisión y de descarga hasta que se obtienen todos los documentos establecidos.

```
>>AranaAFP(x); %Funcion para descargar los documentos de afp.com
%donde x corresponde al número de documentos que se
%pretenden descargar
>>AranaAP(x); %Funcion para descargar los documentos de
%argenpress.info
%donde x corresponde al número de documentos que se
%pretenden descargar
```

Figura 8. Código que invoca las arañas afp.com y argenpress.info.

3.1.3. Selección

En el caso de emol.com y theclinic.cl, se decide trabajar con documentos que traten solo de política; por lo tanto, una vez descargados los archivos, se los revisa de manera manual, uno a uno, para seleccionar aquellos que cumplan con el requisito, hasta cumplir con mínimo de 100 documentos de cada medio, lo que estadísticamente cumple con las observaciones requeridas para obtener estimaciones robustas y para el correcto funcionamiento de la máquina *K-Means*.

3.2. Inicialización de la función

Como se observa en código *K-Means* (Figura 9), al momento de invocar la función de Agrupamiento_KMeans, este debe ir junto con el número de N-Gramas, el número de clústeres "k" con lo que se va a realizar *K-Means* y la opción para realizar o no la gráfica multidimensional, que se activa si la variable ingresada es 'si'.

Figura 9. Código función agrupamiento *K-Means*.

```
function Agrupamiento_KMeans(ngrama,k,mp)
```

3.3. Regularización

Una vez inicializada la función, se comienzan los procesos de regularización de la información, consistentes en dejar la información legible y sin los elementos que entorpezcan la actividad.

3.4. Limpieza PHP

En algunos casos es posible encontrarse con elementos propios de php dentro de los documentos que son extraídos desde Internet, como por ejemplo **"ñ"** correspondiente a la representación de la **"ñ"** en php. También, puede aparecer **"á"** que corresponde a la representación de **"á",** o algún otro carácter. En estos casos, se tiene la función LimpiaPHP encargada de analizar y detectar la existencia de este tipo de caracteres que luego permite transformar en su representación común, como se mencionó en los ejemplos anteriores.

3.5. Regularización de los documentos

El proceso de regularización difiere según el medio de comunicación sobre el cual se va a trabajar, ya que como se menciona con anterioridad, la manera en cómo las páginas administran su información no está reglamentada, por lo que es importante poner atención en qué zonas de los documentos se encuentra la información.

3.5.1. Regularización de EMOL

Al observar los códigos de las noticias de Emol, comparándolos uno a uno, como una tarea previa, se concluye que dentro de Emol la regularización es un proceso simple, según se observa en Figura 10, donde la zona en que el texto se contiene se encuentra separada de los otros elementos como imágenes o videos, encontrándose que la información que se necesita procesar comienza a continuación del div <**div class="EmolText"**> y ubica el texto entre las etiquetas <**p**> y </**p**>, teniendo, además, señalado el término del escrito con <!-- [TW] -->; por tanto, para extraer los datos solo se debe tener en cuenta desde el inicio del div hasta la marca señalada, para luego eliminar todas las etiquetas sobrantes.

Figura 10. Sección HTML que contiene el artículo en emol.com.

```
<div class="EmolText">
    <p>…(Texto Noticioso)…</p>
   <br>
    <p>…(Texto Noticioso)…</p>
   <br>
    <p>…(Texto Noticioso)…</p>
     <!-- [TW] -->
</div>
```

3.5.2. Regularización de The Clinic

De la misma manera que Emol, primero se revisaron los códigos de las noticias y se compararon uno a uno, sin embargo, a diferencia de Emol, The Clinic es algo más complejo, ya que, como se puede observar en el código de la Figura 11, este contiene información que no es necesario registrar, como imágenes o algunos ítems del texto que hacen alusión a artículos pasados. Por este motivo, todo el texto de importancia se encuentra dentro del <**div id="HOTWordsTxt"**>...</**div**> siendo el primer paso la extracción de todo lo que se encuentra en este div. Luego, se quitan los elementos innecesarios, correspondientes a las etiquetas HTML, cuidando la conservación intacta del contenido del texto. Así, en el caso de la línea "<**strong**> <**a href="URL_de_la_referencia"**> Referencia </**a**> </**strong**>" hay que dejar solo **"Referencia"**, borrando lo que se encuentra contenido entre <...>, dejando el texto libre de elementos innecesarios.

Figura 11. Sección HTML que contiene el artículo en theclinic.cl.

```
1 <div id="HOTWordsTxt">
2    <p><a href=" URL/imagen.jpg">
3         <img src="URL/imagen.jpg" alt="" title="…" width="…"height="…" class="…">
4        </a>
5   </p>
6    <p>…(Texto Noticioso)…<strong>
7         <a href="URL_de_la_referencia">Referencia</a>
8         </strong>…(Texto Noticioso)…<br>
9         <span id="more-348361"></span>
10        <br>…(Texto Noticioso)…
11    </p>
12     <p>…(Texto Noticioso)…</p>
13     <p>…(Texto Noticioso)…</p>
14  </div>
```

3.5.3. Regularización de AFP

Como se observa en el código de la Figura 12, el contenido necesario se encuentra ubicado en "<**div id=”release-content”**>...</**div**>", por lo que, lo primero es extraer toda la información contenida en el div. Una situación de importancia en AFP es que en el título, línea 3, se encuentra inserta información *spam* y en el caso de que la palabra "(AFP)" queda escrita junto con el texto, ello podría generar que los elementos se junten, esto se debe a que contienen información que los relaciona; por esta razón, en el caso de afp.com, el título debe ser borrado, conservando solo lo que se encuentra entre los tag <**p**> y </**p**>, de esta forma se deja solo la información importante.

Figura 12. Sección HTML que contiene el artículo en afp.com.

```
1  <div id="release-content">
2      <div>
3     <strong>BERLÍN (AFP)</strong>
4     <p></p>
5     <div style="width:245px" class="leftSide">
6     <img style="width:auto" class="leftSide" src="imagen.jpg">
7     <div class="clearFix"></div>
8     <div class="title"></div>
9     </div>
10        <p>…(Texto Noticioso)…</p>
11        <p>…(Texto Noticioso)…</p>
12        <p>…(Texto Noticioso)…</p>
13    </div>
14  </div>
```

3.5.4. Regularización de ArgenPress

ArgenPress, al igual que el resto de los medios de prensa, tiene la ubicación del texto en la página global, que es lo que se encuentra entre el ... en el código de la Figura 13. Al momento de iniciar la limpieza se aprecia que la información está ubicada dentro de los tag ">...
", donde también existen otros elementos, como es el caso de la línea 2, pues en el título está contenido el nombre del autor junto a información *spam* como: (especial para ARGENPRESS.info), en la misma ubicación que el texto importante, ">...
". Entonces, para extraer la información sin elementos contaminantes, no se toma en cuenta la primera línea debajo del div contenedor y se elimina toda la información que no cumpla con las características mencionadas, como es el caso de las líneas 7 a la 17, a excepción de la línea 9, que contiene texto noticioso.

```
1  <div class="post-body" id="…">
2      Nombre Autor (especial para ARGENPRESS.info)<br>
3      <br>…(Texto Noticioso)…<br>
4      <br>…(Texto Noticioso)…<br>
5      <a href="URL_de_la_referencia.html">
6              URL_de_la_referencia.html
7      </a>
8      <span id="fullpost">
9              …(Texto Noticioso)…
10             <br><br>
11             <br><a href="…(URL)…">
12             <span style="font-size: 78%;">
13                     Haga click aquí para recibir gratis Argenpress…
14             </span></a>
15     </span>
16     <div style="clear: both;"></div>
17 </div>
```

Figura 13. Sección HTML que contiene el artículo en argenpress.info.

3.6. Proceso de tratado de los documentos

3.6.1. FreeLing

Una vez se haya finalizado el proceso de Regularización se inicia el proceso de Procesamiento del Lenguaje Natural, para desarrollar esta tarea se utiliza la función **FreeLing**, la que tiene la capacidad de poder desarrollar varias tareas de NLP, como son el caso de normalización y segmentación. En la figura 14, muestra que el código requiere de la CarpetaA, que es donde se encuentran los documentos regularizados anteriormente, la CarpetaB la cual es "/Datos_Cfreeling", es donde se almacenan los documentos con las tareas de NLP[15] ya aplicadas, y para finalizar necesita que se le determine el idioma en que se encuentra la base documental, la cual en este caso corresponde a español y se le comunica a través de la sentencia 'es' permitiendo desarrollar con éxito los tópicos de NLP.

```
CarpetaA=CarpetaB;
CarpetaB=[Almacenamiento '/Datos_D-mat'];

if exist(CarpetaB)==0
      Txt2mat(CarpetaA,CarpetaB);
end
```

Figura 14. Función FreeLing encargada de realizar los tópicos NLP.

3.7. Construcción de matrices y vectores

Una vez que la información ya fue adaptada, se realiza el proceso de la creación de N-Gramas, donde el objetivo es formar cadenas de palabras con distintas longitudes para lograr realizar el análisis de *K-Means*. Desde aquí se comienza con la generación y transformación de los elementos, *input* con los que las Máquinas de Aprendizaje funcionan, como lo son las matrices y los vectores.

15. Procesamiento lenguaje natural.

3.7.1. Determinación de los grupos

La función Docs (Figura 15) consiste en crear un arreglo para almacenar el nombre del documento original y el medio noticioso al cual corresponde. Para efectuar lo anterior, a través de la variable **Originales**, se le proporciona la ubicación de los archivos que contienen la separación de los distintos medios de comunicación y de los nombres de las noticias en ellos. Esto permite comparar las agrupaciones que realiza *K-Means* con las agrupaciones de datos originales. Por último, el arreglo creado por la función Docs es almacenado en el archivo (Docs.mat), cuya ubicación está en la variable f_final.

```
if exist([CarpetaB '/Docs.mat'])==0

Docs(Originales,f_final);


end
```

Figura 15. Función Docs que realiza la correspondencia de los archivos.

Posteriormente, se construye la Matriz de Incidencia del Término en el Documento, a través de la función **MatrizITD,** que trabaja con los archivos creados por la función **FTDocumento**, donde la ruta de estos archivos se encuentra almacenada en la variable CarpetaDFT. La función MatrizITD (Figura 16) además necesita el valor N-Grama para determinar con que N-Grama se trabaja y, por último, necesita la ubicación donde se almacena el archivo, que en el caso de ngrama=1 tiene el nombre "dic-n1_MITD.mat" y su ruta de almacenamiento se encuentra en la variable f_final.

```
if exist([f_final '/dic-n' num2str(ngrama) '_MITD.mat'])==0
      disp(['/dic-n' num2str(ngrama) '_MITD.mat']);
      MatrizITD(CarpetaDFT,f_final,ngrama);
end
```

Figura 16. La función MatrizITD crea la Matriz de Incidencia del Término en el Documento.

A continuación se aplica la función MatrizFTD, la cual usa el mismo proceso operativo y carpetas que la función MatrizITD, como se muestra en la Figura 17, donde se debe considerar que ngrama=1 se almacena con el nombre "n1-dic_WandF.mat" y posteriormente será almacenada en la variable f_final.

```
if exist([f_final '/n' num2str(ngrama) '-dic_WandF.mat'])==0
disp(['/n' num2str(ngrama) '-dic_WandF.mat']);
MatrizFTD(CarpetaDFT,f_final,ngrama);
end
```

Figura 17. La función MatrizFTD crea la Matriz de la Frecuencia del Término en el Documento.

Finalmente, con este se pasa aplicar el Tf-Idf, donde se requiere de la MatrizITD y MatrizFTD.

3.8. Agrupación K-Means

Al método de agrupamiento *K-Means* se le debe asignar de manera previa el número de clústeres ("k") con los que correrá el modelo. En este caso, como se utilizan vectores multidimensionales, se hace necesario realizar un proceso para graficar dichas dimensiones y generar resultados visibles en un formato plano, la función **MultiPlot** (Figura 18) permite realizar la gráfica descrita. A su vez, dentro de la variable f_final se encuentra la ruta donde almacena el archivo creado por esta función, de nombre **multiplot.mat**, siendo Tf-Idf la variable que contiene los elementos que se desean graficar.

```
if strcmp(mp,'si');
    if exist([f_final '/multiplot.mat'])==0
          eval(['load ' f_final '/n' num2str(ngrama) ...
               '-TfIdf.mat']);
          TfIdf=full(TfIdf);
          MultiPlot(f_final,TfIdf,3,'si');
  else
          TfIdf=0;
          MultiPlot(f_final,TfIdf,3,'no');
   end
   clear TfIdf;
   end
```

Figura 18. La función MultiPlot encargada de realizar un gráfico con multidimensional.

En el código anterior el número 3, que se puede observar en la función MultiPlot, puede tomar también el valor de 2, ya que corresponde al número de dimensiones sobre las que se van a graficar las N dimensiones. El último valor (si o no) indica si existe o si es necesario calcular las diferentes dimensiones. Además es importante recalcar que el cálculo de las diferentes dimensiones puede convertirse en un proceso muy lento, dependiendo del número de dimensiones que se tengan.

En relación con el número de clúster escogidos ("k"), este número dependerá del objetivo del ejercicio. Por ejemplo, si se busca reconocer dos líneas ideológicas en los medios de prensa (derecha e izquierda), k debe ser igual a dos. Por otro lado, si el objetivo es entender el comportamiento de cada uno de los medios por separado, se podrían usar cuatro clúster (cuatro medios escritos). Como siempre los resultados son inciertos, la dispersión de los puntos en torno a los clúster es impredecible, donde los puntos correspondientes a un medio se podrían juntar y/o separar entre sí.

Referente al objetivo de este estudio, en el Cuadro N°1 se muestra un extracto de las noticias y el clúster que le corresponde a cada una de ellas.

Medio	Nombre Artículo	Cluster
AFP	12-soldados-libaneses-mueren-en-combates-contra-radicales-sunies-al-sur-del-pais	1
AFP	al-menos-20-muertos-en-irak-en-varios-atentados-con-coches-bomba	1
AP	amazonia-maravilla-universal-con.html	1
AP	amelia-garcia-logramos-recuperar-el.html	2
Emol	allamand-insiste-en-importancia-de-la-unidad-de-la-alianza-para-derrotar-a-la-nueva-mayoria.html	2
Emol	democracia-cristiana-desarrolla-jornada-de-elecciones-internas.html	1
Thecl	vamos-a-derrotar-a-la-concertacion-en-noviembre	1
Thecl	acusan-a-hinzpeter-de-lavarse-las-manos-en-caso-sobreprecios	1

Cuadro N° 1. Extractos de resultados de los medios de comunicación escritos.

3.9. Resultados

El comportamiento de los datos implica que entre los medios argenpress.info (ap), theclinic.cl (thecl) y afp.com (afp) se agrupan o se encuentren más cercanas sus relaciones, siendo emol.com (emol) el medio con mayor dispersión en relación con los otros medios escritos, asignándole un clúster propio (Figura 19), como se observa en la parte demarcada (círculo verde).

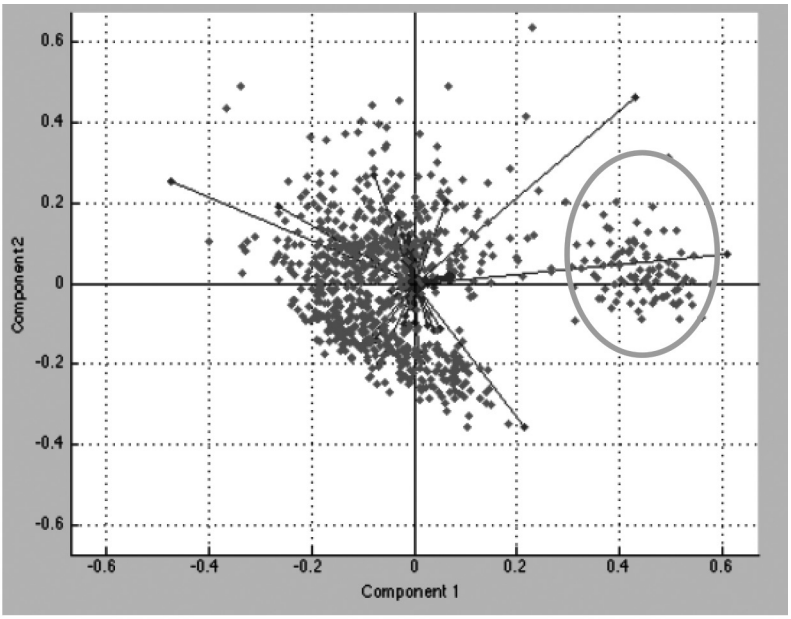


Figura 19. Gráfico multidimensional.

Estos resultados señalan que las ideologías de argenpress.info (ap), theclinic.cl (thecl) y afp.com (afp) son más tendientes a la izquierda. Por el contrario, emol se concentra en pensamientos políticos tendientes a la derecha o mucho más conservadores. Lo importante de estos resultados, además de reconocer los tipos de ideologías, es que por medio de la lingüística que utiliza la prensa digital se puede determinar qué quieren informarnos y de qué forma, además de reconocer a qué público va enfocada la noticia. Por lo tanto, este tipo de metodologías o ejercicios aplicados tiene un alcance mucho más amplios de lo que se muestra en este documento; por ejemplo, en el área de la economía y las finanzas podría inducir a los lectores a enfocarse en alguna estadística en particular o en un instrumento en específico de inversión.

REFERENCIAS BIBLIOGRÁFICAS

Sontag, E. (1972). Temas de Inteligencia Artificial.

López Takeyas, B. (2007). *Introducción a la inteligencia artificial*. Nuevo Laredo, México: Instituto Tecnológico de Nuevo Laredo.

Smola, A. y **Vishwanathan, S. V. N.** (2008). *Introduction to Machine Learning*. Cambridge University Press.

Voronisky, F. y **Martínez, R.** (2012). *Proyecto de aplicación: Sistema multiagente para la simulación de desastres*. Centro Regional de Enseñanza en Ciencia y Tecnología Espacial para América Latina y el Caribe. Tonantzintla, Puebla, México.

Pressmann, R. (2005). *Ingeniería del software. Un enfoque práctico* (7ª edición). México: Interamericana editores.

Thomas, M., Pang, B. y **Lee, L.** (2006). *Get out the vote: Determining support or opposition from Congressional floor-debate transcripts*. Nueva York, Estados Unidos: Department of Computer Science, Cornell University Ithaca.

Bertram, R. (1964). *Sir: A computer program for semantic information retrieval*. MIT press, p. 191.

Klein, P. y **Young, N.** (1999). *Approximation algorithms for NP-hard optimization problems. Capítulo 34, Algorithms and Theory of Computation Handbook*. CRC Press.

Minsky, M. (1968). *Semantic Information Processing*. MIT Press.

Riehle, D. (2000). Framework Design: A Role Modeling Approach. Ph.D.Thesis, Nº 13509. Zürich, Suiza: ETH Zürich.

Brin, S. y **Page L.** (1998). *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. Stanford, Estados Unidos, CA: Computer Science Department, Stanford University.

Evans, M.J. y **Rosenthal, J.** (2004). *Probabilidad y Estadística*. *Capítulo 6, Inferencia basada en la verosimilitud*. Barcelona, España: Editorial Reverté.

Tan, N., Steinbach, M. y **Kumar, V.** (2005). *Introduction to Data Mining*. Nueva Jersey, Estados Unidos: Addison-Wesley, Upper Saddle River.

Hernández, A., Delgado, E. y **Rivera, J.** (2006). Reducción de dimensiones para clasificación de datos multidimensionales usando medidas de información. *Scientia et Technica*. Año XII. (Nº 32). UTP.