Text Mining

YESENIA GONZÁLEZ PEDRAZA

PEARSON
MARKET & OPINION INTELLIGENCE

"**La lengua es nuestra morada vital** [...]
La lengua nos hace y en ella nos hacemos..."
Manuel Alvar, "Vivir en la lengua", en *Por los caminos de nuestra lengua,* Universidad de Alcalá de Henares.

Una aplicación muy popular de *text mining* es el descubrimiento realizado por Don Swanson[1] que extrajo información derivada de colecciones de texto. Teniendo en cuenta que los expertos por lo general sólo pueden leer una pequeña parte de lo que se publica en su campo, no se dan cuenta de los nuevos desarrollos que se suceden en otros. Swanson demostró cómo cadenas de implicaciones causales dentro de la literatura médica pueden conducir a hipótesis para enfermedades poco frecuentes. Investigando las causas de la migraña, dicho investigador extrajo varias piezas de evidencia a partir únicamente de títulos de artículos presentes en la literatura biomédica mediante las siguientes ligas de información:

- El estrés está asociado con la migraña.
- El estrés puede conducir a la pérdida de magnesio.
- Los bloqueadores de canales de calcio previenen algunas migrañas.
- El magnesio es un bloqueador natural del canal de calcio.
- La depresión cortical diseminada (DCD) está implicada en algunas migrañas.
- Los niveles altos de magnesio inhiben la DCD.
- Los pacientes con migraña tienen una alta agregación plaquetaria.
- El magnesio puede suprimir la agregación plaquetaria.

Estas claves sugieren que la deficiencia de magnesio podría representar un papel en algunos tipos de migraña, una hipótesis que no existía en la literatura y que se encontró mediante esas ligas. De acuerdo con Swanson, estudios posteriores han probado experimentalmente esta hipótesis obtenida por *text mining* con buenos resultados.

El objetivo de *text mining* es presentar herramientas que faciliten la gestión y la descripción de *corpus* o textos de gran tamaño y que permitan derivar información de ellos desde el punto de vista estadístico.

Estudios recientes indican que el 80% de la información de una compañía está almacenada en forma de textos o formatos no estructurados. La minería de texto se enfoca en descubrir entre una gran cantidad de información textual: tendencias, desviaciones y asociaciones.

Con *text mining* es fácil saber cómo distintos segmentos utilizan diferentes palabras para expresar la misma idea. Identificar las variaciones en el uso del lenguaje es muy importante si deseamos comunicarnos con nuestros clientes usando sus propias palabras, si queremos saber cómo verbalizan su agrado o desagrado, qué palabras usan para referirse a una campaña, a un comercial, a un programa de gobierno o para describir los usos de nuevas tecnologías o redes sociales.

1. Swanson, D.R.; Smalhaiser, N.R. (1994). *Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease.* Neuroscience research communications. Vol. 15, pág. 1-9.

La búsqueda de una estrategia adecuada depende de los objetivos. ¿Se requiere saber el vocabulario empleado por personas de diversas características con respecto a un producto o servicio? ¿Es necesario conocer cómo se expresa un mismo concepto o idea en diferentes edades?

Unidades léxicas y segmentación del *corpus*

Iniciamos por la unidad del análisis. Esta unidad puede ser un carácter, una palabra, un conjunto de palabras. Si quisiéramos hacer un análisis de los mensajes de texto por celular, una unidad de análisis interesante puede ser el carácter. Tal vez nos interese saber las palabras asociadas a un producto, por ejemplo: *"go"*. Si lo que nos interesa es saber si un concepto está transmitiendo lo que deseamos, la unidad puede ser un segmento[2]: **"La vida es fresca".**

La aplicación de *text mining* en investigación de mercados tiene varias vertientes:

- *Blogs*
- *Focus group*
- Preguntas abiertas

En el caso de las preguntas abiertas, diversos estudios han concluido que la forma de preguntar, ya sea abierta o cerrada, cambia de manera radical los resultados. Un experimento clásico es el realizado por Schuman y Presser (1981)[3], que compara los dos tipos de cuestionamiento. Cuando se preguntó: "¿Cuál es el problema más importante que enfrenta este país [EE.UU.] en la actualidad?", 16% de los estadounidenses mencionan la delincuencia y la violencia de manera abierta, luego se realizó de manera cerrada y se obtuvo 35% de la misma respuesta.

Aun cuando una pregunta se realiza de manera abierta, las respuestas pasan por un procedimiento de codificación que presenta varias desventajas:

- **Mediación del codificador:** que le imprime su sello personal a la codificación.
- **Empobrecimiento del contenido:** cuando la pregunta permite una gran diversidad de respuestas, como en pruebas de concepto o cuando preguntamos qué le comunica cierto mensaje o producto.
- **Las respuestas raras, originales y poco claras se asignan a códigos residuales:** este código residual muchas veces es el código "otros".
- **Destrucción de la forma:** la forma de la información se mutila y a menudo su contenido se empobrece.

Text mining es un conjunto de técnicas de análisis que ayudan a estudiar las respuestas abiertas sin necesidad de codificar.

¿Cómo *text mining* puede resolver problemas en investigación de mercados?

De tres formas principalmente:
1. Identifica vocabulario
2. Encuentra la estructura de asociación de un texto
3. Forma *clusters* semánticos

2. Toda sucesión de palabras no separadas por un signo de puntuación constituye un segmento del *corpus*.
3. Schuman, H. & Presser, F. (1981) *Questions And Answers in Attitude Surveys*, New York, Academic Press.

1 Identificación de campos semánticos por segmentos

La idea es encontrar el espectro de palabras que nuestros clientes usan para describir cierto producto o servicio, hallar las palabras más frecuentemente utilizadas en cada grupo y el contexto de las mismas. El procedimiento consiste en identificar por segmentos **predefinidos** el espectro de palabras que usan hombres y mujeres, por grupos de edad o nivel socioeconómico por ejemplo. Una de las partes más interesantes del análisis es que un grupo se identifica no sólo por el vocabulario que usa sino también por el que **no usa.**

Ejemplo:
Realizamos un análisis de los títulos de las revistas AMAI publicadas desde 1997 hasta enero de 2009. El objetivo es analizar si esta publicación nos podría dar un panorama de la evolución de la investigación de mercados en los últimos 12 años[4]. Segmentamos la variable tiempo en cuatro grupos: 97-99, 00-03, 04-06, 07-09 y analizamos el vocabulario empleado en cada uno.

El ordenamiento de elementos en la tabla 1 pone de manifiesto los rasgos sobresalientes a primera vista. Las palabras repetidas se ordenan en forma decreciente con respecto a los valores de la prueba estadística. En la parte superior están las palabras y segmentos que son más frecuentes en los grupos (especificidad positiva). Al final de la lista están las palabras o segmentos menos repetidos o menos citados (especificidad negativa).[5]

El principio que se utiliza es valorar la probabilidad de que una palabra aparezca en el segmento tomando en cuenta el número de veces que aparece en el *corpus*. La palabra **i** está suprarrepresentada o infrarrepresentada en el segmento contra lo que un modelo de probabilidades (modelo hipergeométrico) dejaba prever.[6]

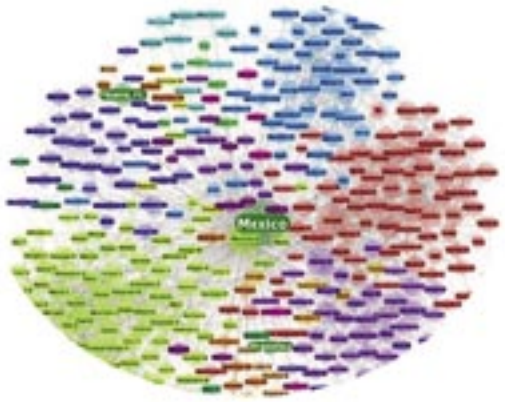
Encontramos que el periodo 97-99 es una etapa de afirmación, de consolidación como asociación. La palabra AMAI tiene especificidad positiva. Es interesante observar que los artículos con el título en forma de pregunta

Tabla 1
Segmento:1997-1999

Characteristic words or segments	*Internal percentage*	*Global frequency*	*Internal frequency*	*Global frequency*	*Test-Value*	*Probability*
Especificidad Positiva						
Entrevista	1.15	0.19	5	6	3.453	0.000
Seminario	1.15	0.22	5	7	3.135	0.001
Eventos	0.69	0.13	3	4	2.337	0.010
Perspectiva	0.69	0.13	3	4	2.337	0.010
Encuesta	1.83	0.80	8	25	2.127	0.017
AMAI	**1.38**	**0.54**	**6**	**17**	**1.999**	**0.023**
Especificidad Negativa						
Importancia	0.00	0.19	0	6	-0.240	0.405
Consumo	0.00	0.19	0	6	-0.240	0.405
Necesidad	0.00	0.19	0	6	-0.240	0.405
Ventas	0.00	0.19	0	6	-0.240	0.405
Fox	0.00	0.19	0	6	-0.240	0.405
¿?	**1.38**	**1.70**	**6**	**53**	**-0.319**	**0.375**

4. El análisis de los títulos de la revista AMAI en el periodo 97-09 es un análisis que requeriría por sí mismo de un artículo. Aquí solo mencionamos un resultado parcial que ayuda a ejemplificar la técnica.
5. L.Lebart, A. Salem, *Statistique Textuelle,* Dunod, 1994.
6. L.Lebart, A. Salem, L. Berry, *Exploring Textual Data,* Kluwer Academic Publisher, Dordrecht, Boston, 1998, 246p.

tienen especificidad negativa, sólo 6 de los 53 títulos formulados como pregunta en los últimos 12 años aparecen en este periodo. ¿Estábamos tan seguros de lo que decíamos que no dudábamos en afirmarlo? ¿Iniciamos el milenio siendo más humildes con respecto a nuestras conclusiones?



2 Estructura de asociación en un texto

El objetivo es visualizar gráficamente la asociación de palabras con segmentos demográficos, actitudinales o de cualquier otro tipo de relevancia para el análisis como la respuesta. Cuando un patrón existe en un texto, algunas técnicas estadísticas pueden detectarlos y exhibirlos.

La herramienta básica para la visualización de asociaciones cuya salida final es un mapa, es bien conocida en investigación de mercados: el análisis de correspondencias. Sin embargo, las tablas a las cuales esta técnica se aplicará son tablas de contingencia, cuyas filas corresponden a las palabras cuya frecuencia en el *corpus* es superior a un umbral dado. Las columnas pertenecen a categorías de locutores: género, nivel socioeconómico, localidad, etcétera.

Como sabemos, el análisis de correspondencias proporciona una imagen de las asociaciones entre palabras y segmentos que se basa en la extracción de subespacios que resumen **lo mejor posible** la información de tabla de partida, por lo que se debe tener cuidado con la calidad de representación de los puntos; las conclusiones deben ser reforzadas con otros métodos.

Ejemplo:
Extraído del estudio "La salud en las ciudades" destinado a conocer los hábitos de vida relacionados con la salud. La pregunta que se analiza es: **"¿Qué es para usted la salud?".**[7]

7. L.Lebart, A. Salem, M. Becué, *Análisis estadístico de textos* Milenio 2000, 215p.

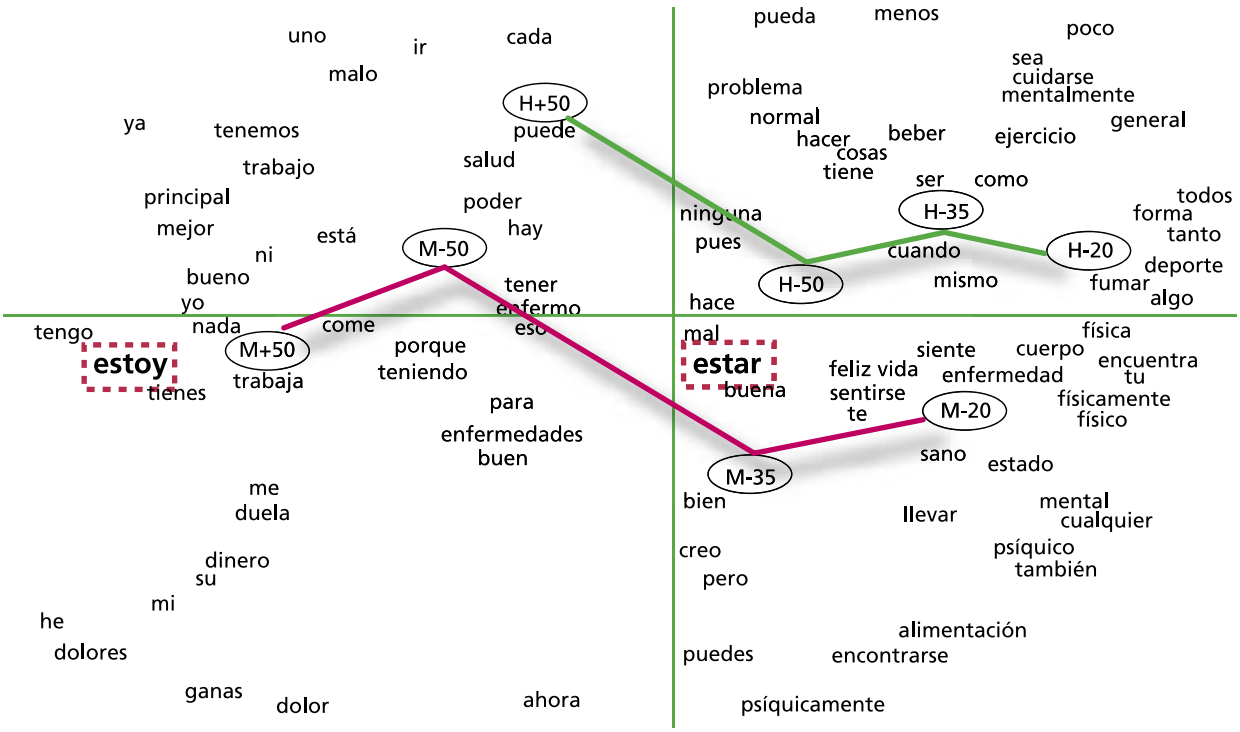
La representación de la figura 1 nos permite ver la asociación entre la utilización de determinadas palabras y las características de los individuos entrevistados:

- **La configuración observada sugiere la existencia de una evolución progresiva del vocabulario con la edad.** Para las mujeres (M) las categorías de edad vienen ordenadas a lo largo del primer eje, desde las más jóvenes a la derecha, hasta las mayores a la izquierda. La transición hacia el empleo de determinadas palabras se hace a edades distintas, a una edad más temprana para ellas.
- **Existe un uso de palabras diferenciado por sexo.** Puede observarse también que las dos trayectorias, la de las mujeres y la de los hombres (H), son paralelas y opuestas sobre el primer eje. Este fenómeno revela la existencia de palabras de uso mayoritariamente femenino y de uso mayoritariamente masculino que se oponen sobre el eje.
- **Conjugación de verbos y uso de adjetivos.** Las dos formas del verbo **estar: estoy** y **estar** se encuentran distantes en el mapa; esta disposición indica que las dos palabras se utilizan en contextos muy diferentes. **Estoy** en las personas mayores hace referencia al propio estado de salud. En un retorno al texto se puede observar que se utiliza en frases como **estoy bien** o **yo no estoy bien,** propias de las personas mayores. Por su parte, la palabra **estar**, al ubicarse cerca del origen, tiene un sentido poco diferenciado por categorías de edad (confirmado por la buena calidad de representación de la palabra en estos ejes).

3 Formación de *clusters* semántico

El objetivo es obtener grupos tan homogéneos como sea posible con respecto a sus opiniones a una pregunta abierta. A diferencia del punto anterior, los grupos se configuran de acuerdo con las palabras que mencionan en común y pueden ser una mezcla de diferentes segmentos.

Figura 1



Por ejemplo, podemos encontrar un grupo específico que se caracteriza por usar determinado vocabulario, conformado por mujeres de clase alta, mayores de 50 años, que a primera vista puede ser casi imposible de encontrar.

Los métodos de clasificación automática se usan para describir las proximidades entre palabras y luego entre las partes de un *corpus*. Como en el análisis de correspondencias, la clasificación jerárquica se aplica a tablas cruzadas o tablas léxicas y usa la misma métrica, la métrica chi2. El principio de la agregación jerárquica parte de un conjunto de **n** elementos con pesos iguales o distintos en los que se han calculado las distancias dos a dos. Primero se agregan los dos elementos más próximos; el par de elementos así agregados constituye un nuevo elemento, cuyo peso es la suma de los pesos agregados. Se calculan las distancias entre este nuevo elemento y cada uno de los elementos que quedan por clasificar. Se reitera este proceso y en la última operación se reagrupa el conjunto de los elementos en el seno de una única clase.

Ejemplo:
Estudio realizado para conocer áreas de oportunidad en el área de atención a clientes de una tienda departamental. Preguntamos: ¿Qué recomendaría al gerente de la tienda para que sus clientes quedaran más satisfechos?

Encontramos dos grupos claramente identificados, el primero conformado por mujeres de nivel alto de 50 años en adelante y otro formado por jóvenes de 18 a 25 años de los niveles C+, C.

Como observamos, las sugerencias de ambos grupos son muy diferentes. Por un lado, los jóvenes de clase media valoran la agilidad del servicio y sugieren para ello más personal, mientras que las mujeres de clase alta valoran un servicio personalizado basado en la capacitación del personal ya existente.

Es interesante analizar la velocidad con la que se agrupan determinadas palabras y los grandes grupos que conforman. El siguiente paso es analizar el contexto en el que aparecen esas palabras en segmentos largos y cuáles resultan de una identidad fortuita.

Como hemos visto, *text mining* nos puede ayudar a dar nuevas luces, nuevos caminos para entender nuestros estudios no sólo en investigación de mercados, sino en general en todas las áreas en donde se desee analizar una gran cantidad de texto.

La información derivada de las preguntas abiertas puede ser analizada de manera rápida y eficiente con este conjunto de técnicas.

En *focus groups,* los resultados han sido muy exitosos y han respaldado los resultados del análisis cualitativo clásico, ayudando a establecer niveles de importancia de variables y formación de *clusters* semánticos.

En *blogs, text mining* ha demostrado su eficacia analizando rápidamente millones de entradas y proporcionando *insights* valiosos para las marcas.

Yesenia González es actuaria por la UNAM. Realizó estudios de posgrado en estadística aplicada en el Instituto de Investigaciones en Matemáticas Aplicadas. Se especializó en investigación de mercados en el ITAM. Tiene nueve años de experiencia en el área de estudios de opinión e investigación de mercados. Yesenia ha sido profesora titular de estadística y ha colaborado en la publicación de artículos y libros en temas relacionados con estadística aplicada. Actualmente se desempeña como directora de unidad de negocio en Pearson y funge como asesora externa de Tesi México.