Centro de Estudios Sociales y de Opinión Pública

Big Data
¿Cómo innovar para alcanzar la "sacudida" que sugiere Carlos Slim?

Documento de trabajo núm. 322

Enero 2020

www.diputados.gob.mx/cesop

CÁMARA DE DIPUTADOS
LXIV LEGISLATURA

CESOP
Centro de Estudios Sociales y de Opinión Pública

Información que fortalece el quehacer legislativo

CÁMARA DE DIPUTADOS

LXIV LEGISLATURA

Información que fortalece el quehacer legislativo

CESOP

Centro de Estudios Sociales y de Opinión Pública

Centro de Estudios Sociales y de Opinión Pública

Organización Interna

Netzahualcóyotl Vázquez Vargas

Director de Estudios Sociales encargado del despacho de la Dirección General del CESOP

Enrique Esquivel Fernández
Asesor General

Ricardo Martínez Rojas Rustrian
Director de Estudios de Desarrollo Regional

Ernesto R. Cavero Pérez
Subdirector de Estudios de Opinión Pública

José Francisco Vázquez Flores
Subdirector de Análisis y Procesamiento de Datos

Katia Berenice Burguete Zúñiga
Coordinadora Técnico

Investigadores

Gabriel Fernández Espejel
José de Jesús González Rodríguez
Roberto Candelas Ramírez
Salvador Moreno Pérez
Felipe de Alba Murrieta
Rafael del Olmo González

Apoyo en Investigación

Luis Ángel Bellota
Natalia Hernández Guerrero
Karen Nallely Tenorio Colón
Ma. Guadalupe S. Morales Núñez
Nora Iliana León Rebollo
Ricardo Ruiz Flores

Alejandro Abascal Nieto
Abigail Espinosa Waldo
Elizabeth Cabrera Robles
Guillermina Blas Damián

Alejandro López Morcillo
Editor

José Olalde Montes de Oca
Asistente Editorial

Big Data

¿Cómo innovar para alcanzar la “sacudida” que sugiere Carlos Slim?

Dr. Felipe de Alba

Juana Martín

Contenido

Big Data: ¿Cómo innovar para alcanzar la “sacudida” que sugiere Carlos Slim? 3

Introducción 3

1. Dimensiones de Big Data 6

2. Algunas herramientas para el manejo de Big Data 8

a) Apache Hadoop 10

b) Apache Storm 10

3. Aprovechamiento del Big Data en redes sociales 12

a) ¿Cómo aplicar esto en la Cámara de Diputados? 12

b) Aplicaciones 14

Comentario general 18

Big Data: ¿Cómo innovar para alcanzar la “sacudida” que sugiere Carlos Slim?

Dr. Felipe de Alba
Juana Martín[1]

Introducción

Hace poco en diversos medios apareció una declaración del empresario mexicano **Carlos Slim** sobre la necesidad de una “sacudida” en México para alcanzar las aspiraciones de cambio del actual gobierno.[2] Un cambio implica transformaciones de fondo, una necesidad de innovar tal como lo afirma el empresario. Esta idea parece por demás pertinente dada la profunda renovación de las prácticas institucionales que tienen lugar hoy en el país. La necesidad no sólo de orden, sino de sistematización. La necesidad de estructurar sistemas de información no solamente es actuar institucionalmente, son rubros identificados aquí como desafíos estratégicos, es decir, centrales en la definición del futuro del país y por supuesto en la modernización del trabajo legislativo. Tal como puede imaginarse, el desafío era analizar la realidad con la mayor precisión. Ahora dicho análisis requiere la mayor diversidad por las diferentes fuentes disponibles. Volveremos sobre este punto.

[1] Felipe de Alba es Doctor en Planeación Urbana por la universidad de Montreal con Postdoctorado en Massachusetts Institute of Technology (MIT, EE.UU.) y de l´École normale supérieure (ENS) de Lyon (Francia). Es Investigador “A” del Centro de Estudios Sociales y Opinión Pública (CESOP) de la Cámara de Diputados. Juana Martín Cerón es licenciada en Estudios Socioterritoriales por la Universidad Autónoma Metropolitana (UAM)- sede Cuajimalpa.

[2] Véase Norte Digital, “Una sacudida y más infraestructura, lo que México necesita: Slim”, 19 de enero de 2019. Disponible en: https://nortedigital.mx/una-sacudida-y-mas-infraestructura-lo-que-mexico-necesita-slim/

Pensando en ello, en este documento hemos desarrollado una reflexión sobre la innovación institucional por medio de herramientas tecnológicas. Se trata del internet o el internet de las cosas o de las redes sociales o del manejo masivo de datos conocido como **Big Data**.

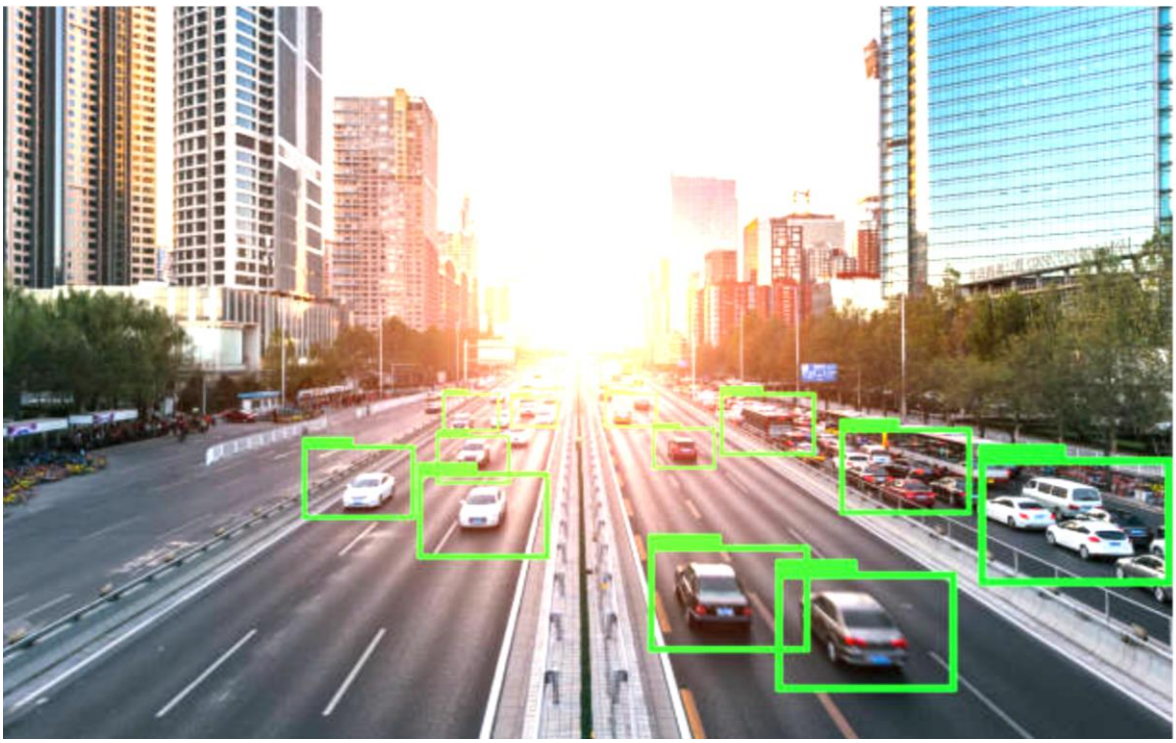
En este sentido, empecemos señalando que el internet en general o específicamente el internet de las cosas (IoT) es una de las ramas de la innovación tecnológica que está generando mundialmente volúmenes masivos de datos estructurados y no estructurados. En su mayoría, estos **datos son heterogéneos y están hospedados en diferentes sistemas**, tanto en bases de datos relacionales como bases de datos de otros tipos. Vivimos en la época con la mayor generación de datos como nunca antes. "*Data is the new science. Big Data holds the answers*".[3]

En dicho sentido (según el país hispanófono donde uno se encuentre) **Big Data** es su definición técnica, es decir, se trata de una serie de recursos para el análisis y el examen de grandes cantidades de datos. Esos datos **tienen una variedad de tipos**, **patrones, tipologías, clasificaciones,** por las cuales hay que descubrir sus patrones ocultos, correlaciones desconocidas y toda otra información útil que contenga. Esta definición –por demás simplificada– aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales.

Frente al incremento de producción masiva de información, los gobiernos, las organizaciones y las empresas privadas comenzaron a almacenar y a procesar datos en todo tipo de formatos y tamaños, además de iniciar a extraer valor de ellos. Es decir, **la obtención de información y su análisis es una oportunidad para aplicar diversas estrategias de acción.** Es allí donde esta reflexión puede ser útil

[3] "Los datos son la nueva ciencia. Big Data tiene las respuestas", traducción propia, afirmación de Pat Gelsinger, Chief Executive Officer (CEO) de la empresa VMware.

para la actividad legislativa. No sobra decir que en algunos otros documentos hemos insistido sobre el valor que dichas herramientas tienen para la práctica legislativa.[4]



Los datos se generan, se producen y se consumen por medio de un sinnúmero de sistemas, redes sociales, aplicaciones, entre otros, que se encuentran no sólo en formatos estructurados y en bases de datos tradicionales (o no) sino también en forma de imágenes, de voz o con referentes al posicionamiento geográfico, etc. Los **formatos de datos** se multiplican, por lo que los **conectores** son cada vez más fundamentales. "*Consumer data will be the biggest differentiator in the next two o three years. Whoever unlocks the reams of data and uses it strategically will win*".[5]

[4] Véase "El Bitcoin, el Blockchain y otras minucias en tiempos del Big Data", julio 2019. Disponible en:http://www5.diputados.gob.mx/index.php/esl/content/download/152004/759739/file/CESOP-IL-72-14-El%20BitcoinyelBolckchain-300719.pdf

[5] "El consumo de datos será el más grande punto de diferencia en los siguientes dos o tres años, quien acumule esos datos será quien ganará las próximas batallas", traducción propia, afirmación de Angela Ahrendts, Senior Vice President of retail en la empresa Apple.

El análisis de la **Big Data** consiste en aplicar herramientas modernas a todo tipo de datos, tanto de programación como en el uso de *software*, particularmente en el caso de datos no estructurados o semiestructurados o estructurados. Igualmente, cuando se trata del procesamiento en lotes (o de grandes volúmenes) o en las transferencias infinitamente rápidas, es decir, en tiempo real.

Entonces el propósito del análisis de **Big Data** es **descubrir información valiosa e irregularidades**, además de comprender mejor el rendimiento o la eficiencia en sus procesos, por ejemplo, de una institución o del comportamiento de sus usuarios o destinatarios (en el argot tecnológico, los clientes).

1. Dimensiones de Big Data

Existen tres características o dimensiones en las que se definen actualmente los datos para su manejo, gestión, tratamiento y análisis. Esas tres características son: **volumen, velocidad y variedad**.[6] Pasemos a los detalles.

Primero, **volumen**: cada día las empresas registran un aumento significativo de sus datos (terabytes, petabytes y exabytes) creados por personas y máquinas. En el 2000 se generaron 800,000 petabytes (PB) de datos almacenados y se espera que esta cifra alcance los 35 zettabytes (ZB) en el presente año (2020). Las redes sociales también generan datos, como es el caso de *twitter* que, por ejemplo, por sí solo genera más de 7 terabytes (TB) diariamente y en caso de *facebook* produce 10 TB de datos cada día. Algunas empresas causan terabytes de datos cada hora de cada día del año.

Segundo, **variedad**: se relaciona con el volumen, pues de acuerdo con éste y con el desarrollo de la tecnología, existen muchas formas de representar los datos (es el caso de datos estructurados y no estructurados). Estos últimos son los que se generan desde páginas web, archivos de búsquedas, redes sociales, foros, correos electrónicos o producto de sensores en diferentes actividades de las personas. Un ejemplo sería convertir 350 mil millones de lecturas de los medidores por año para predecir el consumo de energía.

Tercero, **velocidad**: se refiere a la rapidez con que se crean los datos que son la medida en que aumentan los productos de desarrollo de software (páginas web, archivos de búsquedas, redes sociales, foros, correos electrónicos, entre otros).

Estas tres características tienen coherencia entre sí, por ejemplo, si quisiéramos analizar 500 millones de registros de llamadas telefónicas al día en tiempo real o para predecir la pérdida de clientes por una empresa respecto a otras compañías que compiten en el mercado. Son tres ejes que definen el mundo actual

[6] Juan José Camargo-Vega, Jonathan Felipe Camargo-Ortega y Luis Joyanes-Aguilar, "Conociendo Big Data", *Revista Facultad de Ingeniería*, vol. 24, núm. 38, 2015, pp. 63-77. Disponible en: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-11292015000100006&lng=en&tlng=es (consulta: 20 de enero de 2020).

en su comportamiento productivo, cultural, económico y político. Son tres ejes que hay que valorar para iniciar cualquier proceso de innovación institucional.

2. Algunas herramientas para el manejo de Big Data

Para contar con grandes cantidades de información debemos desarrollar procesos de programación, algoritmos específicos o utilizar software especializado para recolectar, ordenar y analizar dicha información masiva. Es decir, el acceso a los datos y su comprensión suponen contar con habilidades técnicas específicas, lo que representa un importante desafío.

Un gran número de herramientas usadas en **Big Data** tienen la cualidad de ser abiertas (***open source***)[7] y gratuitas, lo que da fe del éxito de este modelo de desarrollo, además de las opciones alternativas de pago.

Detengámonos un momento sobre este término: **open source**. Se trata de una expresión inglesa del ámbito de la informática. Tiene dos distinciones: el software *open source* que dispone de la mencionada característica de presentar su código abierto y el *software* libre (que puede descargarse y distribuirse de manera gratuita).

Los programas más comunes de código abierto son las versiones de **Linux**, **Debían** y **Ubuntu,** pero hay otros. También puede encontrarse el antivirus **Clam Win,** así como **VideoLan** (reproductor de videos). Por último, también puede encontrarse **GIMP**, que es un programa de código libre que permite editar imágenes a nivel profesional. Todos estos programas se pueden descargar de forma gratuita y se podría bajar el código fuente para poder modificarlos. Sugerir modificaciones de un programa a la comunidad de código libre ya se lleva a cabo y si es aceptado se publica una nueva versión del programa con las modificaciones recomendadas. Creo que es posible encontrar incluso en la historia una **open source inniative**,[8] que consiste básicamente en un acuerdo internacional para el desarrollo de alternativas "abiertas" en el mundo de la informática y del desarrollo tecnológico.

Para dar un sentido más práctico a este documento, a continuación se presentan dos herramientas: ***open source,*** que ofrece soluciones para la explotación de información e iniciación de procesos de almacenamiento, procesamiento y análisis. En este caso se trata de las herramientas Apache Hadoop y Apache Storm.

[7] Definición técnica sobre el término "open source". Disponible en: https://es.wikipedia.org/wiki/C%C3%B3digo_abierto
[8] Véase https://www.definicionabc.com/tecnologia/open-source.php

a) Apache Hadoop

Hadoop es **un framework opensource** (marco de trabajo de código abierto) creado con el fin de conseguir una computación segura, escalable y distribuida. **Hadoop** está basado en los documentos de **Google** para **MapReduce** y **Google File System** (**GFS**), y permite el procesamiento distribuido de grandes conjuntos de datos por medio de *clusters* (agrupamientos por sentido, localización o similitud) de computadoras usando simples modelos de programación.

Los dos conceptos en los que se apoya **Hadoop** son, por un lado, la técnica de **MapReduce** y, por otro, el sistema distribuido de archivos **HDFS**:

- **HDFS** (*Hadoop Distributed File System*): es un sistema de archivos distribuido, escalable y portátil típicamente escrito en el lenguaje **JAVA**. Se basa en pequeños y bastos ordenadores en los que cada uno de ellos procesa una parte de información, pero actuando como si fuera uno.

- **MapReduce**: es el modelo de programación utilizado por **Google** para dar soporte a la computación paralela. Trabaja sobre grandes colecciones de datos en grupos de computadoras y sobre *commodity hardware*.

b) Apache Storm

Apache Storm es un sistema de cómputo distribuido en tiempo real gratuito y de código abierto**. Apache Storm** facilita el procesamiento confiable de flujos de datos ilimitados haciendo para el procesamiento en tiempo real lo que **Hadoop** hizo para el procesamiento por lotes. **Apache Storm** es simple, se puede usar muy fácil con cualquier lenguaje de programación.

Apache Storm tiene muchos usos: análisis en tiempo real, aprendizaje automático (*Machine Learning*) en línea, cómputo continuo, etcétera. **Apache Storm** es rápido: un punto de referencia lo registró a más de un millón de *tuplas* procesadas por segundo por nodo. Es escalable, tolerante a fallas, garantiza que sus datos serán procesados y es fácil de configurar y operar.

Apache Storm se integra con las tecnologías de colas y bases de datos que ya usa. Una topología de **Apache Storm** consume flujos de datos y procesa esos flujos de manera arbitrariamente compleja, redistribuyendo los flujos entre cada etapa de la computación según sea necesario.

No obstante, existen muchas otras herramientas que no abordaremos aquí a detalle. Pueden mencionarse **MongoDB, Elasticsearch, Apache Spark, el lenguaje R** o finalmente el **lenguaje de programación Python** con sus infinitas librerías. Este último es considerado uno de los más versátiles y que registra el más alto crecimiento y desarrollo en el mundo de la computación y de la ciencia de los datos (*Data Management*)**.**

Como puede notarse, las herramientas tecnológicas han alcanzado un nivel alto de sofisticación, por lo que a veces parece imposible su acceso al usuario común.

Sin embargo, una ventaja de estas herramientas es que constantemente simplifican sus algoritmos con objeto de hacer fácil el acceso al público en general.

3. Aprovechamiento del Big Data en redes sociales

Con la gran diversidad de herramientas para el manejo de **Big Data** se ha facilitado el análisis de redes sociales al identificar patrones como **sentimientos, interacciones o variables como las franjas de edades, la localización o las preferencias de los usuarios.**

Esa información es gratis –aunque no en todos los casos – y puede extraerse de **twitter, facebook, linkedIn e instagram**, etc. En este sentido, el manejo de **Big Data** no se refiere a alguna cantidad de información o a alguna variable en específico, sino a la variedad de sus formatos.

a) ¿Cómo aplicar esto en la Cámara de Diputados?

El acceso al constante flujo de datos que proporcionan las redes sociales permite identificar, por ejemplo, a los líderes de opinión (*influencers*) en temas relevantes, lo que el público o usuario refiere sobre identidades o gustos (o desagrados o desacuerdos, etcétera).

En el caso concreto de la utilidad para la Cámara de Diputados, la información puede servir para la **toma de decisiones, diseño de políticas públicas y para la acción legislativa pertinente** en general, y de manera particular si se piensa en la sistematización, por ejemplo, de los comentarios vertidos en dichas redes sociales sobre temas coyunturales o el tratamiento de la geolocalización de dichas opiniones o comentarios, y serían identificadas como área de influencia de usuarios o como área de influencia de líderes de opinión (*influencers*). En el caso concreto que nos ocupa se trataría de estudiar opiniones

que existen sobre un problema coyuntural en los distritos de origen de los propios representantes populares. A manera de ejemplo, sólo eso, un ejemplo, aunque podrían citarse muchos, se mencionan tres criterios de análisis para el tratamiento de la información de redes sociales:

1. **Contacto "directo".** Gestión de crisis de reputación y el estudio del sentimiento en relación con acciones o respecto a programas sociales (o "n" atributos: marca, producto, precio, competencia, etcétera). Esto se obtiene al buscar (*request*) palabras claves (*keywords*) relacionadas con las acciones o decisiones de gobierno o acuerdos o desacuerdos sobre programas sociales o en relación con la imagen de funcionarios o de representantes populares, etcétera.

2. **Optimizar el *engagement* de la estrategia de contenidos.** Se obtiene al combinar los contenidos más "exitosos", por ejemplo, de un sitio oficial o la "presencia" en las redes sociales o en los medios de comunicación digitales (por medio de *Google Analytics* + *BuzzSumo,* por ejemplo). Todo ello podría ser combinado con los intereses declarados de los seguidores y sus "me gusta" (*likes) con la* información proveniente de redes como **facebook** o con los temas de conversación de personas que hablan sobre la institución (*Social listening*). Al combinar todos estos recursos tendremos información suficiente no sólo para medir la audiencia, sino también para ofrecer contenidos adecuados cumpliendo con una doble condición: temas de interés y relaciones con el objetivo de la estrategia de difusión (el público objetivo).

3. **Mejorar la relación con destinatarios.** Es posible decir que cuanta mayor información tengamos sobre los destinatarios (información pública, por supuesto), mayor será la posibilidad de ofrecer información y atención institucional pertinente.

Sea de una forma agregada (información-tipo asignada a perfiles de clientes) o información individual (mediante *social login* con herramientas como **Gigya**) se

puede enriquecer el perfil de los destinatarios con sus intereses, su interacción con la institución o funcionario o representante o el nivel de influencia, etcétera.

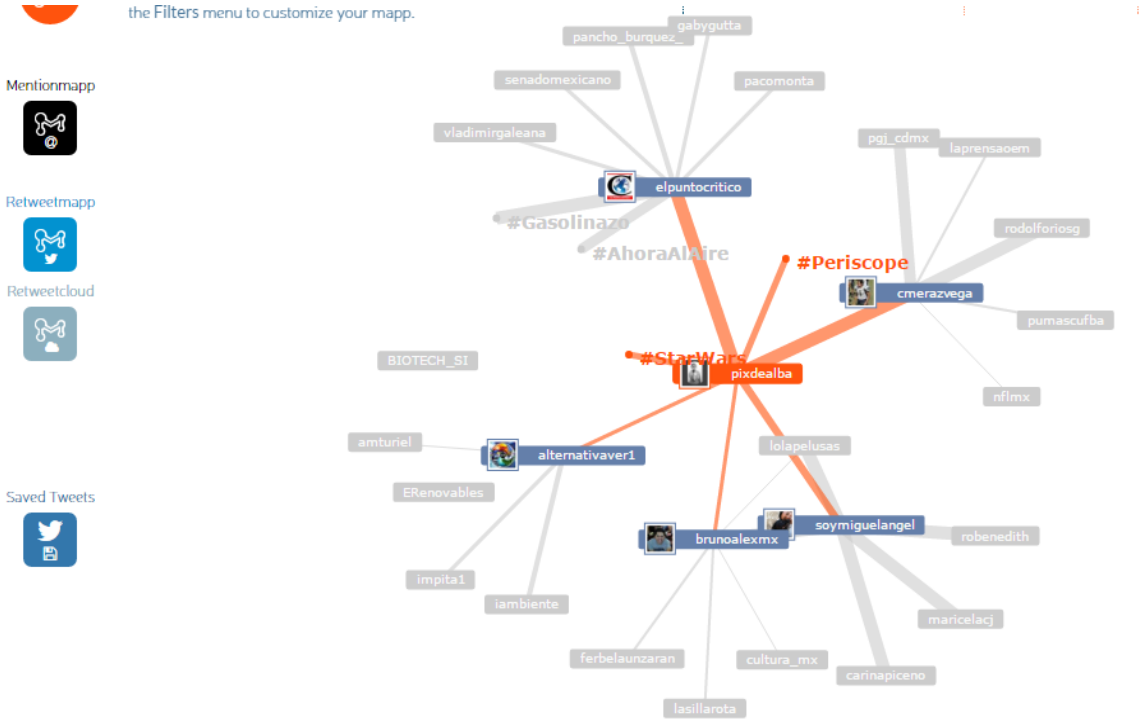
Como puede notarse, esta diversidad de opciones permitiría adelantarse a las necesidades de información para elaborar agendas legislativas pertinentes a un público activo e interactuando con él y poder contar con información instantánea sobre temas relevantes de la agenda pública en los diferentes niveles de intervención del Poder Legislativo.

b) Aplicaciones

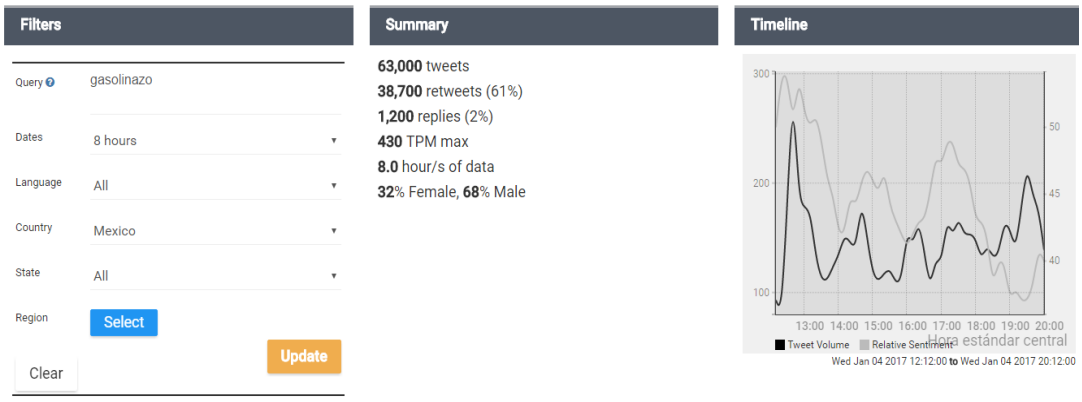
Siguiendo con el ejemplo anterior e ilustrar con más claridad algunas de las herramientas de uso común para el manejo de grandes volúmenes de información en los medios y en las redes sociales, se presenta otra lista breve de herramientas de gran utilidad.

- **Machine Learning** (aprendizaje automático) con cálculo exhaustivo, Inteligencia Artificial (AI) y algoritmos gráficos. Este tipo de herramientas requiere de una amplia explicación que tendremos que hacer en otro documento futuro. Personalmente hemos desarrollado dos libros utilizando estas técnicas.

- **Mentionmappp** (gratuita). Es una herramienta que muestra a los usuarios de **twitter** que han hecho alguna mención en forma de conexiones, generando un mapa interactivo para monitorizar el seguimiento de los usuarios, temas, campañas, menciones, hashtags tanto a nivel particular como de empresa y elaborar un informe en Social Media. Enseguida se presenta un ejemplo de un usuario ficticio.



- **Trendsmap** (de paga). Es una herramienta para conocer en tiempo real qué es lo que está sucediendo en la región, cuáles son los ***hashtags*** más populares y sobre qué tendencias se está hablando en **twitter.** Cuenta con tecnología de **GeoIP** que permite tener cierta precisión sobre la ubicación de un ***hashtag*** en una determinada región.

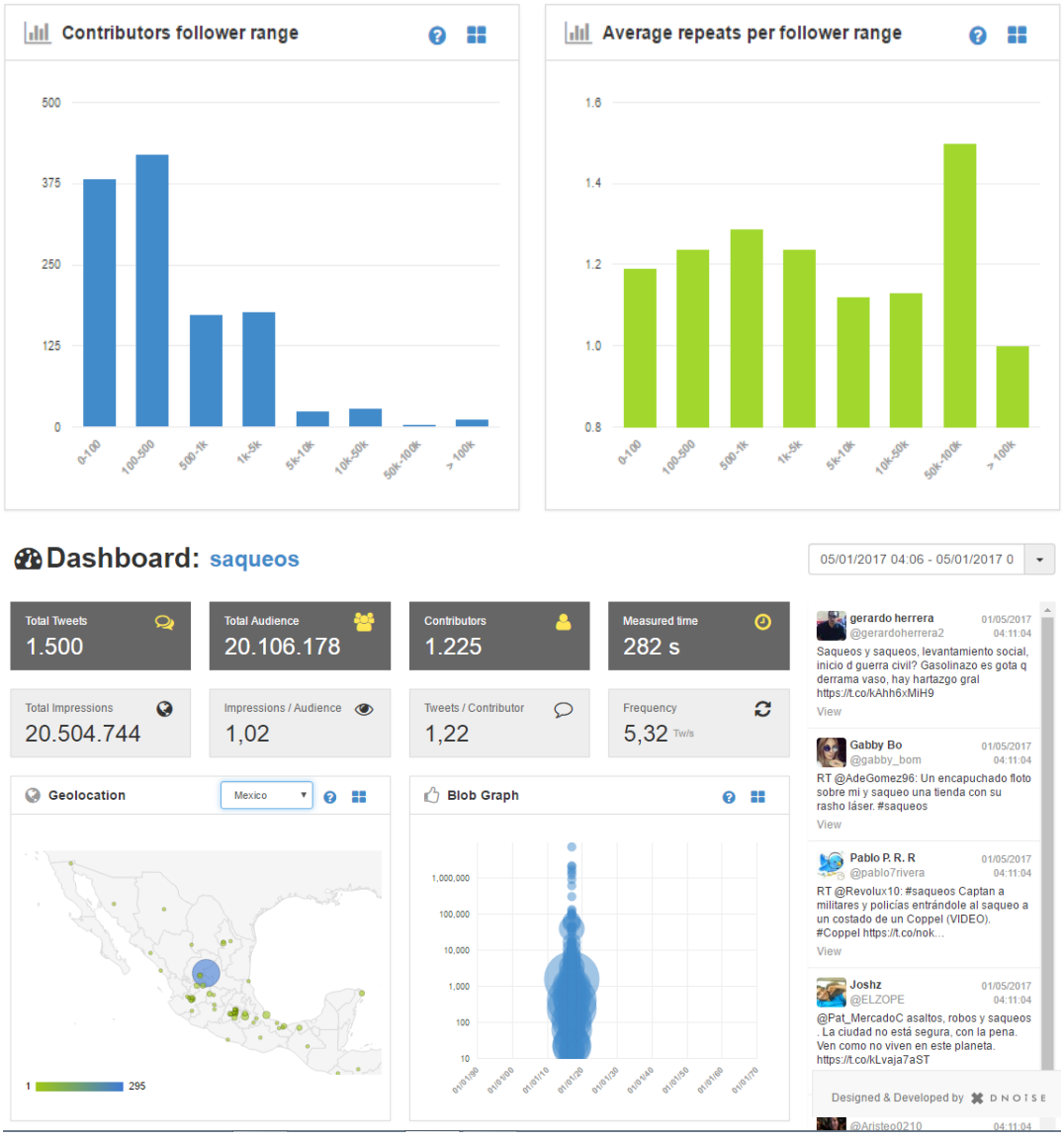


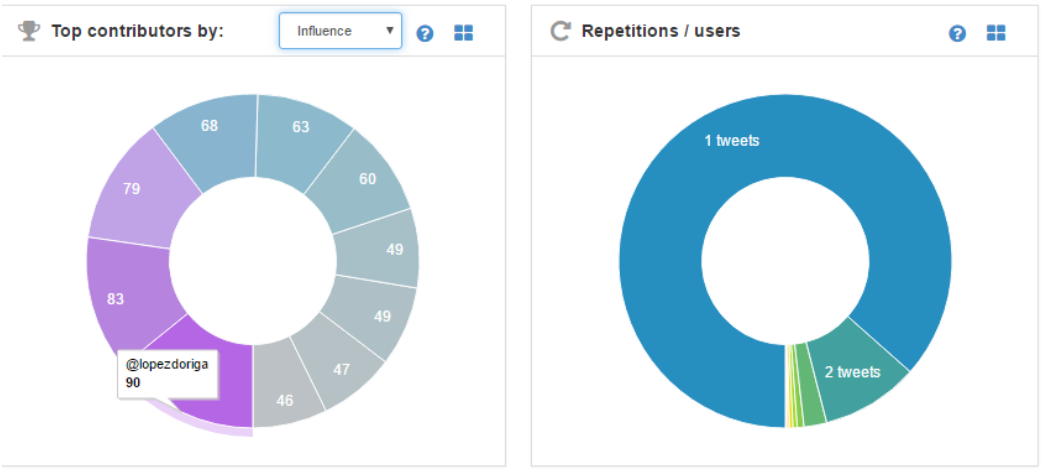
- **BuzzSumo** (gratuita). Es una herramienta que permite encontrar contenidos virales, así como personas influyentes en un sector o tema. Esta **app** permite obtener también los artículos más virales acerca de un concepto, término o palabra clave, independientemente de la web o el blog donde se publicó dicho artículo. El ejemplo en la siguiente imagen es con la palabra clave "gasolinazo".

- **Gigya** (de paga). Es una plataforma que condensa y articula el manejo de redes sociales (**facebook, twitter, linkedin, myspaceID, yahoo, OpenID**) en un solo lugar y proporciona herramientas de análisis para extraer tendencia y lineamientos de los intercambios de información de los sitios sociales. La plataforma **Gigya** ofrece datos y detalles de tráfico por medio de informes con un sistema sencillo de utilizar basado en la web mediante las API.

- **Google Analytics** (gratuita). Se basa en una plataforma de informes el cual puede personalizar. Analiza contenido, redes sociales, publicidad, conversiones y móviles.

- **Follow the hashtag** (gratuita con límites). Con esta herramienta de análisis se puede realizar el seguimiento de un tema en **twitter** que nos pueda resultar de interés como podrían ser los “tweets” que contengan un ***hashtag***, los que contengan una determinada frase, ***url*** (dirección electrónica) o se refieran a un usuario en concreto. La herramienta está predeterminada para ofrecer los últimos 1,500 “tweets” referentes a la búsqueda en los últimos 30 días. A continuación, se ofrecen algunas imágenes para ilustrar esta idea.





Finalmente, puede constatarse que la ventaja principal de este tipo de herramientas es que facilitan análisis con criterios predefinidos para la examinación de secuencias o de regularidades o para establecer una determinada periodicidad (diaria, semanal, mensual, anual o histórica) en la incidencia de un fenómeno o decisión pública. Es decir, estas herramientas son viables para los estudios *ad hoc,* cuando surge una situación inesperada o porque se detecta un punto de relevancia en el trabajo estándar.

Comentario general

Hay varias reflexiones que pueden hacerse respecto a la idea de enfrentar un mundo que produce en cada instante miles de datos. ¿Cuáles son las necesidades institucionales para enfrentar este desafío? Sin ser conclusivo, se presentan aquí cuatro ideas generales.

Primero, si se permite el juego de palabras, las instituciones mexicanas requieren de un proceso de modernización en el que los procesos de recopilación de información, diagnóstico, decisión, análisis, etc., requieren ser transformados; necesitan una "sacudida". Existe un pensamiento estratégico que tiene muchas

aristas obsoletas. En el presente documento se han ofrecido algunas opciones para el cuidado de este aspecto en los procesos institucionales.

Segundo, al observar la lista de alternativas o herramientas tecnológicas presentadas en este documento –una lista no exhaustiva, hay que aclararlo– puede observarse que en su mayoría se trata de herramientas de acceso libre y de código abierto (*Open source*). Los procesos informáticos, a la par de los procesos de toma de decisión, se encaminan desde hace muchos años –aunque ahora con mucha más velocidad– hacia sistemas abiertos.

Tercero, actualmente se observa una producción masiva de información que por su volumen, diversidad y velocidad deja atrás rápidamente a cualquiera que no esté a la altura con las capacidades y habilidades técnicas para su administración, gestión y análisis. Si el desafío antes era analizar la realidad con la mayor precisión, ahora puede completarse diciendo que dicho análisis de la realidad requiere la inclusión de la mayor diversidad de elementos proporcionados por las diferentes fuentes disponibles. Sobre todo, ahora dicho análisis requiere ser "en tiempo real" de acceso rápido, fácil e instantáneo.

Cuarto, en la actualidad el **Big Data** se ha convertido en un activo crucial, que tan sólo pensar en algunas de las grandes empresas tecnológicas del mundo, nos damos cuenta de que gran parte de su valor procede de la cantidad y diversidad de los datos que ofrecen. Pero su principal importancia radica en disponer de datos "limpios" para su procesamiento y si a partir de su análisis se están haciendo las preguntas correctas para la toma de decisiones informadas y se conjeturen comportamientos.

CENTRO DE
ESTUDIOS SOCIALES Y
DE OPINIÓN PÚBLICA

www.diputados.gob.mx/cesop

cesop01

@cesopmx