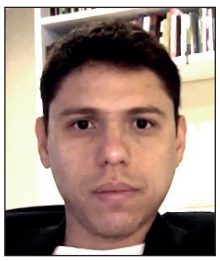
ANALISIS

TÉCNICAS *BIG DATA*: ANÁLISIS DE TEXTOS A GRAN ESCALA PARA LA INVESTIGACIÓN CIENTÍFICA Y PERIODÍSTICA

Big data techniques: Large-scale text analysis for scientific and journalistic research

Carlos Arcila-Calderón, Eduar Barbosa-Caro y Francisco Cabezuelo-Lorenzo



Carlos Arcila-Calderón es profesor de la *Universidad de Salamanca* (España), miembro del *Observatorio de Contenidos Audiovisuales* (*OCA*) y editor de la revista *Disertaciones*. Es doctor europeo en comunicación, cambio social y desarrollo por la *Universidad Complutense de Madrid* y *Máster en Periodismo* por la *Universidad Rey Juan Carlos*. Con anterioridad desarrolló su carrera en la *Universidad del Rosario* y en la *Universidad del Norte* de Colombia, y en la *Universidad de Los Andes* y *Universidad Católica Andrés Bello* de Venezuela.
http://orcid.org/0000-0002-2636-2849

Universidad de Salamanca, Facultad de Ciencias Sociales
Campus Miguel de Unamuno, Edificio FES
Av. Francisco Tomás y Valiente, s/n. 37071 Salamanca, España
carcila@usal.es

Eduar Barbosa-Caro es periodista e investigador universitario. Cuenta con un *Máster en Comunicación* de la *Universidad del Norte* (Colombia). Forma parte del equipo *Colciencias* para el *Grupo de Investigación en Comunicación y Cultura*. Es el editor adjunto del *Anuario Electrónico de Estudios en Comunicación Social Disertaciones* de la *Universidad del Rosario* (Colombia).
http://orcid.org/0000-0001-9009-5581

Universidad del Norte
Vía Puerto de Colombia, km 5. Barranquilla, Colombia
eduarbarbosacc@gmail.com

Francisco Cabezuelo-Lorenzo es profesor en el campus de Segovia de la *Universidad de Valladolid*. Es licenciado en periodismo por la *Universidad Complutense de Madrid* (*UCM*) y licenciado en publicidad y relaciones públicas por la *Universidad Camilo José Cela* (*UCJC*). Está acreditado como profesor titular de universidad y cuenta con un sexenio (2007-2012) reconocido por la *Cneai*. Ha participado en varios programas competitivos de I+D+i de convocatorias autonómicas, estatales y europeas.
http://orcid.org/0000-0002-9380-3552

Universidad de Valladolid, Facultad de Ciencias Sociales, Jurídicas y de la Comunicación
Plaza de la Universidad, 1. 40005 Segovia, España
cabezuelo@hmca.uva.es

Resumen

Este trabajo conceptualiza el término *big data* y describe su importancia en el campo de la investigación científica en ciencias sociales y en las prácticas periodísticas. Se explican técnicas de análisis de datos textuales a gran escala como el análisis automatizado de contenidos, la minería de datos (*data mining*), el aprendizaje automatizado (*machine learning*), el modelamiento de temas (*topic modeling*) y el análisis de sentimientos (*sentiment analysis*), que pueden servir para la generación de conocimiento en ciencias sociales y de noticias en periodismo. Se expone cuál es la infraestructura necesaria para el análisis de *big data* a través del despliegue de centros de cómputo distribuido y se valora el uso de las principales herramientas para la obtención de información a través de software comerciales y de paquetes de programación como *Python* o *R*.

Artículo recibido el 30-03-2016
Aceptación definitiva: 28-06-2016

Palabras clave

Datos; *Big data*; Minería de datos; Aprendizaje automático; Modelamiento de temas; Análisis de sentimientos.

Abstract

This paper conceptualizes the term big data and describes its relevance in social research and journalistic practices. We explain large-scale text analysis techniques such as automated content analysis, data mining, machine learning, topic modeling, and sentiment analysis, which may help scientific discovery in social sciences and news production in journalism. We explain the required e-infrastructure for big data analysis with the use of cloud computing and we asses the use of the main packages and libraries for information retrieval and analysis in commercial software and programming languages such as *Python* or *R*.

Keywords

Data; Big data; Data mining; Machine learning; Topic modeling; Sentiment analysis.

Arcila-Calderón, Carlos; **Barbosa-Caro, Eduar**; **Cabezuelo-Lorenzo, Francisco** (2016). "Técnicas *big data*: análisis de textos a gran escala para la investigación científica y periodística". *El profesional de la información*, v. 25, n. 4, pp. 623-631.

http://dx.doi.org/10.3145/epi.2016.jul.12

1. Introducción

Existe un creciente interés tanto científico como periodístico por la explotación de las grandes cantidades de datos textuales disponibles en internet gracias al uso masivo de los denominados medios sociales (*Facebook*, *Twitter*, blogs, etc.) y de otras fuentes textuales de información (medios de comunicación online, webs oficiales, libros electrónicos, documentos financieros, etc.). Un buen ejemplo de este interés es el caso conocido como *Panama papers* o papeles de Panamá, en el que se han usado técnicas de ciencia de datos (**Woodie**, 2016) para revelar a la opinión pública fraude fiscal y financiero por parte de personajes importantes (jefes de estado, empresarios, políticos, etc.).

Sin embargo, tanto en los campos de las humanidades digitales, la comunicación e información (**Verbeke** *et al.*, 2014), como del periodismo de datos, existe poca claridad y consenso sobre el concepto de *big data* y sobre las técnicas para análisis de textos a gran escala. Este artículo tiene como objetivo sintetizar los principales enfoques existentes sobre *big data* y describir los principales métodos computacionales que científicos sociales y periodistas tienen a su disposición para la explotación y el análisis de información.

> *Big data* se refiere a volúmenes masivos y complejos de información estructurada y no estructurada que requiere de métodos computacionales para extraer conocimiento

A pesar de que se ha intentado vincular el concepto de *big data* sólo con el tamaño de los datos, en términos de *terabytes* o *petabytes* (por ejemplo, en los papeles de Panamá se usaron 2,6 terabytes), esta dimensión es insuficiente para caracterizarlo. El concepto de *big data* se refiere fundamentalmente a volúmenes masivos y complejos de información tanto estructurada como no estructurada, que es recogida durante cierto período de tiempo y que requiere de métodos computacionales para extraer conocimiento.

Otros conceptos importantes ligados al estudio de los *big data* también aluden a su intencionalidad y utilidad (**Murphy**; **Barton**; 2014). El objetivo principal en la generación de datos no contempla generalmente la posibilidad de ser combinados con otros, pues cuando se reúnen grandes cantidades de datos para una finalidad específica, éstos suelen perderse en un mar de información sin pensar en usos secundarios. Por ello se suele sacar el mayor provecho de los datos recogidos sólo a partir de su reutilización básica, su fusión interna y el hallazgo de combinaciones dos por uno (**Mayer-Schönberger**; **Cukier**, 2013), en donde hasta los desechos digitales pueden ser objeto de estudio. Es decir, se realiza una explotación al máximo de los recursos recogidos, pero luego no se suelen reutilizar.

Existen tres retos asociados al fenómeno *big data* (**Nunan**; **Di-Domenico**, 2013) que científicos sociales y periodistas de datos deben tener en cuenta:

- problemas tecnológicos asociados al almacenamiento, seguridad y análisis de los siempre crecientes volúmenes de datos;
- valor comercial que puede ser añadido a través de la generación de *insights* más efectivos;
- impactos sociales, particularmente las implicaciones para la privacidad personal.

Desde un punto de vista académico, estos retos están vinculados a su vez a tres cambios de paradigma:

- mayor importancia de la disponibilidad y acceso de los datos;
- aceptación de niveles de imprecisión y desorden en los datos;
- centrarse más en las correlaciones, en vez de buscar constantemente la causalidad (**Mayer-Schönberger**; **Cukier**, 2013).

Estos cambios, junto con las conceptualizaciones mencionadas, demuestran la inmensa potencialidad que tiene el trabajo con grandes cantidades de datos, pero también dejan ver los problemas tanto técnicos como conceptuales que aún quedan por resolver.

2. Métodos computacionales para el análisis de *big data*

Una vez recogida una importante cantidad de datos textuales (estructurados, semi-estructurados o sin estructura) por medio de procedimientos que van desde la recolección manual (texto por texto) y su digitalización, hasta los más automatizados (como *web scrapping*), y construida una base de datos (relacional, no relacional u orientada a grafos), son necesarios métodos computacionales para realizar un análisis de datos y obtener cierto conocimiento o al menos información relevante y novedosa para la sociedad.

En el caso citado de los papeles de Panamá, se utilizó reconocimiento óptico de caracteres (OCR, por sus siglas en inglés: *optical character recognition*) para la digitalización de 11,5 millones de documentos que contenían el registro de cuatro décadas de negocios de la firma *Mossack Fonseca*. Para realizar búsquedas flexibles a gran escala entre estos documentos no estructurados se utilizó *Apache Solr*, a través de una interface más amigable para los periodistas conocida como *Blacklight Project*. Estos documentos fueron estructurados después en un esquema de relaciones (tipo nodo-arista) para crear una base de datos orientada a grafos usando la tecnología *Neo4j*, lo que finalmente permitió hacer uso de técnicas de análisis de *big data* para encontrar las relaciones entre individuos y datos financieros que destaparon el escándalo mundial.

Las técnicas de ciencia de datos se pueden aplicar, como en el ejemplo anterior, a datos previamente recogidos y trabajados, pero también se puede realizar análisis a gran escala con datos en tiempo real (en *streaming*), lo cual amplía las posibilidades de los científicos.

A continuación se describen y analizan algunas de las técnicas y programas más representativos utilizados en el análisis de grandes cantidades de datos textuales a través de una breve explicación conceptual-metodológica de cada uno.

2.1. Análisis automatizado de contenido

El análisis de contenido es un ejercicio analítico cuyo objetivo es obtener información de cierto conjunto de datos, generalmente textos o grabaciones (**Leetaru**, 2012; **Krippendorff**, 2004). Históricamente, el análisis de contenido se ha servido de otras técnicas que mejoran su alcance y se ha venido aplicando en marcos de investigación cuantitativos, cualitativos y mixtos, mientras "emplea un amplio rango de técnicas analíticas para generar descubrimientos y ponerlos en contexto" (**White**; **Marsh**, 2006, p. 22).

A través de los métodos computacionales y del análisis de contenido automatizado (ACA) se vencen limitaciones que tenían los análisis de contenido tradicionales. Además de mayores muestras y mejor codificación, la confiabilidad alcanzada a través de la tecnología disminuye notablemente los sesgos que puedan desviar la interpretación, con lo que podemos replicar los estudios de manera más acertada y a distintas escalas.

No obstante, estas consideraciones generan argumentos a favor y en contra sobre la fiabilidad y validez de dichos análisis computarizados. **Harwood** y **Garry** (2003) recalcan que al no cumplirse estándares de validez, la generalización de los resultados puede quedar en tela de juicio, mientras que desde otras perspectivas la búsqueda instantánea entre los datos y el incremento en la amplitud de los estudios se configuran como variables que justifican y dan valor al uso de las nuevas tecnologías (**West**, 2001).

Este tipo de análisis automatizados impulsan trabajos de investigación cada vez más diversos. Por ejemplo, **Cheng** *et al.* (2008, p. 2) destacan el incremento en el uso de "la lingüística computacional […] aplicada a dominios como los de la captura de datos de inteligencia, traducción con máquinas, análisis de contenido automatizado y la indexación y recuperación de bases de datos completas".

Estas técnicas son fundamentales para estudiar todo tipo de datos a nuestro alcance, incluyendo los miles de millones de mensajes publicados por medios digitales y redes sociales. Específicamente, se han realizado algunas aplicaciones que permiten la aplicación del análisis de contenido automatizado de forma rápida e intuitiva, entre ellas las más difundidas son *Linguistic inquiry and word count* (*LIWC*), *Hamlet*, *WordStat* y *QDAMiner*. Sin embargo existen cada vez más softwares para el análisis del lenguaje natural para entornos de programación como *Python* o *R*. En el caso de *Python*, destaca una serie de librerías bajo el nombre *NLTK*, y en *R* encontramos la librería *ReadMe*.

En función del tamaño de la información y de las necesidades de los científicos de datos, se pueden generar algoritmos específicos (basados en librerías o con funciones nativas) ejecutados sobre entornos de programación, en especial *Python*. Estos algoritmos o scripts se suelen desarrollar también para análisis a gran escala mediante computación distribuida (ordenadores conectados usualmente en la nube) en entornos como *Hadoop*, *Flink* o *Spark*, requiriendo también programar en otros lenguajes como *Java* o *Scala*. Todas estas tecnologías utilizan la filosofía *Map-Reduce* para distribuir las tareas de análisis en diferentes nodos (*Map*) y luego juntar los resultados en un único archivo (*Reduce*). Además de las *grids* académicas, existen grandes compañías comerciales que proporcionan servicios de computación en la nube (eludiendo muchos problemas técnicos para el usuario final) como *Amazon Web Services* (*AWS*), *Oracle Cloud Computing* o *Microsoft Azure*.

Un ejemplo de un análisis de contenido automatizado a gran escala es el conteo típico de palabras o frecuencias de aparición de un término en conjuntos de datos que se encuentran almacenados y que no pueden ser procesados por los programas comerciales (ejemplo: un *dataset* de 100 GB). Para hacer este conteo se debe escribir un pequeño script *Map-Reduce*, por ejemplo en *Python*, que *tokenice* cada palabra por medio de la fórmula "clave, valor", es decir, "palabra_X, 1". Para ejecutar este script se debe desplegar un cluster de instancias (nodos *master* y *esclavos*) conectadas en la nube en donde se debe subir el archivo con los datos (en un formato también distribuido como *HDFS*, *Hadoop distributed file system*) y el algoritmo que permitirá el análisis. Este centro de cómputo en la nube, usando por ejemplo *Hadoop*, permitiría distribuir las tareas de análisis de forma equilibrada entre los nodos, paralelizando el análisis que en un solo ordenador hubiese sido imposible. Cuando son completadas las tareas de los nodos, se realiza

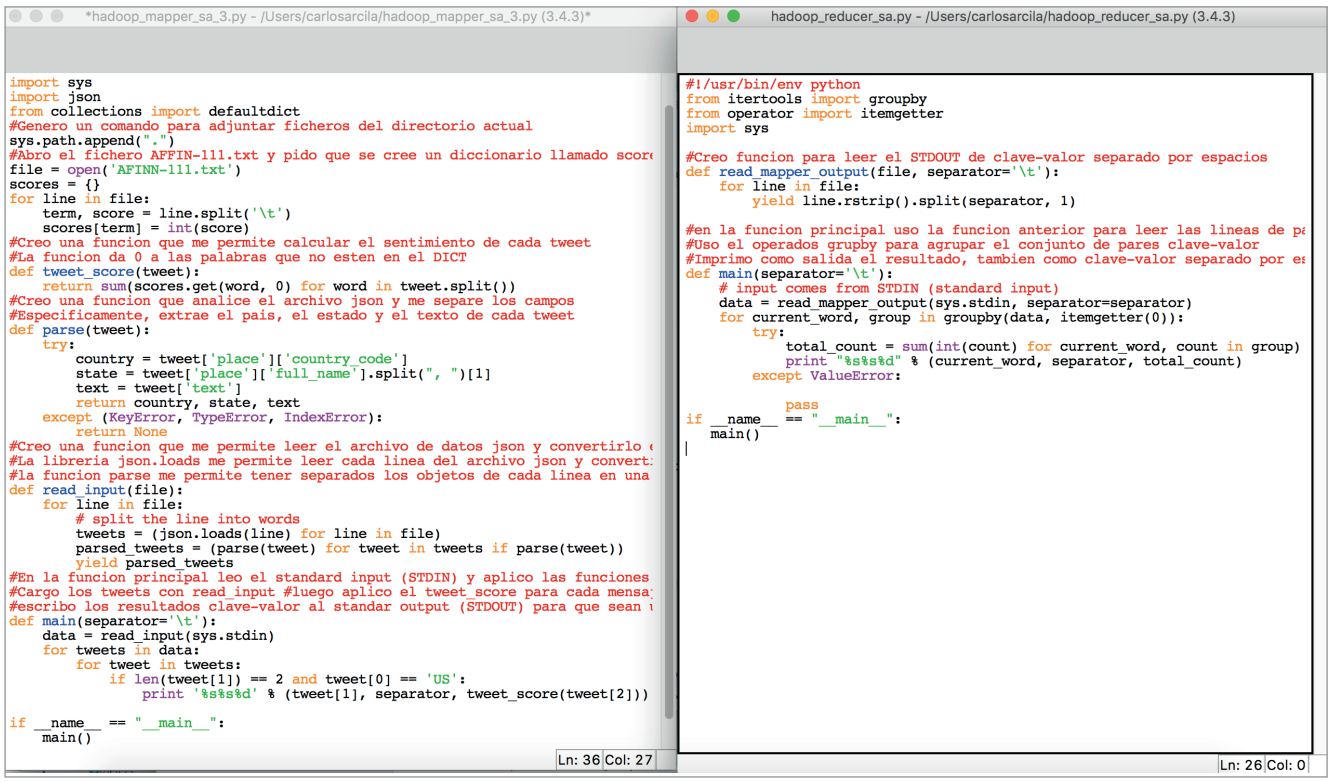


Figura 1. Algoritmo en *Python* para el análisis de sentimientos a textos a gran escala, utilizando *MapReduce* para ser desplegado sobre *Hadoop*

un proceso de resumen de los datos en forma "clave, valor" generados, mediante procedimientos sencillos de suma (SUM) o agrupamiento (GROUP BY). En el caso anterior obtendríamos por ejemplo: "palabra_X, 35", indicando que palabra_X tuvo una frecuencia de 35 apariciones.

2.2. Análisis de sentimiento automatizado

Una de las técnicas aplicadas a grandes cantidades de datos y de mayor interés para científicos sociales y periodistas es probablemente el *sentiment analysis* o análisis de sentimiento. Su objetivo se centra en analizar el vocabulario de un texto con el fin de determinar sus cargas emocionales, haciendo uso de un ordenador que a través de *lexicons* procese, reconozca y evalúe dichos sentimientos (**Leetaru**, 2012), y así saber si los mensajes contienen emociones positivas, negativas o neutras en su estructura (**Feldman**, 2013). *Opinion mining* y *sentiment analysis* son dos cuestiones distintas (**Kechaou**; **Ben Ammar**; **Alimi**, 2013). *Opinion mining* se dirige a la detección de la polaridad, y *sentiment analysis* al reconocimiento de emociones, pero debido a que la identificación de sentimientos es a menudo explotada para la detección de la polaridad, los dos campos se suelen utilizar como sinónimos (**Cambria** *et al.*, 2013).

Destacan estudios tradicionales como el de **Turney** (2002), que aplica el análisis de sentimiento a reseñas (*reviews*) para clasificarlas en recomendadas o no recomendadas; o trabajos como los de **Meena** y **Prabhakar** (2007) que se centran en extraer sentimiento de frases u oraciones. Otros como **Cai** *et al.* (2010) llegan incluso a combinar las técnicas de *sentiment analysis* y *topic modeling* para extraer resultados más concretos sobre estas características (sentimiento y tema) y sus relaciones.

Existen múltiples fuentes de datos a los cuales se puede aplicar análisis de sentimiento, entre las que destacan los blogs, sitios especializados en reseñas, conjuntos de datos ya diseñados y sitios de *microblogging* como *Twitter* (**Vinodhini**; **Chandrasekaran**, 2012). Éste último se ha convertido en el principal reto de los científicos sociales y periodistas, debido a la enorme cantidad de información semi-estructurada en formado *JSON* (*javascript object notation*) que es posible obtener tanto del *streaming* (flujo en directo) como del archivo histórico a partir del uso de las APIs (*Steaming* y *REST*) que *Twitter* ofrece a sus usuarios de forma gratuita. En el mercado existe un gran número de programas comerciales para el análisis de sentimiento (*MeaningCloud*, *Semantria*, *WordStat*, etc.), aunque la mayoría sólo permite análisis de pequeñas cantidades de datos en servidores remotos u ordenadores locales.

> El *sentiment analysis* analiza el vocabulario de un texto para determinar sus cargas emocionales

Un ejemplo de análisis de sentimiento a gran escala es el despliegue de un cluster en *Spark* cuya fuente de datos sea el *streaming* de *Twitter* y que permita monitorizar en tiempo real el tono de los mensajes que se están emitiendo con una etiqueta o *hashtag* (como *#AtentadosParís*), clasificando estos mensajes por zona geográfica para determinar el impacto en varias partes del mundo. Para ello se puede hacer uso de diccionarios de sentimientos, entre los que se encuentran el *Afinn-111* que está disponible tanto en inglés como en castellano y otros idiomas, y que da una valoración a cada palabra ("*love* = +3"; "*war* = -2"). El algoritmo de calificación funciona *tokenizando* palabras con la estructura "*clave*: *valor*" descrita en los párrafos anteriores, pero en vez de contar la aparición

de una palabra, el script debe asignar a cada mensaje un valor a partir de la suma aritmética de los sentimientos detectados (+3 -2 = +1). Esto permite realizar posteriores operaciones (distribuidas usando *MapReduce*) de agrupación (como países con sentimientos más negativos hacia el *hashtag*) y de cruces estadísticos más sofisticados.

2.3. *Data mining*

Implica la extracción de conocimiento a partir de datos masivos y las relaciones subyacentes que pueden existir entre ellos. El *data mining* se originó en 1990 a medida que la tecnología relacional de bases de datos maduró y los procesos de negocio crecieron en automatización (**Dhar**, 2013, p. 67), fomentando la creación de software orientado a aprovechar los datos sobre comportamiento y transacciones, para predecir y planear de manera más acertada. Siguiendo la línea de **Han**, **Kamber** y **Pei** (2006), el *knowledge discovery from data* (término que se ha usado a la par de *data mining*) se puede dividir en siete fases:

- limpieza de datos
- integración de los datos
- selección de datos
- transformación de los datos
- minería de datos
- evaluación de patrones
- presentación del conocimiento.

Desde esta perspectiva, se identifica al *data mining* como sólo uno de varios momentos, si bien de suma importancia, para el conocimiento a partir de los datos, lo que no resta trascendencia a su posición como un instrumento de análisis eficiente de grandes datos. Para **Hand**, **Mannila** y **Smyth** (2001, p. 6) en la minería de datos podemos encontrar datos observacionales que se relacionan con el hecho de que

> "la minería de datos típicamente trata con datos que ya han sido recopilados para algún propósito distinto al del análisis de minería de datos".

Kalina (2013) estima que no se debería concebir la extrapolación de los descubrimientos particulares como finalidad primordial, pues cada conjunto de información habla de ese corpus en particular. Pero debemos tener cuidado y evitar pensar que este tipo de técnicas reemplazan totalmente nuestra labor como investigadores.

> La minería de datos extrae conocimiento a partir de datos masivos y de las relaciones subyacentes que pueden existir entre ellos

Si bien los científicos sociales y los periodistas de datos hacen uso frecuente de software estadístico comercial y de acceso libre (como *Statistical Package for the Social Sciences*, *SPSS*, y su versión libre *PSPP*), para el análisis de datos a gran escala la mayoría de estos paquetes son insuficientes. Existen productos comerciales como *MatLab* que son escalables (permite su ejecución en clusters y nubes), sin embargo, desde la limpieza de datos hasta la visualización final los científicos de datos prefieren la utilización de lenguajes de software libre como *R* o *Python*.

En *R*, el lenguaje de programación estadístico en abierto más extendido, existen cientos de funciones nativas (sin necesidad de librerías adicionales) que permiten el análisis y minería de datos. Se han creado librerías específicas que facilitan y agilizan la minería de datos como *car*, *Hmisc*, *ggplot2*, *dplyr* o *tidyr*. In-

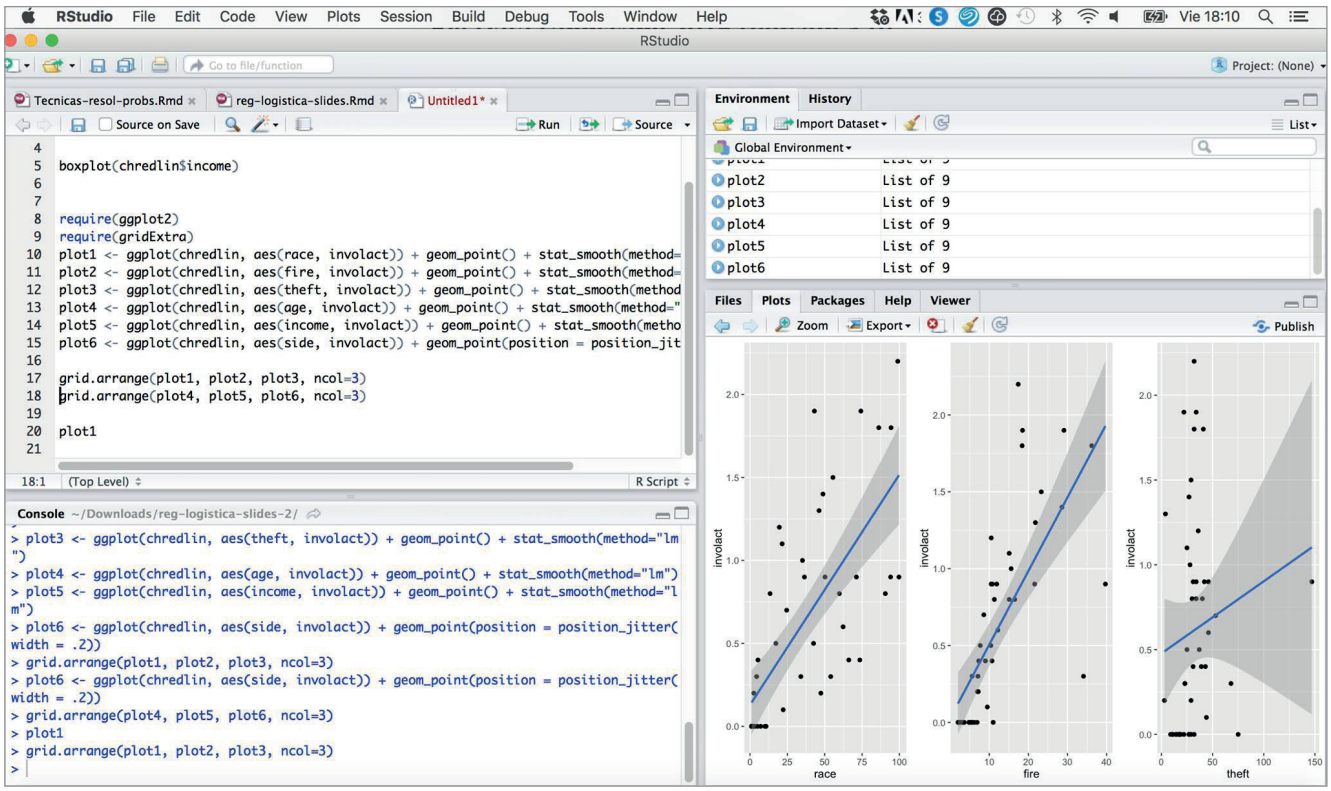


Figura 2. Mineria y visualización de datos en *R*, usando *R Studio*

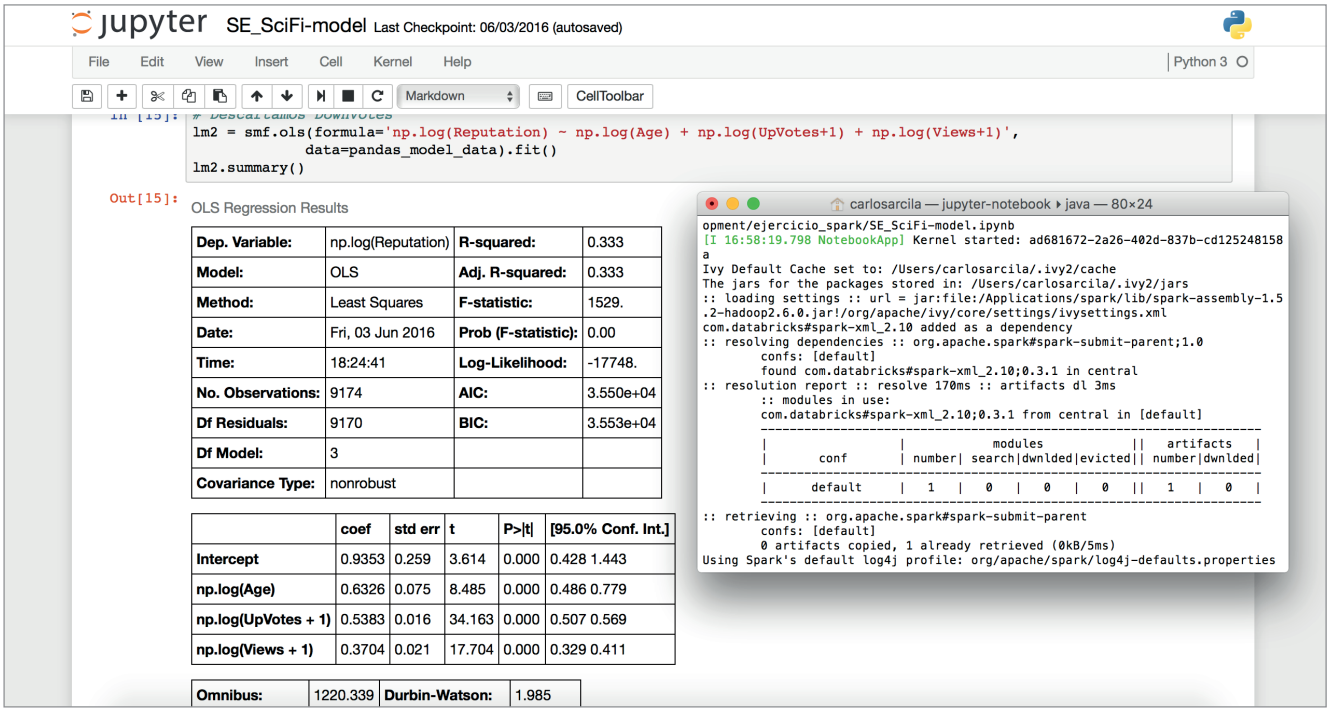


Figura 3. Aplicación de regresión lineal de forma distribuida y en streaming con *Spark*, utilizando *Jupyter* como *iNotebook* de *Python*

cluso se incluyen librerías que proveen la visualización de grafos y el análisis de redes como *igraph*. Una de las limitaciones de la minería de datos de *R* es la paralelización de procesos (el cómputo distribuido en diferentes ordenadores), aunque en las últimas versiones de la plataforma *Spark* se incluye un módulo (aún limitado) de *R* que permite paralelizar los análisis.

Por lo anterior, gran parte de la computación científica requerida para la minería de datos a gran escala se sigue diseñando en scripts para *Java*, *Scala* o *Python*. Este último es probablemente el más usado entre los científicos y periodistas de datos, ya que es un lenguaje de programación interpretado, lo que en cierta medida facilita la sintaxis y la ejecución de las funciones. Además existe un sinfín de librerías en *Python* que facilitan y maximizan funciones típicas y avanzadas de minerías de datos, entre las que destacan *Pandas*, *Numpy*, *Matplotlib* y *SciPy*. *Python* se entiende además en general bastante bien con los principales desarrollos de computación distribuida, y se ha convertido en un estándar dentro de muchas comunidades científicas.

2.4. *Machine learning*

Es un concepto derivado de la propia minería de datos que se refiere al diseño de programas o algoritmos que pueden aprender reglas a partir de datos, adaptarse a cambios y mejorar el rendimiento con la experiencia (**Blum**, 2003). Como campo multidisciplinar en donde confluye la estadística y la complejidad computacional (**Mitchell**, 1997), esta técnica reduce tiempos y costos. También obtiene resultados fiables a través del aprendizaje que realiza la máquina al agregársele parámetros y configuraciones específicas para cada estudio. Un ejemplo de *machine learning* es la clasificación automática de correo electrónico. Para su funcionamiento tenemos en primera instancia unos recursos de texto ya clasificados (ejemplo: correo spam vs. correo no spam) que son cargados al sistema de análisis (datos de entrenamiento o *training*), ya sea a través de una interfaz o una línea de códigos. Este sistema permite entonces generar un conocimiento basado en el corpus introducido previamente, lo que se convierte en un algoritmo con el cual la máquina aprende las reglas subyacentes en dichos documentos (ejemplo: aparición de términos como "lotería", etc.). Tras este paso, estas reglas o patrones son ingresados nuevamente en el sistema de análisis y usados sobre una nueva muestra también clasificada (muestra de prueba o *testing*) para mejorar progresivamente los resultados y su precisión, forjando un análisis cada vez más robusto.

El *machine learning*, usado fundamentalmente para la clasificación y la predicción, se ha aplicado en áreas tan diferentes como las búsquedas en internet y el diseño de medicamentos (**Domingos**, 2012), además de en situaciones puntuales que estimulan el uso de esta técnica, entre las que se puede mencionar, además del volumen de los datos: falta de expertos para resolver un problema a partir de datos, imposibilidad de exponer claramente las reglas de análisis de datos, alta velocidad con que cambia un conjunto de datos, y labores de personalización de grandes conjuntos de información (**Dietterich**, 2003). Este último caso ha sido utilizado para situaciones en las que no existe un algoritmo único, como por ejemplo uno que permita diferenciar automáticamente correos electrónicos no deseados de los legítimos (**Alpaydin**, 2010), tal como se explica en el párrafo anterior.

Se puede dividir el *machine learning* en dos grandes grupos (**Murphy**, 2012):

- aprendizaje supervisado o predictivo, en donde la máquina aprende no sólo de los propios datos finales (*inputs*) sino que es posible darle modelos o datos adicionales ya categorizados (*outputs*) para que el aprendizaje sea mucho más fiable;
- aprendizaje no supervisado o descriptivo en el que sólo se

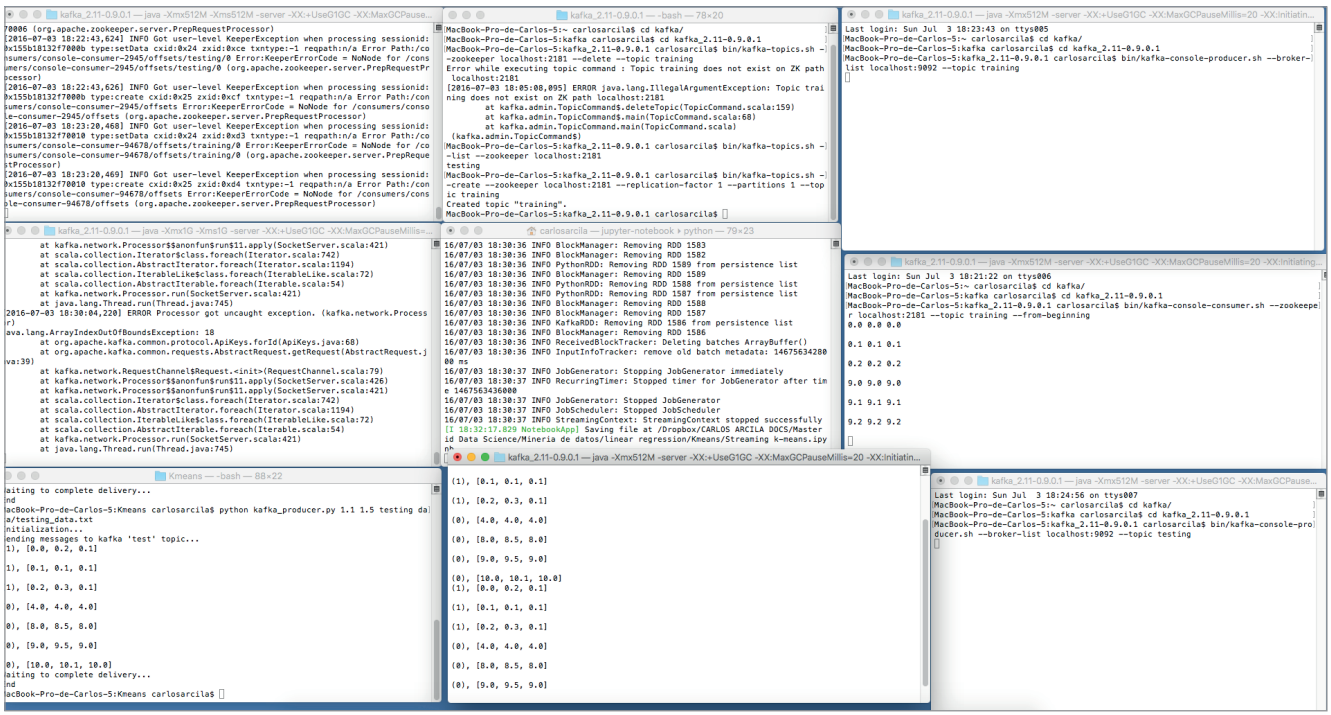


Figura 4. Aplicación del algoritmo *K-Means* para creación de clusters y predicción usando datos en streaming (flujo en tiempo real), utilizando *Spark, Kafka* y *Zookeeper*

dan los *inputs* a la máquina para que encuentre patrones interesantes a partir de los datos.

Desde el punto de vista del algoritmo utilizado (**Kelleher**; **MacNamee**; **D'Arcy**, 2015), las técnicas de *machine learning* se pueden dividir en aprendizaje basado en:

- información: árboles de decisión con algoritmos como *ID3*, métodos de ensamblado como *boosting* o *bagging*, y bosques aleatorios o *random forests*;
- similitud: análisis de cluster no jerárquico con el algoritmo *K-Means*, análisis de cluster jerárquico con extensiones *kernel* usando máquinas de vectores soporte (SVM, por sus siglas en inglés: *support vector machines*) y redes neuronales;
- probabilidad: modelo *Naive Bayes*;
- error: regresión lineal múltiple con método del gradiente.

En la investigación científica de medios podemos mencionar estudios que utilizan algunos de estos algoritmos como los de **Pennacchiotti** y **Popescu** (2011), quienes trabajan en las inmediaciones de los *social media* y el *machine learning* para detectar atributos como la inclinación política, la etnia o afinidades de negocio, o el trabajo de **Téllez-Valero**, **Montes** y **Villaseñor-Pineda** (2009) que aporta conocimientos metodológicos para recopilar y analizar datos sobre reportes noticiosos a través del *machine learning*.

2.4.1. Aprendizaje supervisado

Las aplicaciones de aprendizaje supervisado requieren algoritmos especializados que detecten patrones en los datos. Estos algoritmos pueden implementarse en lenguajes de programación como *Python*, pero al igual que en el análisis de contenido automatizado, si se aplican sobre grandes cantidades de datos requieren plataformas distribuidas para el procesamiento en paralelo. Para superar las dificultades que implica el desarrollo de código y el despliegue de centros de cómputo en la nube, ha prosperado una serie de servicios comerciales que permiten el aprendizaje automático de manera mucho más sencilla. Entre los más extendidos y de relativa facilidad de uso para científicos sociales y periodistas, encontramos la plataforma *AWS* que incluye un módulo llamado *Amazon Machine Learning* (*AML*) que incorpora tanto asistentes como software de visualización, o los servicios de la empresa *Databrick* que basan sus servicios de computación en la nube exclusivamente en *Spark*.

Para el aprendizaje supervisado, modelos como el de predicción basado en máquinas de vectores soporte (SVM) que pasan nuestros datos a un espacio multidimensional, permiten crear algoritmos potentes a partir de datos existentes (un ejemplo: conjuntos de noticias ya separadas por tema) para crear patrones que permitan categorizar automáticamente nuevos conjuntos de textos. Esto es muy útil para la clasificación automática de noticias. Lo mismo sucede para la generación de predicciones de comportamiento de la opinión pública a partir del uso de perfiles creados con encuestas históricas, respondiendo preguntas como cuál es la probabilidad de que un conjunto de ciudadanos que votan al partido X aprueben o desaprueben un tema de agenda emergente.

2.4.2. Aprendizaje no supervisado y *topic modeling*

A diferencia del aprendizaje supervisado, el no supervisado utiliza procedimientos inductivos, extrayendo conocimiento sólo de los datos, como en el caso del análisis de clusters para clasificación. Una de las aplicaciones específicas más útiles para científicos sociales y periodistas es el modelado de temas o *topic modeling*, que comprende la extracción de temáticas a partir de cuerpos de documentos cuya envergadura vence nuestras competencias para obtener manualmente temas, relaciones temporales y patrones en la clasificación (**Arora** *et al.*, 2013). Esta meto-

dología parte de los mismos datos para obtener los temas (nombrados a posteriori por el investigador) en los que luego serán agrupados los documentos (o colecciones de éstos). Para llevar a cabo esta tarea se seleccionan automáticamente palabras del corpus que aparecen frecuentemente, lo que indicaría que podrían pertenecer o no a cierto tema y, observando su presencia en los documentos, podemos buscarlos y clasificarlos sin intervención humana. Ello diversifica el uso que se puede dar a esta técnica y, por ende, los resultados que arroja. Se usa un diccionario de "lista de parada" o *stop list*, cuya función es decirle a un algoritmo qué palabras no deben ser tomadas en cuenta para la creación de clusters.

Existen varios paquetes informáticos para el modelamiento de temas. *Mallet* es uno de los más difundidos, puede descargase gratuitamente y permite la clasificación de documentos, el *clustering* y la extracción de información (**McCallum**, 2002). Uno de los modelos más simples de *topic modeling* es el *Latent Dirichlet Allocation* (*LDA*), que cuenta con dos principios fundamentales:

- patrones implícitos, y
- conjuntos de términos que podríamos llamar *temas* (**Blei**, 2012).

También existen interfaces como *Stanford topic modeling toolbox* para realizar modelado de temas que, a diferencia de *Mallet*, proveen de un entorno gráfico más amigable para llevar a cabo los procesos sin necesidad de conocer la materia en su totalidad. Estos programas, sin embargo, no son escalables, por lo que para analizar datos a gran escala se debe desarrollar código o contratar servicios comerciales, como los de *AWS* o *Microsoft Azure*.

3. Conclusiones

Las grandes cantidades de datos fluyen a través de nuevos canales constituyéndolos en una fuente valiosa de información. Esto da lugar a nuevos retos para las ciencias sociales y el periodismo en lo que a capacidad de procesamiento y análisis se refiere. Si bien es cierto que esta imbricación entre los métodos computacionales y otras disciplinas supone cambios en el quehacer científico, los *big data* y las herramientas relacionadas invitan a repensar las lógicas de investigación social y del propio periodismo desde una perspectiva más amplia, donde se desdibujan aún más los límites entre los campos de estudio y de obtención de información. Las nuevas lógicas implican la necesidad de construir equipos interdisciplinares y centros de análisis de *big data* en las universidades y centros de investigación, que faciliten el desarrollo de proyectos de investigación para explotar el enorme potencial de análisis de estas fuentes para las ciencias sociales y el periodismo.

A partir de los conceptos, teorías y metodologías que se han revisado en este texto, se observa que se necesita mayor profundización (tanto en la teoría como en la práctica), para hacer este campo más accesible (menos requerimientos técnicos y/o facilidad de uso de las plataformas de cómputo distribuido) y que pueda aportar mayor conocimiento en ciencias sociales y en la investigación periodística.

4. Bibliografía

Alpaydin, Ethem (2010). *Introduction to machine learning*. Cambridge/London: The MIT Press. ISBN 978 0262012430

Arora, Sanjeev; **Ge, Rong**; **Halpern, Yoni**; **Mimno, David**; **Moitra, Ankur**; **Sontag, David**; **Wu, Yichen**; **Zhu, Michael** (2013). "A practical algorithm for topic modeling with provable guarantees". En: *30th Intl conf on machine learning*. pp. 280-288.
http://jmlr.org/proceedings/papers/v28/arora13.html

Blei, David M. (2012). "Topic modeling and digital Humanities". *Journal of digital humanities*, v. 2, n. 1, pp. 8-11.
http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei

Blum, Avrim (2003). "Machine learning theory". En: *FOCS 2003 Procs of the 44th Annual IEEE Symposium on foundations of computer science*. Washington DC: IEEE Computer Society, pp. 2-4. ISBN: 0 7695 2040 5

Cai, Keke; **Spangler, Scott**; **Chen, Ying**; **Zhang, Li** (2010). "Leveraging sentiment analysis for topic detection". En: *IEEE/WIC/ACM International Conference on Web Intelligence and Agent Systems: An International Journal*, pp. 265-271.
http://www.csce.uark.edu/~sgauch/5013NLP/S13/hw/Chris.pdf
http://dx.doi.org/10.1109/WIIAT.2008.188

Cambria, Erick; **Schuller, Björn**; **Liu, Bing**; **Wang, Haixun**; **Havasi, Catherine** (2013). "Knowledge-based approaches to concept-level sentiment analysis". *IEEE intelligent systems*, v. 28, n. 2, pp. 12-14.
http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6547971
http://dx.doi.org/10.1109/MIS.2013.45

Cheng, An-Shou; **Fleischmann, Kenneth**; **Wang, Ping**; **Oard, Douglas** (2008). "Advancing social science research by applying computational linguistics". En: *Procs of the American Society for Information Science and Technology*, v. 45, n. 1, pp. 1-12.
http://www.asis.org/Conferences/AM08/proceedings/posters/55_poster.pdf

Dhar, Vasant (2013). "Data science and prediction". *Communications of the ACM*, v. 56, n. 12, pp. 64-73.
https://archive.nyu.edu/bitstream/2451/31553/2/Dhar-DataScience.pdf
http://dx.doi.org/10.1145/2500499

Dietterich, Thomas (2003). "Machine learning". *Nature encyclopedia of cognitive science*. London: Macmillan.
http://eecs.oregonstate.edu/~tgd/publications/nature-ecs-machine-learning.ps.gz

Domingos, Pedro (2012). "A few useful things to know about machine learning". *Communications of the ACM*, v. 55, n. 10, pp. 78-87.
http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf
http://dx.doi.org/10.1145/2347736.2347755

Feldman, Ronen (2013). "Techniques and applications for sentiment analysis". *Communications of the ACM*, v. 56, n. 4, pp. 82-89.
http://dx.doi.org/10.1145/2436256.2436274

Han, Jiawei; **Kamber, Micheline**; **Pei, Jian** (2006). *Data mining. Concepts and techniques*. San Francisco: Morgan Kaufmann Publishers. ISBN: 978 0123814791
http://goo.gl/5zTYb6

Hand, David; **Mannila, Heikki**; **Smyth, Padhraic** (2001). *Principles of data mining*. Cambridge: MIT Press. ISBN: 978 0262082907
ftp://gamma.sbin.org/pub/doc/books/Principles_of_Data_Mining.pdf

Harwood, Tracy; **Garry, Tony** (2003). "An overview of content analysis". *The marketing review*, v. 3, pp. 479-498.
http://dx.doi.org/10.1362/146934703771910080

Kalina, Jan (2013). "Highly robust methods in data mining". *Serbian journal of management*, v. 8, n. 1, pp. 9-24.
http://www.sjm06.com/SJM%20ISSN1452-4864/8_1_2013_May_1_132/8_1_2013_9-24.pdf
http://dx.doi.org/10.5937/sjm8-3226

Kechaou, Zied; **Ben-Ammar, Mohammed**; **Alimi, Adel** (2013). "A multi-agent based system for sentiment analysis of user-generated content". *International journal on artificial intelligence tools*, v. 22, n. 2, pp. 1-28.
http://dx.doi.org/10.1142/S0218213013500048

Kelleher, John D.; **MacNamee, Brian**; **D'Arcy, Aoife** (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. Londres: MIT Press. ISBN: 978 0262029445

Krippendorff, Klaus. (2004). *Content analysis. An introduction to its methodology*. Los Angeles: Sage Publications. ISBN: 978 0761915454

Leetaru, Kalev-Hannes (2011). *Data mining methods for the content analyst: An introduction to the computational analysis of informational center*. New York: Routledge. ISBN: 978 0415895149

Mayer-Schönberger, Viktor; **Cukier, Kenneth** (2013). *Big data. La revolución de los datos masivos*. Madrid: Turner. ISBN: 978 8415832102

McCallum, Andrew-Kachites (2002). *Mallet: A machine learning for language toolkit*.
http://mallet.cs.umass.edu

Meena, Arun; **Prabhakar, T. V.** (2007). Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. En: Amati, Giambattista; Carpineto, Claudio; Romano, Giovanni (eds.). *Advances in information retrieval. 29th European conf on IR research* (*ECIR*), April 2-5, 2007, Rome, Italy, pp. 573-580.
http://dx.doi.org/10.1007/978-3-540-71496-5_53

Mitchell, Tom (1997). *Machine learning.* New York: McGraw-Hill. ISBN: 978 0070428072
http://personal.disco.unimib.it/Vanneschi/McGrawHill_-_Machine_Learning_-Tom_Mitchell.pdf

Murphy, Kevin (2012). *Machine learning. A probabilistic perspective*. Cambridge/London: The MIT Press. ISBN: 978 0262018029

Murphy, Michael; **Barton, John** (2014). "From a sea of data to actionable insights: Big data and what it means for lawyers". *Intellectual property & technology law journal*, v. 26, n. 3, pp. 8-17.
http://www.pillsburylaw.com/publications/from-a-sea-of-data-to-actionable-insights

Nunan, Dan; **Di-Domenico, Maria-Laura** (2013). "Market research and the ethics of big data". *International journal of market research*, v. 55, n. 4, pp. 505-520.
http://dx.doi.org/10.2501/IJMR-2013-015

Pennacchiotti, Marco; **Popescu, Ana-Maria** (2011). "A machine learning approach to Twitter user classification". En: *Procs of the 5th Intl conf on weblogs and social media*. Menlo Park, California: The Association for the Advancement of Artificial Intelligence Press.
https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2886/3262

Téllez-Valero, Alberto; **Montes, Manuel**; **Villaseñor-Pineda, Luis** (2009). "Using machine learning for extracting information from natural disaster news reports". *Computación y sistemas*, v. 13, n. 1, pp. 33-44.
http://www.scielo.org.mx/pdf/cys/v13n1/v13n1a4.pdf

Turney, Peter (2002). *"Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews".* En: *Procs of the 40th Annual meeting of the Association for Computational Linguistics* (ACL), Philadelphia, July 2002, pp. 417-424.
http://www.aclweb.org/anthology/P02-1053.pdf

Verbeke, Mathias; **Berendt, Bettina**; **D'Haenens, Leen**; **Opgenhaffen, Michaël** (2014). "When two disciplines meet, data mining for communication science". En: *64th Annual meeting of International Communication Association (ICA) conf*. Seattle, USA.
https://lirias.kuleuven.be/handle/123456789/436424

Vinodhini, Gopalakrishnan; **Chandrasekaran, Ramaswamy M.** (2012). "Sentiment analysis and opinion mining: A survey". *International journal of advanced research in computer science and software engineering*, v. 2, n. 6, pp. 282-292.
http://www.ijarcsse.com/docs/papers/June2012/Volume_2_issue_6/V2I600263.pdf

West, Mark (2001). *Theory, method, and practice in computer content analysis.* Westport, Connecticut: Ablex Publishing. ISBN: 978 1567505030

White, Marilyn-Domas; **Marsh, Emiliy** (2006). "Content analysis: A flexible methodology". *Library trends*, v. 55, n.1, pp. 22-45.
https://www.ideals.illinois.edu/bitstream/handle/2142/3670/whitemarch551.pdf?sequence=2
http://dx.doi.org/10.1353/lib.2006.0053

Woody, Alex (2016). "Inside the Panama papers: How cloud analytics made it all possible". *Datanami*, 7 April.
http://www.datanami.com/2016/04/07/inside-panama-papers-cloud-analytics-made-possible