Ejemplos de proyectos en el ámbito de la ciencia de datos

PID_00261829

Marçal Mora Cantallops

Tiempo mínimo de dedicación recomendado: 3 horas

Universitat Oberta de Catalunya

Marçal Mora Cantallops

Ingeniero industrial e ingeniero informático por la UPC, máster en Data Science por la UAH y doctorando en Comunicación, Información y Tecnología de la Sociedad en Red por la misma universidad. Investigador en el ámbito de los *game studies*, la ciencia de datos y, en particular, el análisis de redes sociales; está interesado en el uso de estas técnicas para la extracción de conocimiento e información. Ha trabajado en la creación y optimización de modelos estadísticos para logística y planificación de la demanda y actualmente participa en varios proyectos relacionados con la estadística y la ciencia de datos.

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Josep Maria Marco (2019)

Primera edición: febrero 2019

Av. Tibidabo, 39-43, 08035 Barcelona
Diseño: Manel Andreu
Realización editorial: Oberta UOC Publishing, SL

Índice

Introducción..... 5

1. Proyectos de ciencia de datos para el desarrollo y la acción humanitaria..... 7
1.1. La deducción de los desplazamientos diarios de los habitantes de Yakarta a partir de los datos de Twitter 8
1.1.1. La recogida de datos 9
1.1.2. El análisis de los desplazamientos 9
1.2. El uso de los datos masivos para estudiar los patrones de rescate en el mar Mediterráneo 10
1.3. Twitter y la implicación global sobre el cambio climático 12

2. Ciencia de datos para el control de epidemias. El caso del Ébola..... 16
2.1. El caso del Ébola 16
2.2. Metodología 17
2.3. Resultados 18

3. El análisis y la visualización de los recorridos de los taxis de Nueva York..... 21
3.1. Datos 21
3.2. El análisis de la información 22
3.2.1. Variables categóricas 23
3.2.2. Variables numéricas 24
3.2.3. Variables geográficas 26
3.3. La creación de un modelo 27
3.4. La obtención de conocimiento 30

4. Resumen..... 32

Bibliografía..... 33

Introducción

Después de este breve pero intenso recorrido por la ciencia de datos y su entorno, es el momento de presentar algunos ejemplos de aplicaciones, que servirán para entrever las posibilidades que aporta esta disciplina. Los proyectos de ciencia de datos, como veremos, pueden ser muy diversos, desde tareas humanitarias (en las que las Naciones Unidas ha puesto especial atención) hasta situaciones empresariales, pasando por importantes aplicaciones sanitarias o clasificación de especies, entre otros.

En este módulo expondremos algunos de estos casos pero lo haremos desde una perspectiva general. Lo más importante es entender qué tipo de problemas se plantea resolver la ciencia de datos, los procesos que se aplican en su resolución y la forma que tienen las conclusiones o las observaciones que se pueden extraer.

En los tres primeros apartados de este módulo trataremos un total de tres grandes casos (a pesar de que el primero se divida en tres ejemplos más concretos) en los que la ciencia de datos y los datos masivos tienen un papel determinante. Iremos desde una perspectiva más general, a vista de pájaro, hasta un caso completo, que permitirá al estudiante ver el flujo de un pequeño proyecto desde el principio hasta el final.

La estructura del módulo es la siguiente:

1) En primer lugar, veremos algunos de los proyectos más destacados de los disponibles en la página del proyecto Global Pulse de las Naciones Unidas (https://www.unglobalpulse.org/projects).

2) En segundo lugar, veremos una aplicación muy importante: el uso de la ciencia de datos para el control de epidemias. Veremos, en particular, el caso del Ébola.

3) En tercer lugar, veremos un ejemplo que, a pesar de ser planteado de forma muy didáctica, podría ser un caso de actividad empresarial relevante para una empresa de transporte urbano: el análisis de los recorridos de los taxis en Nueva York.

4) Para terminar, repasaremos las ideas más importantes del módulo.

Este módulo es, pues, una oportunidad de ilustrar la ciencia de datos mediante lo que hace y, sobre todo, lo que puede hacer.

1. Proyectos de ciencia de datos para el desarrollo y la acción humanitaria

El proyecto Global Pulse es una iniciativa de innovación de las Naciones Unidas que utiliza la ciencia de datos para contribuir al desarrollo sostenible. Este proyecto consiste en la extracción y el análisis responsable de datos masivos con el objetivo de aportar algo positivo a la humanidad. Su misión es «acelerar el descubrimiento, el desarrollo y la adopción de la innovación en datos masivos para el desarrollo sostenible y la acción humanitaria».

Esta iniciativa, que ya tiene unos años, se estableció basándose en la idea de que los datos nos dan la posibilidad de entender mejor a la humanidad y los cambios en su bienestar, así como de obtener una imagen en tiempo real sobre la respuesta de la humanidad a los cambios políticos, normativos y legales.

Así pues, Global Pulse trabaja para concienciar a los ciudadanos de las posibilidades de la ciencia de datos en el desarrollo humano y la acción humanitaria, pero también busca formar alianzas entre el entorno público y privado para compartir información, generar herramientas analíticas y metodologías que puedan ser utilizadas ampliamente, y extender los avances a toda la red de las Naciones Unidas.

Figura 1. Cómo la ciencia de datos puede contribuir al desarrollo sostenible.

Fuente: United Nations Global Pulse (https://www.unglobalpulse.org/about-newhttps://www.unglobalpulse.org/about-new)

Algunos de los ámbitos en los que se agrupan los proyectos (y que dan una idea clara de su objetivo) son:

- pobreza y hambre
- educación y trabajo de calidad
- igualdad de género
- energía limpia y acción contra el cambio climático
- ciudades sostenibles

El funcionamiento de sus proyectos es el siguiente: una serie de laboratorios de innovación generan ideas y coordinan proyectos, asociados con expertos de cada uno de los ámbitos, gobiernos, académicos y el sector privado. Así, desarrollan e investigan varias aproximaciones para la aplicación de la información digital (habitualmente, en tiempo real o casi real) a los retos que presenta el siglo XXI.

Como veremos en los proyectos que analizaremos a continuación, el planteamiento del proceso es el siguiente:

1) Obtener acceso a los datos, herramientas y experiencia necesarias para descubrir nuevas aplicaciones.

2) Desarrollar las herramientas, aplicaciones y plataformas que puedan mejorar la toma de decisiones o la evaluación de las mismas.

3) Contribuir al desarrollo de marcos regulatorios que aseguren el uso ético de los datos y la privacidad.

4) Involucrar actores clave (gobiernos, empresas, ciudadanos) en las prioridades de innovación y proporcionarles asistencia en la implementación.

1.1. La deducción de los desplazamientos diarios de los habitantes de Yakarta a partir de los datos de Twitter

Dicen que Yakarta, o su área metropolitana, tiene más de 30 millones de habitantes. En la ciudad, el sistema de transporte público gestiona aproximadamente 1,38 millones de desplazamientos diarios. A pesar de que parezcan muchos (y que lo sean), debemos pensar que ciudades como Barcelona o Madrid, con áreas metropolitanas cinco veces más pequeñas, gestionan unos 4 millones de viajes diarios en transporte público, según sus respectivos ayuntamientos. El motivo es simple: estas dos ciudades tienen sistemas de metro centenarios, mientras que Yakarta inauguró su primera línea en 2013, en una ciudad caracterizada por los atascos de tráfico y del transporte rodado (habitualmente, de dos ruedas).

Tanto el gran número de habitantes como el estado de la infraestructura de transporte (y la saturación de la carretera) convierten los desplazamientos diarios en una queja habitual de los residentes. Por este motivo —y por la sostenibilidad de la ciudad— el gobierno local trabaja para mejorar la situación; uno de los caminos elegidos es seguir desarrollando nuevas líneas de metro. Pero

¿cómo se debe decidir el trazado que deben tener? La oficina de estadística de Indonesia optó por la vía habitual: una encuesta. Desde el diseño de la encuesta hasta la obtención de los resultados pasó más de un año. ¿Sería posible ahorrar tiempo, dinero y esfuerzos y obtener resultados similares o igualmente útiles aplicando alguna técnica de ciencia de datos?

1.1.1. La recogida de datos

La importancia de los datos para la planificación urbanística no es única de Yakarta, todas las grandes ciudades del mundo están interesadas en ella. Quizás la forma más fácil de trazar los desplazamientos sea utilizando la información que sale de un dispositivo pequeño que casi todo el mundo lleva en el bolsillo: el teléfono móvil. La inclusión de GPS, sensores y plataformas sociales da mucho juego. En Indonesia, y en Yakarta en particular, las redes sociales están muy extendidas; en 2012, de hecho, fue nombrada como la capital mundial de Twitter, porque era la ciudad con más actividad en la plataforma.

¿Es posible aprovechar la geolocalización de los mensajes enviados a Twitter para dibujar los patrones de desplazamiento en el área metropolitana de Yakarta? El proyecto consistió en extraer millones de mensajes de la red social y analizar las relaciones entre los diez distritos (o ciudades del área) principales. También es importante, sin embargo, tener en cuenta que Twitter puede ser una representación sesgada de la realidad, puesto que no todos los estratos de la población tienen la misma posibilidad de acceso a la red social (pensad, por ejemplo, en la gente mayor o las personas sin teléfono). Por este motivo es necesario calibrar el resultado final para afinarlo.

1.1.2. El análisis de los desplazamientos

Los investigadores recogieron todos los tuits que tenían datos de GPS en el área metropolitana de Yakarta entre el 1 de enero de 2014 y el 30 de mayo del mismo año, coincidiendo con la extracción de la encuesta oficial.

Entonces, para cada usuario:

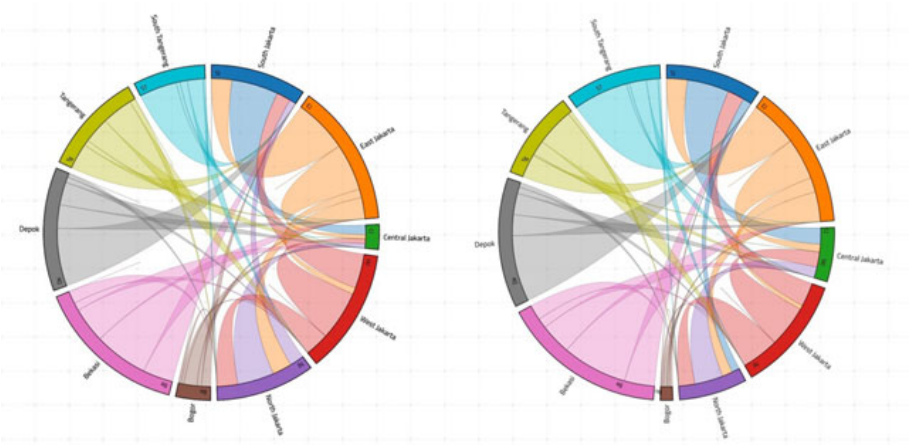
- El origen se establece en el lugar desde el que escribe más mensajes entre las nueve de la noche y las siete de la mañana.

- El destino se establece en el lugar desde el que escribe más durante la semana laboral, excluyendo el origen.

Es importante fijarse en la manera de determinar el origen y el destino porque no son asignaciones directas, sino interpretaciones de patrones de datos. Que el origen de una persona esté donde cena, duerme y se despierta puede tener sentido, del mismo modo que donde escribe la mayoría de mensajes de día (y entre semana) puede ser cerca de su lugar de destino.

Con este método se obtuvo información de más de 300.000 usuarios únicos (a pesar de que se habían recogido datos de aproximadamente millón y medio de usuarios). Para calibrar los resultados se ponderó según la población total de cada distrito, para convertir los datos de Twitter en más proporcionales. En la figura 2 se pueden ver los resultados.

Figura 2. Comparativa entre la encuesta y los datos de Twitter



A la izquierda, los movimientos de la encuesta oficial y a la derecha, los extraídos de Twitter.
Fuente: https://www.unglobalpulse.org/inferring-jakarta-commuting-statistics-twitter

Como se puede comprobar, los resultados obtenidos mediante el análisis de los mensajes en Twitter son muy parecidos a los obtenidos a partir de la encuesta oficial; la ventaja, no obstante, es que la monitorización mediante Twitter puede ser prácticamente en vivo y continua. Además, los resultados se obtienen de forma mucho más rápida y parece que no se pierde calidad. Incluso alguien podría argumentar que los resultados quizás sean más realistas. Al fin y al cabo, ¿quién no ha mentido alguna vez en una encuesta?

1.2. El uso de los datos masivos para estudiar los patrones de rescate en el mar Mediterráneo

Viajes fatales

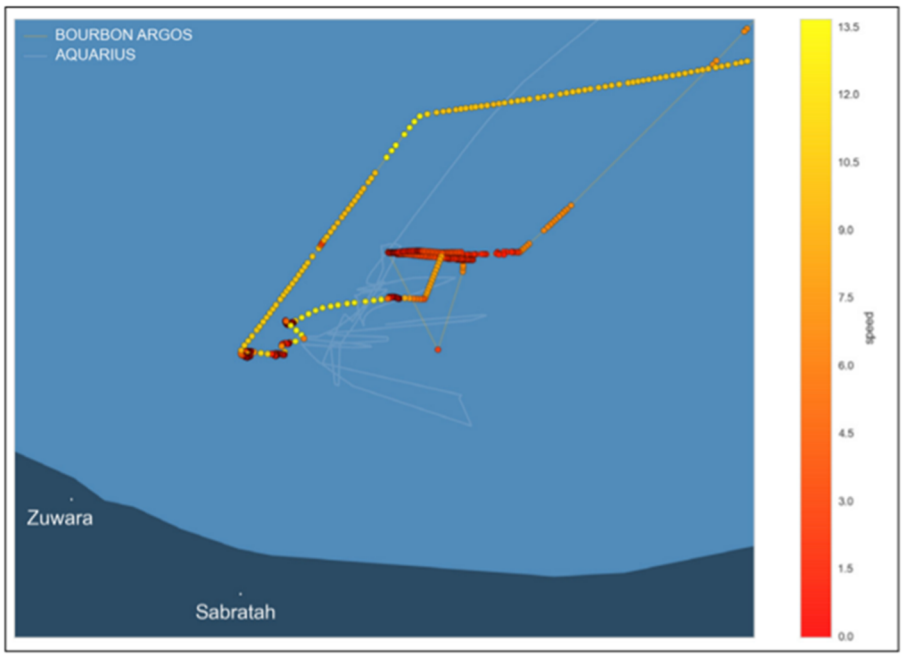
El informe completo, muy interesante para ver también las visualizaciones, está disponible en https://bit.ly/2xrxmmg.

Este proyecto nace en 2017, después de que las Naciones Unidas declarara 2016 como el año con más muertos de migrantes en el mar Mediterráneo. El objetivo era reducir la cifra de víctimas mortales. Pero para conseguirlo era necesario visualizar, primero, el proceso habitual de salvamento que seguían los barcos, mayoritariamente de ONG, para entender las dificultades principales a las que se enfrentaban.

Los científicos de datos utilizaron los datos de localización proporcionados por un sistema llamado AIS y que facilita información sobre la posición y la velocidad de los barcos que envían estos datos (así como su identificador, curso y destino, entre otros) de manera regular, habitualmente cada dos minutos. Es la información que utilizan las autoridades marítimas para, entre otras cosas, evitar colisiones entre naves.

Con esta información es posible, por ejemplo, dibujar visualizaciones como la de la figura 3, que más que un gráfico es una narrativa que explica una historia. En este caso, se trata de una de las operaciones del Aquarius (el barco de las ONG SOS Mediterraneé y Médicos sin Fronteras) y su trayecto rescatando múltiples embarcaciones a la deriva.

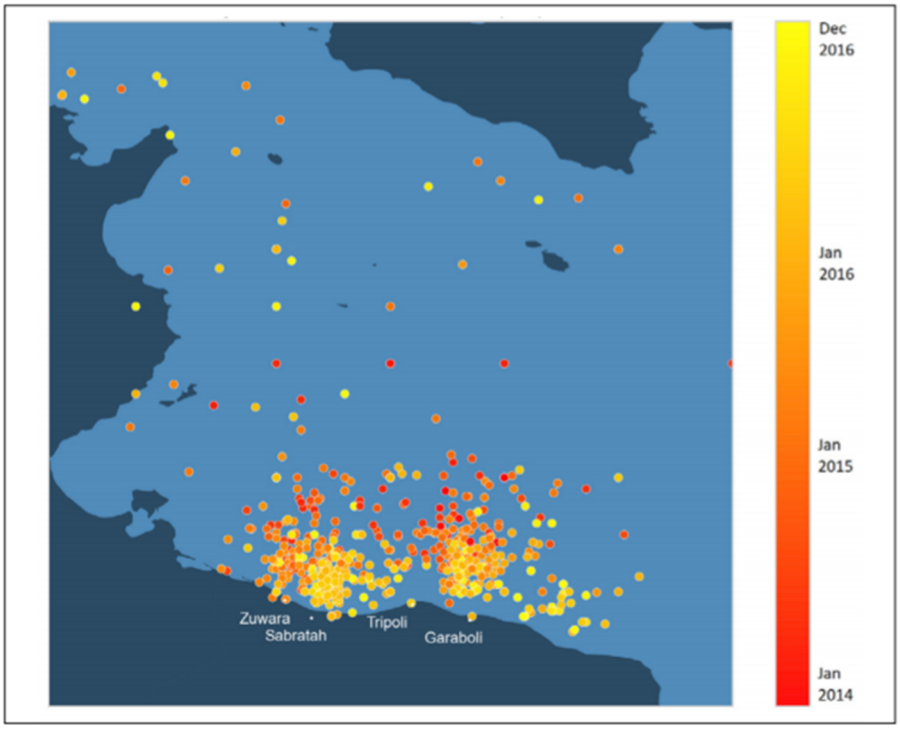
Figura 3. Rescate secuencial de más de un naufragio



Fuente: https://publications.iom.int/system/files/pdf/fatal_journeys_volume_3_part_1.pdf

Del mismo modo que los barcos transmiten información sobre su posición regularmente, existe un sistema de auxilio que sirve para notificar posibles problemas en una región o, más importante aún, para notificar una emergencia a los barcos cercanos (ligándolos legalmente a responder, si es posible). Estos avisos contienen el número de personas estimadas a bordo y la localización aproximada.

Figura 4. Rescates en casos de naufragio



De color rojo, los rescates más antiguos, y de color amarillo, los más recientes. Los rescates tienen lugar cada vez más cerca de la costa. Fuente: https://publications.iom.int/system/files/pdf/fatal_journeys_volume_3_part_1.pdf

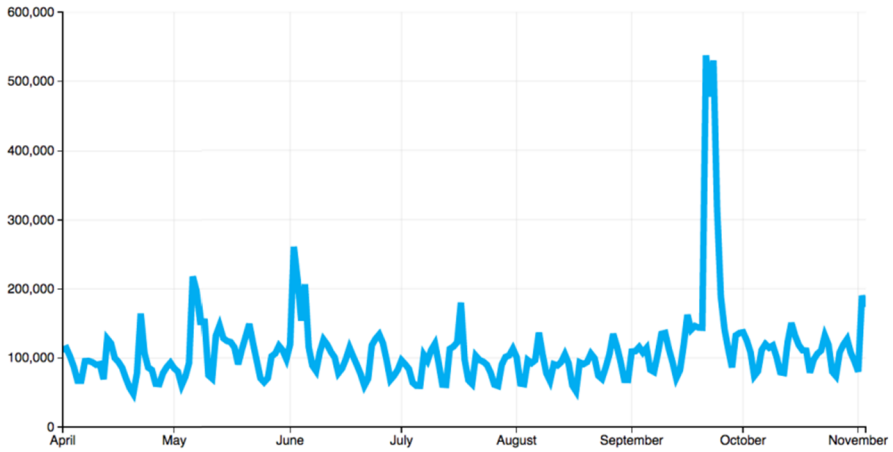
Uno de los hallazgos más importantes (o confirmación de lo que muchos afirmaban basándose en su observación) es que las llamadas de socorro pasaban cada vez más cerca de la costa de Libia (figura 4) y obligaban a las operaciones de rescate a entrar en zonas en las que no solían hacerlo. Así, para salvar al mayor número de personas, es necesario cubrir zonas marítimas cada vez más grandes; si los efectivos o barcos de salvamento son los mismos o cada vez menos, más personas acaban perdiendo la vida en busca de un futuro mejor.

1.3. Twitter y la implicación global sobre el cambio climático

Una de las formas de recoger la opinión o la actividad de la población sobre un tema concreto es monitorizar las redes sociales. Muchas empresas lo hacen para obtener información sobre sus productos, por ejemplo. En este caso, el objetivo de las Naciones Unidas era generar un observatorio en tiempo real sobre el discurso de los usuarios de Twitter en el ámbito global (sí, a escala mundial) respecto al cambio climático antes, durante y después de la cumbre del clima de 2014.

La idea es que Twitter puede actuar como representante del interés público, puesto que el diálogo que se genera en las redes se puede equiparar a la conversación pública general. El volumen se muestra en la figura 5.

Figura 5. Volumen diario de tuits en inglés sobre el cambio climático



El pico corresponde a los dos días en los que se celebró la cumbre. Fuente: http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_Climate_Monitor_2015_0.pdf

Además, el carácter textual de los mensajes proporciona una opción que raramente se encuentra en otros medios: el análisis de contenido y de sentimiento. A pesar de que el estudio se centra en los tuits en inglés, castellano y francés, es importante destacar que el análisis automático de temas solo se puede realizar de forma adecuada en el contenido en inglés. La ciencia de datos avanza muy rápidamente, pero el análisis del lenguaje natural requiere un esfuerzo de entrenamiento y de categorización manual que, hoy en día, solo es funcional para el inglés.

NLTK

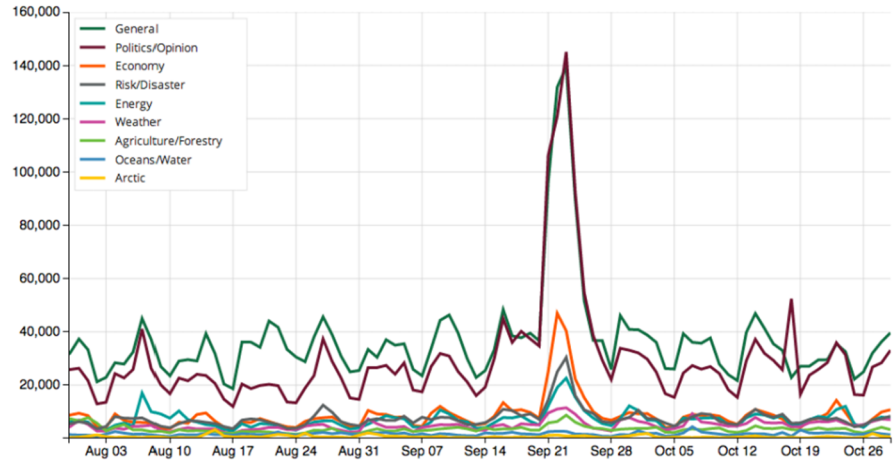
La librería de Python más extensa y utilizada para tratar el lenguaje natural es Natural Language Toolkit (http://www.nltk.org/), pero muchas de sus funciones solo son útiles para el inglés.

Los investigadores de este caso clasificaron los mensajes en nueve categorías o temáticas:

- general
- política/opinión
- energía
- economía
- riesgos/catástrofes
- agricultura/bosques
- tiempo (meteorológico)
- ártico
- océanos/agua

Cada mensaje podía clasificarse en más de un tema, si contenía referencias. Por ejemplo, «la lucha contra el cambio climático empieza protegiendo los océanos y los bosques» podría pertenecer a hasta tres categorías (general, océanos y bosques). El resultado gráfico del análisis se muestra en la figura 6.

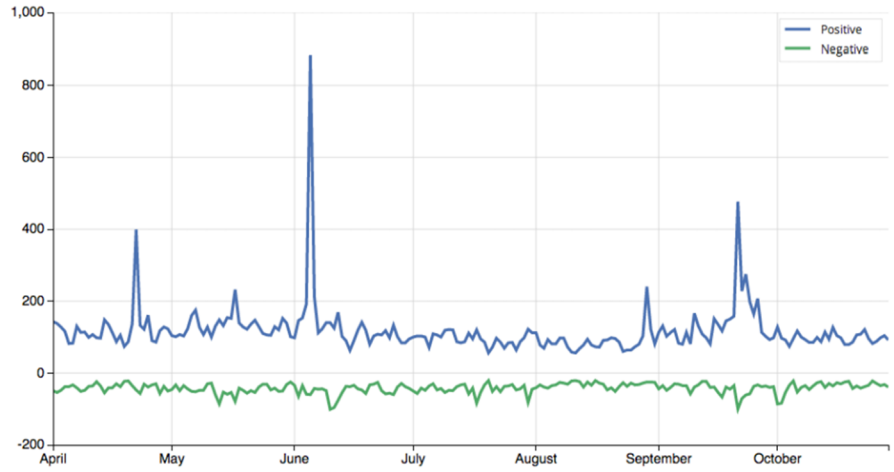
Figura 6. Volumen diario de tuits en inglés según el tema



Fuente: http://www.unglobalpulse.org/sites/default/files/ungp_projectseries_climate_monitor_2015_0.pdf

Las técnicas de análisis de lenguaje natural no solo nos permiten clasificar temáticas, sino que también posibilitan, por ejemplo, deducir el sentimiento asociado a los mensajes. No es una ciencia exacta ni fácil: las personas escriben con estilos diferentes, con sentidos claros o irónicos, con elisiones, con frases complejas o simples. Así, el análisis del sentimiento es un intento estadístico de cuantificarlo. En la figura 7 se muestra el análisis del sentimiento. Por un lado, en positivo, los que hablan a favor de actuar contra el cambio climático y, por el otro, los mensajes que se oponen a ello.

Figura 7. Tuits positivos y negativos sobre el cambio climático



Fuente: http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_Climate_Monitor_2015_0.pdf

Combinando ambos gráficos se pueden extraer conclusiones o percepciones durante el periodo de estudio:

- De media, cada día se escribían unos 140.000 mensajes relacionados con el cambio climático. La gran excepción se produce el mes de septiembre, durante la celebración de la cumbre, con más de 400.000 mensajes (multiplicando por tres la media).

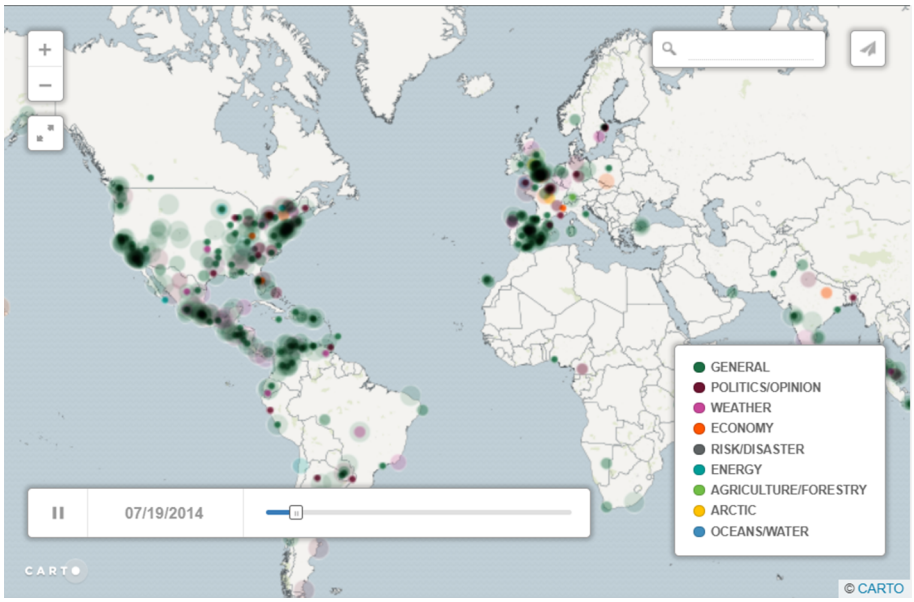
- La cumbre despertó interés durante el acontecimiento, pero también después. Durante el mes siguiente se detectaron entre el 10 % y el 15 % de conversaciones adicionales.

- La política y la economía son los temas que se comentan más. Los grandes influenciadores de la conversación son, pues, políticos y empresarios de renombre. Estas observaciones permiten, por ejemplo, decidir cuáles deberían ser las personalidades con quienes se tendría que contactar para conseguir que el mensaje tuviera más impacto.

- El sentimiento experimenta dos grandes picos en los mensajes polarizados: el primero, en el mes de junio, que coincide con la celebración del Día del Medio Ambiente, y el segundo, en septiembre, con la cumbre.

Así pues, se comprueba de forma metódica que los acontecimientos (en este caso, una cumbre) pueden afectar al discurso y a la opinión pública. La monitorización completa de la figura 8, con mapa incluido, se puede ver (y permite la interacción) en http://www.unglobalpulse.net/climate/.

Figura 8. Captura de la animación de los mensajes en el mapa



Fuente: http://www.unglobalpulse.net/climate/

2. Ciencia de datos para el control de epidemias. El caso del Ébola

Las aplicaciones de la ciencia de datos para salvar vidas siguen madurando y lo hacen guiadas tanto por los avances del sector como por los nuevos retos que surgen. Uno de ellos es la gestión de las enfermedades infecciosas o epidemias, en la que la ciencia de datos empieza a ofrecer posibilidades tanto a agencias humanitarias como a ONG, ya sea para ver tendencias o correlaciones, como para ayudar a la toma de decisiones.

Y es que la gran cantidad de datos derivados del uso de las tecnologías de la información y la comunicación (TIC) muestra un potencial igualmente elevado para enfrentarse a estos retos. La huella digital que deja el uso de servicios en línea, teléfonos y otras transacciones digitales puede ser tratada, analizada y utilizada para mejorar las decisiones y para proveer servicios individualizados con información personalizada. En los países del Tercer Mundo, donde las infraestructuras pueden presentar deficiencias, la expansión del uso del teléfono móvil proporciona un valor particularmente elevado, especialmente en las emergencias.

Uno de estos proyectos ha sido desarrollado por la Unión Internacional de Telecomunicaciones (ITU), la agencia de las Naciones Unidas especializada en TIC. El trabajo de la ITU consistió en utilizar los datos masivos para ayudar a seguir la evolución de una emergencia sanitaria respetando la privacidad de los usuarios.

ITU

El objetivo de la Unión Internacional de Telecomunicaciones es, literalmente, «conectar toda la gente del mundo».

2.1. El caso del Ébola

El año 2014 fue el año de la expansión del virus del Ébola por el este de África, que mató a miles de personas. Los países más afectados fueron Liberia, Guinea y Sierra Leona, pero el pánico y el miedo se extendieron por todo el mundo. Las agencias sanitarias, además, se encontraban con la dificultad de trazar y contener el virus mortal, especialmente debido al largo periodo de incubación de la enfermedad y de los rituales funerarios de algunos países.

La respuesta de la ITU fue el lanzamiento de la aplicación móvil Ebola-Info-Sharing el 19 de diciembre de 2014. Disponible en inglés y francés, esta aplicación gratuita servía para distribuir información oficial entre los usuarios y las organizaciones para facilitar la comunicación en el terreno.

Además, la información espaciotemporal que proporciona la población con el uso del teléfono móvil es clave para la intervención en el control del Ébola. El motivo es simple: el Ébola es una enfermedad que se transmite por contacto y, por eso, responde a la movilidad humana. La globalización ha facilitado los

desplazamientos y, por lo tanto, también los contagios. Por otro lado, la tecnología ha facilitado las comunicaciones. En consecuencia, los datos sobre las llamadas (CDR, Call Detail Record) atraen cada vez más la atención de políticos e investigadores de todas partes por su capacidad de capturar los patrones de desplazamiento humano.

Los CDR extraídos de las redes móviles posibilitan la elaboración de mapas espaciales, que son un reflejo de la vida diaria de las personas de un país o una zona, de sus dinámicas temporales y de sus desplazamientos, y se convierten así en una forma de extraer o identificar problemas poco evidentes. Sin embargo, una de las limitaciones es el límite nacional de la mayoría de los operadores, que hace difícil analizar los movimientos transnacionales.

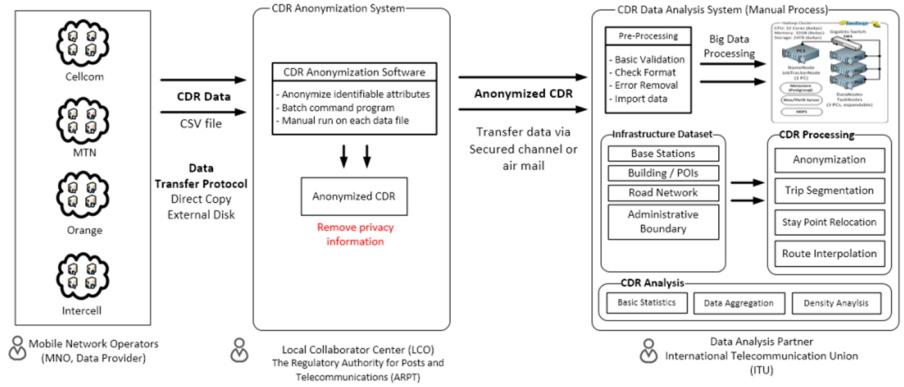
En este caso, los CDR se aplican para visualizar y analizar una información difícil de tratar de otro modo: los movimientos de las personas que se han movido desde una zona afectada por un brote de la enfermedad hasta otras zonas. A partir de esta información se puede, pues, prever dónde es probable que haya un nuevo brote, pero también entender las dinámicas propias de una población y contribuir a un mejor plan de acción para retos futuros.

2.2. Metodología

Nos fijaremos en el caso particular de Guinea[1]. El proceso de análisis de los CDR necesita múltiples pasos que se muestran a continuación (figura 9):

(1)https://bit.ly/2PIG2bJ

Figura 9. Proceso de análisis de datos de la información de los teléfonos móviles



Fuente: https://bit.ly/2PIG2bJ

- Operadores de telefonía móvil: la información se recoge mediante los operadores. En este caso, la información de posicionamiento se obtiene de los datos que tiene el operador de cada terminal en sus servidores. Este conjunto de datos se exporta a los centros colaboradores locales y normalmente se obtienen en formato CSV (separado por comas). En este caso, las transferencias se hicieron de forma manual con copias en discos duros externos. ¡No todo es alta tecnología!

CSV

El formato CSV (ficheros separados por comas) es un formato habitual de organización de grandes cantidades de datos, especialmente en empresas, aunque empieza a extenderse el uso de los JSON (Javascript Object Notation), mucho más flexible y legible.

- Centros colaboradores locales: un paso importante, especialmente para la ética y la privacidad, es la eliminación de los datos personales. Estas unidades se encargan de hacer la limpieza y son habitualmente lideradas por el mismo regulador, que se encarga de pasar datos anonimizados (y, por lo tanto, no identificables) a la unidad que se encargará del análisis de datos.

- Centro de análisis de datos: el rol de este centro o colaborador es el de almacenar, mantener y tratar los datos obtenidos. Los datos se pasan a un entorno de datos masivos (véase la arquitectura en la figura 10). El análisis posterior consiste en conseguir:
  - Dividir los viajes y los desplazamientos de los usuarios en segmentos, de forma que se puedan identificar los trayectos en movimiento y los lugares de parada.

  - Buscar los puntos de estancia. Puesto que los datos obtenidos proporcionan la posición de la torre más cercana, no son bastante exactos para situar al usuario, pero si se trabaja con las localizaciones en una zona concreta es posible determinar la posición más probable del sujeto a partir de la triangulación. Con los datos de varias torres, en cambio, es posible encontrar más puntos de referencia y, por lo tanto, limitar la zona donde es posible que se encuentre el usuario. Así se consigue una localización más realista.

  - Interpolar las rutas, puesto que una vez determinados los puntos de estancia se debe dibujar la ruta más probable entre las dos, que no siempre coincide con el camino más corto.

  - Agregar los resultados en una tabla. En este caso, por hora y por kilómetro cuadrado (no se necesita más granularidad).

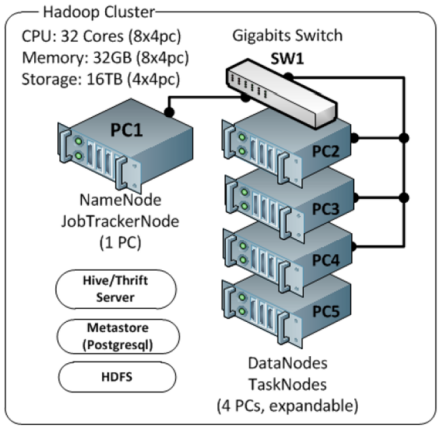
  - Visualizar los datos. En este caso, se hizo un mapa animado utilizando Mobmap[2].

[(2)]Podéis probar de cargar unos datos de muestra sobre Tokio en: https://bit.ly/2z6AnWw.

2.3. Resultados

Como resultado del estudio se extrajeron muchos datos y conclusiones interesantes, desde una aproximación de censo de teléfonos móviles en el país hasta una correlación entre usuarios, zonas más habitadas y desplazamientos más habituales. También se vio, por ejemplo, cuáles eran los recorridos que se solían hacer entre pueblos. Pero para el caso que nos ocupa nos centraremos en una parte más interesante: el dibujo del Ébola.

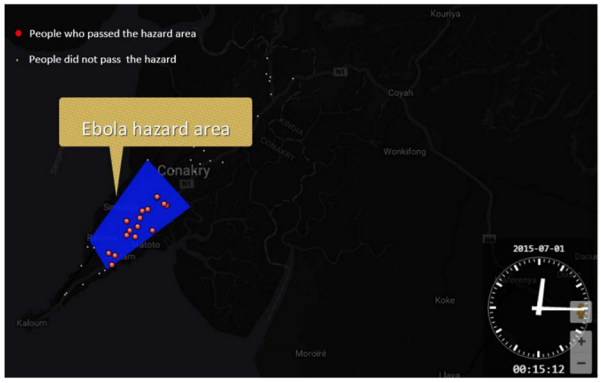
Figura 10. La arquitectura de los datos masivos utilizada. ¡Hadoop no es ciencia ficción!



Fuente: https://www.itu.int/en/ITU-D/Emergency-Telecommunications/Documents/2017/Reports/GN/EN/D012A0000D03301PDFE.pdf

Mediante el análisis y la visualización es posible hacer cosas como las de las figuras 11 y 12. En un primer vistazo se ve una zona afectada por la epidemia y que, por lo tanto, se considera zona de peligro. Los puntos rojos indican personas que se encontraban dentro de la zona en el momento de peligro, mientras que los puntos pequeños y blancos son otras personas que se consideran a salvo.

Figura 11. Personas que se encuentran en una zona de riesgo



Fuente: https://www.itu.int/en/ITU-D/Emergency-Telecommunications/Documents/2017/Reports/GN/EN/D012A0000D03301PDFE.pdf

Con esta información (y marcando los puntos que tienen una probabilidad más alta de haber entrado en contacto con la enfermedad) se puede trazar la evolución del Ébola a lo largo del tiempo. La figura 12, por ejemplo, muestra la evolución y los caminos hechos por estos puntos 48 horas después de haber pasado por la zona de peligro. Fijaos qué lejos han llegado algunos de los puntos. En particular hay tres que ya están en los extremos del país, amenazando un área enorme de Guinea en su recorrido.

Figura 12. Desplazamientos de las personas analizadas 48 horas después de pasar por la zona de peligro

Fuente: https://www.itu.int/en/ITU-D/Emergency-Telecommunications/Documents/2017/Reports/GN/EN/D012A0000D03301PDFE.pdf

Si bien la falta de información inmediata y exacta sobre los movimientos de las personas en una catástrofe natural, emergencia o epidemia puede limitar seriamente la efectividad de la respuesta humanitaria, hemos visto que la naturaleza ubicua de los teléfonos móviles se puede convertir en una oportunidad. La actividad de llamadas, mensajes y el uso de las torres de comunicación en general proporciona información valiosa de la actividad de las personas (acumulaciones y desplazamientos después o durante un acontecimiento), para así mejorar los avisos a la población y la gestión de emergencias. Casos como el Ébola, además, son epidemias difíciles de controlar, fácilmente contagiables y que sus portadores pueden extender rápidamente a escala nacional o internacional, así que se deben tener formas para actuar rápidamente. Como acabamos de ver, la ciencia de datos puede tener un papel destacado en este asunto.

3. El análisis y la visualización de los recorridos de los taxis de Nueva York

Hasta ahora hemos visto casos reales de aplicación de la ciencia de datos pero no hemos podido seguir su proceso completo. Es el momento de ver un ejemplo completo que, además, se basa en un conjunto de datos accesible y del día a día: los recorridos que hacen los taxis en la ciudad más famosa del mundo, Nueva York.

3.1. Datos

Cada día se acumulan más y más datos en las organizaciones; la mayoría quedan en manos privadas (lo que es normal, puesto que la información se puede considerar una ventaja competitiva), pero también hay conjuntos de datos que se ponen a disposición del público general. Hay incluso organizaciones que optan por el concepto de datos abiertos y permiten el uso y aprovechamiento de sus datos por parte de cualquier persona interesada.

En el caso de los viajes en taxi que nos ocupan, no se trata exactamente de datos abiertos; una ley del estado de Nueva York, no obstante, permite pedir datos bajo ciertos supuestos. Con una petición de estas características se obtuvo un conjunto de datos muy detallado que contiene todos los viajes realizados por taxis en 2013[3]. Entre los datos se incluye el punto de recogida (inicio del viaje) y el punto final (o destino), el tiempo del trayecto, la distancia recorrida y el coste.

[3]Podéis leer la historia entera en la página web de su autor: https://chriswhong.com/open-data/foil_nyc_taxi/.

Notebook

Se recomienda seguir la explicación en paralelo en el Notebook, disponible en: https://bit.ly/2IxZKnP.

El hecho de que sean datos reales (y grandes) se hace patente, sobre todo, en dos aspectos. El primero, que ocupa más de 19 GB repartidos en varios ficheros separados por comas[4] y con 14 millones de registros por fichero. El segundo, que hay una gran cantidad de registros incompletos, errores, columnas superfluas, etc.

[4]De hecho, los datos están disponibles aquí: www.andresmh.com/nyctaxitrips/.

En este módulo seguiremos un proceso típico de los proyectos de ciencia de datos: analizaremos la información, construiremos un modelo que evaluaremos e intentaremos hacer predicciones. Y decimos que lo intentaremos porque, quién sabe, quizás este caso esconde un secreto más oscuro de lo que parece...

3.2. El análisis de la información

En primer lugar, hay que saber de qué datos disponemos. Para echar un primer vistazo, lo más sencillo es cargar los datos y verlos en forma de tabla. En el caso de los taxis, por ejemplo, las primeras filas y columnas tienen la forma que se muestra en la tabla 1.

Tabla 1

	medallion	***hack-license***	***vendor-id***	***rate-code***	***...***	***trip_dist***	***pickup_lat***	***...***
0	89D227...	BA96D...	CMT	1	...	1.0	40.757977	...
1	0BD7C8...	9FD8F...	CMT	1	...	1.5	40.731781	...
2	0BD7C8...	9FD8F...	CMT	1	...	1.1	40.737770	...
3	DFD220...	51EE8...	CMT	1	...	0.7	40.759945	...
4	DFD220...	51EE8...	CMT	1	...	2.1	40.748528	...

Fuente: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Esto es solo una muestra para hacerse una idea, puesto que la tabla es mucho más grande (y extensa). Vayamos por partes:

- Las dos primeras columnas (*medallion* y *hack-license*) identifican la licencia del taxi y, por lo tanto, no son muy interesantes a la hora de construir el modelo.

- Después hallamos columnas como *vendor-id* o *rate-code*, que parecen ser variables categóricas. Suele ser interesante representar estas variables gráficamente para entender las distribuciones, como veremos a continuación. En este conjunto de datos hay variables para identificar a qué grupo empresarial pertenece el taxi, qué código de tarifa se aplicó o qué medio de pago se utilizó, por ejemplo.

- También tenemos una serie de columnas con unos valores mucho más familiares: números. Estas columnas tienen datos sobre la distancia recorrida en el viaje, el coste o la duración. Con las variables numéricas también es interesante ver si se correlacionan entre ellas; los gráficos más habituales en este caso son las nubes de puntos o *scatterplots*.

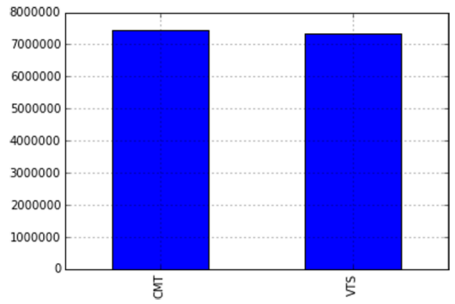
- Finalmente, el conjunto de datos tiene una serie de cifras que, a pesar de ser numéricas, corresponden a un dominio muy concreto: las coordenadas geográficas de los puntos de salida y de llegada, en latitud y longitud. La particularidad de estas columnas es que sus valores se limitan a la geografía estudiada, pero, sobre todo, que se pueden representar en un mapa, de manera que aportan información adicional.

A continuación nos centraremos en algunas de las variables para entender el proceso inicial.

3.2.1. Variables categóricas

Una de las primeras cosas que se pueden hacer es visualizar la distribución de las variables categóricas, para intentar ver la relevancia. La primera, por ejemplo, es el proveedor del sistema (figura 13).

Figura 13. Representación de la frecuencia de cada *vendor*

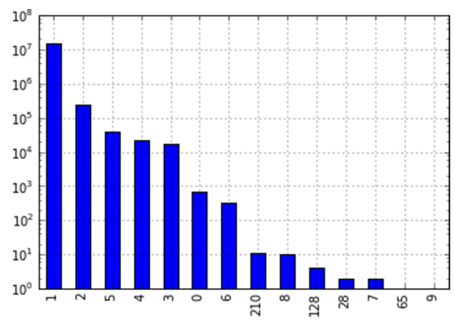


Fuente: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Como se puede observar, el conjunto de datos analizado es solo una parte del total, que contiene algo más de 14 millones de registros. De estos, más o menos la mitad pertenecen a cada uno de los servicios (CMT y VTS).

Es más interesante ver las zonas tarifarias. Del gráfico de barras de la figura se puede extraer que la zona 1 está desproporcionadamente presente. De hecho, hay que fijarse en que, para una correcta representación, la escalera del eje vertical (que representa la frecuencia absoluta de aparición de cada tarifa) se ha debido hacer logarítmica. Así, una primera deducción es que, teniendo en cuenta que es Nueva York, es más que probable que la zona 1 corresponda a la isla de Manhattan.

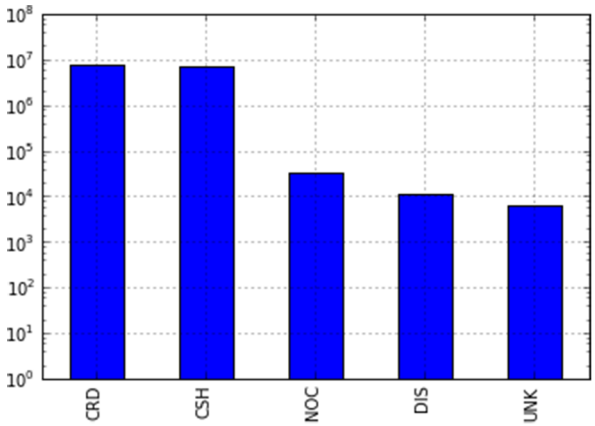
Figura 14. Zonas tarifarias de los viajes del conjunto de datos



Fuente: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Otra observación interesante corresponde al método de pago, representado en la figura 15. Y es interesante no tanto porque vemos que los dos métodos más habituales, con diferencia, son la tarjeta de crédito (CRD) y el efectivo (CSH), sino porque podemos detectar que, por ejemplo, hay cierta cantidad de datos desconocidos (UNK). En este caso son unos pocos miles de líneas que, en una muestra de 14 millones, pueden ser poco significativas, pero sirven como ejemplo para plantearnos si estos datos se tendrían que limpiar antes de seguir elaborando un modelo.

Figura 15. Métodos de pago

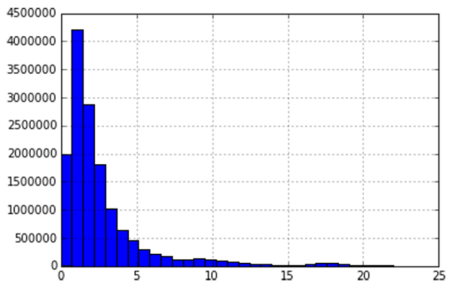


Fuente: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

3.2.2. Variables numéricas

Para las variables numéricas ya no haremos gráficos de barras, sino que mostraremos su versión continua: histogramas. Si tenemos en cuenta que estamos analizando viajes en taxi, ¿qué puede ser interesante? Pues la distribución de las distancias recorridas, por ejemplo (figura 16).

Figura 16. Histograma de las distancias recorridas

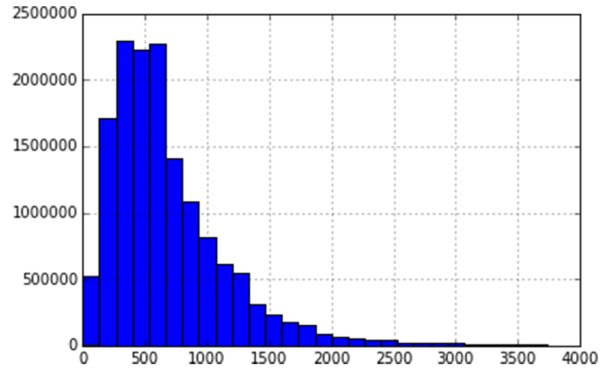


Fuente: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Así, comprobamos que la distribución tiende hacia viajes cortos, pero fijaos que los trayectos que una persona puede hacer a pie se hacen poco en taxi. La mayoría de los trayectos son en torno a un kilómetro de distancia, a pesar de que los hay mucho más largos.

Siguiendo la misma línea podemos comprobar los tiempos de trayecto. En la figura 17 podemos apreciar la distribución, que muestra que los trayectos más habituales se sitúan en torno a los 500-700 segundos, más o menos 10 minutos de viaje.

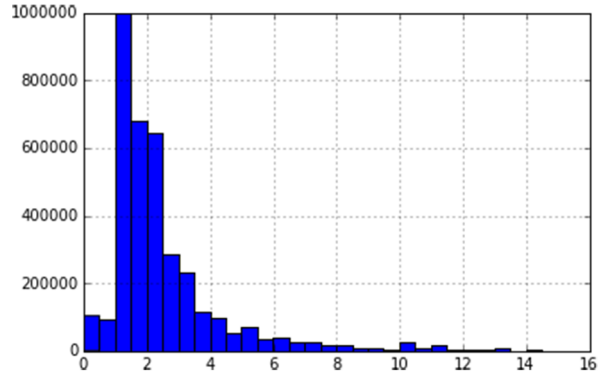
Figura 17. Histograma de la duración de los viajes



Fuente: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Hasta ahora hemos visto variables muy descriptivas (y no dejan de ser interesantes), pero a continuación nos fijaremos en una que, como taxistas, quizás nos resulte más importante: las propinas. Imaginemos por un momento que nuestro objetivo final es entender qué motiva a un cliente a dejar propina y de qué cantidad. Podemos empezar por visualizar las propinas en los viajes en los que esta se da (que no son todos).

Figura 18. Distribución de las propinas



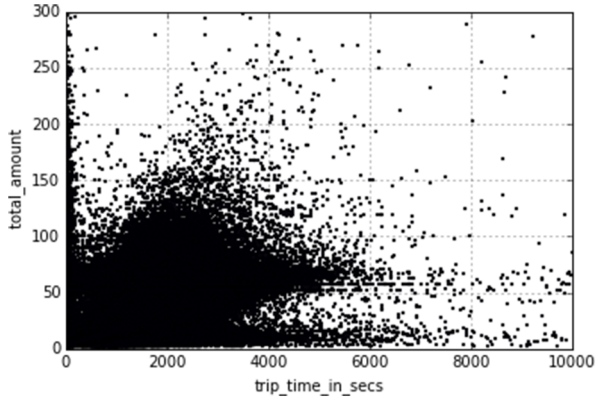
Fuente: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

De la figura 18 se extrae que la propina más habitual está entre uno y dos dólares (por lo que intuimos que esto también está relacionado con el hecho de que la mayoría de los viajes sean cortos). No es mucha información, pero como mínimo vemos que cada dólar cuenta y que, por lo tanto, hay que ser eficiente en los viajes.

Antes hemos indicado que las variables cuantitativas permiten ver sus relaciones. En la figura 19, por ejemplo, se muestra la nube de puntos de la relación entre el tiempo de viaje y el coste. Fijaos en varias cosas importantes. En primer lugar, se ve cierta correlación (que se espera) entre el tiempo y el coste (la idea

es que cuanto más tiempo, más caro es un viaje). No obstante, la dispersión es bastante grande y, por lo tanto, la relación no es clara. Pero un gráfico como este también nos permite ver otras cosas. ¿Os habéis fijado que sobre el eje vertical hay muchos puntos? Parece ser que nuestro conjunto de datos tiene muchos viajes de una duración de cero segundos. No tiene mucho sentido, ¿verdad? Pues todavía tiene menos cuando algunos de estos viajes de poca duración tienen, además, costes muy altos. Es un claro indicador de problemas en algunos registros, que pueden contener errores y necesitan ser eliminados del conjunto antes de seguir.

Figura 19. Relación entre la propina y el tiempo de viaje

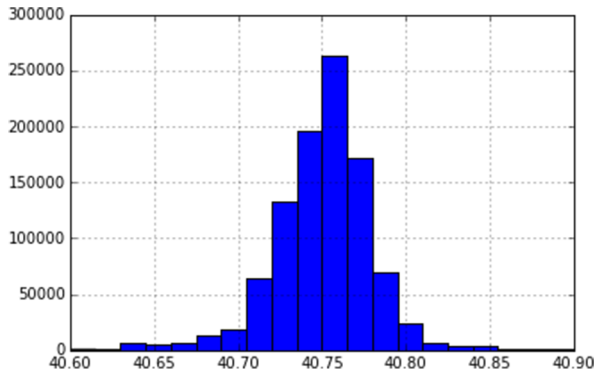


Fuente: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

3.2.3. Variables geográficas

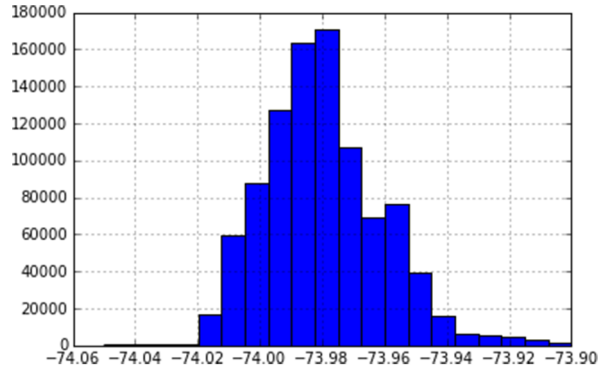
Durante las observaciones iniciales también hemos detectado toda una serie de variables geográficas. Estas incluían la latitud y la longitud, tanto de recogida como de destino. Si representamos el histograma de las latitudes y longitudes de los puntos de destino, por ejemplo, obtendremos los histogramas de las figuras 20 y 21, que muestran una distribución similar a una normal.

Figura 20. Latitud de los puntos de destino



Fuente: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

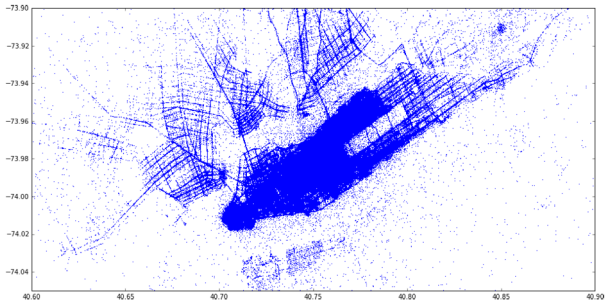
Figura 21. Longitud de los puntos de destino



Fuente: http://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

No obstante, las coordenadas geográficas tienen más información que un simple histograma. Las podemos dibujar o situar en un mapa. O, en un caso con tantos datos como este, utilizarlas para que sean los mismos puntos los que hagan un patrón y construyan el mapa. Fijaos qué gráfico más bonito (e informativo) se obtiene al representar las parejas longitud-latitud de los puntos de recogida (figura 22).

Figura 22. Mapa generado a partir de los puntos de recogida



Fuente: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

No hemos obtenido solo un mapa de Nueva York, sino que además son fáciles de detectar las zonas con más actividad, marcadas en un color más intenso y denso. El perfil de Manhattan se intuye a la perfección y el color se va difuminando al alejarse del centro de la ciudad (es decir, se cogen más taxis en el centro que en la periferia).

3.3. La creación de un modelo

Imaginemos ahora que somos taxistas. Después de visualizar y analizar el conjunto de datos nos hemos percatado de una columna que nos despierta el interés: la propina. Durante nuestro trabajo diario hemos visto que algunos clientes no dejan propina y otros sí, que a veces dejan una cantidad pequeña y otras, una más sustancial. Más de una vez hemos analizado los factores que hacen que un usuario deje una cantidad determinada de propina, pero nunca hemos tenido suficientes datos para intentarlo entender. Nunca... hasta ahora.

Una vez llegados a este punto no es necesario que frenemos la ambición. Si fuéramos capaces de conocer los factores de influencia, quizás también seríamos capaces de prever qué clientes son más propensos a dejar buenas propinas y, por lo tanto, podríamos intentar recoger siempre a los mejores clientes. Incluso podríamos instalarnos una aplicación en el teléfono móvil que nos avisara de los clientes tacaños antes de que fuera demasiado tarde y ya hubieran subido al coche.

Así, los objetivos podrían ser:

- Entender los factores que influyen sobre el hecho de dejar o no propina en los viajes en taxi en Nueva York.

- Utilizar este conocimiento para predecir la propina que recibiremos y, por lo tanto, evitar las situaciones sin propina.

Este caso, sin embargo, incluye una lección importante. Si intentarais construir un modelo con el conjunto de datos completo obtendríais unos resultados muy buenos en las métricas, pero cuando intentarais aplicarlo a vuestro taxi los resultados serían un desastre. El motivo es que las predicciones demasiado buenas para ser verdad no existen.

Una regla no escrita de la ciencia de datos es que si se obtiene una precisión mayor que la esperada, lo más probable es que el modelo esté haciendo algo inesperado. El mundo es complejo y difícil, sería extraño que se pudiera modelar de forma asequible. En este caso, el modelo inicial aparece con una variable dominante en la predicción de la propina: la forma de pago.

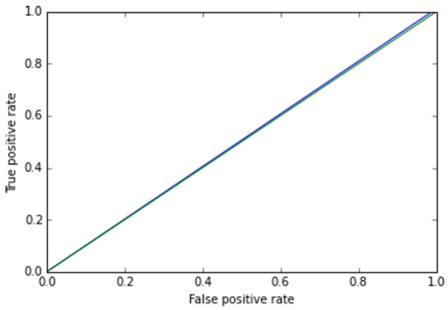
Como usuarios ocasionales del taxi podríamos pensar que lo normal son las situaciones siguientes: los clientes que pagan con tarjeta es más probable que utilicen la moneda electrónica para pagar el coste exacto y no dejar propina; las personas que pagan en metálico, en cambio, suelen redondear el resultado del taxímetro y, por lo tanto, siempre dejan algo de propina. Tiene sentido, ¿verdad? Pues no.

Si cogemos el conjunto de datos y hacemos las pruebas, veremos que la inmensa mayoría de los usuarios que pagan con tarjeta dejan propina. ¿Y qué pasa con los usuarios que pagan en metálico? Nuestro conjunto de datos dice que... ¡nadie ha dejado propina! ¿Cómo puede ser? Pues, simplemente, no es posible. La explicación es sencilla: cuando los clientes dejan propina en metálico, el conductor no lo registra adecuadamente para que aparezca en los datos. ¡Tantas ganas de saber si un cliente deja o no propina y al final parece ser que lo que acabamos de descubrir es un claro caso de fraude a la hacienda pública!

Dejando las curiosidades a un lado, ante un caso como este no hay medias tintas: si los datos están comprometidos, no se pueden utilizar. Así, podemos rehacer la pregunta e intentar buscar los motivos que llevan a los clientes **que no pagan en efectivo** a dejar propina. Nos molesta eliminar la mitad de los registros, sí, pero no tenemos más remedio.

Con los datos limpios ya podemos pasar a construir el modelo. Una forma fácil de empezar es con un modelo de regresión logística para intentar clasificar los viajes según la propina. No entraremos en el detalle de la implementación (podéis consultarla en Notebook), pero sí veremos los resultados de la curva ROC (figura 23). Esta curva representa la precisión del modelo y, con un valor de 0.5, aproximadamente, indica que el modelo no es mejor que una predicción aleatoria del resultado. Es decir, si eligiéramos los viajes a suertes acertaríamos igual que el modelo. ¡Pero no nos desanimemos!

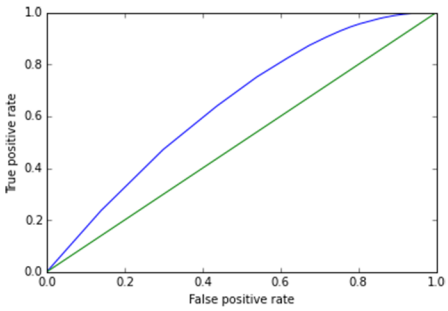
Figura 23. Curva ROC del clasificador lineal



Fuente: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Viendo los malos resultados del clasificador lineal, podemos pasar a otras opciones. La más directa es buscar un clasificador que no sea lineal. En este caso se utiliza un modelo de bosque aleatorio (*random forest*), que, como hemos visto en módulos anteriores, no es más que un conjunto de árboles de decisión.

Figura 24. Curva ROC del clasificador no lineal



Fuente: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Este modelo obtiene la curva de la figura 24. Lo más importante a tener en cuenta ahora mismo es que el área que queda bajo la curva es del 0.64 y mejora notablemente el rendimiento del modelo anterior. Todavía faltaría mucho camino, pero por ahora lo dejaremos aquí. Este modelo ya nos puede ser útil,

puesto que nos indica que hay cierta tendencia —es decir, que hay algunas variables que tienen influencia sobre la propina— y nos puede ayudar a identificarlas. Y es que, según este modelo, las variables más importantes son las que se muestran en la tabla.

Tabla 2

	Característica	Importancia relativa
0	dropoff_latitude	0.165411
1	dropoff_longitude	0.163337
2	pickup_latitude	0.163068
3	pickup_longitude	0.160285
4	trip_time_in_secs	0.122214
5	trip_distance	0.112020
6	fare_amount	0.067795

Fuente: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Fijaos que parece que el origen y el destino del usuario son la parte más relevante, seguidos de tres variables que podrían resultar más obvias: la duración, la distancia y el coste del viaje. Hasta ahora, sin embargo, no hemos introducido variables categóricas en el modelo; este sería el momento de empezar a introducir factores adicionales. Podríamos, por ejemplo, introducir una diferenciación en función de la zona tarifaria o del método de pago (siempre recordando que hemos excluido el pago en metálico). Y esto no es todo: en ciencia de datos también es muy habitual hacer ingeniería de características o *feature engineering*. Esta ingeniería consiste en generar nuevas características a partir de las ya existentes. Podríamos crear, por ejemplo, una medida de velocidad media (dividiendo la distancia por el tiempo) o aprovechar las variables de fecha y tiempo para extraer el día de la semana (o si es día laborable o festivo) o la hora (por si es hora punta). Las posibilidades son infinitas, pero siempre hay que hacerlo con sensatez, las nuevas variables deben tener sentido y tenemos que intentar que no estén relacionadas directamente con las variables del conjunto. En caso contrario, sea por relación o por exceso, podemos tener problemas de sobreentrenamiento con facilidad.

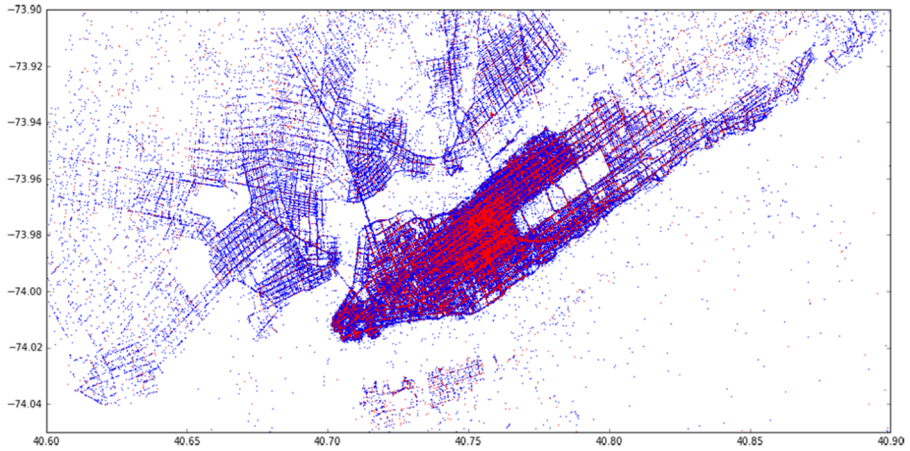
3.4. La obtención de conocimiento

A pesar de que la obtención del modelo haya resultado un fracaso relativo, el simple acto de construirlo nos ha aportado conocimiento. De hecho, la conclusión más importante todavía no la hemos condensado de una forma visible. En la lista de características relevantes de la tabla 2 han aparecido las coordenadas como las variables más importantes. Es decir, parece ser que las propinas dependen, sobre todo, de su distribución geográfica. Quizás aquí te-

nemos una pista importante si somos taxistas; parece que hay lugares donde, si recogemos a un cliente, es más fácil que nos dejen propina. ¿Cuáles serán? ¿Times Square? A continuación lo descubriremos.

La figura 25 muestra los puntos de destino con la peculiaridad de que pintamos de rojo los puntos de los viajeros que no dejan propina y de azul los de los que sí dejan propina. ¿Veis algo interesante?

Figura 25. Distribución geográfica de los destinos



De color rojo, los que no dejan propina y de color azul, los que sí la dejan. Fuente: https://github.com/brinkar/realworld-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

La conclusión es clara: los viajeros que acaban su trayecto en el centro de la ciudad no suelen dejar propina. ¿Y por qué ocurre esto? Hay varias posibilidades:

- Puede ser que la congestión del tráfico haga que los pasajeros se desesperen porque llegan tarde, los conductores hagan sonar el claxon y se estresen y, en general, que el viaje sea menos agradable.

- Pero también puede ser por culpa de los turistas. La mayoría de los taxis del centro son utilizados por visitantes extranjeros que, en gran medida, vienen de países europeos (que raramente dejan propina) o asiáticos (que todavía dejan menos). No es que sean más tacaños, sino que es una cuestión cultural.

Así, un taxista que quiera más propinas debe dirigirse a las afueras a recoger pasajeros. El problema será que, probablemente, no recoja tantos viajeros, pero aquí buscábamos información sobre la propina, no sobre el beneficio neto. Al final, aquello importante es ver que los datos generados en el mundo real nos pueden servir para explicar fenómenos sobre el mismo mundo real y las personas que generan esta información.

4. Resumen

En este módulo hemos visto varios casos de aplicación de la ciencia de datos, casos que extraemos del mundo real y que intentan hallar respuestas a problemas relevantes. Hemos repasado, también, el proceso de un proyecto completo de ciencia de datos, a pesar de que no se haya entrado en el detalle del código. Lo más importante es que queden claros los puntos siguientes:

- Cada día hay más organizaciones e individuos produciendo cantidades de datos cada vez más grandes.

- Algunos de estos conjuntos de datos están disponibles públicamente y otros son privados. Tienen varios objetivos: desde mejorar la vida de las personas hasta obtener beneficio, incluyendo la obtención de conocimiento sobre algunos fenómenos y otros casos.

- Los datos del mundo real suelen ser complejos, poco limpios e incompletos. Visualizarlos nos sirve de gran ayuda, al igual que conocer su contexto.

- Los resultados que son demasiados buenos para ser ciertos probablemente no lo sean.

- Normalmente se empieza por modelos simples, a los que se va añadiendo complejidad para mejorarlos. Es importante tener claro el motivo que se encuentra detrás de cada decisión.

- No hay ningún paso inútil, incluso un modelo fracasado puede servir para obtener conocimiento.

Bibliografía

Brink, Henrik; Richards, Joseph W.; Fetherolf, Mark (2016). *Real-World Machine Learning*. Nueva York: Manning Publications Co.

UN Global Pulse (2017). *Inferring Jakarta Commuting Statistics from Twitter*. <https://www.unglobalpulse.org/inferring-jakarta-commuting-statistics-twitter>.

UN Global Pulse (2017). *Using Big Data to Study Rescue Patterns in the Mediterranean*. https://www.unglobalpulse.org/projects/using-big-data-study-rescue-patterns-mediterranean>.

UN Global Pulse (2015). *Using Twitter to Measure Global Engagement on Climate*. <https://www.unglobalpulse.org/projects/twitter-climate-change>.

Zavazava, Cosmas (2015). *How Big Data will help fight global epidemics*. <https://news.itu.int/big-data-will-help-fight-global-epidemics/>.