GOOGLE NGRAM VIEWER Y LA ANALÍTICA CULTURAL

¿Pueden las herramientas de análisis algorítmico de la cultura confirmar hipótesis sobre hechos históricos?

Francesc Llorens. Abril 2013
@FrancescLlorens – fllorensc@uoc.edu

Se conceptualice como era del *petabyte*, era del *Big Data*, era de los *algoritmos* o como *neocuantitativismo*, lo cierto es que, desde que la evolución de la tecnología ha convertido en despreciable el coste de almacenamiento de la información y ha vuelto más barato guardar todos los datos que decidir cuáles se guardan y cuáles no, estamos asistiendo a una revolución, quizás aún subterránea para la mayoría, pero muy evidente para las empresas tecnológicas y para determinados paradigmas de explicación de la cultura que, como la *analítica cultural*, se plantean una revisión, o en todo caso una extensión de los modos en que puede ser generado el conocimiento que comúnmente llamamos científico.

Chris Anderson[1] declaró muerto el método científico tradicional, allá por 2008, a manos de los nuevos sistemas algorítmicos de obtención de patrones y modelos predictivos. Concordemos o disintamos, las *humanidades digitales*, la investigación semiótica, la historia de la literatura, la teoría de la comunicación, la geografía, las visualizaciones de datos, el diseño gráfico y un amplio abanico de disciplinas, tradicionales o embrionarias, se apropian a velocidad creciente de métodos de investigación cuantitativos, que recurren generalmente a visibilizar relaciones numéricas que permanecían ocultas en la producción gigantesca de *raw data* (datos brutos) en Internet.

Un repaso, siquiera somero, a los proyectos de Lev Manovich[2], a las líneas de trabajo de Alejandro Piscitelli[3] o Carlos Escolari[4], a la reflexión crítico-filosófica de Pierre Lévy[5] o a los estudios numéricos "distantes" de Franco Moretti[6] bastarán para darnos cuenta de que estamos ante algo más que una moda o un producto colateral de las tecnologías de tratamiento de datos. Posiblemente debamos reconsiderar el papel de los modelos positivistas en la formación de determinado tipo de conocimiento futuro (y, obviamente, su más que probable ligazón con la *ideología*), revisar nuestro concepto de determinismo y prepararnos para recibir, combatir, o ambas cosas, una epistemología imparable que, por cierto, procede en algún sentido de aunar lo consciente y lo inconsciente, lo verdadero y lo falso, y que aún está por ver a qué tipo de exigencia validatoria someterá sus hipótesis o qué tolerancia a errores considerará aceptable. A toda teoría de la verdad debe yuxtaponerse la correspondiente teoría del error.

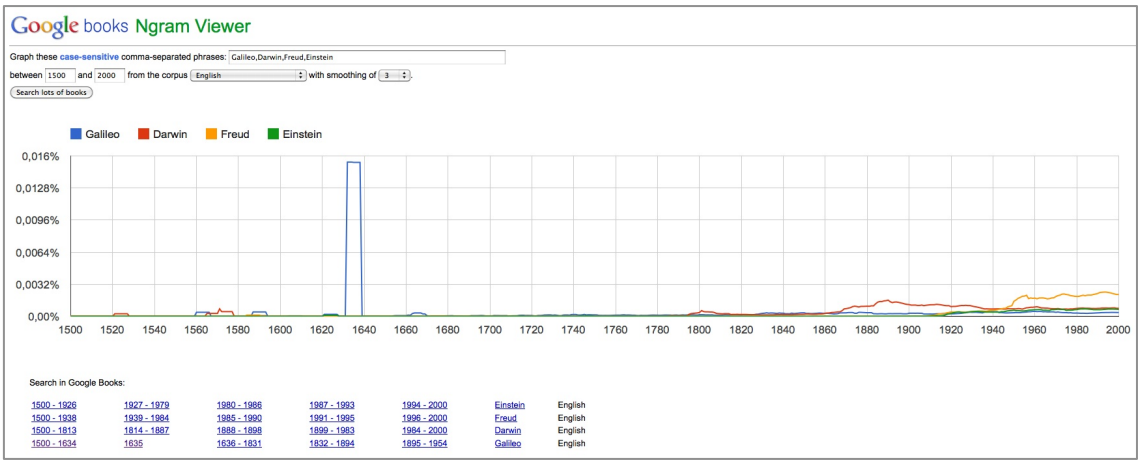
Este artículo, sin embargo, no explorará el marco teórico que subyace a la generalización de los modelos algorítmicos, ni algún aspecto filosófico particularmente asociado a él. En cambio, mi objetivo es alcanzar ciertas reflexiones generales a partir del estudio de un instrumento de análisis de tendencias basadas en ingente información cuantitativa sobre libros digitalizados[7]. Tal instrumento es el software de análisis de referencias bibliográficas de Google denominado *Ngram Viewer*[8].

Galileo y el *grama*: visualizaciones cuantitativas basadas en ocurrencias bibliográficas

Google ha desarrollado, entre otras aplicaciones de laboratorio orientadas a la visualización de datos, *Ngram Viewer*. En breve, Ngram es un algoritmo que bucea en el conjunto de libros indexados por Google y, apoyándose en la potencia de los metadatos utilizados en esa indexación, realiza comparaciones de "apariciones" de términos, llamados *gramas* a lo largo del tiempo. Un grama es una cadena, no necesariamente una palabra en sentido convencional. Lo que hace Ngram es, pues, en primer lugar, *contar*. Ngram cuenta ocurrencias de gramas. Luego, aplica otro tipo de fórmulas estadísticas y funciones booleanas, de agrupación, etc. La base de datos de todos los libros indexados por Google se denomina "corpora" (plural del latinismo *corpus*). Google ha realizado dos corpora, es decir, dos compilaciones o preparaciones de su gigantesca base de datos bibliográfica: una en 2009 y la más reciente en julio de 2012. Ambos corpora difieren entre sí, siendo el último mucho más exacto, dice Google, con respecto a los resultados devueltos. Reparemos en que realizar estas agrupaciones supone que *han debido tomarse determinadas decisiones* iniciales sobre el modo de preparar los datos para que el algoritmo los inspeccione[9]. Con Ngram es fácil realizar una comparación del mismo concepto en los dos corpora, 2009 y 2012, y por regla general se observa que el número de ocurrencias del grama buscado se afina en el corpora más reciente. Ello debe ser interpretado en términos de una mayor precisión de la búsqueda lo que a su vez confirmaría, en sentido contrario, la mayor fiabilidad y depuración del algoritmo.

El conjunto de los libros referenciados por Google presenta un primer nivel de organización: el idioma. En su uso básico, introducimos una o varios gramas en la caja de búsqueda y seleccionamos el idioma y los años de inicio y de final de la búsqueda. Ngram nos devuelve en un *gráfico de línea* la totalidad de "citas" del grama para el idioma y el periodo de tiempo elegidos. En un uso avanzado se pueden refinar las búsquedas y su combinatoria hasta un nivel de complejidad considerable, que incluye sumas y diferencias de gramas, cadenas dentro de cadenas, etc.

Como ejemplo inicial de la potencia de Ngram, y apenas accedemos a su página principal, Google muestra la comparación de apariciones de los términos "Albert Einstein", "Sherlock Holmes" y "Frankenstein" para el corpora 2012 en lenguaje inglés y entre los años 1800 y 2000. Sin embargo, en este artículo voy a ocuparme de otro ejemplo que circula por la web[10] y en el que se han creído encontrar conclusiones confirmatorias con respecto a hechos históricos conocidos. Para ello, tomemos, como hace el ejemplo de referencia, los gramas "Galileo", "Darwin", "Freud" y "Einstein". La imagen inferior muestra la comparativa de estos gramas para el lenguaje inglés entre los años 1500 y 2000.



Prescindiendo de cuestiones tales como la forma de las crestas y mesetas, que tiene su razón de ser, puede verse que la línea azul del gráfico, correspondiente a "Galileo", presenta una elevación exagerada en el periodo que va de 1633 a 1638. Ante este resultado se ha echado

mano de los libros de historia para ver qué pudo acontecer allí que justifique tal enervamiento de la curva asociada al científico de Pisa.

En los años finales de su vida Galileo publicó sus dos obras más importantes, por lo que hace a la sistematización de su pensamiento y a la trascendencia para el futuro de la ciencia del universo. En 1633 aparece el *Dialogo sopra i due massimi sistemi del mondo tolemaico e copernicano*, su obra fundamental. En ella Galileo presta claro apoyo al sistema heliocéntrico de Copérnico a la vez que ridiculiza al anticuado geocentrismo ptolemaico. En 1638, cuatro años antes de morir, ve la luz su crucial *Discorsi e dimostrazioni matematiche, intorno a due nuove scienze attenenti alla meccanica & i movimenti locali*. Obra que, al afirmar que la teoría copernicana era mecánica y no sólo geométricamente correcta, constituye el origen de la física experimental moderna.

El resultado arrojado por Ngram, y así se sugiere en la interpretación citada, confirmaría la tesis de que el pico del gráfico para los años mencionados, esto es, la "forma de la curva" es el *espejo* de un acontecimiento crucial para la historia del pensamiento científico. Si la hipótesis fuera correcta, pues, además del hallazgo que supone el refrendo empírico obtenido por vía algorítmica, también se estaría avalando —y he aquí lo realmente importante desde el punto de vista epistemológico— la idoneidad de esta herramienta y otras similares para confirmar o desmentir acontecimientos históricos. Sustancialmente, se plantearía la siguiente cuestión: ¿bajo qué condiciones, y hasta qué punto, las herramientas cuantitativas son capaces de "revelar" explicaciones de fenómenos sociales e históricos? Cuestion *crucis* sería también establecer las diferencias, para una revelación algorítmica dada, entre *confirmación* de hechos conocidos y *descubrimiento* (o "desvelación") de nuevos hechos. Este extremo queda, de momento, solamente apuntado.

Una vez repuestos del efecto fascinante que supone la observación de las curvas algorítmicamente producidas, es momento de profundizar un poco más en los resultados que las avalan. Google es honesto en su advertencia de que Ngram puede arrojar resultados engañosos cuando se trabaja con pocas muestras. Antes de 1800, por ejemplo, apenas se había publicado medio millón de libros en inglés, por lo que, ante tal escasez, el hallazgo de un grama que satisface la condición impuesta en la búsqueda puede disparar su puntuación en la gráfica. En la visualización *online*, cuando pasamos el ratón por los picos de las líneas se muestra información porcentual de las ocurrencias del término representado. Por ejemplo, el valor de ocurrencias para "Galileo" en 1633 se sitúa en el 0,015%. Esto debe ser leído así: de todos los gramas disponibles en el corpora 2012 de Google para el lenguaje inglés y el año 1633, el 0,015% corresponde al grama "Galileo".

En este punto, se estaría tentado de considerar que disponemos de una evidencia empírica nítida. La evidencia de que disponemos, sin embargo, no sólo no es nítida, sino que es de todo punto insuficiente para fundamentar una explicación plausible de alguna cosa. Para empezar, la información relativa a la ocurrencia del grama "Galileo" entre 1633 y 1638 es engañosa y por sí misma poco significativa, por lo cual debe considerarse conjuntamente con otras evidencias. Pues, si observamos el valor porcentual correspondiente al año 2000, el resultado es que sólo un 0,00037% de los libros del mismo corpora contienen el grama buscado. Y, aunque este dato es compatible con el anterior —ya que hablamos de porcentajes en años diferentes—, sucede que las exageradas diferencias de proporciones relativas disparan la curva para el periodo 1633-1638, dando la impresión de que se prestó una atención casi revolucionaria a algún fenómeno durante esos años, cuando podría haber otras explicaciones, por ejemplo, suponer que se publicaron decenas de miles de libros más en el año 2000 que en 1633 (lo cual es más que una suposición, obviamente). Google advierte de que en estas circunstancias necesitamos considerar los datos normalizados para cada segmento temporal, para evitar el sesgo.

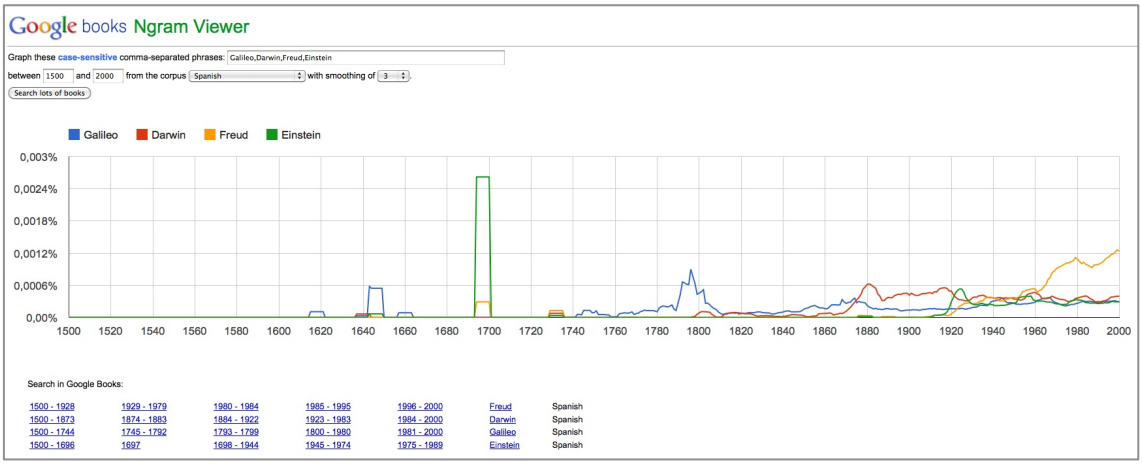
Así pues, para concluir si la visualización revela un efecto histórico *causado por la fiabilidad del instrumento de análisis*, no es suficiente con la comparación porcentual de ocurrencias. Aún debemos considerar otros datos a nuestra disposición. Particularmente, debemos atender a la

composición real de la muestra de gramas. En este punto, Google arroja luz al respecto, pero sólo relativamente: otra de las *decisiones* en la organización de los datos consiste en agruparlos por segmentos temporales (por rangos de años). Tales rangos pueden ser consultados en la web de Ngram, debajo de cada gráfico generado. Lo que Google dice de ellos es que las ocurrencias se agrupan en series anuales por "interestingness" algo así como por alto interés o importancia, lo que no es muy clarificador. Dado que los rangos no son iguales para cada grama, probablemente se obtienen a través de una subrutina del algoritmo, ignorada por nosotros, que realiza nuevas suposiciones sobre la segmentación temporal. A veces, como en el caso de Galileo, en lugar de un rango de años puede aparecer un año solo (1635). Ello se debe, dice Google, a la frecuencia suficiente con que el grama buscado aparece en el corpora ese año. Nada más.

Procedamos, pues, a inspeccionar qué hay detrás de las series temporales. Si hacemos clic en el año 1635, Ngram conecta con Google Books y enseña, o eso debemos suponer, la muestra utilizada para construir el gráfico en ese año. Aunque el modo en que Ngram y Google Books construyen el conjunto de datos retornados es distinto, ello no afecta a los objetivos de este análisis. Google Books retorna tres libros fechados en 1635. A simple vista advertimos que ahí existe información inconsistente. De los tres resultados devueltos, uno corresponde a la traducción al latín que Matthias Bernegger hizo, a instancias del propio Galileo, del *Dialogo*, otro a una obra de 1921, que contiene el valor "1635" probablemente en un metadato, y en la que el grama "Galileo" aparece una sola vez, en la página 323. Y la tercera a un famoso libro de Mary Wollstonecraft hija (Mary Shelley), publicado en Londres en 1835, en la que la autora de *Frankenstein* escribe diversas biografías históricas (aunque precisamente la de Galileo no era obra suya). Hasta donde es posible descender en los resultados, la evidencia empírica devuelta por Google Books no presenta suficientes garantías en tanto soporte confirmatorio de la interpretación propuesta.

Límites en el algoritmo y límites en la interpretación: errar es humano (y también mecánico)

En una variación del experimento anterior, produzcamos una nueva visualización alterando algún constructor, por ejemplo, el *idioma* y veamos qué sucede. Manteniendo los mismos gramas, ahora elegiremos el español como lengua de publicación. El resultado puede verse en la figura siguiente:



En una observación superficial, este gráfico y el anterior presentan un rasgo semejante. Una elevación extrema, en forma de meseta, para uno de los gramas. Ahora es el grama "Einstein" el que muestra valores elevados entre los años 1694 y 1700.

Establecer cualquier inducción a partir de este único efecto es precipitado. Pero, nótese que un principio del método científico consiste en atribuir, *para un mismo contexto experimental y en ausencia de otras consideraciones*, a idénticos efectos idénticas causas. Esta puntualización no tiene tanto que ver con el proceder del instrumento (Ngram) cuanto con el proceder de las personas: dado que el primer resultado arrojado fue interpretado en términos de concordancia entre datos cuantitativos y procesos históricos, con la misma lógica en la mano se estaría tentado de ir a buscar también hechos históricos relevantes tras las crestas del gráfico. Sin embargo, de proceder así nos llevaríamos una tremenda decepción, pues al contrastar un hecho histórico destacable relativo a Albert Einstein nos encontraríamos con el vacío. En 1694 aún faltaban casi 200 años para que Einstein naciera. Precisamos, pues, igual que en el caso de Galileo, acudir a las muestras tras las series para aclarar la situación.

Pero las muestras devueltas por Google Books no resultan aclaratorias, al menos en un sentido positivo. Al contrario, sólo revelan la existencia de inexactitudes que ningún procedimiento metódico de investigación aceptaría. Después de revisar los rangos de datos e investigar las muestras individuales, podemos identificar la existencia de varios tipos de errores en el conjunto del proceso hermenéutico. Estos errores también se producían en el caso inicial en lengua inglesa. De todos modos, para ser rigurosos debemos hablar de *supuestos* errores, dado que no sabemos con total certeza cómo trata el algoritmo las cadenas de datos. Mantengamos el término "error" como un modo de llamar la atención sobre irregularidades cuya procedencia es heterogénea, cuando no desconocida. Detectamos errores de *precisión* en la identificación de muestras, de *pertinencia* en la selección de muestras y de *perplejidad*, o desconocimiento del modo en que una muestra confirma o desmiente una hipótesis o supuesta correlación entre la evidencia algorítmica y los hechos históricos.

Precisión

Al comprobar la serie del año 1697, que aparece en solitario supuestamente porque hay un alto número de correspondencias en la indexación del corpora, Google Books nos devuelve un único libro:

He aquí un caso curioso. La fecha de esta referencia es 1697. Ahora bien, es evidente que esto no es posible. Este libro no fue editado ese año, sino varios siglos después. El error podría derivar de una anotación equivocada de la fecha en el registro, lo que, a su vez, podría tener un origen humano o mecánico. Sin embargo, en este caso no se trata de un error en el contenido del campo 'fecha de edición', sino de la utilización errónea de otro metadato: la dirección física de la editorial que editó el libro, que contiene como parte de la cadena el número 1697, como se ve en la captura tomada de su página de créditos:

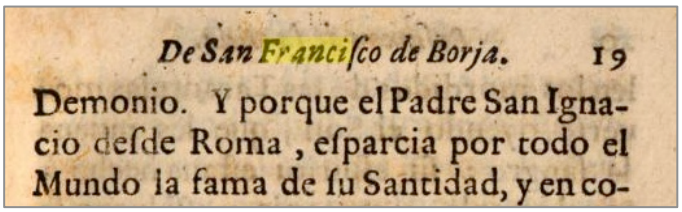
El algoritmo ha encontrado la cadena "1697" *en algún lugar* de una referencia y ha mostrado el ítem como evidencia empírica.

Consideremos ahora el segmento 1500-1696, aún más lejano de la época en que vivió Einstein. Este segmento lista 8 libros, todos ellos ediciones en castellano antiguo, digitalizadas por Google. Al analizar cualquiera de ellos, el sistema resalta la ocurrencia del grama "Einstein". Tomemos las *Excellencias y primacias del apostol Santiago el Major unico patron de Espana,* escrito en 1657 por Antonio Calderón (arzobispo de Granada) y Gerónimo Pardo.


Patron de España, y Capitan General
de las armas Catholicas.
y de todas las guerras contra Infieles
CON
VN COMPENDIO HISTORICO DE LA
vida hechas, y muerte del glorioso Apostol, en
que sumariamente se refieren las grandezas d

Como puede observarse, el algoritmo ha destacado el término 'Infieles' que, evidentemente, ha sido confundido con 'Einstein' por el subsistema de reconocimiento óptico de caracteres (el OCR). La misma confusión se da en los otros siete libros listados, lo que es absolutamente lógico, dado que una ocurrencia del grama "Einstein" en esta época, aún siendo *real* —esto es, aún coincidiendo con la cadena de búsqueda introducida—, es obvio que no podría referirse en ningún caso al autor de la Teoría de la Relatividad.

Otro tanto sucede con el grama "Freud", que también presenta una elevación inaudita, aunque bastante menor que la de Galileo, curiosamente en la misma época. Al verificar las muestras que soportan la visualización en el anormalmente extenso rango 1500-1928, encontramos que la representación gráfica se apoya en 145 libros. Tomemos el primero que aparece, *Resumen de la vida y milagros de San Francisco de Borja* de Scipio Sgambata, publicado en 1641, doscientos años antes del nacimiento del psicoanalista. Al comprobar la aparición del grama "Freud" en él, damos con esto:


De San Francisco de Borja. 19
Demonio. Y porque el Padre San Igna-
cio desde Roma, esparcia por todo el
Mundo la fama de su Santidad, y en co-

Ahora, el OCR, subsistema de Google Books con el que el algoritmo de Ngram trabaja en conjunción, confunde "Freud" con "Franci", provocando el listado erróneo de la muestra. Y así, para la mayoría de muestras antiguas.

Toda la evidencia presentada por Google Books para el grama "Einstein" y la serie de datos entre 1500 y 1696, así como mucha de la evidencia de los otros gramas para los periodos más antiguos, procede de errores en la rutina de digitalización. Google nos advierte también contra la posibilidad de este tipo de errores de reconocimiento. Para la mayoría de ejemplares antiguos, el escaneado es muy defectuoso e incluso transparece la cara opuesta de cada hoja, como se aprecia en la imagen anterior. Google asegura que estos errores eran más frecuentes en el corpora de 2009 y que su OCR ha mejorado sustancialmente en el corpora de 2012. Pero parece que aún debe mejorar un poco más, o bien despreciar mayor número de muestras.

Pertinencia

Una cuestión relacionada con la cantidad de soporte empírico que se encuentra tras la visualización de un rango de datos es la de la *pertinencia* de la muestra. En otras palabras: cuántos ejemplares, del conjunto seleccionado por el algoritmo, constituyen muestras relevantes o pertinentes para avalar la hipótesis explicativa. Pues no es suficiente que Ngram nos devuelva mil muestras en un rango. Es necesario discriminar, de esas mil, las que apuntan de verdad al objeto investigado.

De los 187 libros que componen el rango 1500-1873 para el grama "Darwin", sólo 14 (un exiguo 7,5%) se refieren a Charles Darwin, autor de la *El Origen de las Especies*. La proporción va a peor en los primeros rangos de los otros gramas. Para el rango inicial del grama "Freud", Google Books lista 163 ejemplares, de los que sólo 6 (el 3,7%) contienen referencias al verdadero Sigmund Freud. Ya se señaló que para el rango 1500-1696 del grama "Einstein" los resultados arrojaban el 0% de muestras pertinentes. Si nos las habemos con sistemas de proceso masivo de datos, la primera conclusión quizás sea que los datos disponibles para este periodo no alcanzan a poner a prueba la potencia del algoritmo, sin duda, pero la segunda conclusión es que los datos seleccionados contienen referencias irrelevantes de cuya contabilidad hay que desconfiar, ante cuya presencia hemos de ser extremadamente juiciosos y que deben ser separadas con claridad de las referencias correctas. Experimentalmente hablando, en el caso de estudio existe una proporción de error inaceptable, por imposible de justificar, para presumir que las muestras algorítmicamente listadas constituyen un soporte válido confirmatorio de cualquier explicación teórica propuesta.

Aún en el caso de no producirse errores algorítmicos, como sucede con los libros digitalizados actuales, o con aquellos de los que se poseen fragmentos de texto, la pertinencia de la muestra es cuestionada entonces por la posibilidad de que el grama buscado responda a más de un elemento existente en el mundo real o histórico y relevante en algún contexto imaginable. Por ejemplo, Freud puede referirse a Sigmund, el psicoanalista, o a su nieto Lucien, el pintor. Darwin puede ser Charles, el naturalista, o Erasmo, su abuelo, el botánico y filósofo del siglo XVIII.

Apenas en el lugar 23 de los libros listados para el grama "Darwin" y el periodo 1984-2000 aparece este ejemplar, que se refiere precisamente a Erasmus Darwin, y no a Charles:

En algunas circunstancias, y hasta cierto punto, estos equívocos podrán ser corregidos mediante una búsqueda avanzada de cadenas, utilizando operadores y modificadores. Por tanto, es necesario considerar e investigar esta posibilidad antes de asumir apresuradamente un conjunto de datos retornados como el universo cuantitativamente válido de datos.

Perplejidad

Si no es aceptable tomar en cuenta muestras cuya referencia al objeto investigado es irrelevante, aún lo es menos considerar como parte de la evidencia otras que no presentan, hasta donde el investigador alcanza, ninguna relación con dicho objeto. Aunque parezca raro, tales muestras existen y son seleccionadas por el algoritmo de Ngram. En otras palabras: el sistema proporciona un tipo de evidencia que, con la información disponible, *no sabemos de qué modo presta soporte confirmatorio* a las hipótesis que pudieran deducirse de la visualización de los datos. Veamos: si se toma un rango cualquiera, al azar, de un grama cualquiera, se encontrarán con mucha probabilidad muestras de las que es difícil, o imposible, suponer cómo han llegado a ser seleccionadas por el algoritmo. Consideremos la siguiente muestra, listada en el grama "Darwin" para el rango 1500-1873:

O esta otra, listada en el grama "Galileo" para el rango de años 1981-2000:

O ésta, para el grama "Freud" y el segmento 1985-1995:

No existe conexión semántica, temática, espacio-temporal o de cualquier índole relevante en un contexto de investigación controlado, que indique que estas muestras deben formar parte de la visualización. Pero son contabilizadas algorítmicamente. Desde un punto de vista experimental, y siempre que se desee mantener una vigilancia humana del procedimiento, no podemos aceptar un elemento de una muestra del que no sabemos *cómo afecta a la verificación de la hipótesis*, esto es, si la confirma, la desmiente o es absolutamente independiente de ella.

Por último, ¿qué sucede con las crestas y mesetas superpuestas en el gráfico de referencia para los rangos históricos en los que aún no habían nacido los autores? ¿Por qué esa superposición para los años 1645, 1700 o 1730? Bien podría suceder que la digitalización de libros de esas épocas haya sido realizada a partir de fuentes muy parciales, como bibliotecas o instituciones que dispusieran sólo de ejemplares de esos periodos. Y bien podrían existir otros ejemplares en instituciones que no hayan firmado un acuerdo de digitalización con Google, luego a todos los efectos es como si no existieran. Si aceptamos como indicativo de un efecto histórico la cresta del grama "Galileo" alrededor de 1635, deberemos explicar también la desaparición de toda referencia en los siguientes 20 años o su casi inexistencia en los anteriores 70. Quizás la explicación correcta sea la más sencilla: no tenemos corpus de esos años ciegos y, por motivos que no tienen que ver ni con el algoritmo ni con la hipótesis de trabajo, sí disponemos de corpus para algunos otros momentos puntuales de la historia.

Lo cuantitativo y lo cualitativo: en puertas de una nueva epistemología del dato masivo

El optimismo cuantificacional debe moderarse, o modularse. La analítica cultural necesita de una ingente cantidad de datos para producir correlaciones y evidenciar patrones que puedan llegar a ser tenidos en cuenta, en ausencia de otros criterios, por un sistema de calidad científico. Esto no tiene nada de misterioso y no es más que la extensión de una propiedad esencial de los estadísticos básicos: un estadístico básico es más estable y confiable cuanto mayor es el tamaño de la muestra a partir de la cual se obtiene. Los algoritmos contienen muchas funciones de tipo estadístico y probabilístico[11]. A enormes conjuntos de datos, las magnitudes estadísticas tienden a la estabilidad y ello permite descartar determinado número de casos por irrelevantes, inconsistentes o erróneos, y que el poder confirmatorio del procedimiento se mantenga en esencia inalterable.

No hay muestras de datos, ni universos de datos, *per se*. Hay muestras y universos *pasados* y *presentes*[12]. Y hay muestras y universos asumiendo condiciones teóricas previas, como algún tipo de segmentación de los datos, un área geográfica dada, un idioma preseleccionado.... Suponiendo un universo de datos para un determinado momento (esto es, sin considerar su extensión futura con la incorporación de nueva evidencia) entonces, cuando muestra y universo coinciden, los estadísticos básicos ya no *predicen* comportamientos poblacionales (ya no confirman o desmienten hipótesis), sino que *describen* propiedades. Y éste es el paso filosófico fundamental (y el que la ciencia realiza cuando concede carta de natalidad a una ley de la naturaleza): una confirmación de tipo inductivo es un procedimiento cuantitativo pero una descripción de las propiedades de un universo de datos es una *descripción de una parte de la realidad*; por lo tanto es una atribución ontológica, *ergo* cualitativa. Este supuesto crucial, no explícito, de los procesos algorítmicos de tipo Big Data debe ser visibilizado: para conjuntos ingentes de datos (para la computación ideal de *todos* los datos), el porcentaje de error se volvería tan despreciable que un patrón o correlación estable no se interpretaría sólo como el apoyo inductivo de una *tendencia* en la población, sino como la *descripción objetiva* de una característica constitutiva suya, esto es, como una propiedad ontológica de una parte de la realidad.

Sería un grave error desconocer, menospreciar o criticar desde posiciones interpretivistas irreductibles la potencia de la visibilización de patrones y tendencias basadas en el análisis automático de datos cuantitativos. Como lo sería suponer que cualquiera de esas críticas va a frenar el avance de lo que se preconiza como un nuevo paradigma tecnocultural. Pero no lo sería menos deificar el producto de este *tracking* algorítmico de la cultura sin triangularlo con otros sistemas racionales de formulación de hipótesis, comprensión de fenómenos y explicación de resultados. Y aún lo sería más esconder que los procedimientos numéricos, aparentemente formales y desprovistos de coloración, encierran en diferentes sentidos un poderoso componente teórico e ideológico que oscila desde las decisiones sobre la selección y el formateo de los *raw data* hasta los objetivos extranuméricos a que los datos deben servir (la *metateoría*). Todas las objeciones clásicas a la neutralidad científica procedentes de la sociología de la ciencia siguen siendo válidas en la era de las metaheurísticas.

En otras palabras, aunque es perfectamente legítimo, y quizás irreversible, revisar nuestras concepciones clásicas sobre aspectos del método científico (en particular sobre la naturaleza de la confirmación, la dialéctica hipótesis-datos y las relaciones entre explicación y predicción) aún está por demostrar que la producción algorítmica de teoría pueda justificar sistemáticamente la relevancia, y por ende la aceptación, de un dato *porque* haya sido seleccionado por un algoritmo.

Con independencia de la dificultad de reducir cualquier noción de cultura a la cantidad de veces que se producen "datos" culturales, deben constatarse los siguientes extremos que será perentorio incorporar a toda reflexión holística y seria sobre la analítica cultural:

1. El formateo de los datos brutos, esto es, la simple preparación para su procesamiento algorítmico, presupone la asunción de supuestos teóricos. Las condiciones bajo las cuales un dato es interpretado como relevante y, por tanto, no es excluido, o la depuración gramatical a que se somete el dato *a priori*, ya son condiciones teóricas. Pierre Lévy lo expone con claridad en *Le médium algorithmique* (véase nota 5). Por ejemplo, Ngram Viewer asume ciertas reglas para la división de palabras (el "token") que no son las mismas en todos los idiomas y que tienen efectos en los resultados. No conocemos estas reglas, han variado entre los corpora de 2009 y 2012 e incluso no existen para el idioma chino (para el que se utiliza un sistema estadístico de segmentación) lo que significa que la descomposición interna del algoritmo ante la misma búsqueda es distinta para cada lengua. Es evidente que la confusión del OCR entre "Freud" y "Franci" no se produciría en los textos en lengua china o hebrea (en los que, de hecho, no hay muestras). Ello es natural y se controla definiendo el universo de aplicación del algoritmo[13]; pero pone en evidencia el componente teórico en el formateo de los *raw data*.

2. Jamás conocemos en profundidad el algoritmo en un sistema de análisis masivo propiedad de terceros. Los algoritmos poseen partes públicas y partes privadas, y además se protegen mediante secreto de patente, pues proporcionan ventajas competitivas a las empresas. Aún cuando se produzca un resultado confirmatorio sobre una asunción teórica previa, no siempre es fácil de conocer la razón de fondo y el por qué de este ajuste. Podemos constatar una correlación con fidelidad, pero el modo en que los datos contribuyen a fundamentar la explicación derivada del análisis de esa correlación no siempre puede ser justificado (error de perplejidad). Por ejemplo, desconocemos por qué Ngram selecciona determinados libros en los que al realizar idéntica búsqueda desde Google Books no aparece el término buscado.
3. En el caso de Ngram existen dos problemas de fiabilidad que son independientes de los procedimientos algorítmicos. En primer lugar, el error humano en el formateo de los datos, que debe acotarse y descartarse. *Idealiter*, una investigación debería ser capaz de eliminar estos errores. En segundo lugar, el problema del OCR, esto es, del filtro que en definitiva localiza el patrón de búsqueda. De la importancia de este factor da cuenta el hecho de que Google haya reconstruido radicalmente los corpora sobre los que trabaja el algoritmo de Ngram Viewer.
4. La producción algorítmica de teoría no es la producción algorítmica de metateoría, es decir, de elementos subyacentes, justificativos o legitimantes, incluyendo los ideológicos. Ambas cosas no pueden mezclarse y debe quedar claro donde termina una y empieza, si empieza, la otra. La idea de que el Big Data debería denominarse *smart data* o *predictive analytics*[14] es sugerente y llamativa desde el punto de vista comercial, pero la predicción de tendencias puede ser interesadamente utilizada para decretar *de facto* que una realidad ha dejado de existir y ha sido sustituida por otra. Hace unos pocos meses se popularizó un estudio (llevado a cabo en el seno de la Cornell University y en el marco de una investigación doctoral[15]) basado en datos masivos de Twitter que dictaminaba en qué momentos del día y qué días de la semana somos más felices los humanos[16]. Así, si la humanidad es más feliz, pongamos por caso, los días laborables entre las 8 y las 9 horas y los sábados, ello, a su vez, servirá para tomar decisiones (publicitarias, comerciales, de segmentación informativa…) que sí afectan a la vida de las personas, comenzando por la elemental que consiste en sentirnos *anómalos* si no nos reconocemos dentro la campana de Gauss de la felicidad (concepto metateórico en sí mismo). Una *anomalía inferencial* puede ser tratada como un simple error (muestral, por ejemplo) pero una *anomalía descriptiva* no es un mero rasgo teórico, sino metateórico, pues supone discriminar facetas o sectores de realidad y *distribuir* a las personas en función de ciertos sesgos psicológicos de su carácter derivados de un constructo algorítmico previo. Poco importa, a efectos de una multitud de decisiones que finalmente impactarán sobre mí, que me sienta feliz, si el algoritmo dice que no lo soy. Y al contrario. Esta forma de proceder no es en absoluto novedosa, pero su impacto *masivo* es la gran atracción circense del futuro de la cuantificación.
5. Los factores externalistas en la ciencia (en el sentido de Michel Foucault[17]) adquieren tanto o más peso que los factores internalistas: las expectativas de un programa de investigación, las decisiones políticas, los intereses económicos e incluso el control académico sobre lo que debe o no ser investigado son elementos decisivos en la configuración de algoritmos de análisis cuantitativo. Quien fuera director científico de Google hace unos años, Krishna Bharat[18], se hizo famoso por afirmar que los algoritmos no tienen ideología. Pero, podríamos contestarle, las personas que los escriben y las empresas para las que los escriben sí.

Cualquiera que sea su bondad, matemáticamente hablando, el algoritmo sufre la coerción de su construcción probabilística: no puede decir más de lo que permiten sus límites. El resto es un proceder humano, común y corriente. Como cualquier escala de medida, el algoritmo posee errores intrínsecos. No se trata de permanecer impermeables por principio a la analítica cultural ni de descartar los algoritmos (no creo que ello sea posible, ni deseable), sino de conocer bien sus mecanismos, los intereses para los que han sido creados y obrar con cautela. Probablemente

no podemos concebir ya la generación no algorítmica de cualquier suceso. Pero, por la misma razón, tenemos la exigencia de conocer mejor que nunca las rutinas operacionales por las cuales se decreta una nueva realidad, un nuevo saber y un nuevo método científico. Pues la libertad procede del conocimiento profundo, no de la evasión.

NOTAS

[1] Chris Anderson. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

[2] Lev Manovich. (2012). The meaning of statistics and digital humanities. http://lab.softwarestudies.com/2012/11/the-meaning-of-statistics-and-digital.html Y, en general, sus proyectos y publicaciones: http://lab.softwarestudies.com/p/publications.html

[3] Alejandro Piscitelli. (2013). Las humanidades digitales y la fusión entre arte y ciencia. http://conectarlab.com.ar/las-humanidades-digitales-y-la-fusion-entre-arte-y-ciencia/

[4] Carlos A. Scolari. (2012). Occupy Semiotics (Hacia una semiótica del Big Data). http://hipermediaciones.com/2012/12/16/occupy-semiotics-big-data/

[5] Pierre Lévy. (2013). Le médium algorithmique. http://pierrelevyblog.com/2013/02/17/le-medium-algorithmique/

[6] Franco Moretti. (2007). *La literatura vista desde lejos*. Editorial Marbot. Barcelona. http://www.marbotediciones.com/es/inicio/catalogo/la-literatura-vista-desde-lejos/item/la-literatura-vista-desde-lejos

[7] Jean-Baptiste Michel*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden*. (2010). *Quantitative Analysis of Culture Using Millions of Digitized Books*. Science (Published online ahead of print: 12/16/2010)

[8] http://books.google.com/ngrams/ O, para descarga de la aplicación, datos, y otras muy interesantes advertencias http://www.culturomics.org/home

[9] Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, William Brockman, Slav Petrov. (2012). *Syntactic Annotations for the Google Books Ngram Corpus*. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Volume 2: Demo Papers (ACL '12).

[10] Véase http://agorarosario.blogspot.com.ar/2013/02/apostillas-sobre-analitica-cultural.html

[11] Por ejemplo, el tratamiento de gramas puede incorporar procedimientos probabilísticos de tipo cadena de markov https://www.ibm.com/developerworks/mydeveloperworks/blogs/nlp/entry/the_chain_rule_of_probability?lang=en

[12] Ello es lo que lleva a Clayton Chistensen (http://www.claytonchristensen.com/) a afirmar: "Data is only available about the past. If we base all our decisions on data, we will only act when it's too late" Nieman Foundation for Journalism. Harvard. http://www.nieman.harvard.edu/assets/Image/microsites/disruptor/splash/index.html En streaming el 27 de febrero 2013.

[13] De hecho, es el procedimiento sugerido en http://www.culturomics.org/Resources/A-users-guide-to-culturomics

[14] Es lo que propone John De Goes. Véase http://venturebeat.com/2013/02/22/big-data-is-dead-whats-next/

[15] Investigación de Scott Golder del Departamento de Sociología de la Cornell University. Véase la aplicación en http://timeu.se/ y el *paper* en http://redlog.net/timeuse/

[16] Véase http://abcnews.go.com/blogs/technology/2011/09/twitter-used-to-track-the-worlds-mood/

[17] Michel Foucault distinguió, en relación con la historia de la ciencia, entre historia interna y externa, a fin de esclarecer las enormes diferencias que existen entre "decir la verdad" y "estar en la verdad" de cada época (Michel Foucault. *El orden del discurso*. Tusquets. Barcelona, 1970). Tal distinción, además, se halla históricamente ligada a la definición de los criterios de normalidad y exclusión.

[18] Citado por F.F. Campillo. *Modelos de verdad en la cultura digital.* www.uv.es/=demopode/libro1/FrancescFelipe.pdf