Anonimización de datos no estructurados a través del reconocimiento de entidades nominadas

Fadi Hassan, Josep Domingo-Ferrer, Jordi Soria-Comas
Universitat Rovira i Virgili
Department d'Enginyeria Informàtica i Matemàtiques
Cátedra UNESCO de privacidad de datos
CYBERCAT-Centre de Recerca en Ciberseguretat de Catalunya
Av. Països Catalans 26, 43007 Tarragona
E-mail {fadi.hassan, josep.domingo, jordi.soria}@urv.cat

***Resumen*—La anonimización de datos estructurados ha sido ampliamente estudiada en el pasado reciente. Sin embargo, anonimizar datos no estructurados (típicamente documentos de texto) sigue siendo una tarea manual que necesita más atención por parte de los investigadores. La principal dificultad a la hora de tratar datos no estructurados es que no existe un esquema de base de datos que pueda usarse para medir riesgos en la privacidad. De hecho, datos confidenciales y valores cuasi-identificadores pueden estar repartidos a lo largo de los documentos a anonimizar. En este trabajo proponemos usar un marcador de reconocimiento de entidades nominadas basado en aprendizaje automático. La finalidad es crear un sistema capaz de detectar todos los atributos que tienen implicaciones en la privacidad (identificadores, cuasi-identificadores y atributos confidenciales). Concretamente, presentamos una prueba de concepto centrada en la detección de atributos confidenciales. Consideramos un caso de estudio en el cual los valores confidenciales a detectar sean nombres de enfermedades en diagnósticos médicos. Una vez detectados estos valores de atributos confidenciales, se pueden usar técnicas estándar de control de revelación estadística para datos estructurados con el objetivo de controlar el riesgo de revelación.**

***Index Terms*—Anonimización, datos no estructurados, reconocimiento de entidades nominadas, campos condicionados aleatorios.**

I. Introducción

Hoy en día, se están recogiendo grandes cantidades de datos de fuentes muy diversas, a menudo sin que los individuos afectados tengan constancia de ello. Una recogida de datos tan sistemática, acoplada a nuevas técnicas de análisis de datos, ha dado paso a lo que se conoce como *big data* o *megadatos*. Aunque a veces se califiquen como moda, los megadatos implican un cambio significativo en la forma de manipular los datos. En este trabajo, nos preocupamos de las implicaciones de privacidad de los megadatos, concretamente de los megadatos no estructurados.

En el entorno tradicional, los datos eran recogidos principalmente a través de encuestas u otras fuentes de datos administrativos. Como resultado, solían tener una naturaleza estructuradas (una tabla). La amplia variedad de fuentes de datos en el contexto actual de megadatos (por ejemplo, correos electrónicos enviados y recibidos, participación en redes sociales, etc.) nos fuerza a considerar otros tipos de datos, como datos semiestructurados o no estructurados (texto libre). Ya en 2005, se comentaba en [9] que un 80 % de los datos comerciales y médicos se guardaban de forma no estructurada. En el contexto sanitario, el uso adecuado de estos datos es crítico para la investigación y la creación de políticas, además de ser útil para industrias relacionadas, como seguros de salud.

La nueva Regulación General de Protección de Datos europea (en inglés, *General Data Protection Regulation* o GDPR, [8]) afirma que es necesario el consentimiento explícito de los individuos afectados para usar información identificable personalmente (en inglés, *personally identifiable information* o PII) para fines secundarios (diferentes a los fines primarios que motivaron su recogida, como la sanidad o la facturación de un servicio). Idealmente, el recolector de datos debería esforzarse en adquirir dicho consentimiento. Sin embargo, esto puede no ser posible en la práctica. Puede ser difícil contactar con los individuos para obtener su consentimiento. Además, los individuos con condiciones anómalas suelen preocuparse más por su privacidad, lo que los hace menos propensos a dar consentimiento para usar sus datos. Debido a estos inconvenientes, es probable que la PII obtenida mediante consentimiento sea algo sesgada.

Para evitar la necesidad del consentimiento, los datos usados para fines secundarios no deberían identificar al individuo. La anonimización, también conocida como control de revelación estadística (en inglés, *statistical disclosure control* o SDC), proporciona una manera de transformar la PII en información que no puede asociarse a un individuo concreto identificado y, por lo tanto, no está sujeta a regulaciones de privacidad.

Existe una gran cantidad de literatura sobre SDC para el caso de datos estructurados [10], [5], [4]. Los datos estructurados son aquellos que pueden describirse como un conjunto de registros que corresponden a un individuo y contienen los valores de un conjunto fijo de atributos para ese individuo. La anonimización de datos estructurados consiste normalmente en eliminar los atributos que son identificadores y enmascarar los atributos cuasi-identificadores. Estos últimos son atributos que no son identificadores por sí solos, pero que conjuntamente pueden permitir enlazar el registro con alguna fuente de datos externa que contenga identificadores y, por lo tanto, podrían reidentificar el individuo al que pertenece el registro. Alternativamente, en vez de (o además de) enmascarar los cuasi-identificadores, se pueden enmascarar los atributos confidenciales para introducir incertidumbre sobre sus valores.

Una vez decididos qué atributos son cuasi-identificadores y cuáles son confidenciales, la anonimización de datos estruc-

turados puede automatizarse (cierto es que, en algunos casos, la decisión puede no ser del todo clara, ya que depende de la información de contexto que se supone disponible para el intruso). Sin embargo, la automatización de la anonimización de datos no estructurados es mucho más compleja, ya que no hay un esquema de base de datos con el que clasificar los datos en atributos identificadores, cuasi-identificadores y confidenciales. Como resultado, anonimizar datos no estructurados sigue siendo a día de hoy una tarea en gran parte manual.

De hecho, se puede argumentar que los datos no estructurados en forma de texto son los más complejos a que debe enfrentarse la anonimización. Otros tipos de datos que pueden parecer más complejos a primera vista o bien pueden reducirse a texto no estructurado usando herramientas automatizadas de extracción semántica (como ocurre con vídeos o audios) o bien es mejor no anonimizarlos debido a la comprensión todavía incompleta de su semántica (como es el caso de los datos genéticos).

Contribución y estructura de este artículo

El propósito de este trabajo es automatizar la extracción de atributos cuasi-identificadores y/o confidenciales de datos no estructurados en forma de texto. Es decir, queremos ser capaces de identificar, de forma automática, atributos como número de pasaporte, nombre, localización, edad, fecha de nacimiento, etc. Para ser más concisos, en este trabajo nos enfocaremos en documentos de diagnósticos médicos. Una vez se complete este proceso de identificación automática de atributos, podremos aplicar algunos métodos diseñados para anonimizar datos estructurados. Para identificar atributos, usaremos un marcador de reconocimiento de entidades nominadas (en inglés, *named-entity recognition* o NER) [7].

En la sección II, presentamos brevemente algunos conceptos importantes para entender este trabajo. En la sección III, se recuerdan trabajos previos relacionados con la anonimización de documentos. En la sección IV, describimos nuestra propuesta. Los experimentos se presentan en la sección V y las conclusiones e ideas para trabajos futuros se recogen en la sección VI.

II. Antecedentes

II-A. Reconocimiento de entidades nominadas

El reconocimiento de entidades nominadas es la tarea de localizar y categorizar términos importantes de un texto [18]. Este reconocimiento es una fuente de información para diferentes aplicaciones de procesamiento de lenguaje natural. El NER ha sido utilizado para mejorar el rendimiento de varias aplicaciones, como la respuesta de preguntas [12], la traducción automática de texto [1], la recuperación de información [24], y el análisis de los sentimientos de los tweets [11].

El NER también es útil para la anonimización de datos no estructurados (p.e. documentos de texto libre). En particular, puede detectar los términos que pueden usarse para reidentificar a un individuo y aquellos términos que contienen información sensible. Una vez localizados estos términos, constituyen información estructurada que puede ser anonimizada usando métodos de SDC (tales como generalización, supresión, etc.) para mantener bajo control el riesgo de revelación.

Hay muchos esquemas de marcado para NER. En este trabajo, usamos el esquema de marcado IOB2 [22]. En IOB2, cada palabra en el texto se marca usando una de las tres posibles etiquetas: I, O o B, que indican si la palabra está dentro (inside), fuera (outside) o al principio (beginning) de una entidad nominada.

II-B. Campos condicionados aleatorios

En procesamiento de lenguaje natural, hay dos modelos comunes para resolver tareas de NER: modelos de Markov ocultos (en inglés, *hidden Markov models* o HMMs), usados en trabajos como [17], [28], y campos condicionados aleatorios (en inglés, *Conditional Random Fields* o CRFs), usados en trabajos como [3], [6], [11]. Los NER basados en CRFs son ampliamente usados y aplicados, y normalmente dan mejores resultados en varios campos [16]; por ello, en este trabajo diseñamos nuestro modelo usando CRFs.

Los CRFs [15] son modelos de grafos no dirigidos entrenados de forma condicionada que a menudo se aplican en el reconocimiento de patrones. Estos modelos se usan para calcular la probabilidad condicionada de valores en determinados nodos de salida dados los valores asignados a otros nodos de entrada.

III. Trabajo relacionado

Se han propuesto varias técnicas de anonimización para datos no estructurados en forma de texto. Muchas de ellas pueden clasificarse en una de las siguientes categorías: técnicas basadas en diccionarios y técnicas de aprendizaje automático [20].

En el pasado, la anonimización de documentos se hacía de forma manual, buscando y reemplazando las entidades nominadas. Sweeney [25] propuso el método de fregado que se basa en la definición de plantillas para las entidades nominadas, como la localización, el nombre y el país. Una vez encontradas estas entidades, se enmascara el valor relacionado.

Neamatullah et al. [19] propusieron un software para anonimizar documentos que usaba tablas de búsqueda lexicográfica, expresiones regulares y heurísticas simples, realizando comprobaciones en el contexto para localizar entidades nominadas. Tras la búsqueda, se cambiaban estas entidades por valores de categorías no indexadas (por ejemplo, reemplazar "Nueva York" por "[**Localización**]").

Vico y Calegari [27] proponen una arquitectura de software para anonimizar documentos. La idea principal es reconocer las entidades nominadas con una arquitectura de múltiples herramientas de procesamiento de lenguaje natural. Después del reconocimiento, reemplazan las entidades sensibles por un valor genérico de una categoría indexada (p.e. reemplazar "Fiebre" por "termino_generico_1").

En 2016, el United Kingdom Data Archive (UKDA) lanzó una herramienta de ayuda para la anonimización de texto [23]. Esta herramienta identifica los números y las palabras que empiezan con letra mayúscula y los reemplaza por "XXX".

Kleinberg et al. [13] diseñaron Netanos, una herramienta que permite a los investigadores anonimizar textos largos. Usan aprendizaje automático para reconocer entidades nominadas (por ejemplo, personas, localizaciones, tiempos y fechas) y las sustituyen por valores de categoría indexadas

que preservan la privacidad (tales como “Localización_1”, “Persona_1”).

IV. Metodología

El objetivo de este trabajo es localizar términos en un texto no estructurado que puedan tener implicaciones en la privacidad, ya sea porque pueden ser usados para reindentificar a un individuo o porque contienen información confidencial.

IV-A. Enfoque general

Formalmente, dada una colección de documentos de texto $D_1, \ldots, D_n$, queremos localizar los atributos que tienen relevancia para la privacidad. Específicamente, queremos obtener los atributos identificadores $\mathcal{ID} = \{ID_1, \ldots, ID_p\}$, los atributos cuasi-identificadores $\mathcal{QID} = \{QID_1, \ldots, QID_q\}$, y los atributos confidenciales $\mathcal{C} = \{C_1, \ldots, C_r\}$. El conjunto $\mathcal{ID}$ debería contener los atributos identificadores que aparecen en al menos uno de los documentos; por ejemplo, $\mathcal{ID}$ contendrá “Num Pasaporte” si al menos uno de los documentos contiene el número de pasaporte (incluso si los otros documentos no lo contienen). De forma similar, el conjunto $\mathcal{QID}$ debería contener los atributos cuasi-identificadores que aparecen en al menos un documento, y el conjunto $\mathcal{C}$, los atributos confidenciales que aparecen en al menos un documento.

Una vez determinados estos conjuntos, la colección de documentos puede considerarse como un fichero *estructurado* con registros $D_1, \ldots D_n$ y atributos que son elementos de $\mathcal{ID} \cup \mathcal{QID} \cup \mathcal{C}$. Obviamente, este fichero estructurado probablemente sea escaso, ya que no todos los atributos toman valores en todos los documentos. Para anonimizar este fichero, procedemos como en el caso de ficheros estructurados. Los valores de los atributos en $\mathcal{ID}$ deberían suprimirse de todos los registros/documentos y se deberían enmascarar los atributos en $\mathcal{QID}$ y/o $\mathcal{C}$. Dependiendo del tipo de enmascaramiento usado, puede ser necesario tratar primero los atributos que falten en algunos documentos; una posibilidad es imputarlos usando síntesis parcial [5], [10].

De esta manera, el problema de anonimizar datos no estructurados se reduce a localizar las apariciones de los atributos relevantes para la privacidad en la colección de documentos y anonimizar el conjunto de datos estructurado resultante. Podemos abordar la tarea de localizar las particiones de atributos creando varios modelos de aprendizaje automático, cada uno destinado a reconocer un tipo diferente de entidad nominada. Por ejemplo, un primer modelo para reconocer atributos identificadores (tales como número de pasaporte, número de la seguridad social, etc.), otro para reconocer atributos cuasi-identificadores (tales como localización, fecha de nacimiento, edad, código postal, etc.), y otro más para reconocer atributos confidenciales (como nombres de enfermedades).

IV-B. Prueba de concepto

Como prueba de concepto, nos centramos en localizar datos confidenciales en diagnósticos médicos. Proponemos un modelo basado en CRFs para extraer los nombres de enfermedades de una historia clínica. Dado un texto, este modelo predice una secuencia de etiquetas IOB2.

Una vez predicha la secuencia de etiquetas IOB2 para cada parte de la historia clínica, podemos interpretar esta secuencia de etiquetas y extraer la entidad “enfermedad”. Por ejemplo, si tenemos la frase “La retinopatía fue evaluada por oftalmoscopia” y la secuencia de etiquetas IOB2 {B-DIS, O, O, O}, nos movemos por la secuencia y cada palabra correspondiente a una etiqueta B-DIS se considera el principio de una entidad enfermedad y cada palabra correspondiente a una etiqueta I-DIS se considera dentro de una entidad enfermedad. Por lo tanto, una palabra B-DIS junto con las siguientes palabras I-DIS forman una entidad enfermedad. De hecho, las etiquetas B-DIS e I-DIS tienen la misma finalidad, pero B-DIS se encarga de distinguir dos entidades enfermedad consecutivas.

Tabla I
Extracción de características.

Característica	Descripción
Raíz de la palabra	Por ejemplo, la raíz de “enfermedad” es “enfermo”. Extraemos las raíces usando “SnowballStemmer”, de la librería nltk [2].
Longitud de palabra	Cantidad de letras de la palabra.
Forma de la palabra	Forma de la palabra, la cual puede ser ’minúsculas’, ’mayúsculas’, ’primera mayúscula’, ’combinada’.
Tipo de palabra	Tipo al que pertenece la palabra (nombre, adverbio, adjetivo, etc.). Usamos el marcador “Stanford POS” para extraer esta característica [26].

La Figura 1 muestra la estructura del modelo propuesto para reconocer nombres de enfermedades. Consiste en tres pasos:

- El primer paso es el análisis, que divide una frase en fichas (*tokens*).
- El segundo paso es el extractor de características; en este paso, usamos una ventana de tres palabras (palabra actual, previa y siguiente) y extraemos las características de esas palabras. La Tabla I describe las características que consideramos.
- El tercer paso utiliza un modelo CRF, que toma las características del segundo paso y produce una secuencia de etiquetas para la frase completa.

V. Resultados Experimentales

En esta sección describimos los resultados experimentales de la prueba de concepto mencionada anteriormente. Hemos programado los experimentos en Python y hemos usado “sklearn-crfsuite” para CRF [14] y “SnowballStemmer” para la extracción de raíces de palabras [2].

V-A. Datos

En nuestros experimentos, hemos utilizado textos médicos etiquetados para estudiar la relación entre enfermedades y tratamientos. Estos archivos fueron obtenidos de MEDLINE 2001 usando los 100 primeros títulos y los 40 primeros resúmenes de 59 archivos medline01n*.xml, disponibles en [21].

Estos datos contienen 3645 frases etiquetadas. Las etiquetas son: “DISONLY”, “TREATONLY”, “TREAT PREV”, “DIS PREV”, “TREAT SIDE EFF”, “DIS SIDE EFF”, “DIS VAG”, “TREAT VAG”, “TREAT NO” y “DIS NO”. Como únicamente estábamos interesados en las enfermedades, mantuvimos sólo las 629 frases con la etiqueta “DISONLY”.

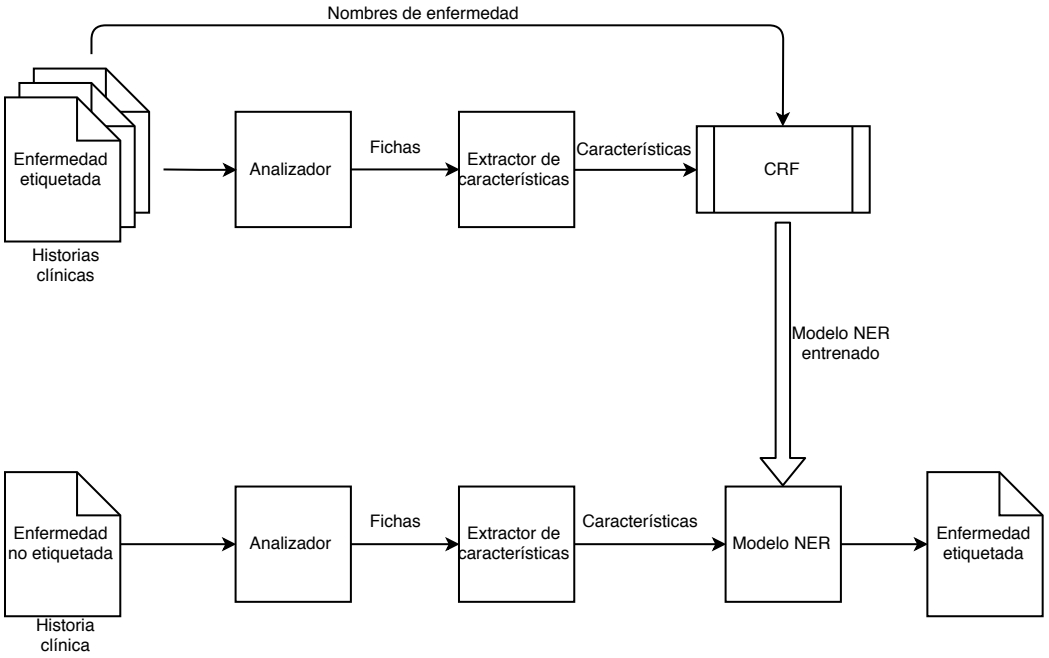

Figura 1. Arquitectura del marcador de reconocimiento de entidades nominadas

V-B. Métricas de evaluación

Usamos tres métricas para evaluar el rendimiento del modelo propuesto para el reconocimiento de enfermedades:

- *Precisión*. Número de enfermedades identificadas correctamente por el clasificador dividido por el número total de enfermedades identificadas:

$$Precision = \frac{|S \cap T|}{|S|},$$

donde S es el conjunto de enfermedades identificadas por el clasificador y T es el conjunto de enfermedades correctas de acuerdo con el fichero original.

- *Recuerdo*. Número de enfermedades identificadas correctamente por el clasificador dividido por el número de enfermedades correctas en el fichero original:

$$Recuerdo = \frac{|S \cap T|}{|T|}.$$

- *F1*. Media armónica de precisión y recuerdo:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recuerdo}}{\text{Precision} + \text{Recuerdo}}.$$

V-C. Resultados y discusión

Realizamos la evaluación experimental en dos fases: entrenamiento del modelo y prueba del modelo. De las 629 muestras de frases etiquetadas, 503 fueron usadas para el entrenamiento del modelo (80 % de las muestras), y 126 fueron usadas para la prueba (20 % de las muestras).

La fase de entrenamiento se llevó a cabo usando 10 veces una validación cruzada, explicada a continuación. Partimos el fichero destinado al entrenamiento en 10 submuestras del mismo tamaño. De las 10 submuestras, una la mantuvimos como datos de validación para probar el modelo durante esta fase, y las nueve restantes se usaron para entrenar el modelo.

Tabla II
EVALUACIÓN DEL MODELO EN EL FICHERO DE PRUEBA A NIVEL DE PALABRA

	Precisión	Recuerdo	Puntuación F1
B-DIS	0.766	0.677	0.719
I-DIS	0.789	0.709	0.747
promedio / total	0.778	0.693	0.733

Aunque muchas de las palabras del fichero fueron etiquetadas como O (fuera de enfermedad), estábamos interesados en palabras etiquetadas como B-DIS (principio de enfermedad) e I-DIS (dentro de enfermedad). Por lo tanto, calculamos la precisión, el recuerdo y la puntuación F1 sólo para B-DIS e I-DIS. Por ejemplo, si tenemos la frase "Evaluación diagnóstica del paciente con presión sanguínea elevada", las partes de la frase son {"Evaluación", "diagnóstica", "del", "paciente", "con", "presión", "sanguínea", "elevada"} y sus correspondientes etiquetas {O, O, O, O, O, B-DIS, I-DIS, I-DIS}. La entidad nominada contiene tres palabras "presión sanguínea elevada". La Tabla II muestra la evaluación de la predicción de etiquetas comparada con las etiquetas correctas a nivel de palabra (de forma separada para cada palabra). En cambio, la Tabla III muestra las mismas métricas de evaluación para entidades enteras. Es decir, en el ejemplo anterior, la Tabla **??** se refiere por separado a las tres palabras "presión", "sanguínea" y "elevada", mientras que la Tabla III se refiere a la entidad "presión sanguínea elevada"; en este último caso, a menos que *las tres palabras por completo* de la entidad fuesen correctamente etiquetadas, la entidad entera se consideraría desclasificada.

De acuerdo con la Tabla III, nuestro modelo funcionó significativamente mejor en precisión que en recuerdo. Es muy posible que el valor de recuerdo pueda incrementarse

Tabla III
EVALUACIÓN DEL MODELO EN EL FICHERO DE PRUEBA A NIVEL DE ENTIDAD

	Precisión	Recuerdo	Puntuación F1
Entidad Enfermedad	0.742	0.660	0.698

usando más muestras de entrenamiento. Consideramos que los resultados anteriores son prometedores, ya que se acercan a los del etiquetado manual de los datos usados.

VI. CONCLUSIONES Y TRABAJO FUTURO

En este trabajo, hemos tratado la anonimización de datos no estructurados en forma de texto. Como prueba de concepto, nos hemos centrado en localizar nombres de enfermedades (es decir, atributos confidenciales) en historias clínicas. Una vez localizados, estos atributos confidenciales pueden protegerse usando técnicas estándar de SDC para datos estructurados.

La principal contribución de este trabajo se basa en la arquitectura del reconocedor de entidades nominadas. El modelo propuesto está basado en aprendizaje automático y es mejor que los enfoques de NER basados en diccionarios. Específicamente, evita el problema que surge cuando las entidades que se quiere localizar no aparecen en el diccionario usado.

Como trabajo futuro, planeamos extender la prueba de concepto presentada para la detección de identificadores y cuasi-identificadores. Esto conllevará un esfuerzo considerable para generar ficheros que incluyan notas para atributos como nombres, localizaciones, edades, etc. Estos ficheros serán usados posteriormente para entrenar los modelos de detección de identificadores y cuasi-identificadores.

AGRADECIMIENTOS

Agradecemos los fondos recibidos de las siguientes organizaciones: Comisión Europea (proyectos H2020-644024 “CLARUS” y H2020-700540 “CANVAS”), Generalitat de Cataluña (Premio ICREA-Acadèmia para J. Domingo-Ferrer) y Gobierno de España (proyectos TIN2014-57364-C2-1-R “SmartGlacis” y TIN2015-70054-REDC). Las opiniones de este documento son de los autores y no reflejan necesariamente las opiniones de UNESCO o cualquiera de los financiadores.

REFERENCIAS

[1] B. Babych y A. Hartley. Improving machine translation quality with automatic named entity recognition. En *Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT Through Other Language Technology Tools: Resources and Tools for Building MT (EAMT ’03)*, pp. 1–8. Association for Computational Linguistics, 2003.

[2] S. Bird, E. Klein y E. Loper. *Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit*. O’Reilly, 2009. El Natural Language Tooking software (nltk) está disponible en: https://www.nltk.org

[3] A. Culotta, R. Bekkerman y A. McCallum. *Extracting Social Networks and Contact Information from Email and the Web.* Computer Science Department Faculty Publication Series, no. 33. University of Massachusetts-Amherst, 2004.

[4] J. Domingo-Ferrer, D. Sánchez y J. Soria-Comas. *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Morgan & Claypool, 2016.

[5] J. Drechsler. *Synthetic Datasets for Statistical Disclosure Control*. LNS 201. Springer, 2011.

[6] A. Ekbal, R. Haque y S. Bandyopadhyay. Bengali part of speech tagging using conditional random field. En *Proceedings of the Seventh International Symposium on Natural Language Processing (SNLP-2007)*, 2007.

[7] J. R. Finkel, T. Grenager y C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. En *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL ’05)*, pp. 363–370. Association for Computational Linguistics, 2005.

[8] EU General Data Protection Regulation, 2016/679. https://gdpr-info.eu

[9] S. Grimes. Structure, models and meaning. *Intelligent Enterprise*, Mar. 2005.

[10] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte-Nordholt, K. Spicer y P.P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.

[11] M. Jabreel, F. Hassan y A. Moreno. Target-dependent sentiment analysis of tweets using bidirectional gated recurrent neural networks. In *Advances in Hybridization of Intelligent Methods*, pp. 39–55. Springer, 2018.

[12] M. A. Khalid, V. Jijkoun y M. De Rijke. The impact of named entity normalization on information retrieval for question answering. En *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval (ECIR’08)*, pp. 705–710. LNCS 4956, Springer, 2008.

[13] B. Kleinberg, M. Mozes, Y. van der Toolen y B. Verschuere. *Netanos - Named Entity-based Text Anonymization for Open Science*. Open Science Framework, Jan. 31, 2018. https://osf.io/w9nhb

[14] M. Korobov. sklearn-crfsuite, 2015. https://sklearn-crfsuite.readthedocs.io/en/latest/

[15] J. Lafferty, A. McCallum y F. C. N. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. En *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, pp. 282-289. ACM, 2001.

[16] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang y H. Xu. A comprehensive study of named entity recognition in Chinese clinical text. *Journal of the American Medical Informatics Association*, 21(5):808–814, 2013.

[17] S. Morwal, N. Jahan y D. Chopra. Named entity recognition using hidden Markov model (HMM). *International Journal on Natural Language Computing*, 1(4):15–23, 2012.

[18] D. Nadeau y S. Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.

[19] I. Neamatullah, M. M. Douglass, L. H. Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark y G. D. Clifford. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):32, 2008.

[20] R. Pérez-Laínez, A. Iglesias y C. de Pablo-Sánchez. Anonimytext: anonymization of unstructured documents. Universidad Carlos III de Madrid, 2009. https://e-archivo.uc3m.es/handle/10016/19829

[21] B. Rosario y M. A. Hearst. Classifying semantic relations in bioscience texts. En *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*. Association for Computational Linguistics 2004. Datos disponibles en: http://biotext.berkeley.edu/dis_treat_data.html

[22] E. F. Sang y J. Veenstra. Representing text chunks. En *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 173–179. Association for Computational Linguistics, 1999.

[23] United Kingdom Data Service. *Text Anonymization Helper Tool*. Último acceso: Mar. 24, 2018. https://bitbucket.org/ukda/ukds.tools.textanonhelper/wiki/Home

[24] B. M. Sundheim. Overview of results of the MUC-6 evaluation. En *Proceedings of the TIPSTER Text Program: Phase II*, pp. 423–442. Association for Computational Linguistics, 1996.

[25] L. Sweeney. Replacing personally-identifying information in medical records, the Scrub system. En *Proceedings of the AMIA Annual Fall Symposium*, p. 333. American Medical Informatics Association, 1996.

[26] K. Toutanova, D. Klein, C. Manning y Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. En *Proceedings of HLT-NAACL 2003*, pages 252-259. 2003.

[27] H. Vico y D. Calegari. Software architecture for document anonymization. *Electron. Notes Theor. Comput. Sci.*, 314(C):83–100, 2015.

[28] G. Zhou y J. Su. Named entity recognition using an HMM-based chunk tagger. En *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL’02)*, pp. 473–480. Association for Computational Linguistics, 2002.