(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 969 343**

(51) Int. Cl.:

G06F 21/62 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **23.12.2020 PCT/EP2020/087816**

(87) Fecha y número de publicación internacional: **01.07.2021 WO21130337**

(96) Fecha de presentación y número de la solicitud europea: **23.12.2020 E 20833918 (4)**

(97) Fecha y número de publicación de la concesión europea: **18.10.2023 EP 4081924**

(54) Título: **Conservación de la privacidad en una base de datos consultable construida a partir de textos no estructurados**

(30) Prioridad:

23.12.2019 EP 19383189

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
17.05.2024

(73) Titular/es:

MEDSAVANA S.L. (100.0%)
Calle Larra 12, bajo izquierda
28004 Madrid, ES

(72) Inventor/es:

TELLO GUIJARRO, JORGE;
LUMBRERAS SANCHO, SARA;
FERNÁNDEZ GARCÍA, JAVIER y
MARCHESSEAU, STEPHANIE

(74) Agente/Representante:

BERTRÁN VALLS, Silvia

DESCRIPCIÓN

Conservación de la privacidad en una base de datos consultable construida a partir de textos no estructurados

Campo técnico

La presente divulgación se refiere a un método implementado por ordenador de generación de una base de datos consultable originada a partir de textos no estructurados en el que se conserva la privacidad individual. La presente divulgación también se refiere a un producto de programa informático para generar la base de datos y a un aparato de procesamiento de datos para generar la base de datos.

Antecedentes

El análisis de datos recopilados en un campo es una herramienta útil para identificar tendencias y patrones que pueden usarse para formar una comprensión más precisa de campo. Pueden recopilarse y analizarse datos, almacenándose los datos analizados en una base de datos consultable. Puede proporcionarse información agregada por la base de datos consultable con el fin de proporcionar una herramienta para análisis estadístico de los datos para identificar las tendencias y patrones en el campo.

Para bases de datos consultables basadas en datos referentes a individuos en las que las consultas automatizadas son posibles, un atacante puede descubrir información sobre un individuo mediante una combinación de diferentes consultas agregadas. Más específicamente, un atacante que realiza una combinación de diferentes consultas en la base de datos puede deducir información específica sobre un individuo, por ejemplo sumando y restando los resultados de las consultas.

Como tal, un atacante todavía puede usar bases de datos consultables que proporcionan datos agregados para identificar información referente a un individuo.

Una medida existente para ayudar a mantener la confidencialidad de datos en una base de datos consultable es mediante el uso de privacidad diferencial. El principio de la privacidad diferencial es proporcionar a cada individuo aproximadamente la misma cantidad de privacidad que resultaría de la eliminación de sus datos. Esta privacidad se proporciona añadiendo ruido aleatorio, usando un algoritmo de generación de ruido, a las consultas lo cual hace imposible reconstruir información sobre los individuos. Hay muchas técnicas de este tipo disponibles, tales como el mecanismo de Laplace u otros mecanismos similares.

Sin embargo, tales técnicas todavía plantean un problema de seguridad ya que la generación de ruido es determinista; es decir, el ruido generado se genera artificialmente a partir de los datos en la base de datos. Por tanto, un atacante puede realizar un alto número de consultas en la base de datos lo cual revela el algoritmo de generación de ruido usado. Dicho de otro modo, el ruido puede someterse a ingeniería inversa a partir de un número indefinido de consultas realizadas en la base de datos. Una vez que el atacante conoce características del algoritmo de generación de ruido, entonces esto puede usarse para descubrir información adicional referente a los datos en la base de datos (dado que el ruido generado depende de los datos subyacentes). Alternativamente, un atacante puede descubrir información referente al algoritmo de generación de ruido a partir de los programadores que participan en la creación del algoritmo de generación de ruido.

Por tanto, existe la necesidad de proporcionar una cantidad similar de privacidad a individuos de una manera más segura (es decir, usando una técnica de potenciación de la privacidad que sea menos propensa al ataque en sí misma).

El documento US 2014/278409 A1 divulga un método para anonimizar textos no censurados modificando vectores de características asociados con los textos no censurados.

Sumario

Según un primer aspecto, se proporciona un método implementado por ordenador de generación de una base de datos consultable, que comprende: recibir un corpus de documentos en texto libre que contienen datos confidenciales, estando los documentos en texto libre relacionados con el mismo dominio; asignar, por un sistema de procesamiento de lenguaje natural (NLP) entrenado, una o más entidades nombradas resumidas a cada documento en texto libre en el corpus; y almacenar las entidades nombradas resumidas de cada documento en texto libre asignadas por el sistema de NLP en una base de datos consultable configurada para proporcionar información agregada referente a las entidades nombradas; en el que el sistema de NLP está configurado de tal manera que las entidades nombradas resumidas se reconocen y desambiguan con una precisión de entre 0,75 y menos de 1 y una exhaustividad de entre 0,75 y menos de 1, y en el que la razón de precisión y exhaustividad es de entre 0,7 y 1,3; en el que la base de datos consultable está libre de la adición de ruido artificial por un algoritmo de generación de ruido artificial. Mantener los valores de precisión y exhaustividad por debajo de 1 añade privacidad a la base de datos consultable haciendo que sea menos probable que ataques a partir de un alto número de consultas agregadas sean satisfactorios. Mantener la precisión y exhaustividad por encima de 0,75 garantiza un nivel aceptable de exactitud de los resultados agregados.

Ventajosamente, los errores en la precisión y exhaustividad surgen de los errores de lectura del propio sistema de NLP, en vez de generarse artificialmente. Por tanto, el método da como resultado una base de datos consultable que proporciona información agregada exacta al tiempo que garantiza la privacidad individual, en el que el procedimiento de generación de ruido no puede descubrirse por un atacante debido a que la generación de ruido surge de manera natural a partir del procedimiento de entrenamiento del sistema de NLP, y a partir de las ambigüedades inherentes a la comunicación de lenguaje natural, en vez de insertarse artificialmente. Dicho de otro modo, el método da como resultado una base de datos consultable más segura ya que el propio método de generación de ruido no es artificial. Es decir, el método da como resultado una base de datos consultable más segura ya que el propio método de generación de ruido no es aleatorio, ni se diseña de manera artificial de un modo determinista. Garantizando que la privacidad surge a partir de errores probabilísticos en el reconocimiento y la desambiguación, en vez de a partir de un algoritmo de generación de ruido artificial, un atacante no puede aprender el tipo de generación de ruido a partir de varias consultas de base de datos. La seguridad aumentada surge independientemente del contenido de los datos en la base de datos ya que se basa en la precisión y exhaustividad del sistema de NLP.

Por tanto, puede considerarse que el método garantiza una “privacidad natural” ya que la privacidad surge del error de lectura del sistema de NLP, en vez de surgir a partir de ruido artificial inyectado en la base de datos, como en el caso de métodos tradicionales de privacidad diferencial.

La precisión puede ser de entre 0,75 y 0,95. La exhaustividad puede ser de entre 0,75 y 0,95. Más preferiblemente, la precisión puede ser de entre 0,85 y 0,95 y la exhaustividad puede ser de entre 0,85 y 0,95. Se ha encontrado que estos valores garantizan un nivel de privacidad similar a los métodos tradicionales de privacidad diferencial. La razón de precisión y exhaustividad puede ser de entre 0,8 y 1,2, y preferiblemente de 0,9 y 1,1. El número de documentos en texto libre recibidos puede ser de al menos 49, preferiblemente al menos 1000 y más preferiblemente al menos 39.000.

El sistema de NLP puede comprender uno o más algoritmos de aprendizaje automático. El método puede comprender las etapas de entrenar cada uno del uno o más algoritmos de aprendizaje automático mediante: seleccionar uno o más subconjuntos de documentos en texto libre en el dominio; asignar la una o más entidades nombradas resumidas a los documentos en el uno o más subconjuntos para formar uno o más conjuntos de entrenamiento; entrenar el uno o más algoritmos de aprendizaje automático usando el uno o más conjuntos de entrenamiento; seleccionar un segundo subconjunto de documentos en texto libre en el dominio; introducir el segundo subconjunto del corpus de documentos en texto libre en el sistema de NLP; evaluar si el sistema de NLP reconoce y desambigua las entidades nombradas resumidas con una precisión de entre 0,75 y menos de 1 y una exhaustividad de entre 0,75 y menos de 1, y en el que la razón de precisión y exhaustividad es de entre 0,7 y 1,3; y si no es así, volver a entrenar el uno o más algoritmos de aprendizaje automático de tal manera que la precisión, exhaustividad y razón de precisión y exhaustividad están dentro de los intervalos requeridos.

El entrenamiento puede realizarse de manera iterativa. Este método proporciona un procedimiento de entrenamiento dinámico para formar un sistema de NLP, en el que la privacidad diferencial se garantiza mediante el error de lectura del uno o más algoritmos de aprendizaje automático.

Si tras la evaluación se requiere reducir la precisión, entonces el uno o más algoritmos de aprendizaje automático pueden volver a entrenarse proporcionando datos de entrenamiento que contienen entidades nombradas resumidas asignadas a cadenas incorrectas de texto. Además, si tras la evaluación se requiere reducir la exhaustividad, entonces el uno o más algoritmos de aprendizaje automático vuelven a entrenarse proporcionando datos de entrenamiento que contienen texto relacionado con una entidad nombrada resumida a la que no se le ha asignado esa entidad nombrada resumida y/o a la que se le ha asignado una entidad nombrada resumida diferente. El sistema de NLP puede entrenarse deliberadamente para tener un rendimiento peor con el fin de garantizar el nivel requerido de privacidad.

Los conjuntos de entrenamiento pueden generarse manualmente por uno o más usuarios o usando un segundo sistema de NLP que tiene una precisión y exhaustividad mayores de 0,85 y preferiblemente mayores de 0,95. La precisión y exhaustividad del sistema de NLP pueden evaluarse manualmente pro un usuario o comparando la salida del sistema de NLP con la de un segundo sistema de NLP que tiene una precisión y exhaustividad mayores de 0,85 y preferiblemente mayores de 0,95.

El sistema de NLP puede comprender uno o más algoritmos basados en reglas. El uso de algoritmos basados en reglas con texto no estructurado libre puede ser una fuente de generación de ruido. Por ejemplo, los algoritmos basados en reglas pueden no ser capaces de detectar erratas en el texto, dando como resultado un número superior de falsos negativos. Por tanto, los falsos negativos pueden considerarse como verdaderamente aleatorios.

Los documentos en texto libre pueden ser historias clínicas y las entidades nombradas resumidas pueden comprender información del paciente y terminología médica.

La entidad resumida nombrada y un término de desambiguación asociado pueden almacenarse en la base de datos. Los documentos en texto libre pueden ser historias clínicas y el sistema de NLP puede entrenarse para asignar uno o más de los siguientes términos de desambiguación a las entidades nombradas resumidas: información del paciente,

antecedentes médicos, antecedentes médicos familiares, antecedentes farmacéuticos, antecedentes de tratamiento, síntomas, resultados de pruebas, evoluciones y notas. Alternativamente, los documentos en texto libre pueden ser registros de seguro y el sistema de NLP puede entrenarse para asignar uno o más de los siguientes términos de desambiguación a las entidades nombradas resumidas: cobertura por pérdida o daño, riesgo derivado, riesgo, contenido relacionado legal, figura jurídica, acción de póliza, acontecimiento en el tiempo, requisito legal.

Según un segundo aspecto, se proporciona un producto de programa informático que comprende instrucciones que, cuando se ejecutan por un ordenador, hacen que el ordenador: reciba un corpus de documentos en texto libre que contienen datos confidenciales, estando los documentos en texto libre relacionados con el mismo dominio; asigne, por el sistema de procesamiento de lenguaje natural (NLP) entrenado, una o más entidades nombradas resumidas a cada documento en texto libre en el corpus; y almacene las entidades nombradas resumidas de cada documento en texto libre asignadas por el sistema de NLP en una base de datos consultable configurada para proporcionar datos agregados referentes a las entidades nombradas; en el que el sistema de NLP está configurado de tal manera que las entidades nombradas resumidas se reconocen y desambiguan con una precisión de entre 0,75 y menos de 1 y una exhaustividad de entre 0,75 y menos de 1, y en el que la razón de precisión y exhaustividad es de entre 0,7 y 1,3; en el que la base de datos consultable está libre de la adición de ruido artificial por un algoritmo de generación de ruido artificial.

Según un tercer aspecto, se proporciona un aparato de procesamiento de datos para generar una base de datos consultable, que comprende un sistema de procesamiento de lenguaje natural (NLP) entrenado y configurado para: recibir un corpus de documentos en texto libre que contienen datos confidenciales, estando los documentos en texto libre relacionados con el mismo dominio; asignar, por el sistema de procesamiento de lenguaje natural (NLP) entrenado, una o más entidades nombradas resumidas a cada documento en texto libre en el corpus; y almacenar las entidades nombradas resumidas de cada documento en texto libre asignadas por el sistema de NLP en una base de datos consultable configurada para proporcionar datos agregados referentes a las entidades nombradas; en el que el sistema de NLP está configurado de tal manera que las entidades nombradas resumidas se reconocen y desambiguan con una precisión de entre 0,75 y menos de 1 y una exhaustividad de entre 0,75 y menos de 1, y en el que la razón de precisión y exhaustividad es de entre 0,7 y 1,3; en el que la base de datos consultable está libre de la adición de ruido artificial por un algoritmo de generación de ruido artificial.

Breve descripción de los dibujos

Para permitir una mejor comprensión de la presente divulgación, y para mostrar cómo puede llevarse a cabo la misma, ahora se hará referencia, únicamente a modo de ejemplo, a los dibujos adjuntos, en los que:

la figura 1 muestra un diagrama de bloques de un sistema informático a modo de ejemplo para generar una base de datos consultable según la presente divulgación;

la figura 2 muestra un método para generar una base de datos consultable según la presente divulgación;

la figura 3 muestra un diagrama de bloques de un sistema de NLP según la presente divulgación;

la figura 4 muestra un diagrama de flujo que ilustra cómo puede entrenarse un sistema de NLP según la presente divulgación;

las figuras 5A a 5E ilustran la salida del sistema de NLP en el ejemplo 1; y

las figuras 6A a 6E ilustran la salida del sistema de NLP en el ejemplo 2.

Descripción detallada

A lo largo de esta divulgación, el término “dominio” puede referirse a un concepto o grupo de conceptos que se refieren a una única disciplina o campo. El dominio puede representarse por una única ontología específica de dominio, que puede ser una ontología convencional en la industria, una ontología generada no convencional en la industria o una ontología generada por usuario. Los ejemplos de un dominio incluyen medicina y seguros.

A lo largo de esta divulgación, el término “precisión” de un sistema se refiere a la fracción de entidades nombradas resumidas correctamente identificadas y desambiguadas por el sistema en comparación con el número total de entidades nombradas resumidas identificadas por el sistema. El valor se define como TP/(TP+FP), donde TP es el número de verdaderos positivos y FP es el número de falsos positivos.

A lo largo de esta divulgación, el término “exhaustividad” de un sistema se refiere a la fracción de entidades nombradas resumidas correctamente identificadas y desambiguadas por el sistema, en comparación con el número total de entidades nombradas resumidas introducidas en el sistema. El valor se define como TP/(TP+FN), donde TP es el número de verdaderos positivos y FP es el número de falsos positivos.

A lo largo de esta divulgación, el término “verdadero positivo” se refiere a una entidad nombrada resumida correctamente identificada y desambiguada. Dicho de otro modo, se refiere a un resultado que indica que una entidad nombrada resumida está presente cuando realmente está presente. Por ejemplo, en el contexto de un documento, cuando una identidad resumida nombrada en el texto del documento se identifica y desambigua correctamente, esta identidad resumida nombrada se considera un “verdadero positivo”.

A lo largo de esta divulgación, el término “falso positivo” se refiere a una entidad nombrada resumida incorrectamente identificada y desambiguada. Dicho de otro modo, se refiere a un resultado que indica que una entidad nombrada resumida está presente cuando realmente no está presente. Por ejemplo, en el contexto de un documento, cuando una identidad resumida nombrada en el texto del documento se identifica y desambigua incorrectamente, esta identidad resumida nombrada se considera un “falso positivo”.

A lo largo de esta divulgación, el término “verdadero negativo” se refiere a un fragmento de texto correctamente ignorado. Dicho de otro modo, se refiere a un resultado que indica que una identidad nombrada resumida no está presente cuando realmente no está presente. Por ejemplo, en el contexto de un documento, si un fragmento de texto en el documento se ignora correctamente, esto puede considerarse un “verdadero negativo”.

A lo largo de esta divulgación, el término “falso negativo” se refiere a una identidad nombrada resumida que se ignora incorrectamente. Dicho de otro modo, se refiere a un resultado que indica que una identidad nombrada resumida no está presente cuando realmente está presente. Por ejemplo, en el contexto de un documento, cuando se encuentra que una identidad nombrada resumida no está presente cuando realmente está presente en el documento, esto puede considerarse un “falso negativo”.

Para una entidad nombrada resumida dada, un documento puede considerarse un “verdadero positivo” si la entidad nombrada resumida se identifica y desambigua correctamente para ese documento. De manera similar, un documento puede considerarse un “falso positivo” si la entidad nombrada resumida particular se identifica y desambigua incorrectamente para ese documento. Además, un documento puede considerarse un “verdadero negativo” si la entidad nombrada resumida particular no se identifica y desambigua para el documento y el texto de documento no contiene la entidad nombrada resumida. Finalmente, un documento puede considerarse un “falso negativo” si la entidad nombrada resumida particular no se identifica y desambigua para el documento y el texto de documento sí contiene la entidad nombrada resumida.

A lo largo de esta divulgación, el término “texto libre” se refiere a texto que no está organizado de una manera predefinida, en contraposición a texto estructurado que puede ser texto almacenado, por ejemplo, en forma de campos.

A lo largo de esta divulgación, el término “sistema de procesamiento de lenguaje natural” o “sistema de NLP” se refiere a cualquier algoritmo o grupo de algoritmos configurado para procesar datos de lenguaje natural. Un sistema de NLP puede comprender un único algoritmo, tal como un algoritmo de aprendizaje automático o algoritmo basado en reglas o puede comprender una colección de algoritmos configurados para ejecutarse en paralelo o en serie. Por ejemplo, un sistema de NLP puede comprender un algoritmo configurado para identificar y desambiguar un conjunto de entidades nombradas resumidas, y un algoritmo adicional para identificar y desambiguar un segundo conjunto de entidades nombradas resumidas. El sistema de NLP puede comprender además, por ejemplo, un algoritmo basado en reglas o de aprendizaje automático para identificar abreviaturas en el texto libre y/o un algoritmo para identificar un contexto para cada entidad nombrada resumida identificada.

Los ejemplos de algoritmos basados en reglas que pueden usarse en el sistema de NLP incluyen:

- árboles de decisión diseñados manualmente basados en expresiones regulares, para identificar partes de texto fácilmente predecibles (número de seguridad social, resultados cuantitativos de pruebas, dosificación de medicamentos, …); y

- coincidencia de palabras a partir de un diccionario de palabras tales como una ontología.

Los ejemplos de algoritmos de aprendizaje automático que pueden usarse en el sistema de NLP incluyen:

- algoritmos de clasificación (tales como aprendizaje profundo, bosque aleatorio, árboles de decisión, K-NN o similares);

- métodos de regresión (por ejemplo árboles de decisión, regresión logística, máquinas de vectores de soporte o similares);

- análisis de agrupación (tal como medias k, detección de anomalía o similares);

- reducción de características (tal como análisis de componentes principales, análisis discriminante lineal o similares); y

- extracción de características (tal como análisis de componentes independientes, análisis de covarianza, clasificación recursiva o similares).

Los métodos de aprendizaje de los algoritmos de aprendizaje automático pueden estar supervisados, semisupervisados o no supervisados.

A lo largo de esta divulgación, el término "entidad nombrada resumida" puede referirse a cualquier término que pertenece a la ontología específica de dominio del dominio.

Tal como se comentó anteriormente, cuando pueden realizarse consultas automatizadas en una base de datos que proporciona información agregada referente a los datos almacenados en la base de datos, existe el riesgo de que un atacante descubra información no divulgada, tal como datos personales, por medio de una combinación de un alto número de consultas agregadas. La privacidad diferencial es una técnica bien conocida de potenciación de la privacidad, que se basa en el principio de que la privacidad de una persona no puede verse comprometida si un observador que ve la información agregada de la base de datos no puede saber si se usaron los datos de un individuo particular. Esto se garantiza mediante la adición de ruido aleatorio a la base de datos (por ejemplo, mediante la adición de falsos positivos o falsos negativos a los resultados agregados), lo cual inhibe la capacidad del observador para reconstruir la información sobre cualquier individuo contenida en la base de datos a partir de la combinación de consultas.

En los métodos divulgados en el presente documento, hay falsos positivos y falsos negativos presentes de manera similar en los datos guardados en la base de datos, lo que significa que se garantiza la privacidad de la misma manera que para privacidad diferencial.

La justificación formal para los métodos divulgados en el presente documento puede resumirse de la siguiente manera.

El uso de un sistema de procesamiento de lenguaje natural (NLP) para identificar (es decir, reconocer y desambiguar) identidades resumidas nombradas que aparecen en un texto es propenso a errores. El sistema de NLP puede cometer dos tipos de errores: puede detectar una entidad nombrada resumida cuando no aparece (un falso positivo) o, a la inversa, puede no lograr identificar una entidad nombrada resumida que aparece en el texto (un falso negativo).

La exactitud del sistema de NLP se determina mediante estos dos tipos de errores y puede resumirse mediante dos parámetros, precisión y exhaustividad, que se definen de la siguiente manera:

$$Precisión = \frac{TP}{TP + FP}$$

$$Exhaustividad = \frac{TP}{TP + FN}$$

Donde TP es el número de verdaderos positivos, FP es el número de falsos positivos y FN es el número de falsos negativos. Por ejemplo, para una entidad nombrada resumida particular, TP puede ser el número de documentos verdaderos positivos, FP el número de documentos falsos positivos y FN el número de documentos falsos negativos.

La frecuencia verdadera (P*) (por ejemplo, prevalencia) en la que aparece un concepto nombrado resumido en un texto es la suma del número de verdaderos positivos y falsos negativos (es decir el número de veces que aparece realmente la entidad nombrada resumida en los textos, o el número de documentos en los que aparece realmente una entidad nombrada resumida particular). Esto puede resumirse de la siguiente manera:

$$P^* = TP + FN = P * \frac{Precisión}{Exhaustividad} = (TP + FP) * \frac{Precisión}{Exhaustividad}$$

Donde P es el número total de veces que se identifica la entidad nombrada resumida por el sistema de NLP (es decir TP + FP). Más particularmente, P es el número total de documentos en los que se identifica la entidad nombrada resumida por el sistema de NLP. Por tanto, se desprende que si la precisión y la exhaustividad adoptan valores similares, la frecuencia de entidades nombradas resumidas identificadas por el sistema de NLP es un valor similar al número real de veces que aparece la entidad nombrada resumida en los textos. Dicho de otro modo, precisión/exhaustividad ≈ 1.

Tal como se explica en más detalle a continuación, no se necesita que la razón de precisión y exhaustividad sea exactamente 1. La razón de estos valores puede adoptar cualquier valor dentro de un intervalo de valores alrededor de 1, dependiendo de la exactitud requerida de los resultados. Por ejemplo, en algunas aplicaciones puede ser aceptable una razón de entre aproximadamente 0,7 y 1,3. En aplicaciones en las que se requiere una exactitud superior, el límite inferior de la razón puede ser de 0,8, 0,9 ó 0,95, y el límite superior de la razón puede ser de 1,2, 1,1 ó 1,05 (contemplándose cualquier combinación de los valores de límite superior y valores de límite inferior).

Dado que cualquier estimación de precisión y exhaustividad es propensa a errores (debido a que el análisis de precisión y exhaustividad se basa en un subconjunto de un corpus de textos), diferirán del valor verdadero en los errores ε1 y ε2:

$$\frac{Precisión\ real}{Exhaustividad\ real} = \frac{Precisión\ medida + \varepsilon 1}{Exhaustividad\ medida + \varepsilon 2}$$

Por tanto, los valores medidos de precisión y exhaustividad serán exactos si los errores ε1 y ε2 son bajos en comparación con la precisión y exhaustividad medidas. Por tanto, la precisión y exhaustividad deben ser relativamente altas (es decir, positivas y razonablemente próximas a uno, por ejemplo al menos 0,70 y más preferiblemente al menos 0,75). También se apreciará por el experto en la técnica que debe seleccionarse un subconjunto de tamaño adecuado del corpus que sea estadísticamente representativo del corpus, de tal manera que los errores ε1 y ε2 sean suficientemente bajos, tal como se conoce en la técnica.

La probabilidad de errores cuando se identifica una entidad nombrada resumida puede resumirse de la siguiente manera:

$$Error_1 = \frac{FP}{n} = \frac{TP}{n} \cdot \frac{1 - Precisión}{Precisión}$$

$$Error_2 = \frac{FN}{n} = \frac{TP}{n} \cdot \frac{1 - Exhaustividad}{Exhaustividad}$$

Donde *FP* y *FN* son las tasas de falsos positivos y negativos, respectivamente, *TP* es el número de verdaderos positivos y *n* es el número total de documentos.

Para probabilidades de error aceptables, los límites para precisión y exhaustividad pueden derivarse basándose en la proporción de positivos (la proporción promedio de textos en los que aparece el concepto particular, *TP/n*). Por tanto, los valores objetivo para la precisión y exhaustividad basándose en los valores objetivo para los errores en función de la estimación de verdaderos positivos se resumen de la siguiente manera:

$$Precisión\ objetivo = \frac{1}{\frac{Error\ objetivo_1}{TP/n} + 1}$$

$$Exhaustividad\ objetivo = \frac{1}{\frac{Error\ objetivo_2}{TP/n} + 1}$$

Para errores objetivo del 5% y una razón de positivos del 50%, los valores objetivo de precisión y exhaustividad son del 90,90%. Para un error objetivo del 10%, los valores objetivo de precisión y exhaustividad son del 83,33%.

A partir de las definiciones de P, P* y precisión y exhaustividad facilitadas anteriormente, puede derivarse que el número total de positivos en el corpus (por ejemplo, el número de documentos con la entidad nombrada resumida detectada) es:

$$P = (TP + FP) = P^* * \left(1 + \frac{Exhaustividad - Precisión}{Precisión}\right) = P^* * \left(\frac{Exhaustividad}{Precisión}\right)$$

Por tanto, puede suponerse que el número promedio de positivos por documento (es decir, probabilidad de que en un documento se haya detectado la entidad nombrada resumida) es una distribución de Bernoulli con una probabilidad:

$$p = \frac{P^*}{n} * \left(1 + \frac{Exhaustividad - Precisión}{Precisión}\right)$$

Donde n es el número total de documentos. Por tanto, agregando resultados a lo largo del corpus puede modelarse la distribución como un binomio, en el que los parámetros del binomio son:

$$n_{BINOMIO} = n$$

$$p_{BINOMIO} = p$$

El binomio puede considerarse como una distribución normal para n suficientemente grande, por ejemplo si n*p>5 y n(1-p)>5, con una distribución normal resumida como:

$$N(np, \sqrt{np(1-p)})$$

Por tanto, la proporción de documentos en el corpus en los que se ha identificado una entidad nombrada resumida puede modelarse por medio de una distribución normal de los siguientes parámetros:

$$\mu = np = P^* \cdot \left(\frac{Exhaustividad}{Precisión}\right)$$

$$\sigma = \sqrt{\frac{1}{n} \cdot \left(P^* \cdot \frac{Exhaustividad}{Precisión}\right)\left(1 - P^* \cdot \frac{Exhaustividad}{Precisión}\right)}$$

Sabiendo esto, la frecuencia de concepto tal como se lee por el sistema de NLP estará contenida en los intervalos de confianza $[\mu - z \cdot \sigma, \mu + z \cdot \sigma]$, donde z es un parámetro (por ejemplo, $z = 2$ producirá $Prob(P \in [\mu - z \cdot \sigma, \mu + z \cdot \sigma])$=0,95.

Con el fin de tener una estimación centrada alrededor del valor verdadero de P^* (es decir, con el fin de hacer que la desviación estándar adopte un valor bajo), es necesario tener valores similares de precisión y exhaustividad:

$$\frac{Precisión}{Exhaustividad} \approx 1$$

Esto significa que, si se considera que un error o intervalo de confianza del 10% (medido como proporción de prevalencia verdadera P*) es aceptable, la razón entre precisión y exhaustividad será:

$$0{,}90 < \frac{Precisión}{Exhaustividad} < 1{,}10$$

En diferentes aplicaciones, pueden ser aceptables diferentes valores de error, en cuyo caso la razón de precisión y exhaustividad puede adoptar un intervalo más amplio (o más estrecho) de valores aceptables.

La mitad de la anchura del intervalo de confianza expresada como proporción de su promedio será:

$$hw = \frac{z \cdot \sigma}{\mu} = z \cdot \sqrt{\frac{1}{n} \cdot \frac{\left(1 - P^* \cdot \frac{Exhaustividad}{Precisión}\right)}{\left(P^* \cdot \frac{Exhaustividad}{Precisión}\right)}}$$

$$n = \frac{z^2}{hw^2} \cdot \frac{\left(1 - P^* \cdot \frac{Exhaustividad}{Precisión}\right)}{\left(P^* \cdot \frac{Exhaustividad}{Precisión}\right)}$$

Para una mitad de la anchura del 10% de la media para un nivel de confianza del 95% (z = 2), suponiendo una proporción del 50% de documentos en los que aparece la entidad leída (es decir P*/n) y una razón $\frac{Precisión}{Exhaustividad}$ = 0,90, se obtiene un número necesario de documentos n = 49. El parámetro más sensible es la prevalencia. En condiciones en las que P^*/n = 10%, el número de documentos será de n = 39.600. En resumen, el número de documentos en el corpus puede seleccionarse basándose en las circunstancias particular de la aplicación (es decir, la naturaleza de las entidades que están leyéndose y el nivel de confianza deseado).

Los mecanismos de privacidad diferencial limitan la cantidad de información sobre un sujeto particular que puede extraerse a partir de diferentes consultas en la base de datos. Esta sección calcula el impacto de la precisión y exhaustividad en un ejemplo específico estilizado. El ejemplo es únicamente a modo de ejemplo, para representar de manera resumida los beneficios del ruido introducido en los métodos de NLP divulgados en el presente documento.

Se supone que el atacante conoce el valor de $f + 1$ variables diferentes para el sujeto objetivo y que la frecuencia promedio de estos términos es p. La base de datos contiene N sujetos en total. El número de sujetos que aparecerán en una consulta que contiene f filtros es:

$$N \cdot p^f$$

El atacante acumula dos consultas en las que la diferencia en los sujetos filtrados es lo más baja posible, de manera ideal uno. Esto significa:

$$N \cdot p^f - N \cdot p^{f+1} \approx 1$$

El número de variables que necesita conocer previamente el atacante sobre el sujeto es mayor para frecuencias más grandes y menor para las más pequeñas. Por ejemplo, si los términos que conoce el atacante sobre el sujeto son relativamente poco frecuentes (por ejemplo, con $p = 5\%$) y no están correlacionados, con N = 1.000.000 sujetos será suficiente conocer 6 campos. Si los términos son más frecuentes, necesitará conocer que el sujeto tiene una lectura positiva para 20 variables.

Los errores en el procedimiento de lectura de NLP hacen más difícil usar la información para construir el ataque. Esto puede conceptualizarse de la siguiente manera:

La probabilidad de que un verdadero positivo sigua siéndolo (es decir, la probabilidad de que el sujeto objetivo siga estando en la consulta después de introducirse ruido) es:

$$Exhaustividad^f$$

La probabilidad de que la consulta no obtenga nuevos sujetos (a partir de falsos positivos) es:

$$\left(\frac{1-p}{1-p+p\cdot\left(\frac{1}{Precisión}-1\right)}\right)^f$$

Esto significa que la probabilidad de que las dos consultas todavía contengan el sujeto objetivo y tengan un diferencial de un sujeto es:

$$\left(\frac{Exhaustividad\,(1-p)}{1-p+p\cdot\left(\frac{1}{Precisión}-1\right)}\right)^f$$

En los siguientes ejemplos, esto significa que, con $Precisión = Exhaustividad = 0{,}85$ (lo cual, tal como se expuso anteriormente, mantiene la exactitud de resultados agregados), el ataque en el caso para $p = 5\%$, el ataque sólo necesitará ser satisfactorio en el 42% de los casos.

En el otro caso, con $p = 50\%$, el ataque será satisfactorio con una probabilidad de tan sólo el 0,2%.

Si $Precisión = Exhaustividad = 0{,}75$, los ataques serán satisfactorio con una probabilidad del 22% y 1,78E-5.

Teniendo en cuenta el análisis anterior, se entenderá que un sistema de NLP que tiene al menos una de la precisión y la exhaustividad por debajo de 1 y ambas por encima de 0,75, y que tiene el intervalo requerido de valores de razón, puede usarse para generar una base de datos consultable en la que se requiere privacidad diferencial, en vez de añadiendo ruido artificial a una base de datos como en métodos tradicionales de privacidad diferencial. Los valores particulares para precisión, exhaustividad y la razón entre estos valores pueden variar dependiendo de los requisitos de cualquier aplicación particular. Se ha encontrado que, cuando la precisión y exhaustividad adoptan valores de entre 0,75 y 0,95, y preferiblemente de 0,85 y 0,95, y la razón de precisión con respecto a exhaustividad es de entre 0,7 y 1,3, y preferiblemente entre 0,9 y 1,1, el sistema de NLP conserva la privacidad individual al tiempo que también proporciona resultados agregados exactos. El número de documentos en el corpus es preferiblemente mayor de 49, preferiblemente mayor de 1000 e incluso más preferiblemente mayor de 39.000.

La figura 1 muestra un diagrama de bloques de un sistema 100 informático a modo de ejemplo (es decir, un aparato de procesamiento de datos) para generar una base de datos consultable según la presente divulgación.

El experto apreciará que el sistema 100 informático mostrado en la figura 1 es uno de muchos sistemas que pueden usarse para implementar los métodos divulgados en el presente documento. En otros ejemplos, el sistema informático puede ser un sistema informático distribuido, por ejemplo que comprende uno o más servidores o sistemas informáticos clientes, que usa cualquier técnica informática distribuida conocida. En algunos ejemplos, puede usarse un ordenador de uso general o cualquier otro sistema de procesamiento para implementar los métodos divulgados en el presente documento. Además, el sistema 104 de NLP puede implementarse en software, hardware o cualquier combinación de los mismos para lograr los métodos divulgados en el presente documento.

El sistema 100 informático comprende una memoria 102, un elemento 110 de visualización, un procesador 112, una o más conexiones 114 de red y una o más interfaces 116 de usuario. La memoria 102 comprende un sistema 104 de procesamiento de lenguaje natural (sistema de NLP), uno o más programas 106, uno o más repositorios 108 de datos y una base 109 de datos consultable. Aunque se muestra que el sistema 104 de NLP está almacenado en la memoria 102, parte o la totalidad del mismo puede estar almacenado en otra parte, tal como en otros medios legibles por ordenador (no mostrados). El sistema 104 de NLP y uno o más programas 106 pueden ejecutarse en el procesador

112 para realizar los métodos divulgados en el presente documento. El elemento 110 de visualización y una o más interfaces 116 de usuario (tales como una o más interfaces de entrada) pueden usarse por un usuario para hacer funcionar uno o más programas 106 para realizar las etapas manuales de los métodos divulgados en el presente documento, tales como las etapas de entrenamiento manual.

El experto apreciará que el sistema 100 informático puede alimentarse mediante cualquier medio de alimentación adecuado. La memoria 102 puede comprender uno o más dispositivos de memoria volátil o no volátil, tales como DRAM, SRAM, memoria flash, memoria de sólo lectura, RAM ferroeléctrica, unidades de disco duro, discos flexibles, cinta magnética, discos ópticos o similares. Asimismo, el procesador 112 puede comprender una o más unidades de procesamiento, tales como un microprocesador, GPU, CPU, procesador de múltiples núcleos o similar. Las conexiones 114 de red pueden ser cableadas, tales como ópticas, de fibra óptica, Ethernet o similares, o cualquier comunicación inalámbrica adecuada.

En otros ejemplos, algunos componentes mostrados en la figura 1 pueden no estar presentes. Por ejemplo, en los métodos divulgados en el presente documento que no requieren una entrada de usuario manual, el elemento 110 de visualización y algunas de la una o más interfaces 106 de usuario pueden no requerirse. En ejemplos adicionales, la entrada de usuario puede proporcionarse de manera externa en vez de a través de una conexión 114 de red del sistema 100 informático. Puede haber componentes adicionales presentes que no se muestran en la figura 1. Por ejemplo, el sistema 100 informático puede comprender una pluralidad de procesadores 112.

El sistema 104 de NLP comprende uno o más componentes que juntos funcionan para implementar los métodos divulgados en el presente documento. Los componentes pueden comprender o más algoritmos de aprendizaje automático de NLP y/o algoritmos basados en reglas de NLP, que pueden ejecutarse en serie o en paralelo, dependiendo de la naturaleza del texto que va a analizarse en el método. El experto apreciará que el sistema 104 de NLP puede implementarse usando técnicas de programación convencionales y que puede usarse cualquier lenguaje de programación adecuado para implementar los métodos divulgados en el presente documento. Por ejemplo, el sistema 104 de NLP puede ser un archivo ejecutable que se ejecuta en el sistema 100 informático o puede implementarse usando cualquier técnica de programación alternativa conocida en la técnica. Por ejemplo, cuando el sistema 100 de NLP comprende varios componentes, cada componente puede ejecutarse de manera independiente en procesadores o sistemas informáticos independientes, o bien en serie o bien en paralelo según se requiera usando cualquier técnica informática distribuida.

Además, la base 109 de datos consultable puede almacenarse de manera independiente del sistema 100 informático, transmitiendo el procesador 112 la salida del sistema 104 de NLP a la base de datos a través de la una o más conexiones 114 de red o a través de otra conexión de datos.

El sistema 100 informático está configurado para comunicarse con una red 118 a través de una o más conexiones 114 de red. El sistema 100 informático es capaz de recibir un corpus de documentos 120 en texto libre a través de la red 118. El corpus de documentos en texto libre recibidos se analizan por el sistema 104 de NLP y los resultados se almacenan en la base 109 de datos. Además, el sistema 100 informático está configurado para recibir entradas y transmitir salidas a una interfaz 122 de cliente a través de la red 118. Por ejemplo, la interfaz 122 de cliente puede ser cualquier interfaz adecuada tal como una interfaz de web u otra interfaz de búsqueda adecuada, a través de la cual un cliente es capaz de realizar búsquedas para obtener información agregada a partir de la base 109 de datos. El cliente puede realizar una consulta, que se envía al sistema 100 informático a través de la red 118. La consulta puede recibirse por uno o más programas 106 almacenados en la memoria 102, lo cual hace que el procesador 112 busque en la base 109 de datos en consecuencia. Entonces el procesador 122 hace que los resultados de la búsqueda, es decir información agregada referente a la base 109 de datos, a la interfaz 122 de cliente a través de la red 118. Alternativa o adicionalmente, pueden realizarse consultas agregadas similares a través de la una o más interfaces 116 de usuario del sistema 100 informático.

La red 118 puede ser cualquier red adecuada para transmitir datos entre los componentes de la red, tal como una red inalámbrica o cableada. La interfaz 122 de cliente puede ser una aplicación web accesible a través de Internet, por ejemplo.

El experto apreciará que esta divulgación contempla muchas otras variaciones del sistema 110 informático, la red 118 y la interfaz 122 de cliente adecuadas para implementar los métodos divulgados en el presente documento.

La figura 2 muestra un método 200 para generar una base de datos consultable según la presente divulgación.

En una primera etapa 202, se recibe un corpus de documentos en texto libre que contienen datos confidenciales. El corpus de documentos en texto libre pueden comprender más de 49 documentos, más de 1000, más de 39.000 o más. Por ejemplo, el sistema 100 informático puede recibir un corpus de documentos 120 en texto libre a través de una red 118. Los documentos en texto libre están relacionados con el mismo dominio. Por ejemplo, los documentos en texto libre pueden estar en el campo de los seguros (por ejemplo, informes de reclamaciones de seguro) o el campo de la medicina (por ejemplo, historias clínicas).

Posteriormente, en la etapa 204, a cada uno de los documentos en texto libre se les asigna una o más entidades nombradas resumidas por un sistema de NLP. Por ejemplo, el sistema 104 de NLP se ejecuta por el procesador 112 para procesar el texto contenido en cada documento, identificándose entidades nombradas resumidas en cada documento. El sistema de NLP está entrenado de tal manera que las entidades nombradas resumidas se reconocen y desambiguan con una precisión de entre 0,75 y menos de 1 (preferiblemente entre 0,75 y 0,95 y más preferiblemente entre 0,85 y 0,95) y una exhaustividad de entre 0,75 y menos de 1 (preferiblemente entre 0,75 y 0,95 y más preferiblemente entre 0,85 y 0,95), y de tal manera que la razón de precisión y exhaustividad es de entre 0,7 y 1,3, y preferiblemente entre 0,8 y 1,2, y más preferiblemente entre 0,9 y 1,1. Cada entidad nombrada resumida asignada puede asociarse con una porción de texto en los documentos en texto libre. Tal como se comentó anteriormente, que la exhaustividad y precisión sean menores de 1 garantiza un nivel de privacidad individual, garantizando los valores mínimos de exhaustividad y precisión y su razón que la información agregada todavía sea exacta. El sistema de NLP puede entrenarse según los métodos de entrenamiento divulgados en el presente documento.

Finalmente, en la etapa 206, las entidades nombradas resumidas asignadas para cada documento se almacenan en una base de datos consultable tal como la base 109 de datos.

Una vez completado el método, un usuario puede realizar consultas agregadas a la base de datos consultable, por ejemplo a partir de la interfaz 122 de cliente a través de la red 118 o a partir de una interfaz 116 de usuario, procesándose la base de datos basándose en la consulta para generar la información agregada pedida (por ejemplo, mediante un programa 106 ejecutado por un procesador 112) y proporcionarla al usuario.

La figura 3 muestra un diagrama de bloques de un sistema 104 de NLP a modo de ejemplo. El sistema 104 de NLP comprende un motor 302 de procesamiento previo, un motor 304 de reconocimiento y un motor 306 de desambiguación. Aunque se presentan como componentes independientes en el sistema 104 de NLP, resultará evidente para un experto que pueden estar comprendidos en un único módulo o de otro modo.

El motor 302 de procesamiento previo está configurado para recibir un corpus de documentos en texto libre y procesar previamente los documentos en preparación para los procedimientos de reconocimiento y desambiguación. Por ejemplo, el motor 302 de procesamiento previo recibe el corpus de documentos 120 en texto libre a través de la red 118. Entonces, el motor 302 de procesamiento previo distingue el texto libre en frases y divide cada frase en símbolos. Finalmente, el motor 302 de procesamiento previo convierte las frases en un vector de símbolos.

Por ejemplo, el motor 302 de procesamiento previo puede recibir el siguiente texto:

Esta mujer blanca de 23 años de edad acude por quejas de alergia. Solía tener alergia cuando vivía en Seattle pero cree que ha empeorado aquí.

Entonces, el motor 302 de procesamiento previo distingue el texto en frases y símbolos de la siguiente manera:

Frase 1: [Esta] [mujer] [blanca] [de] [23] [años] [de] [edad] [acude] [por] [quejas] [de] [alergia][.]

Frase 2: [Solía] [tener] [alergia] [cuando] [vivía] [en] [Seattle] [pero] [cree] [que] [ha] [empeorado] [aquí][.]

En las que [*texto*] indica un símbolo. Después se convierten las frases en vectores de símbolos de la siguiente manera:

Frase 1: [“Esta”, “mujer”, “blanca”, “de”, “23”, “años”, “de”, “edad”, “acude”, “por”, “quejas”, “de”, “alergia”, “.”]

Frase 2: [“Solía”, “tener”, “alergia”, “cuando”, “vivía”, “en”, “Seattle”, “pero”, “cree”, “que”, “ha”, “empeorado”, “aquí”, “.”]

El motor 304 de reconocimiento está configurado para analizar vectores de símbolos para reconocer las entidades nombradas resumidas en el texto. El motor 304 de reconocimiento puede comprender uno o más algoritmos basados en reglas y algoritmos de aprendizaje automático tal como se describe en el presente documento. Por ejemplo, el motor 302 de reconocimiento recibe los vectores de símbolos de las frases en cada texto del corpus a partir del motor 302 de procesamiento previo y detecta y clasifica entidades nombradas resumidas en los textos.

Por ejemplo, el motor de reconocimiento recibe los vectores de símbolos de la frase 1 y la frase 2 anteriores y reconoce las siguientes entidades:

Frase 1: [[“23”, “años”, “de”, “edad”], [“mujer”], [“alergia”]]

Frase 2: [[“solía”, “tener”, “alergia”]]

El motor 306 de desambiguación está configurado para normalizar entidades reconocidas con respecto a una terminología existente (es decir, a partir de una ontología). El motor 306 de desambiguación está configurado además para aplicar alguna información de contexto a partir del vector de símbolos con el fin de ayudar a desambiguar las entidades. Por ejemplo, el motor 306 de desambiguación recibe las entidades nombradas resumidas a partir del motor

304 de reconocimiento y las desambigua. El motor 306 de desambiguación puede comprender uno o más algoritmos basados en reglas y algoritmos de aprendizaje automático tal como se describe en el presente documento.

Por ejemplo, el motor 206 de desambiguación recibió las entidades de la frase 1 y la frase 2 anteriores y las desambigua de la siguiente manera:

Frase 1: [Edad – 23], [Sexo – femenino] [Diagnóstico - alergia]

Frase 2: [Antecedentes personales - alergia]

Resultará evidente para un experto que pueden usarse otros sistemas de NLP, dependiendo del contenido del texto no estructurado que va a analizarse. En particular, puede usarse cualquier sistema 104 de NLP que sea capaz de analizar el corpus de documentos en texto libre de tal manera que las entidades nombradas resumidas se reconozcan y desambigüen a los niveles de precisión y exhaustividad requeridos.

Por ejemplo, puede usarse cualquier procedimiento de conversión en símbolos y vectorización de texto adecuado para el motor 302 de procesamiento previo. Se indica que el módulo 302 de procesamiento previo de texto también puede realizar cualquier forma de procesamiento previo de texto según se requiera, tal como formateo previo de texto, eliminación de signos de puntuación o similar.

La figura 4 muestra un método 400 para entrenar un sistema de NLP que comprende uno o más algoritmos de aprendizaje automático según la presente divulgación.

En una primera etapa 402, se selecciona uno o más subconjuntos de documentos en texto libre en el dominio. Estos subconjuntos pueden proporcionarse a partir del corpus de documentos en texto libre que van a analizarse o proporcionarse de manera independiente.

En segundo lugar, en la etapa 404, la una o más entidades nombradas resumidas se asignan a cada documento en el uno o más subconjuntos para formar uno o más conjuntos de entrenamiento, en las que las entidades nombradas resumidas asignadas en cada subconjunto son las salidas objetivo para cada algoritmo de aprendizaje automático respectivo.

En la etapa 406, el uno o más algoritmos de aprendizaje automático se entrenan usando el uno o más conjuntos de entrenamiento. Dicho de otro modo, los conjuntos de entrenamiento se proporcionan a los algoritmos de aprendizaje automático y los algoritmos de aprendizaje automático analizan los documentos respectivos y sus entidades nombradas resumidas asignadas para entrenar el algoritmo de aprendizaje automático.

A continuación, se selecciona un segundo subconjunto de documentos en texto libre en el dominio en la etapa 408. El segundo subconjunto puede tomarse a partir del corpus de documentos de texto o pueden obtenerse de manera independiente.

Posteriormente, en la etapa 410, se introduce el segundo subconjunto de documentos en texto libre en el sistema de NLP. Después se evalúa el sistema de NLP para determinar si el sistema de NLP reconoce y desambigua las entidades nombradas resumidas con una precisión de entre 0,75 y menos de 1 (o entre 0,75 y 0,95, o entre 0,85 y 0,95) y una exhaustividad de entre 0,75 y menos de 1 (o entre 0,75 y 0,95, o entre 0,85 y 0,95), y en el que la razón de precisión y exhaustividad es de entre 0,7 y 1,3 (o entre 0,8 y 1,2, o entre 0,9 y 1,1). Si se cumplen estos requisitos, entonces el procedimiento de entrenamiento termina (etapa 418). Si no es así, entonces el uno o más algoritmos de aprendizaje automático vuelven a entrenarse (etapa 416), por ejemplo generando nuevos conjuntos de entrenamiento y usando los nuevos conjuntos de entrenamiento para entrenar los algoritmos. En algunos ejemplos, el método mostrado en la figura 4 se realiza de manera iterativa hasta que se cumplen los valores de exhaustividad, precisión y su razón (es decir, si la respuesta en la etapa 414 es “no”, entonces el método vuelve a la etapa 402).

Se indica en particular que los valores de precisión y exhaustividad pueden ser demasiado altos como para garantizar un nivel satisfactorio de privacidad individual. En ese caso, pueden proporcionarse datos de entrenamiento adicionales a los modelos de entrenamiento que contienen un número aumentado de falsos negativos y/o falsos positivos. Por ejemplo, si debe reducirse la precisión, entonces pueden proporcionarse datos de entrenamiento que contienen entidades nombradas resumidas asignadas a cadenas incorrectas de texto. Si debe reducirse la exhaustividad, entonces pueden proporcionarse datos de entrenamiento que contienen texto relacionado con una entidad nombrada resumida a la que no se le ha asignado esa entidad nombrada resumida y/o a la que se le ha asignado una entidad nombrada resumida diferente.

Cuando el sistema de NLP comprende una pluralidad de algoritmos de aprendizaje automático, los algoritmos pueden entrenarse usando el mismo subconjunto de documentos o subconjuntos diferentes.

En algunos ejemplos, los conjuntos de entrenamiento pueden generarse manualmente por uno o más usuarios. En particular, el/los usuario(s) puede(n) asignar manualmente las entidades nombradas resumidas a los documentos en

texto libre según un conjunto de directrices. Por ejemplo, si los documentos en texto libre son historias clínicas, el/los usuario(s) puede(n) ser profesionales médicos entrenados para asignar las entidades nombradas resumidas relevantes a los documentos en texto libre. La precisión y exhaustividad pueden evaluarse de manera similar con la ayuda del/de los usuario(s), en las que el/los usuario(s) asigna(n) correctamente las entidades nombradas resumidas al segundo subconjunto de documentos, comparándose la asignación de usuario de las entidades nombradas resumidas con la salida del sistema de NLP con el segundo subconjunto. En la evaluación, se considera que las asignaciones de usuario son las asignaciones correctas para calcular el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos del sistema de NLP.

Alternativamente, los conjuntos de entrenamiento pueden generarse introduciendo el/los primer(os) subconjunto(s) de documentos en un segundo sistema de NLP que tiene una precisión mayor de 0,85 (preferiblemente mayor de 0,95) y una exhaustividad mayor de 0,85 (preferiblemente mayor de 0,95) (se considera que un sistema de NLP que tiene valores de precisión y exhaustividad mayores que estos umbrales tiene una tasa de error similar a una tasa de error humana). Entonces pueden evaluarse la precisión y exhaustividad del primer sistema de NLP introduciendo el segundo subconjunto de documentos en el segundo sistema de NLP y considerando que las asignaciones de entidades nombradas resumidas del segundo sistema de NLP son correctas. Entonces puede compararse la salida del primer sistema de NLP con las asignaciones realizadas por el segundo sistema de NLP para calcular el número de verdaderos positivos, verdaderos negativos, falso positivo y falsos negativos y por tanto calcular la precisión y exhaustividad.

El experto apreciará que el método de entrenamiento anterior es a modo de ejemplo y que puede usarse cualquier método adecuado de entrenar el sistema de NLP, que se conocen bien en la técnica.

Los métodos divulgados en el presente documento se comentarán ahora con referencia a ejemplos de corpus de documentos en texto libre.

Ejemplo 1 – Historia clínica

Los siguientes documentos son ejemplos de historias clínicas que forman un corpus de documentos en texto libre.

Documento 1 (sujeto A)

Esta mujer blanca de 23 años de edad acude por quejas de alergia. Solía tener alergia cuando vivía en Seattle pero cree que ha empeorado aquí. En el pasado ha probado con loratadina y cetirizina. Ambas funcionaron durante poco tiempo pero luego parecieron perder eficacia. También ha usado fexofenadina. La usó el verano pasado y empezó a usarla de nuevo hace dos semanas. No parece que esté funcionando muy bien. Ha usado pulverizaciones sin receta pero no pulverizaciones nasales con receta. Tiene asma pero no requiere medicamento diario para esto y no cree que esté empeorando., MEDICAMENTOS:, Su único medicamento actual es norgestimato y fexofenadina.

Documento 2 (sujeto B)

Tiene enfermedad por reflujo gastroesofágico., ANTECEDENTES QUIRÚRGICOS ANTERIORES:, Incluye cirugía reparadora en su mano derecha hace 13 años., ANTECEDENTES SOCIALES:, Actualmente está soltero. Bebe aproximadamente diez veces al año. Fumaba significativamente hasta hace varios meses. Ahora fuma menos de tres cigarrillos al día., ANTECEDENTES FAMILIARES:, Cardiopatía en ambos abuelos, abuela con accidente cerebrovascular y una abuela con diabetes. Niega obesidad e hipertensión en otros miembros de la familia., MEDICAMENTOS ACTUALES:, Ninguno., ALERGIA:, Es alérgico a la penicilina.

Documento 3 (sujeto C)

ENFERMEDAD ACTUAL:, Hoy he visto a ABC. Es un hombre muy agradable de 42 años de edad y 344 libras. Mide 5'9". Tiene un IMC de 51. Ha tenido sobrepeso desde hace diez años desde que tenía 33 años, como máximo ha llegado a pesar 358 libras y como mínimo 260. Está sometiéndose a intentos quirúrgicos de pérdida de peso para sentirse bien, volverse sano y empezar a hacer ejercicio de nuevo. Estuvo seis meses sin beber alcohol y sin ingerir demasiadas calorías. Ha seguido múltiples programas de pérdida de peso comerciales incluyendo Slim Fast durante un mes hace un año y dieta Atkins durante un mes hace dos años.

Documento 4 (sujeto D)

2-D, MODO M:, ,1. Dilatación de aurícula izquierda con un diámetro de aurícula izquierda de 4,7 cm., 2. Ventrículos izquierdo y derecho de tamaño normal., 3. Función sistólica de LV normal con fracción de eyección de ventrículo izquierdo del 51%., 4. Función diastólica de LV normal., 5. Morfología normal de válvula aórtica, válvula mitral, válvula tricúspide y válvula pulmonar.

Documento 5 (sujeto E)

1. El tamaño de cavidad y el grosor de tabique del ventrículo izquierdo parecen normales. El movimiento de tabique y

la función sistólica del ventrículo izquierdo parecen hiperdinámicos con una fracción de eyección estimada del 70% al 75%. Se observa casi una obliteración de la cavidad. También parece haber un gradiente de tracto de flujo de salida del ventrículo izquierdo aumentado a nivel del centro de la cavidad compatible con función sistólica del ventrículo izquierdo hiperdinámica. Se observa un patrón de relajación del ventrículo izquierdo anómalo.

Los documentos anteriores pueden proporcionarse a un sistema de NLP que está entrenado para reconocer y desambiguar las identidades nombradas resumidas con la precisión y exhaustividad requeridas. La salida del sistema de NLP se ilustra en las figuras 5A a 5E.

Se observará que la salida del sistema de NLP tiene una mezcla de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.

Después, el sistema de NLP guarda los resultados de la salida (es decir, los verdaderos y falsos positivos) en una base de datos consultable de la siguiente manera (la columna que indica si el registro es un verdadero positivo o un falso positivo es únicamente con propósitos de ilustración).

Almacenamiento de muestra de historias clínicas

Sujeto	Término	Entidad nombrada resumida	¿TP o FP?
A	[síntoma - alergia]	alergia	TP
A	[antecedentes personales – tratamiento]	loratadina	TP
A	[antecedentes personales – tratamiento]	cetirizina	TP
A	[antecedentes personales – tratamiento]	fexofenadina	TP
A	[síntoma - cardiopatía]	asma	FP
B	[síntoma – reflujo gastroesofágico]	enfermedad por reflujo gastroesofágico	TP
B	[síntoma – cirugía reparadora]	cirugía reparadora	FP
B	[síntoma – alergia – penicilina]	alérgico a penicilina	TP
C	[prueba – IMC – 51]	IMC de 51	TP
C	[síntoma – asma]	ejercicio	FP
D	[síntoma – dilatación de aurícula izquierda]	Dilatación de aurícula izquierda	TP
D	[síntoma – fracción de eyección ventricular - 51%]	Fracción de eyección ventricular del 51%	TP
D	[síntoma – función diastólica de LV normal]	Función diastólica de LV normal	TP
E	[síntoma – tamaño de ventrículo izquierdo normal]	Tamaño de cavidad y grosor de tabique de ventrículo izquierdo parecen normales	TP
E	[síntoma – tamaño de cavidad de ventrículo izquierdo normal]	Tamaño de cavidad y grosor de tabique de ventrículo izquierdo parecen normales	TP
E	[síntoma – grosor de tabique de ventrículo izquierdo normal]	Tamaño de cavidad y grosor de tabique de ventrículo izquierdo parecen normales	TP

Entonces pueden realizarse consultas agregadas en la base de datos consultable. Por ejemplo, pueden realizarse las siguientes consultas:

de pacientes con alergia: 40% (2)

de pacientes con grosor de tabique de ventrículo izquierdo normal: 20% (1)

de pacientes con asma: 20% (1).

Ejemplo 2 – Seguro

Los siguientes documentos son ejemplos de registros de seguro que forman un corpus de documentos en texto libre.

Documento 1 (sujeto A)

Contaminación radiactiva

*Esta Póliza no cubre ninguna pérdida o daño que surja directa o indirectamente de reacción nuclear, radiación nuclear o contaminación radiactiva, independientemente de cómo pueda haberse causado tal reacción nuclear, radiación nuclear o contaminación radiactiva * No obstante, si el incendio es un riesgo asegurado y surge un incendio directa o indirectamente de reacción nuclear, radiación nuclear o contaminación radiactiva, cualquier pérdida o daño que surja directamente de tal incendio estará cubierto (sujeto a las disposiciones de esta Póliza)*

Documento 2 (sujeto B)

Enfermedad infecciosa

Sin perjuicio de cualquier contenido en el sentido contrario en la Póliza, la cobertura en virtud de la presente no se extiende para incluir lesión, enfermedad o muerte de una persona asegurada o cualquier responsabilidad vinculada al Asegurado por la pérdida o daño de propiedad de tercero, lesión, enfermedad o muerte de un tercero como resultado de reclamaciones que surjan directa o indirectamente de, provocadas por, que se produzcan a través de, en consecuencia de o de cualquier manera atribuibles a enfermedad infecciosa, gripe aviar o cualquier enfermedad que se haya declarado epidémica por la Organización Mundial de la Salud.

Documento 3 (sujeto C)

Arbitraje

Cualquier disputa que surja de esta Póliza será remitida para la decisión de un Árbitro que se designará por ambas partes o, si no pueden acordar un único árbitro, para la decisión de dos árbitros, uno designado por escrito por cada parte (dentro del plazo de un mes después de que cualquier parte requiera por escrito que se haga). Los dos árbitros nombrarán entonces mutuamente un tercer árbitro al que deberán nombrar por escrito los árbitros. El tercer árbitro se sentará con los árbitros y presidirá sus reuniones. La toma de una decisión por el árbitro, los árbitros o el tercer árbitro será una condición precedente para cualquier derecho de acción contra Nosotros.

Documento 4 (sujeto D)

Hipoteca

Por el presente se acuerda que, en el caso de cualquier pérdida o daño que esté asegurado por el presente, Nosotros pagaremos a los Acreedores hipotecados o a dichos Cesionarios tal como se menciona en el Anexo en la medida de su interés y que este seguro, sólo en la medida en que se refiere al interés en el mismo de los Acreedores hipotecarios o dichos Cesionarios, no quedará invalidado por ninguna acción o negligencia del Deudor hipotecario o Propietario de los Edificios.

Documento 5 (sujeto E)

Cancelación por Nosotros

Tenemos derecho a cancelar esta Póliza otorgándole a Usted siete (7) días por notificación por correo certificado por escrito a Su última dirección conocida. Si se ha realizado una reclamación, o se ha notificado un incidente que puede dar lugar a una reclamación, entonces no se reembolsará la prima. Si no se ha realizado ninguna reclamación, entonces le reembolsaremos una prima prorrateada en proporción a la cantidad de tiempo que haya estado en vigor Su Póliza.

Los documentos anteriores pueden proporcionarse a un sistema de NLP que está entrenado para reconocer y desambiguar las identidades nombradas resumidas con la precisión y exhaustividad requeridas. La salida del sistema de NLP se ilustra en las figuras 6A a 6E.

Se observará que la salida del sistema de NLP tiene una mezcla de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.

Después, el sistema de NLP guarda los resultados de la salida (es decir, los verdaderos y falsos positivos) en una base de datos consultable de la siguiente manera (la columna que indica si el registro es un verdadero positivo o un falso positivo es únicamente con propósitos de ilustración).

Almacenamiento de muestra de registros de seguro

Sujeto	Término	Entidad nombrada resumida	¿TP o FP?
A	[cobertura por pérdida o	no cubre ninguna pérdida o daño	TP

	daño – negada – reacción nuclear, radiación nuclear]	que surja directa o indirectamente de reacción nuclear, radiación nuclear	
A	[riesgo derivado – incendio provocado por reacción nuclear]	incendio surge directa o indirectamente de reacción nuclear, radiación nuclear	TP
A	[riesgo – contaminación radiactiva]	Contaminación radiactiva	TP
B	[riesgo – enfermedad infecciosa]	Enfermedad infecciosa	TP
B	[enfermedad infecciosa – gripe aviar]	Gripe aviar	TP
B	[riesgo – riesgo relacionado con la salud]	Organización Mundial de la Salud	FP
C	[contenido relacionado legal – arbitraje]	Arbitraje	TP
C	[contenido relacionado legal – disputa]	disputa	TP
C	[figura jurídica – árbitro]	Árbitro	TP
C	[figura jurídica – árbitro]	tercer árbitro	FP
D	[riesgo – hipoteca]	Hipoteca	TP
D	[ejecución de riesgo – pago de hipoteca]	pago a los Acreedores hipotecarios	TP
D	[figura jurídica – deudor hipotecario]	Deudor hipotecario	TP
E	[acción de póliza – cancelación unilateral]	Cancelación por Nosotros	TP
E	[acontecimiento en el tiempo – días – 7]	siete (7) días	TP
E	[requisito legal – certificación]	correo certificado	FP
E	[acción de póliza – negada – reembolso]	sin reembolso	TP

Entonces pueden realizarse consultas agregadas en la base de datos consultable. Por ejemplo, pueden realizarse las siguientes consultas:

de pólizas con reembolsos: 0% (0)

de pólizas con acciones de póliza: 40% (2)

de pólizas relacionadas con hipotecas: 20% (1)

La exactitud de los resultados agregados en ambos ejemplos puede garantizarse al tiempo que se conserva la privacidad cuando los valores de precisión y exhaustividad del sistema de NLP adoptan los valores tal como se especifican en esta divulgación.

Todo lo anterior se encuentra completamente dentro del alcance de la presente divulgación y se considera que forma la base para realizaciones alternativas en las que se aplican una o más combinaciones de las características anteriormente descritas, sin limitación a la combinación específica divulgada anteriormente.

A la vista de esto, habrá muchas alternativas que implementen las enseñanzas de la presente divulgación. Se espera que un experto en la técnica será capaz de modificar y adaptar la divulgación anterior para adaptarse a sus propias circunstancias y requisitos dentro del alcance de la presente divulgación, al tiempo que conserve algunos o todos los efectos técnicos de la misma, o bien divulgados o bien derivables a partir de lo anterior, a la vista de su conocimiento general común en esta técnica. Todos de tales equivalentes, modificaciones o adaptaciones se encuentran dentro del alcance de la presente divulgación. El alcance de la invención está definido por las reivindicaciones independientes.

REIVINDICACIONES

1. Método implementado por ordenador de generación de una base de datos consultable que tiene privacidad diferencial, que comprende:

recibir un corpus de documentos en texto libre que contienen datos confidenciales, estando los documentos en texto libre relacionados con el mismo dominio;

asignar, por un sistema de procesamiento de lenguaje natural (NLP) entrenado, una o más entidades nombradas resumidas a cada documento en texto libre en el corpus; y

almacenar las entidades nombradas resumidas de cada documento en texto libre asignada por el sistema de NLP en la base de datos consultable configurada para proporcionar información agregada referente a las entidades nombradas;

en el que el sistema de NLP está configurado de tal manera que las entidades nombradas resumidas se reconocen y desambiguan con una precisión de entre 0,75 y menos de 1 y una exhaustividad de entre 0,75 y menos de 1, y en el que la razón de precisión y exhaustividad es de entre 0,7 y 1,3;

en el que la base de datos consultable está libre de la adición de ruido artificial por un algoritmo de generación de ruido artificial y la privacidad diferencial de la base de datos consultable surge de la precisión, exhaustividad y razón de precisión y exhaustividad del sistema de NLP.

2. Método según la reivindicación 1, en el que la precisión es de entre 0,75 y 0,95 y la exhaustividad es de entre 0,75 y 0,95, preferiblemente en el que la precisión es de entre 0,85 y 0,95 y la exhaustividad es de entre 0,85 y 0,95.

3. Método según la reivindicación 1 ó 2, en el que la razón de precisión y exhaustividad es de entre 0,8 y 1,2, preferiblemente en el que la razón de precisión y exhaustividad es de entre 0,9 y 1,1.

4. Método según cualquier reivindicación anterior, en el que el número de documentos en texto libre recibidos es mayor de 49, preferiblemente mayor de 1000 y más preferiblemente mayor de 39.000.

5. Método según cualquier reivindicación anterior, comprendiendo el sistema de NLP uno o más algoritmos de aprendizaje automático, comprendiendo el método las etapas de entrenar cada uno del uno o más algoritmos de aprendizaje automático mediante:

seleccionar uno o más subconjuntos de documentos en texto libre en el dominio;

asignar la una o más entidades nombradas resumidas a los documentos en el uno o más subconjuntos para formar uno o más conjuntos de entrenamiento;

entrenar el uno o más algoritmos de aprendizaje automático usando el uno o más conjuntos de entrenamiento;

seleccionar un segundo subconjunto de documentos en texto libre en el dominio;

introducir el segundo subconjunto del corpus de documentos en texto libre en el sistema de NLP;

evaluar si el sistema de NLP reconoce y desambigua las entidades nombradas resumidas con una precisión de entre 0,75 y menos de 1 y una exhaustividad de entre 0,75 y menos de 1, y en el que la razón de precisión y exhaustividad es de entre 0,7 y 1,3; y

si no es así, volver a entrenar el uno o más algoritmos de aprendizaje automático de tal manera que la precisión, exhaustividad y razón de precisión y exhaustividad están dentro de los intervalos requeridos.

6. Método según la reivindicación 5, en el que el entrenamiento del uno o más algoritmos de aprendizaje automático se realiza de manera iterativa según la reivindicación 5 hasta que la precisión, exhaustividad y razón de precisión y exhaustividad están dentro de los intervalos requeridos.

7. Método según la reivindicación 5 ó 6, en el que, si tras la evaluación se requiere reducir la precisión, entonces el uno o más algoritmos de aprendizaje automático vuelven a entrenarse proporcionando datos de entrenamiento que contienen entidades nombradas resumidas asignadas a cadenas incorrectas de texto; y/o en el que, si tras la evaluación se requiere reducir la exhaustividad, entonces el uno o más algoritmos de aprendizaje automático vuelven a entrenarse proporcionando datos de entrenamiento que contienen texto relacionado con una entidad nombrada resumida a la que no se le ha asignado esa entidad nombrada resumida y/o a la que se le ha asignado una entidad nombrada resumida diferente.

8. Método según cualquiera de las reivindicaciones 5 a 7, en el que el uno o más conjuntos de entrenamiento se forman asignando manualmente las entidades nombradas resumidas al uno o más subconjuntos, por uno o más usuarios; o

en el que el uno o más conjuntos de entrenamiento se forman asignando las entidades nombradas resumidas al uno o más subconjuntos por un segundo sistema de NLP que tiene una precisión mayor de 0,85 y una exhaustividad mayor de 0,85, y preferiblemente una precisión mayor de 0,95 y exhaustividad mayor de 0,95.

9. Método según cualquiera de las reivindicaciones 5 a 8, en el que la precisión, exhaustividad y razón de precisión y exhaustividad se evalúan asignando manualmente las entidades nombradas resumidas al segundo subconjunto de documentos por uno o más usuarios, y comparando la asignación de usuario de entidades nombradas resumidas al segundo subconjunto con la salida del sistema de NLP; o

en el que la precisión, exhaustividad y razón de precisión y exhaustividad se evalúan asignando las entidades nombradas resumidas al segundo subconjunto de documentos por un segundo sistema de NLP que tiene una precisión mayor de 0,85 y una exhaustividad mayor de 0,85, preferiblemente una precisión mayor de 0,95 y exhaustividad mayor de 0,95, y comparando la salida del segundo sistema de NLP con la salida del sistema de NLP.

10. Método según cualquier reivindicación anterior, en el que el sistema de NLP comprende uno o más algoritmos basados en reglas.

11. Método según cualquier reivindicación anterior, en el que los documentos en texto libre son historias clínicas, y las entidades nombradas resumidas comprenden información del paciente y terminología médica.

12. Método según cualquier reivindicación anterior, en el que la entidad resumida nombrada y un término de desambiguación asociado se almacenan en la base de datos.

13. Método según la reivindicación 12, en el que:

los documentos en texto libre son historias clínicas y el sistema de NLP se entrena para asignar uno o más de los siguientes términos de desambiguación a las entidades nombradas resumidas: información del paciente, antecedentes médicos, antecedentes médicos familiares, antecedentes farmacéuticos, antecedentes de tratamiento, síntomas, resultados de pruebas, evoluciones y notas; o

los documentos en texto libre son registros de seguro y el sistema de NLP se entrena para asignar uno o más de los siguientes términos de desambiguación a las entidades nombradas resumidas: cobertura por pérdida o daño, riesgo derivado, riesgo, contenido relacionado legal, figura jurídica, acción de póliza, acontecimiento en el tiempo, requisito legal.

14. Producto de programa informático para generar una base de datos consultable que tiene privacidad diferencial, que comprende instrucciones que, cuando se ejecutan por un ordenador, hacen que el ordenador:

reciba un corpus de documentos en texto libre que contienen datos confidenciales, estando los documentos en texto libre relacionados con el mismo dominio;

asigne, por el sistema de procesamiento de lenguaje natural (NLP) entrenado, una o más entidades nombradas resumidas a cada documento en texto libre en el corpus; y

almacene las entidades nombradas resumidas de cada documento en texto libre asignadas por el sistema de NLP en la base de datos consultable configurada para proporcionar datos agregados referentes a las entidades nombradas;

en el que el sistema de NLP está configurado de tal manera que las entidades nombradas resumidas se reconocen y desambiguan con una precisión de entre 0,75 y menos de 1 y una exhaustividad de entre 0,75 y menos de 1, y en el que la razón de precisión y exhaustividad es de entre 0,7 y 1,3;

y en el que la base de datos consultable está libre de la adición de ruido artificial por un algoritmo de generación de ruido artificial y la privacidad diferencial de la base de datos consultable surge de la precisión, exhaustividad y razón de precisión y exhaustividad del sistema de NLP.

15. Aparato de procesamiento de datos para generar una base de datos consultable que tiene privacidad diferencial, que comprende un sistema de procesamiento de lenguaje natural (NLP) entrenado y configurado para:

recibir un corpus de documentos en texto libre que contienen datos confidenciales, estando los documentos en texto libre relacionados con el mismo dominio;

asignar, por el sistema de procesamiento de lenguaje natural (NLP) entrenado, una o más entidades nombradas resumidas a cada documento en texto libre en el corpus; y

almacenar las entidades nombradas resumidas de cada documento en texto libre asignadas por el sistema de NLP en la base de datos consultable configurada para proporcionar datos agregados referentes a las entidades nombradas;

en el que el sistema de NLP está configurado de tal manera que las entidades nombradas resumidas se reconocen y desambiguan con una precisión de entre 0,75 y menos de 1 y una exhaustividad de entre 0,75 y menos de 1, y en el que la razón de precisión y exhaustividad es de entre 0,7 y 1,3;

en el que la base de datos consultable está libre de la adición de ruido artificial por un algoritmo de generación de ruido artificial y la privacidad diferencial de la base de datos consultable surge de la precisión, exhaustividad y razón de precisión y exhaustividad del sistema de NLP.

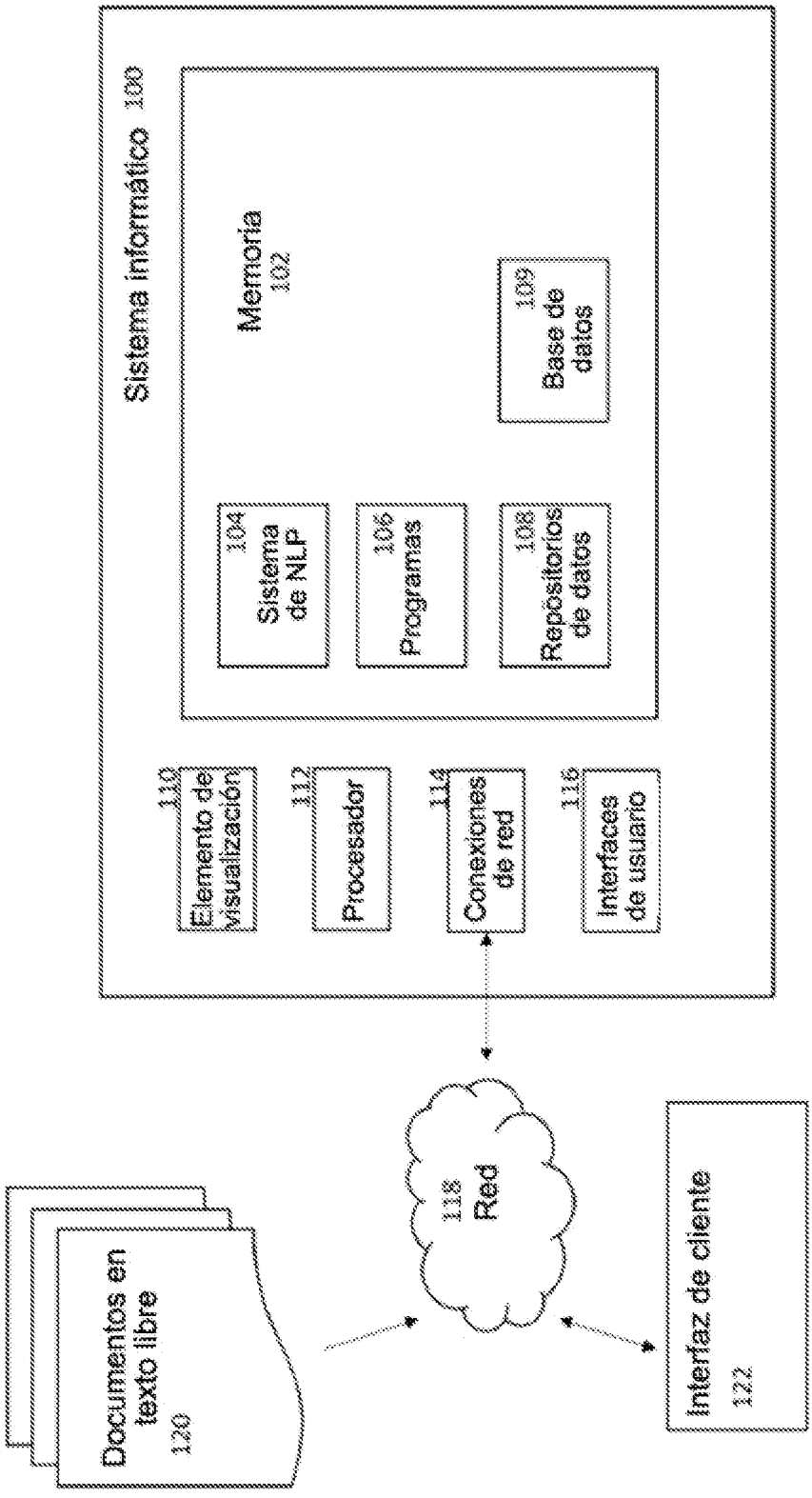
Sistema informático 100
Memoria 102
104 Sistema de NLP
106 Programas
108 Repositorios de datos
109 Base de datos
110 Elemento de visualización
112 Procesador
114 Conexiones de red
116 Interfaces de usuario
118 Red
Documentos en texto libre 120
Interfaz de cliente 122

FIG. 1

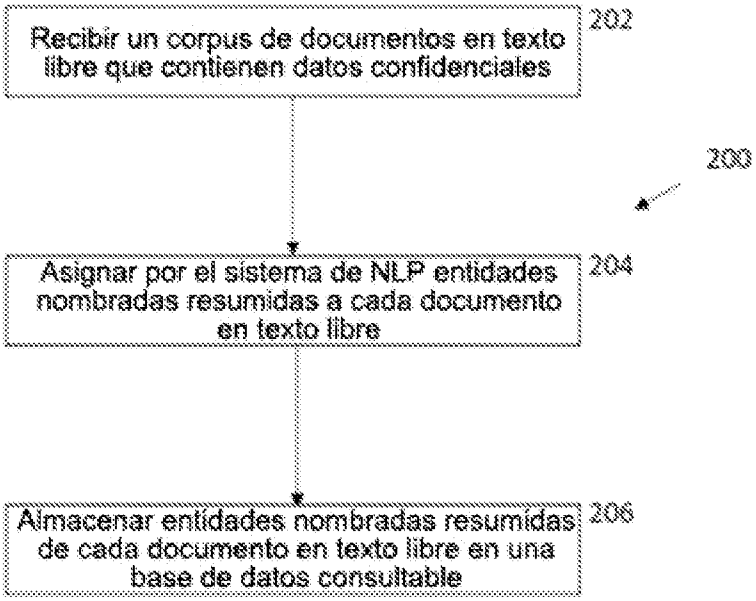
Recibir un corpus de documentos en texto libre que contienen datos confidenciales
202
200
Asignar por el sistema de NLP entidades nombradas resumidas a cada documento en texto libre
204
Almacenar entidades nombradas resumidas de cada documento en texto libre en una base de datos consultable
206

FIG. 2

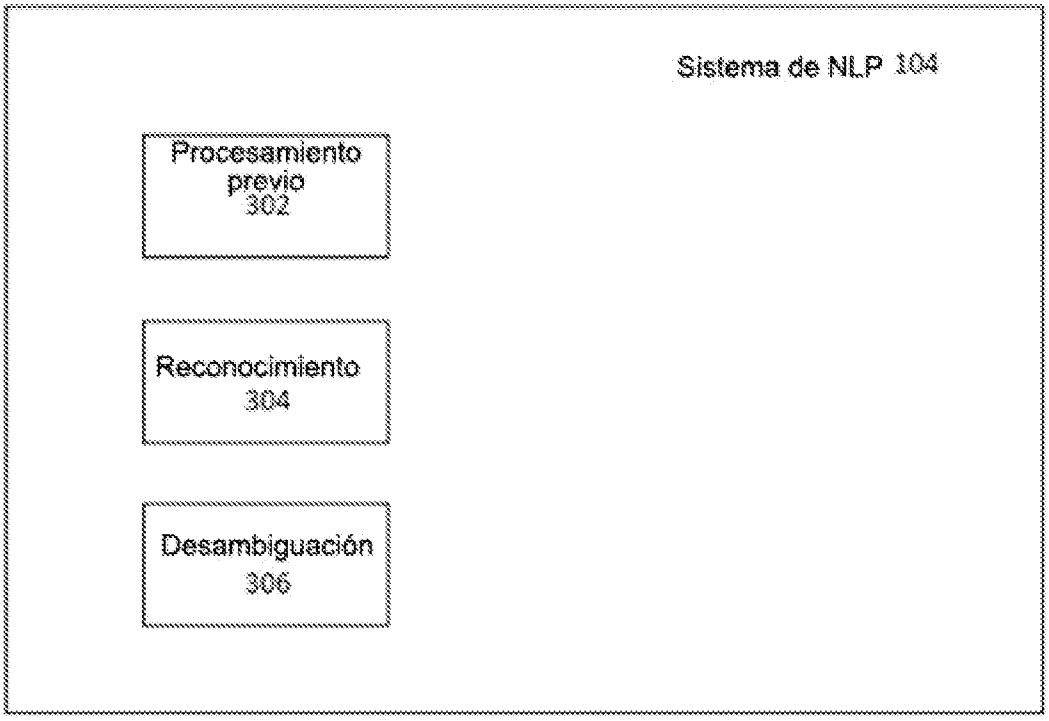
Sistema de NLP 104
Procesamiento previo 302
Reconocimiento 304
Desambiguación 306

FIG. 3

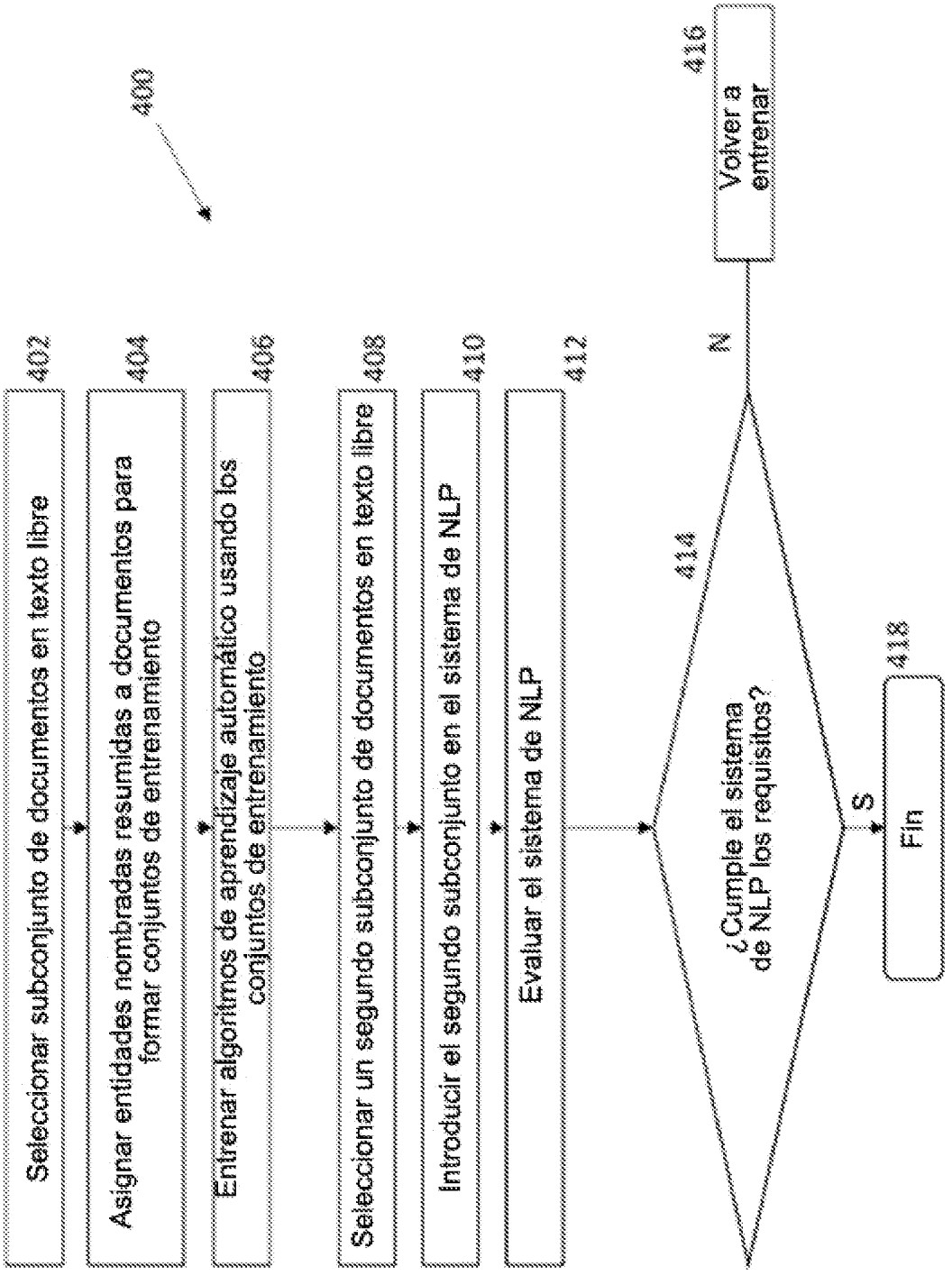
400
Seleccionar subconjunto de documentos en texto libre
402
Asignar entidades nombradas resumidas a documentos para formar conjuntos de entrenamiento
404
Entrenar algoritmos de aprendizaje automático usando los conjuntos de entrenamiento
406
Seleccionar un segundo subconjunto de documentos en texto libre
408
Introducir el segundo subconjunto en el sistema de NLP
410
Evaluar el sistema de NLP
412
¿Cumple el sistema de NLP los requisitos?
414
N
Volver a entrenar
416
S
Fin
418

FIG. 4

Documento 1

Esta mujer blanca de 23 años de edad acude por quejas de [síntoma - alergia] alergia[1]. Solía tener [antecedentes personales - alergia] alergia[1] cuando vivía en Seattle pero cree que ha empeorado aquí. En el pasado ha probado [antecedentes personales - tratamiento] con loratadina[1] y [antecedentes personales - tratamiento] cetirizina[1]. Ambas funcionaron durante poco tiempo pero luego parecieron perder eficacia. También ha usado [antecedentes personales - tratamiento] fexofenadina[1]. La usó el verano pasado y empezó a usarla de nuevo hace dos semanas. No parece que esté funcionando muy bien. Ha usado pulverizaciones sin receta[4] pero no pulverizaciones nasales con receta. Tiene [síntoma - cardiopatía] asma[2] pero no requiere medicamento diario para esto y no cree que esté empeorando.,MEDICAMENTOS:, Su único medicamento actual es norgestimato[3] y [antecedentes personales - tratamiento] fexofenadina[1].

FIG. 5A

Documento 2

Tiene [síntoma - reflujo gastroesofágico] enfermedad por reflujo gastroesofágico[1]., ANTECEDENTES QUIRÚRGICOS ANTERIORES:, Incluye [síntoma - cirugía reparadora] cirugía reparadora[2] en su mano derecha hace 13 años., ANTECEDENTES SOCIALES:, Actualmente está soltero. Bebe aproximadamente diez veces al año. Fumaba[3] significativamente hasta hace varios meses. Ahora fuma menos de tres cigarrillos al día., ANTECEDENTES FAMILIARES:, Cardiopatía[3] en ambos abuelos, abuela con accidente cerebrovascular[3] y una abuela con diabetes[3]. Niega obesidad y hipertensión en otros miembros de la familia., MEDICAMENTOS ACTUALES:, Ninguno., ALERGIA:, Es [síntoma - alergia - penicilina] alérgico a la penicilina[1].

FIG. 5B

Leyenda	
[]	Término de desambiguación
xxx	Identidad nombrada resumida
[1]	Ejemplo de verdadero positivo
[2]	Ejemplo de falso positivo
[3]	Ejemplo de falso negativo
[4]	Ejemplo de verdadero negativo

ES 2 969 343 T3

Documento 3

ENFERMEDAD ACTUAL:, Hoy he visto a ABC. Es un hombre muy agradable de 42 años de edad y 344 libras. Mide 5'9". Tiene [prueba - IMC - 51] un IMC de 51[1]. Ha tenido sobrepeso[3] desde hace diez años desde que tenía 33 años, como máximo ha llegado a pesar 358 libras y como mínimo 260. Está sometiéndose [intervención - cirugia] a intentos quirúrgicos[2] de pérdida de peso para sentirse bien, volverse sano y empezar a hacer [síntoma - asma] ejercicio[2] de nuevo. Estuvo seis meses [antecedentes personales - beber alcohol]sin beber alcohol[2] y sin ingerir demasiadas calorías. Ha seguido múltiples programas de pérdida de peso comerciales[4] incluyendo Slim Fast durante un mes hace un año y dieta Atkins durante un mes hace dos años.

FIG. 5C

Documento 4

2-D, MODO M:, ,1. [síntoma - dilatación de aurícula izquierda] Dilatación de aurícula izquierda[1] con un diámetro de aurícula izquierda de 4,7 cm[3].,2. Ventrículos izquierdo y derecho de tamaño normal[3]., 3. Función sistólica de LV normal[1] con [síntoma - fracción de eyección ventricular - 51%] fracción de eyección de ventrículo izquierdo del 51%[1]., 4. [síntoma - función diastólica de LV normal] Función diastólica de LV normal[1]., 5. Morfología normal de válvula aórtica[3], válvula mitral[3], válvula tricúspide[3] y válvula pulmonar[3].

FIG. 5D

ES 2 969 343 T3

Documento 5

1. El [síntoma - tamaño de ventrículo izquierdo normal][síntoma - tamaño de cavidad de ventrículo izquierdo normal][síntoma - grosor de tabique de ventrículo izquierdo normal] tamaño de cavidad y el grosor de tabique del ventrículo izquierdo parecen normales[1]. El movimiento de tabique y la función sistólica del ventrículo izquierdo parecen hiperdinámicos con una fracción de eyección estimada del 70% al 75%[3]. Se observa [síntoma - obstrucción pulmonar] casi una obliteración de la cavidad[2]. También parece haber [síntoma - aumento de flujo de salida ventricular] un gradiente de tracto de flujo de salida del ventrículo izquierdo aumentado[1] a nivel del centro de la cavidad compatible con función sistólica del ventrículo izquierdo hiperdinámica[3]. Se observa un patrón de relajación del ventrículo izquierdo anómalo[3].

FIG. 5E

Documento 1

[riesgo - contaminación radiactiva] Contaminación radiactiva[1]

Esta Póliza [cobertura por pérdida o daño - negada - reacción nuclear, radiación nuclear] no cubre ninguna pérdida o daño que surja directa o indirectamente de reacción nuclear, radiación nuclear[1] o contaminación radiactiva[3], independientemente de cómo pueda haberse causado tal reacción nuclear, radiación nuclear o contaminación radiactiva * No obstante, si el incendio es un riesgo asegurado y [riesgo derivado - incendio provocado por reacción nuclear] surge un incendio directa o indirectamente de reacción nuclear, radiación nuclear[3] o contaminación radiactiva[3], cualquier pérdida o daño que surja directamente de tal incendio estará cubierto (sujeto a las disposiciones de esta Póliza[4])

FIG. 6A

Documento 2

[riesgo - enfermedad infecciosa] Enfermedad infecciosa[1]

Sin perjuicio de cualquier contenido en el sentido contrario en la Póliza, la cobertura en virtud de la presente no se extiende para incluir lesión[3], enfermedad[3] o muerte[3] de una persona asegurada o cualquier responsabilidad vinculada al Asegurado por la pérdida o daño de propiedad de tercero, lesión, enfermedad o muerte de un tercero como resultado de reclamaciones que surjan directa o indirectamente de, provocadas por, que se produzcan a través de[4], en consecuencia de o de cualquier manera atribuibles a enfermedad infecciosa, [enfermedad infecciosa - gripe aviar] gripe aviar[1] o cualquier enfermedad que se haya declarado epidémica por la [riesgo - riesgo relacionado con la salud] Organización Mundial de la Salud[2].

FIG. 6B

ES 2 969 343 T3

Documento 3

[contenido relacionado legal – arbitraje] Arbitraje[1]

Cualquier [contenido relacionado legal – disputa] disputa[1] que surja de esta Póliza será remitida para la decisión de un [figura jurídica – árbitro] Árbitro[1] que se designará por ambas partes o, si no pueden acordar un único árbitro[1], para la decisión de dos árbitros[1], uno designado por escrito por cada parte (dentro del plazo de un mes[4] después de que cualquier parte requiera por escrito que se haga). Los dos árbitros nombrarán entonces mutuamente un [figura jurídica - árbitro] tercer árbitro[2] al que deberán nombrar por escrito los árbitros[3]. El [figura jurídica - árbitro] tercer árbitro[2] se sentará con los árbitros[3] y presidirá sus reuniones[4]. La toma de una decisión por el árbitro[3], los árbitros[3] o el [figura jurídica - árbitro] tercer árbitro[2] será una condición precedente para cualquier derecho[4] de acción contra Nosotros.

FIG. 6C

Documento 4

[riesgo – hipoteca] Hipoteca[1]

Por el presente se acuerda que, en el caso de cualquier pérdida o daño[4] que esté asegurado por el presente, Nosotros [ejecución de riesgo – pago de hipoteca] pagaremos a los Acreedores hipotecados[1] o a dichos Cesionarios[3] tal como se menciona en el Anexo en la medida de su interés[4] y que este seguro, sólo en la medida en que se refiere al interés en el mismo de los Acreedores hipotecarios o dichos Cesionarios[3], no quedará invalidado por ninguna acción o negligencia del [figura jurídica – deudor hipotecario] Deudor hipotecario[1] o Propietario de los Edificios.

FIG. 6D

ES 2 969 343 T3

Documento 5

[acción de póliza – cancelación unilateral] Cancelación por Nosotros[1]

Tenemos derecho a cancelar esta Póliza otorgándole a Usted [acontecimiento en el tiempo – días - 7] siente (7) días[1] por notificación por [requisito legal – certificación] correo certificado[2] por escrito a Su última dirección conocida. Si se ha realizado una reclamación[1], o se ha notificado[4] un incidente[3] que puede dar lugar a una reclamación, entonces [acción de póliza – negada - reembolso] no se reembolsará[1] la prima. Si no se ha realizado ninguna reclamación[1], entonces le reembolsaremos[3] una prima prorrateada en proporción a la cantidad[4] de tiempo que haya estado en vigor Su Póliza.

FIG. 6E

ES 2 969 343 T3