

Estudio de identificación de autoría

Ciencias de la Computación e Inteligencia Artificial



12 de marzo de 2019

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

Tutores: Luis de Campos Ibáñez y Juan F. Miguel Huete

Índice

[**1.** **Introducción** 2](#_Toc39450630)

[**1.1** **Propósito** 2](#_Toc39450631)

[**1.2** **Partes** 3](#_Toc39450632)

[**1.3** **Justificación** 3](#_Toc39450633)

[**1.4** **Ámbito** 4](#_Toc39450634)

[**1.5** **Relevancia teórica y práctica de la investigación** 4](#_Toc39450635)

[**1.6** **Estado de la cuestión** 5](#_Toc39450636)

[**1.7** **Objetivos, exposición del problema y preguntas de investigación** 6](#_Toc39450637)

[**1.8** **Breve descripción de la metodología de la investigación** 6](#_Toc39450638)

[**1.9** **Esquema** 6](#_Toc39450639)

[**2** **Profundizar sobre la autoría** 7](#_Toc39450640)

[**3** **Objetivo** 7](#_Toc39450641)

[4 Datos 7](#_Toc39450642)

[5 Preprocesamiento 8](#_Toc39450643)

[6 Clasificación 9](#_Toc39450644)

[7 Comparativa 13](#_Toc39450645)

[8 Estudio 13](#_Toc39450646)

[9 Conclusión 13](#_Toc39450647)

[10 Bibliografía 13](#_Toc39450648)

[Iniciativas parlamentarias 13](#_Toc39450649)

[Una Encuesta de Métodos Modernos de Atribución de Autoría 13](#_Toc39450650)

[El efecto del tamaño del conjunto de autores y el tamaño de los datos en la atribución de autoría 13](#_Toc39450651)

[Artificial Intelligence a Modern Approach 13](#_Toc39450652)

# **Introducción**

## **Propósito**

Vivimos un contexto tecnológico sin precedentes conocido como la era de la **información**. Cada vez somos más conscientes del poder que nos proporciona adquirirla y usarla. Por contraposición, al mismo tiempo vivimos una realidad en la que se nos hace más difícil responder de forma eficiente al volumen de datos que se crea, estudios hablan de que solo se almacena menos de un 0,4% de la información que se produce. Datos cada vez más complejos e interconectados, datos que requieren de un preprocesamiento no trivial y en última instancia datos que deben ser inferidos a partir de otros.

Es en este último punto donde el aprendizaje automático, más conocido como *Machine Learning* adquiere un papel clave al enlazarse y apoyarse con otras áreas del conocimiento. Nos permite responder a preguntar que o, por un lado, requerían del esfuerzo y presencia del ser humano o ni siquiera se sabía una respuesta.

Es indiscutible que muchos modelos de *Machine Learning* no tienen competidores actualmente que asemejen en términos de eficiencia y eficacia. Algunos de los ejemplos como:

* Diagnósticos médicos
* Procesamiento del lenguaje natural o *Natural Language Processing*.
* Búsqueda online
* Coches inteligentes
* …

A medida que ha pasado el tiempo la lista de aplicaciones se ha hecho interminable al mismo tiempo que la lista de publicaciones científicas o *papers* relacionadas con el tema ha crecido considerablemente. Este fenómeno, principalmente, se debe a dos motivos:

* Existe un componente **económico** que ha decido apostar fuertemente por una industria relacionada con modelos predictivos. La necesidad de automatizar trabajos siempre ha tenido un rol fundamental en una empresa.
* El acceso a la **tecnología** para implementar, entrenar y validar modelos está prácticamente al alcance de todo el primer mundo. Esto se debe a la apuesta por igual que se ha hecho por la nube o más conocida como la *cloud*, que permite no disponer en local de los recursos en *hardware* necesarios para realizar estas tareas*.* Existe también un componente de desarrollo *software*, que ha posibilitado esta situación, dejando atrás la barrera técnica que encontraban muchos investigadores para testear sus soluciones.

Es una etapa dorada para la aplicación de todos estos conceptos que se han venido desarrollando en el último siglo de forma teórica y que empiezan a ver sus primeros frutos en el presente.

Existen infinidad de formas en las que nos podemos encontrar la información. Es sabido, que la dificultad de la predicción o respuesta a la pregunta que plantee el problema estará sumamente relacionada con el perfil de información que poseamos. Pues aun siendo la misma pregunta el problema es distinto si la entrada es una matriz de características, una imagen o un texto. Para cada una de estas posibilidades se proponen diferentes estrategias para proceder.

En el caso que nos atañe en este trabajo, la información está representada por texto etiquetado por un autor. De las muestras disponibles que hay para un autor se debe inferir las características que lo definen. De este modo, llegado un texto nuevo el modelo predictivo debe discernir en base a las características extraídas a que autor pertenece. Este problema es conocido como identificación de la autoría o *authorship attribution problem*.

A lo largo de la historia se han producido numerosos debates sobre la autoría de obras transcendentales para el conocimiento humano. El hecho de conocer el autor de un contenido da un peso conceptual extra a sus palabras que, apoyado por su biografía y su circunstancia crean un marco decisivo desde el que poder abarcar cualquier estudio. Esto es debido a que a veces podría ser más conclusivo responder a la pregunta, ¿quién desarrolló un contenido? Que el contenido mismo.

El propósito de este trabajo es automatizar la identificación del autor de un documento sobre de un conjunto de autores previamente definido. Se desarrollar y argumentará un estudio completo sobre que es diferencial y que no para etiquetar el autor de una obra textual.

## **Partes**

El trabajo constará de 10 secciones definidas en el índice inicial y fácilmente consultables. El formato electrónico permite navegabilidad sobre el mismo gracias a la inserción de vínculos sobre las entradas. A lo largo del trabajo se pueden encontrar referencias en forma de vínculos al apartado bibliográfico que se encuentra en la última sección de este documento.

## **Justificación**

He decido abordar este trabajo tras la lectura de [una Encuesta de Métodos Modernos de Atribución de Autoría](#_Una_Encuesta_de) (*A Survey of Modern Authorship Atribution Methods*, Universidad de Aegean) y El [efecto del tamaño del conjunto de autores y el tamaño de los datos en la atribución de autoría](#_El_efecto_del) (*The effect of autor set size and data size in Authorship Attribution*). Ambas investigaciones fueron sugeridas por los tutores del proyecto. La lectura del capítulo V y VI del libro Inteligencia Artificial un Enfoque Moderno (Artificial Intelligence A Modern Approach por Stuart Russell y Peter Norvig).

También existe un componente previo personal que me ha llevado a aceptar esta temática. Mi afición por la lectura ha ocupado gran parte de mi vida, haciendo énfasis un especial énfasis en las obras filosóficas, y unido a mi interés por la psicología humana sobre la cual daremos algunas pincelas a lo largo del trabajo.

Desde un punto de vista laboral, mi trabajo actual como Científico de Datos para la Prevención del Fraude y el Crimen Organizado en Deloitte comparte muchas áreas de conocimiento con el tema que nos ocupa. Desde la identificación de nombres sobre listas sancionadas (Watch List Filtering) haciendo uso de emparejamiento por lógica difusa, hasta la identificación y clasificación de alertas sospechosas en los conceptos de las transferencias bancarias. Pues como ya se ha mencionado antes, las aplicaciones de *Machine Learning* inundan el mercado actual. Manteniendo arquitecturas de modelos semejantes en problema aparentemente distintos. *ML* ha dado un paso de gigante en lo que abstracción de problemas se refiere.

Desde el punto de vista de las ciencias de la computación y la inteligencia artificial rama del conocimiento que estudio, este problema hace uso de muchos de los conceptos y herramientas que en ellas se explican. Debido a que ha sido una decisión propia el estudio de esta especialidad, es razonable la elección del proyecto.

## **Ámbito**

El problema de la Identificación de Autoría o la Autoría de Documentos es una cuestión interdisciplinar que comparten actualmente lingüística, recuperación de información o *information retrieval* e inteligencia artificial. Siendo estos dos últimos desde los que se va abordar el trabajo. Nos adentrándonos en la subrama aprendizaje automático perteneciente a la inteligencia artificial.

Podemos decir que los tres ámbitos específicos que predominan en este trabajo son el procesamiento del lenguaje natural, la recuperación de información y el aprendizaje automático. A continuación, daremos una definición básica de ambos conceptos:

El procesamiento del lenguaje natural (NLP) es un rango teóricamente motivado de técnicas computacionales para analizar y representar textos que ocurren naturalmente en uno o más niveles de análisis lingüístico con el fin de lograr el procesamiento del lenguaje humano en una variedad de tareas o aplicaciones.

La Recuperación de información es la conversión de grandes volúmenes de texto en estructuras simplificadas y comprensibles para su uso posterior.

El Aprendizaje automático es el estudio de algoritmos de computación que mejoran automáticamente por medio de la experiencia.

Si quisiéramos profundizar más, existen números subcategorías dentro del problema al que nos enfrentamos, dependiendo del registro del lenguaje, permisibilidad de faltas de ortografía, tamaño de los textos, número de autores, número de muestras… Más adelante concretaremos estos conceptos en el apartado 2.

## **Relevancia teórica y práctica de la investigación**

[resumen del artículo de la encuesta y algo más]

Relevancia teórica

{Explicar los posibles avances dentro del campo }

A continuación, se presentan algunas aplicaciones que tiene el problema:

* Desde la literatura nos encontramos con debates actuales sobre la identificación de autores de obras anónimas o puesta en duda de obras que fueron atribuidas sin estudio previo. Pero la realidad es que muchas obras continúan anónimas actualmente ya sea por deterioro de la misma o falta de candidatos.
* Un ejemplo de la primera es la conocida obra del Lazarillo de Tormes que tras una lista de candidatos se resolvió su autoría gracias a un estudio lingüista reconociendo a Sebastián de Horozco como escritor.
* Un ejemplo de la segunda es el debate que hubo en torno a las obras de Shakespeare debido a su estilo impropio de la cuna del autor y a las lagunas de su biografía.
* Desde la criminología, derecho, psicología y psiquiatría nos encontramos considerable número de problemas.
* Peritaje de conversaciones electrónicas.
* Autoría de cartas de suicidio.
* Falsificación en las relaciones laborales.
* Falsificación de estudios (TFGs y TFMs).
* Identificación de trastornos.
* Psicología evolutiva y del aprendizaje.
* Desde la automatización de tareas, como es nuestro caso, tenemos algunos ejemplos.
* Documentación automática (nuestro caso)
* Detección de suplantaciones de identidad

## **Estado de la cuestión**

El primer estudio que se realizó sobre la materia fue en 1887 sobre las obras de Shakespeare, publicado por Thomas Corwin Mendenhall. Seguido medio siglo después por los trabajos estadísticos de Tule (1938:1944) and Zipf (1932). Aunque sin duda el trabajo más notorio y reconocido es el estudio realizado por Mosteller and Wallace (1964) construido a partir de ‘The Federalist Papers’, un conjunto de 146 documentos de longitud variada escritos 3 autores diferentes. El método usado en este último estudio fue un modelo Bayesiano estadístico centrado sobre un grupo de palabras comunes en inglés.

Antes de este estudio, la capacidad de diferenciar autores se veía desde un punto de vista lingüístico dependiente del estilo literario de cada uno. Se definió un conjunto heterogéneo de mediciones, aproximadamente llegaron a ser unas 1000, sobre propiedades concretas y triviales del uso del lenguaje. Algunas de estas como la frecuencia de palabras por frase, la frecuencia de caracteres por frase o el uso de palabras poco frecuentes.

La metodología de trabajo estaba bastante limitada tanto por los medios disponibles en el momento como por los problemas que se planteaban. La mayor parte de estos problemas cumplían las siguientes características:

* El documento a analizar usualmente se trataba de una obra completa o libro.
* El número de autores sobre los que se realizaba el estudio era pequeño. Aproximadamente 2 o 3 autores.
* Existía un alto componente subjetivo en la evaluación de los métodos propuestos.
* La decisión de que método era el más apropiado para un problema partía de una ausencia notable de un banco de problemas resueltos.

A partir de la revolución de Internet la metodología tradicional aplicada sobre los nuevos problemas se quedó obsoleta debido a la diversificación y volumen de problemas. Es en este punto, cuando el procesamiento del lenguaje natural (natural language processing) en conjunto con aprendizaje automático y recuperación de información se imponen como áreas del conocimiento y modelos de trabajo para la identificación de la autoría.

Desde la recuperación de información se desarrollaron técnicas eficientes para la representación y clasificación de grandes volúmenes de información.

Desde el aprendizaje automático con el desarrollo de algoritmo capaces de trabajar con problemas de dimensión. En 1992 con la publicación realizada por E. Boser, Isabelle M. Guyon y Vladimir N. Vapnik en la que sugirieron la aplicación del modelo machina de vectores soporte sobre kernel no lineales.

Desde el procesamiento del lenguaje natural con el desarrollo de herramientas eficientes que analizaran características del lenguaje.

## **Objetivos, exposición del problema y preguntas de investigación**

## **Breve descripción de la metodología de la investigación**

## **Esquema**

Que es el problema de la Autoría de Documentos (Autorships).

La iniciativa parlamentaria permite a un diputado o a un senador presentar al parlamento un proyecto de artículo constitucional, de ley o de decreto. Este proyecto puede ser redactado de modo completo o formulado en términos generales. La comisión de la cámara donde ha sido depositada la iniciativa decide si es necesario darle curso. Por ejemplo, una iniciativa sobre un tema que ya está en discusión en el parlamento no será declarada válida. Si la comisión considera que la iniciativa puede ser acogida, el proyecto sigue el itinerario legislativo clásico (examen por parte de la comisión de la otra cámara, procedimiento de consulta, cámaras del parlamento, etc.)

Desde 1978 se recogen las iniciativas parlamentarias, pero nosotros nos centraremos 2008.

Fuente de las iniciativas parlamentarias:

Lenguaje parlamentario, corpus

Lenguaje oral cuidado

Lenguaje político de controversia, trufado por citas generales (con caracteres atractivos) función de alusión al otro. Estructura: argumentación, muy estructurado.

Extracción de características:

1. Características generales
2. Características lingüísticas
3. Características temáticas
4. Características a nivel de palabras

Explicar matriz dispersa

Explicar función logarítmica para la

# **Profundizar sobre la autoría**

# **Objetivo**

Estos son los objetivos del proyecto:

1. Realizar un **preprocesamiento** correcto de los **datos** para que puedan ser utilizados como entrada en un modelo de aprendizaje automático.
2. Proponer diversos **modelos** entrenados de clasificación multietiqueta que presenten buenos resultados.
3. Validar los modelos y compararlos entre sí.
4. Definir un criterio justificado para elegir el mejor modelo y presentarlo en un ambiente de producción.

# Datos

La fuente de datos utilizada son la Iniciativas Parlamentarias del Senado durante el año 2008. Estas iniciativas están recogidas en 5260 ficheros en formato XML. El formato XML presenta la siguiente estructura:

…

…

…

El

El **nombre**

En el esquema explicado se puede observar que

En nuestro caso los datos los hemos extraídos de las [iniciativas parlamentarias del congreso de los diputados del año 2008](#_Iniciativas_parlamentarias) que presentan mayormente un registro formal, una baja permisibilidad de faltas de ortografía, un tamaño medio reducido de los textos, entorno autores y con muestras.

Estrictas reglas

Géneros parlamentarios

Subjetividad (propaganda) VS Información (Debate intrínseco)

Extraccion del tema

Que el texto sea argumentativo significa que usa palabras del contrario en las respuestas. Para demostrar esto dar la matriz de intersección

# Preprocesamiento

Explicar preprocesamiento de los XML

-Eliminar palabras especiales

-Agrupar párrafos en un solo documento

# Clasificación

**Bolsa de Palabras**

La estructura de datos o representación Bolsa de Palabras (Bag of Words) simplifica de forma eficiente el uso de texto en Aprendizaje Automático. Simplemente se base en realizar un conteo de las apariciones de una palabra a lo largo del corpus. Esto descarta gran parte de la complejidad que aportan las estructuras del lenguaje como capítulos o párrafos. Por lo tanto, llamaremos bolsa a la distribución de las frecuencias absolutas de las palabras en el corpus.

probabilidad de la secuencia de caracteres en desde hasta

Para computar la bolsa de palabras se siguen los siguientes pasos:

1. Token: definir la expresión que reconocerá una palabra para nuestro corpus. Ejemplo: una palabra es un conjunto de caracteres rodeados por espacios.
2. Tokenización: dividir cada documento en función del criterio que se haya definido en el token.
3. Construcción del vocabulario: realizar la unión algebraica de todos los conjuntos extraídos por la tokenización de cada documento.
4. Codificación: realizar el conteo de las apariciones de las palabras del vocabulario sobre los documentos.

Usualmente para aumentar la probabilidad de emparejamiento de palabras y reducir la bolsa de palabras se siguen algunas pautas:

* Unificación de mayúsculas minúsculas: se realiza una transformación sobre los caracteres del corpus pasándolos todos a minúscula, por ejemplo. En algunos corpus en los que abunden muchos nombres propios o de organizaciones puede llevar a un encarecimiento de la información extraía. El apellido “Pino” y el árbol “pino” pasarían a ser el mismo token dentro de la bolsa.
* Lematizador: mediante un algoritmo se extrae la palabra de la que deriva la original. Algoritmos como *WordNetLematizer* o *Lancaster* entre otros nos ayudan a realizar esa función. Este tipo de algoritmos parten de la idea de que existen agrupaciones de palabras que describen grafos dirigidos con grandes sumideros. Aprovechar esos sumideros a costa de perder cierto grado de información puede ser interesante si aumenta el número de emparejamientos.
* Derivación: es una operación con el mismo objetivo que el lematizador. La principal deferencia, es que en este caso se realiza una extracción de la raíz.
* Eliminar acentuación: es una práctica muy común en los idiomas que provienen del latín.

Con el objetivo de eliminar palabras extremadamente poco frecuentes y muy frecuentes se definen dos umbrales:

* Frecuencia de documento mínima: se trata de la frecuencia relativa mínima que admitimos para considerar una palabra dentro de la bolsa de palabras.
* Frecuencia de documento máxima: se trata de la frecuencia relativa máxima que admitimos para considerar una palabra dentro de la bolsa de palabras.

Un tipo de umbral con el mismo propósito es el que limita el tamaño del vocabulario. Usualmente se seleccionan las tokens más frecuentes encontrados.

Una de las principales desventajas que tiene esta estructura es que perdemos la información de la posición que tiene cada palabra con respecto al resto. Esta propiedad tiene una relación directa con el lenguaje, pues puede no tiene el mismo significado una palabra al principio de una frase, en medio o al final de la misma. Este significado dependiente de la posición debemos trasladárselo al modelo. Una de las formas más frecuente de hacerlo es partir del parámetro *N-gram*. Consideras un conjunto de palabras extra que cumple las mismas propiedades anteriormente descritas pero que define su token como unión de 2 o más palabras dependiendo del rango que fijemos. Los valores frecuentes para el rango de *N-gram* van desde *1-gram* hasta *3-gram* inclusive. Esto se debe a que se experimenta una fuerte disminución de la frecuencia cuando se usan valores mayores de 3.

Este último punto ha adquirido tal importancia que ha conformado un concepto de modelos dentro de la práctica.

Aplicación

Para el conjunto de datos que nos atañe siempre se ha seguido un análisis o tokenización a nivel de palabras. En la siguiente tabla se puede ver las 10 más frecuentes de todo el corpus:

|  |  |
| --- | --- |
| Palabra | Frecuencia absoluta |
| *de* | 700.842 |
| *que* | 518.938 |
| *la* | 389.484 |
| *en* | 321.489 |
| *el* | 272.735 |
| *se* | 223.706 |
| *los* | 158.813 |
| *no* | 141.399 |
| *es* | 128.702 |
| *del* | 116.744 |

Para este análisis solo se han eliminado los acentos y se han pasado todas las palabras a minúscula, dando un total de 64.259 palabras o características *1-gram*. Se puede ver a continuación un diagrama de barras de las 25 primeras ordenadas por frecuencia:

**Modelos *N-gram***

Un *N-gram* se puede definir como una cadena de Markov de orden , puesto que la probabilidad de sobre el carácter depende exclusivamente de los caracteres que le preceden. Por lo tanto, para un *3-gram*:

Los modelos *N-gram* son usualmente aplicados de 3 formas dependiendo la tokenización que se realice sobre el corpus:

* Realizado a nivel de **palabras** que se dividen por espacios o signos de puntuación. Los valores usuales no sobrepasan *3-gram*.
* Realizado a nivel de caracteres. En estos casos el rango está más abierto dependiendo de la aplicación que se le dé al modelo.
* Realizado a nivel de caracteres dentro de una misma palabra. Este caso es más frecuente y común que el anterior, ya que, no suele tener mucho sentido que la cadera del vocabulario atraviese dos palabras diferentes de un texto.

Normalmente dentro de estos modelos basados en bolsas de palabras si se da el caso de un *N-gram*, los *(N-i)-gram* que lo contienen también forman parte del vocabulario. Dando la posibilidad de detectar token concretos de forma separada, pero sabiendo que no se presentan juntos.

**Vectorización de documentos**

El proceso de vectorización de un conjunto de documentos se resume en los siguientes puntos:

1. Construcción de un vocabulario a partir del corpus.
2. Siendo el número de componentes del vocabulario , para cada documento extraemos su vector correspondiente que recoge las frecuencias de cada uno de los token en dicho documento.

Un ejemplo de aplicación de este proceso sobre la siguiente frase:

Construcción del vocabulario:

|  |  |
| --- | --- |
| Palabra | Frecuencia |
| *que* | 2 |
| *es* | 1 |
| *lo* | 1 |
| *quieres* | 1 |

Vectorización del documento:

Cuando se realiza este proceso sobre un número ingente de documentos el vocabulario que se obtiene genera vectores individuales con una alta frecuencia de valores 0. El conjunto de estos vectores se conforman en una matriz siendo el número de documentos y el tamaño del vocabulario. Esta matriz para ser almacenable en memoria requiere de una estructura de datos especial llamada **matriz dispersa**. Las matrices dispersas son un conjunto de ternas donde son los índices de posición de la matriz y el valor que toma esa posición. El resto de combinación de índices no recogida se consideras 0.

El uso de matrices dispersas tiene como ventajas:

* Disminución en el uso de memoria
* Disminución del tiempo consumido de cómputo para ciertas operaciones

Pero también presenta una desventaja notable:

* Muchos de los modelos de Aprendizaje Automático no aceptan representaciones dispersas de las características, obligándonos a realizar entrenamientos parciales o incrementales tras una posterior transformación a matrices densas. Algunos modelos tampoco aceptan entrenamientos parciales.

**Función TF-IDF**

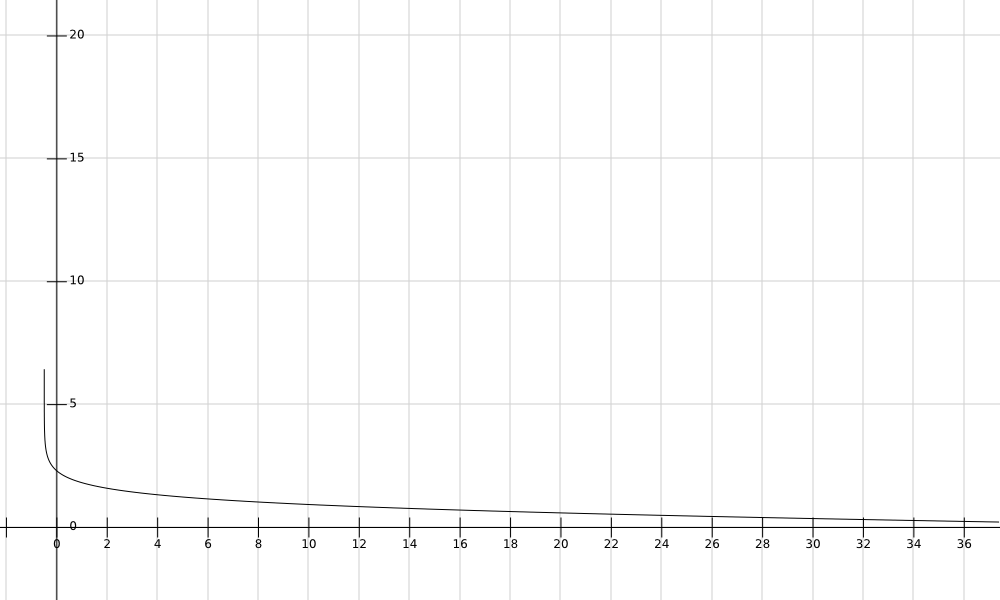
Es un uso más que extendido dentro del proceso de extracción de características sobre texto la aplicación de TF-IDF tras la vectorización de nuestro corpus.

TF hace alusión a la frecuencia de un término dentro del corpus mientas que IDF refleja la inversa de frecuencia sobre el documento.

El objetivo TF-IDF es mezclar en una misma coordenada del documento la importancia del token sobre el corpus y sobre el documento. Hasta ahora el proceso de vectorización solo recogía propiedades elegidas por el corpus con información exclusivamente del documento.

Definición de IDF:

Donde es el número de documentos en el corpus y el número de documento que contienen . Analicemos el comportamiento de esta función mediante la siguiente gráfica para un valor de :



Es observable que conforme aumenta el número de documentos que contienen el valor de IDF cae logarítmicamente decayendo su importancia.

La función TF-IDF es un función en dos variables que se aplica sobre el documento el token y que tiene la siguiente expresión:

Donde es el número de veces que aparece el token en el documento , es la media de tokens que aparecen por documento, y usualmente.

Un punto interesante de esta función es asignar el valor 0 donde el proceso de vectorización asigne un 0. De otro modo, la eficiencia en memoria de la representación dispersa de la matriz se volvería en nuestra contra. Esto fácilmente comprobable gracias a la multiplicación del numerador de la fracción.

**ANOVA F-Valor**

Se trata de un método de selección de características numéricas estadístico que requiere etiquetado. ANOVA compara la variabilidad entre grupos para determinar si hay alguno que sea significativamente diferente.

Este test se realiza para cada una de las características extraídas por el modelo *N-gram* obteniendo las que presente mejores resultados en el test.

**Modelo MLP**

Uno de los modelos más conocidos dentro de Aprendizaje Automático son las redes neuronales artificiales. Un modelo inspirado en el comportamiento del cerebro humano que pretenden simular el intercambio de información entre las **neuronas**.

Las redes neuronales se estructuran en capas que forman un grafo dirigido que parte desde los inputs de las características hasta los outputs del etiquetado. Las capas intermedias que no son ni inputs ni outputs se les llama capas ocultas. Las capas están formadas por neuronas. Cada capa puede tener un número deferente de neuronas.

Usualmente a los nodos del grafo se les llama unidades. Al enlace que une la unidad con la unidad se le asocia un peso . Todas la unidades están conectadas con una

La estructura de datos que encapsula este modelo es el grafo dirigido. Realizando una gran abstracción existen 3 tipos de neuronas:

* Neurona entrada: es aquella que transmite una característica y actúa como entrada de información al modelo.
* Neurona salida: es aquella que etiqueta el resultado final de la predicción en función de la información recibida de un subconjunto de neuronas.
* Neurona oculta: es aquella que tras recibir información de un subconjunto de las neuronas y realizar operaciones con dicha información transmite a un subconjunto nuevo de neuronas.

La función de activación

Dropout

**Modelo SVM**

**Reducción LSD**

**Capas Embebidas**

**Capas LSTM**

**Capas Convolucionales Separables 1D**

**Capas Agrupamiento Máximo 1D**

Explicar 2 modelos:

* SGDClassifier
* Random forest classifier
* Support Vector Machine

Grafica de separación para cada uno generada

# Comparativa

Definir score

Balanceados y no balanceado

Ajuste de hiperparamentros

Graficas sobre hiperparametros

# Estudio

# Conclusión

# Bibliografía

## [Iniciativas parlamentarias](http://www.congreso.es/consti/constitucion/indice/sinopsis/sinopsis.jsp?art=87&tipo=2)

## [Una Encuesta de Métodos Modernos de Atribución de Autoría](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.440.1634&rep=rep1&type=pdf)

## [El efecto del tamaño del conjunto de autores y el tamaño de los datos en la atribución de autoría](https://www.researchgate.net/publication/220675445_The_effect_of_author_set_size_and_data_size_in_authorship_attribution)

## Artificial Intelligence a Modern Approach

<https://surface.syr.edu/cgi/viewcontent.cgi?referer=https://scholar.google.es/&httpsredir=1&article=1019&context=cnlp>

Boser, Bernhard E.; Guyon, Isabelle M.; Vapnik, Vladimir N. (1992). ["A training algorithm for optimal margin classifiers"](http://www.clopinet.com/isabelle/Papers/colt92.ps.Z). *Proceedings of the fifth annual workshop on Computational learning theory – COLT '92*. p. 144. [CiteSeerX](https://en.wikipedia.org/wiki/CiteSeerX_(identifier)) [10.1.1.21.3818](https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.3818). [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1145/130385.130401](https://doi.org/10.1145%2F130385.130401). [ISBN](https://en.wikipedia.org/wiki/ISBN_(identifier)) [978-0897914970](https://en.wikipedia.org/wiki/Special:BookSources/978-0897914970).