Reconocimiento autómatico de autor vía aprendizaje de máquina

Dr. Gerardo Eugenio Sierra Martínez[1], Tonatiuh Hernández García[1,2]

[1]Grupo de Ingeniería Lingüística, Instituto de Ingeniería, UNAM
[2]Lengua y Literatura Hispánicas, UNAM

1. Introducción

La detección automática de autor es un desafío para el Procesamiento de Lenguaje Natural: es útil para detectar plagios y atribuir la autoría de textos anónimos. Actualmente, se requiere medir hasta 26 características estilométricas (SAUTEE, UNAM) para detectar la similitud entre textos, un proceso largo y computacionalmente demandante.

Proponemos un sistema automático de reconocimiento de autor que utiliza seis categorías estilométricas para identificar un número variable de autores con una precisión del 98%, mejorando el estado del arte de 95% de (Grieve, 2007).

Nuestro sistema es escalable, permite detección multiautor, reduce el número de características y los recursos computacionales empleados, y no es afectado por la temática de los textos.

2. Extracción de características: la huella digital escrita

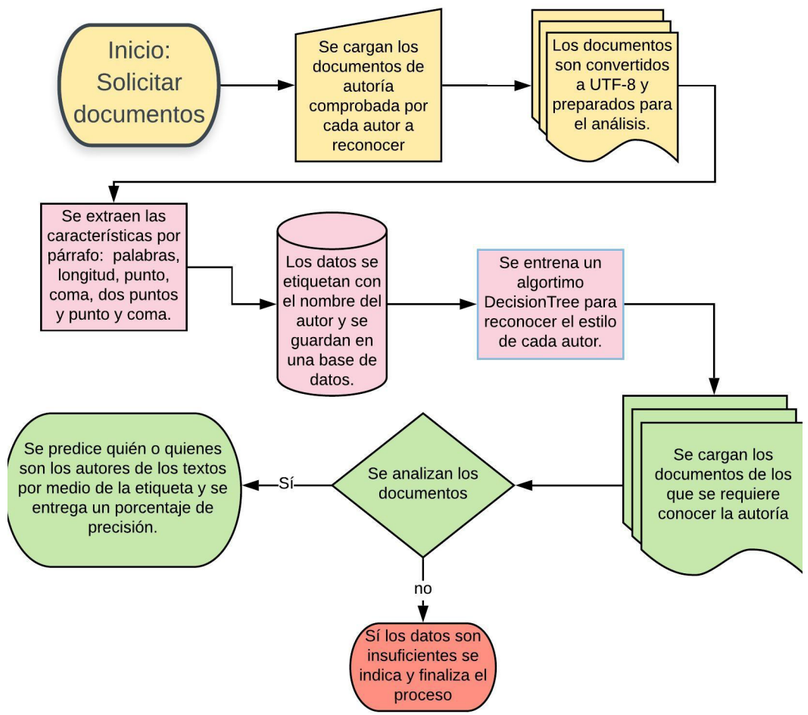
Se utilizan textos de autoría comprobada pertenecientes a los autores por detectar. Entre más textos y mas variedad de géneros, más exacto será el reconocimiento. Nuestro algoritmo aprovecha la segmentación natural del texto en párrafos para crear una unidad de análisis.

Las características que hemos seleccionado son frecuentes, fácilmente cuantificables, inmunes al control consciente del autor, y ultra precisos, funcionan como una huella digital escrita multidimensional que es única para cada persona.

La extracción de características es sencilla: de los textos de autoría comprobada se obtienen las ocurrencias por párrafo de cada característica, se etiqueta con el nombre del autor y se almacenan los datos.

Características por párrafo (conteo de ocurrencias)					
.	,	:	;	Longitud	Palabras

3. Flujo de trabajo

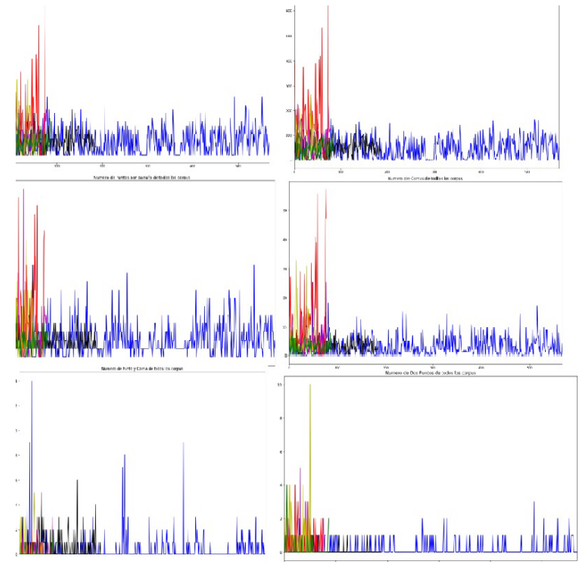


4. Entrenamiento

Con los datos por autor, entrenamos un algoritmo Decision Tree para reconocer la huella digital escrita de cada autor, un espacio de multimensional único para cada persona, pues la combinación de elementos estilométricos que usamos reflejan estructuras sintácticas inconsientes.

5. Experimento

Para poner a prueba nuestro método, hicimos experimentos con ¡El Móndrigo! un texto anónimo escrito por la policía para desprestigiar el movimiento estudiantil del 68. Utilizamos textos de autoría probada de los cinco autores sospechos para correr nuestro método.



6. Resultados

En las gráficas, visualizamos el estilo escrito, una por cada característica: en azul vemos el texto anónimo, en negro el autor responsable de su escritura, en otros colores el resto de los autores.
El algoritmo identificó al autor con el 98% de precisión, su funcionamiento se visualiza abajo. Este es un excelente método de identificación de autor.