Tarea 3 - Identificación de autores por técnicas de agrupamiento

Claudia Lissette Gutiérrez Díaz
Licenciada en Ciencias Computacionales, FCFM, UANL

1. Introducción

A través de los reportes anteriores hemos ido descubriendo el cómo se le puede sacar provecho a los textos, analizar a profundidad el modo de escritura de un autor con uno o varios de sus libros y convertirlo a una estadística descriptiva de tal manera que nos pueda orientar en una decisión de si adquirir o no un libro.

Lo que abordaremos en este reporte es el uso de la técnica de aprendizaje automático para poder identificar al autor por las estadísticas descriptivas que los caracterizan. Se analizarán los capítulos individuales de cuatro libros de diferentes autores y temáticas, se extraerán las características descriptivas de cada uno y se realizará aprendizaje automático de entrenamiento y prueba para poder identificar, por medio de las estadísticas obtenidas, a qué autor pertenece.

2. Descripción de los datos

Para efectos de este reporte, hemos decidido usar cuatro libros de diferentes autores, dos de ellos, cubren la misma temática de fantasía y aventura, los otros dos, tienen temática diferente.

El primero de ellos es el que hemos utilizado en reportes anteriores, *El Reino del Fuego*, por el autor *Neil D'Arc Pridh*, caracterizado por sus libros de fantasía y aventura. El segundo, de temática similar, añadido un poco de misterio, es el libro de *Siete esqueletos decapitados*, escrito por *Antonio Malpica*. Pusimos un libro de temática similar con la intención de observar el comportamiento del agrupamiento, de si, por ser de la misma temática, las estadísticas son similares. También añadimos que el autor es mexicano, por lo que la forma de escritura será interesante de analizar. El tercer libro, cambia totalmente la temática ya que se trata de un thriller psicológico, hablamos de *El hombre equivocado*, por el autor estadounidense *John Katzenbach* y traducido por el español *Rafael Marín Trechera*, con este libro añadimos el cambio de temática, el factor traducción de un libro y que este traductor no sea mexicano, esperamos que al tener tantas diferencias con respecto a los dos anteriores el modelo pueda identificarlo con facilidad. Por último, el cuarto libro corresponde a la temática de realismo narrativo, hablamos de *Cien años de soledad*, por el autor colombiano *Gabriel García Márquez*, con este libro no abordamos el factor de la traducción, pero sí el factor de no ser mexicano, por lo que el uso de las palabras propias de la región, al igual que con el traductor español, pueden hacer diferencia con respecto a los dos autores mexicanos, también consideramos el cambio de temática, por lo que la unión de estos dos factores pueden ayudar a identificar fácilmente el libro.

A continuación, incluimos la sinopsis de los cuatro libros mencionados:

- ***El Reino del Fuego***: Diversos eventos provocan que una misteriosa mujer pelirroja pierda sus memorias. Al despertar, se encuentra en un mundo el cual ella, a pesar de no recordar absolutamente nada, sabe que es totalmente ajeno al lugar a donde pertenece. Aventurándose a una nueva tierra desconocida, deberá enfrentar extravagantes retos para recuperar sus recuerdos y conocer su nuevo hogar el cual constantemente parece amenazarla.
- ***Siete esqueletos decapitados***: ¿Cuánto miedo puedes soportar, Mendoza? Sergio no lo sabe. Pronto descubrirá que es necesario conocer el verdadero terror para resolver el misterio de unos horribles asesinatos, comprender su destino y, a la vez, salvar su propia vida.
- ***El hombre equivocado***: Ashley Freeman, estudiante de Historia del Arte de la Universidad de Boston, tiene una relación de una noche con un desconocido llamado Michael O'Connell, que será el inicio de una pesadilla. El encuentro resultará funesto, ya que O'Connell demostrará pronto ser un psicópata obsesionado con controlar la vida de Ashley. La novela se centra en gran medida en la ingeniosa batalla de Ashley y su familia para acabar con O'Connell.
- ***Cien años de soledad***: Entre la boda de José Arcadio Buendía con Amelia Iguarán hasta la maldición de Aureliano Babilonia transcurre todo un siglo. Cien años de soledad para una estirpe única, fantástica, capaz de fundar una ciudad tan especial como Macondo y de engendrar niños con cola de cerdo. En medio, una larga docena de personajes dejarán su impronta a las generaciones venideras, que tendrán que lidiar con un mundo tan complejo como sencillo.

3. Metodología y resultados

Basándonos en los reportes anteriores en el buen resultado que obtuvimos, seguiremos usando la librería `Spacy` para el análisis de los textos. Para añadir más características que entren en la agrupación, se añade la librería `textdescriptives`, utilizado comúnmente para extraer estadísticas descriptivas de textos, es más usado en el idioma inglés, pero para nuestros propósitos nos será de mucha utilidad.

A los cuatro libros se les hizo pre-procesamiento, se les quitó el contenido inicial, prólogo y agradecimientos, al no analizar propiamente el texto, no se excluyeron palabras ni se hizo lematización de éstas. Una vez limpiados, se dividieron por capítulos y se les asignó un grupo, el cual correspondía al número de libro. Este agrupamiento inicial servirá para las validaciones finales y no con el propósito de etiquetar los registros.

Para cada capítulo se extrajeron 49 características utilizando las librerías antes mencionadas, entre ellos, el sentimiento positivo, negativo o neutral, el tipo de sentimiento como felicidad, enojo, disgusto, tristeza, sorpresa, etc., la cantidad de frases, palabras, palabras únicas, proporción de palabras únicas, letras; media, mediana y desviación estándar de enunciados, sílabas, palabras; también se usó como medición el coeficiente de facilidad de lectura, las dependencias de palabras, proporciones de tipos de palabras, como adjetivos, verbos, adverbios, números, pronombres, etc.; así como el nivel de coherencia de los capítulos, entre otros.

El modelo de clasificación que utilizaremos es el `K-Medias`, ya que nuestro propósito es saber diferenciar a los autores sin saber exactamente a qué autor corresponde el capítulo del libro. Asignamos la cantidad de cuatro grupos directamente, así que omitimos la parte de calcular el número óptimo de agrupaciones. También para efectos visuales usamos el `Análisis de componentes principales`, para re-dimensionar el contenido y pueda ser observado en gráficas.

En la figura 1 y figura 2, podemos comparar, el resultado de aplicar el modelo de `K-Medias`, a los capítulos ordenados aleatoriamente y sin etiquetar, contra la realidad, respectivamente. Observamos a simple vista que la categorización hecha por el modelo y la realidad son bastante similares.

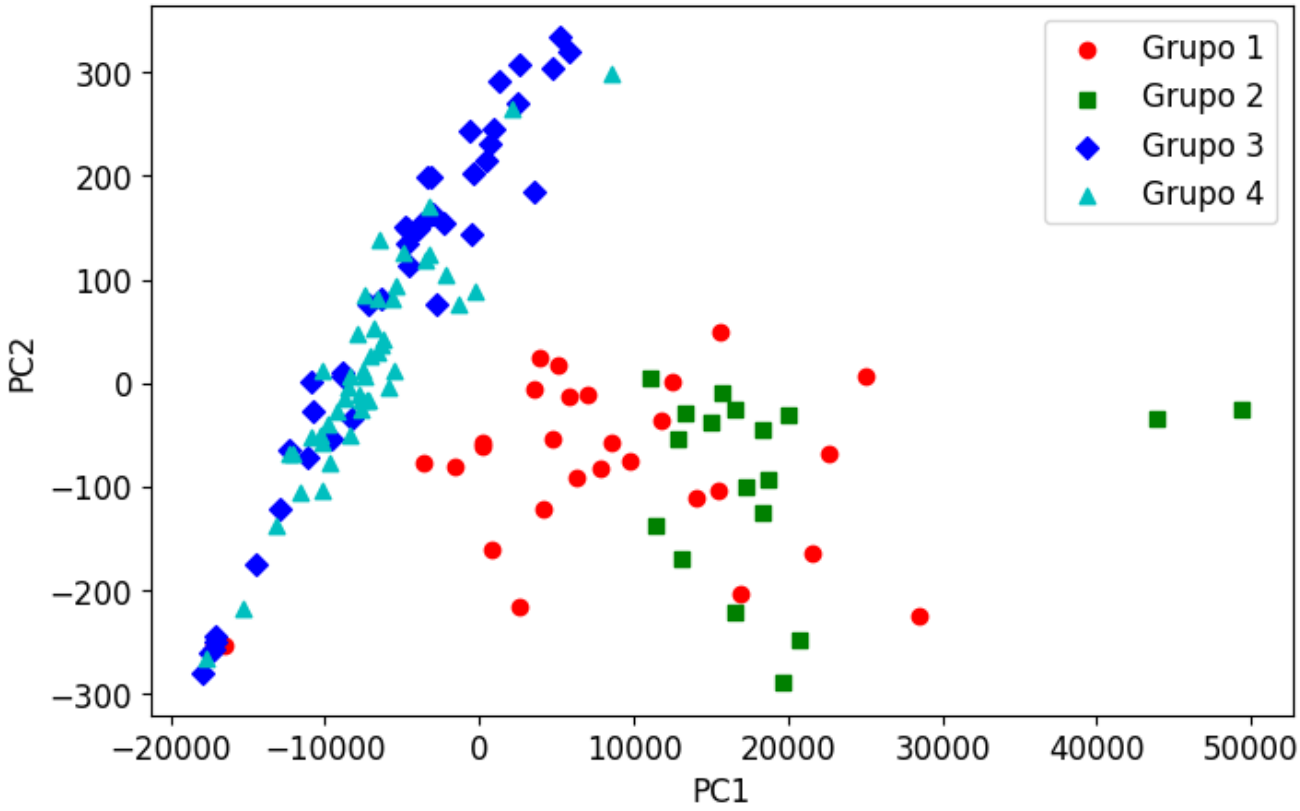

Figura 1. Distribución de grupos por medio de `K-Medias`.

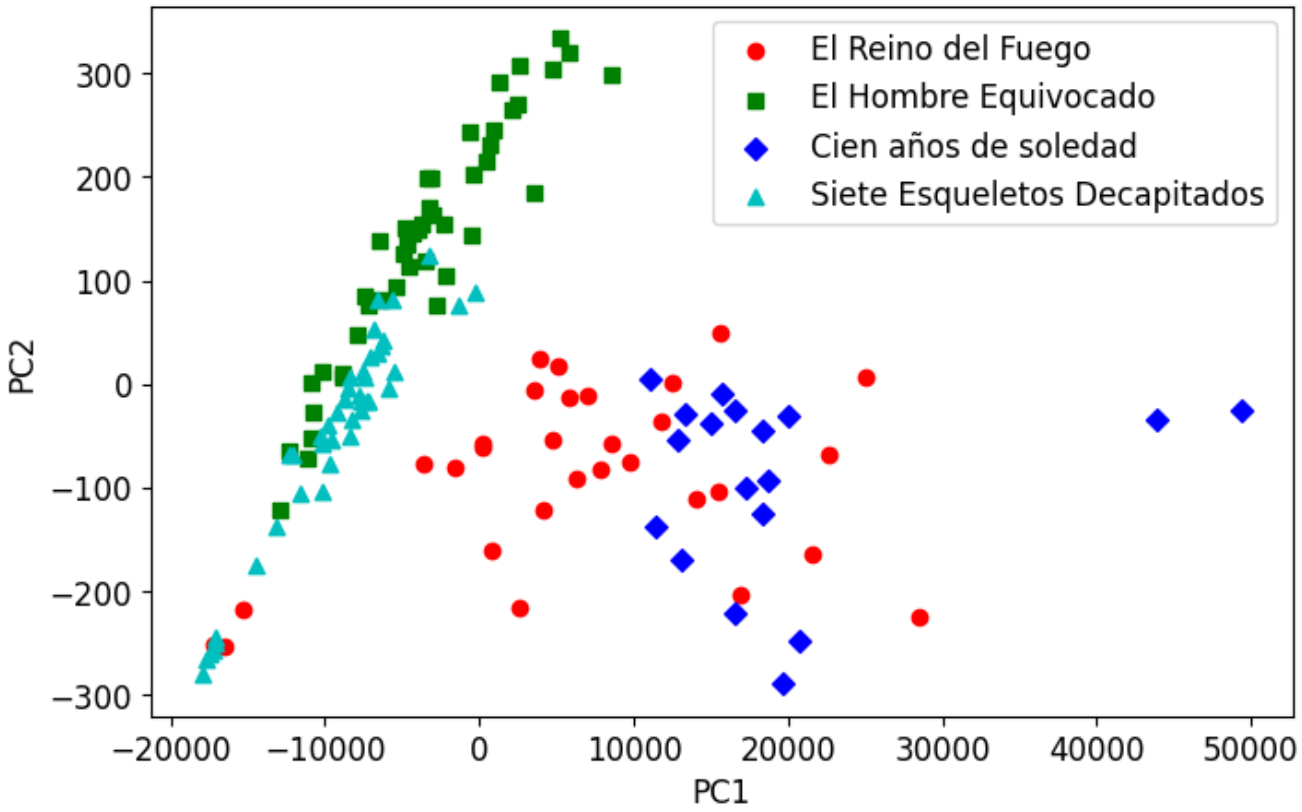

Figura 2. Distribución de grupos con el libro real asignado a cada capítulo

En la tabla 1, se puede observar con mejor detalle la asignación de cada capítulo de los libros entre los diferentes grupos, vemos que todos los capítulos de *Cien años de soledad*, fueron asignados a un solo grupo, así como el libro *El Reino del Fuego*, salvo por un capítulo, fueron asignados al grupo 1. Mientras que los libros *El hombre equivocado* y *Siete Esqueletos Decapitados*, fueron mezclados entre los grupos 2 y 3, por lo que podemos intuir que la temática es incorrecta o que el traductor y el escritor tienen formas de escribir similares. Observando la sinopsis del libro del mexicano, vemos que trata de asesinatos, temática similar a la del autor *John Katzenbach*.

Grupo\Libro	El Reino del Fuego	El hombre equivocado	Cien años de Soledad	Siete esqueletos decapitados
1	29	0	0	0
2	0	0	18	0
3	0	34	0	8
4	1	12	0	34

Cuadro 1. Pureza de la asignación de grupos

En la figura 3, podemos observar de una manera más amigable lo explicado por la tabla.

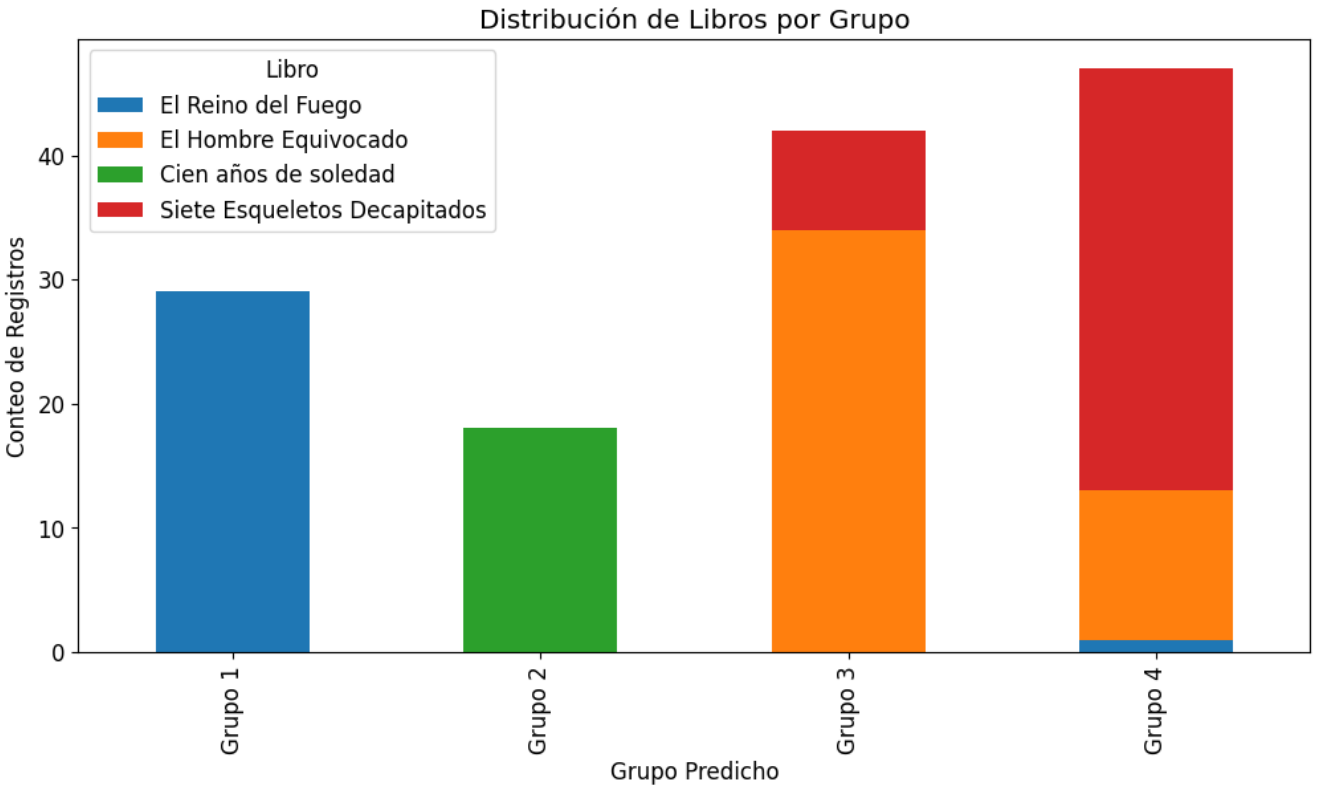

Figura 3. Distribución de los capítulos del libro por grupo.

4. Conclusión

El estudio demuestra que es posible utilizar técnicas de aprendizaje automático para identificar el autor de un libro basándose en estadísticas descriptivas de los textos. Aunque la temática y el estilo de escritura pueden influir en la precisión de la clasificación, el modelo de `K-Medias` mostró ser una herramienta efectiva para este propósito. Los resultados sugieren que, con un conjunto de características adecuadamente seleccionado, es factible automatizar el proceso de identificación de autores, lo cual podría ser útil en la toma de decisiones editoriales y en estudios literarios.