



Práctica Final

Predicción del género de libros

INTELIGENCIA ARTIFICIAL EN LAS ORGANIZACIONES

GRUPO 83-1

Miguel Gutiérrez Pérez

100383537@alumnos.uc3m.es

Mario Lozano Cortés

100383511@alumnos.uc3m.es

Alba Reinders Sánchez

100383444@alumnos.uc3m.es

Alejandro Valverde Mahou

100383383@alumnos.uc3m.es

GitHub: ***InteligenciaArtificialOrganizaciones***

December 10, 2020

Índice:

1 Introducción

El objetivo de esta práctica consiste en abordar una clasificación sobre resúmenes de libros para determinar su género literario. Las razones que llevan a la elección de este problema tienen que ver con que actualmente cualquier persona con la dedicación suficiente puede escribir un libro sin la necesidad del patrocinio de una editorial, lo que conlleva una **explosión en el número de nuevos libros generados**. Por consiguiente, las librerías y bibliotecas necesitan catalogar una gran cantidad de escritos, lo cual, les lleva a necesitar de métodos de clasificación automática. Por ello, se plantea el uso de resúmenes y metadatos de los libros puesto que la tarea de clasificación **debe poder realizarse con el menor número de datos posible**, puesto que no todos los libros que llegan a estas entidades disponen de todos los datos completos.

La primera cuestión imprescindible que surge al conocer el problema propuesto es qué técnica de Inteligencia Artificial emplear. Dado que se realiza un análisis sobre diferentes textos, la opción evidente es la **Minería de Texto**, la cual es una técnica de minería de datos que busca extraer **información útil y relevante de documentos de texto** de diferentes fuentes diferentes, como puede ser páginas web, correos electrónicos, periódicos o redes sociales. Para ello, se hace una identificación de patrones en los datos, como puede ser la repetición de palabras o conjuntos de palabras, estructuras sintácticas que se repitan a lo largo de los datos, etc. Esta minería de texto tiene numerosas aplicaciones, y en esta práctica se van a desarrollar una clasificación en función de unas categorías que serán definidas gracias a la elección de un dataset apropiado.

A continuación se ofrece un **esquema del funcionamiento de la tarea propuesta** en donde un libro sin catalogación llega a alguna de estas entidades que necesitan catalogar su género a partir de la información más reducida posible (generalmente título y argumento). Inicialmente se plantea la distinción de un único género, sin embargo, **es bien sabido que un escrito no tiene por qué adscribirse a un único género** y por lo tanto se debe considerar como futura **ampliación** catalogar tantos como sea posible.



Title: Animal Farm

Plot: *Animal Farm is a novel about a group of animals who take control of the farm they live on. The animals get fed up of their master, Farmer Jones, so they kick him out. Once they are free of the tyrant Jones, life on the farm is good for a while and there is hope for a happier future of less work, better education and more food. However, trouble brews as the pigs, Napoleon and Snowball, fight for the hearts and minds of the other animals on the farm. Napoleon seizes power by force and ends up exploiting the animals just as Farmer Jones had done. The novel ends with the pigs behaving and even dressing like the humans the animals tried to get rid of in the first place.*

⇒ **Genre:** Political satire

Figure 1: Esquema de la tarea

2 Conjunto de datos

En esta sección se describirá el **conjunto de datos** utilizado en su forma original, así como todos los cambios que se vayan a hacer como parte del **preprocesado** junto con las razones que han llevado a su realización.

2.1 Estructura original

El conjunto de datos ha sido obtenido de **Kaggle** [INSERTAR], y dispone de un total de **54283 libros y 12 columna** con información sobre ellos. En la Figura X se puede apreciar visualmente la forma de las instancias de este conjunto de datos.

Author(s)	Description	Edition	Format	ISBN	No. Pages	Avg. Rating	No. Ratings	No. Reviews	Title	Genres	Cover Image
William Golding	At the dawn of the next world war, a plane crashes on an uncharted island, stranding a group of...	Penguin Great Books of the 20th Century	Paperback	9.78E+12	182	3.66	1840595	30634	Lord of the Flies	Classics Fiction Young Adult Academic School Literature	Link

Figure 2: Instancia del conjunto de datos original

A continuación se va a explicar el **contenido** de cada columna en mas detalle:

- **Author(s):** Cadena de caracteres que indica el o los autores del libro. En caso de ser mas de uno, cada autor aparece separado por '|’.
- **Description:** Cadena de caracteres que indica el resumen o descripción del libro.
- **Edition:** Cadena de caracteres que indica la edición del libro.
- **Format:** Cadena de caracteres que indica el formato del libro, como por ejemplo *handcover* o *paperback*.
- **ISBN:** Valor numérico que indica el ISBN del libro. Debido a que se utiliza notación científica para su representación, no se puede leer ni usar correctamente, pues faltan números.
- **No. Pages:** Valor numérico que indica el número de páginas que tiene el libro.
- **Avg. Rating:** Valor numérico que indica la valoración media que los usuarios han dado al libro.
- **No. Ratings:** Valor numérico que indica al cantidad de valoraciones de usuario que ha recibido el libro.
- **No. Reviews:** Valor numérico que indica la cantidad de críticas que ha recibido el libro.
- **Title:** Cadena de caracteres que indica el título del libro.
- **Genres:** Cadena de caracteres que indica los géneros a los que pertenece el libro. Los diferentes géneros aparecen separados por '|’.
- **Cover Image:** Imagen que muestra la portada del libro. Se indica un enlace que lleva hasta dicha imagen, aunque por motivos de espacio no se ha incluido en la figura de arriba.

Inicialmente se consideró usar el conjunto de datos **CMU Book Summary Dataset** [INSERTAR], pero la manera en la que estaban dispuestos los géneros, así como el resto de atributos, daba lugar a un **procesado más complejo** para obtener sus valores. Además, se consideró que

los géneros que se utilizaban no eran muy acertados, como por ejemplo *Speculative fiction* o *Postmodernism*. Posteriormente se encontró el conjunto de datos que se ha descrito mas arriba, el cual tenía muchas mas instancias (54,283 frente a 16,559), proporcionaba mucha más información (tenía mas atributos), y, lo mas importante, era mucho mas sencillo obtener los valores de sus instancias.

2.2 Preprocesado

En toda tarea de minería de texto existe algún tipo de **preprocesado**, ya sea obtener solo los valores que se necesiten, eliminar instancias con errores, vectorizar el texto, etc. Esta vez no va a ser diferente, e incluso se puede afirmar que ha sido una **parte importante del trabajo**, y que ha ocupado una cantidad considerable de tiempo. A continuación se detallan todas las **modificaciones** que se han llevado a cabo:

- **Selección de las columnas útiles:** De las 12 columnas que tiene el conjunto de datos seleccionado, solo son necesarias dos, *description* y *genres*, que contenían, respectivamente, el resumen o descripción del libro y los géneros que se le atribuyen. Por ello, solo solo se mantendrán dichas columnas eliminando el resto.
- **Eliminación de instancias con descripciones en un lenguaje diferente al ingles:** Los resúmenes del conjunto de datos se encuentran escritos no solo en inglés, si no también en árabe, italiano, chino, coreano, japonés, portugués...Se ha decidido por razones evidentes, mantener solo aquellos libros cuyos resúmenes estén en ingles. Para la detección del idioma se ha utilizado la librería *langdetect* [<https://pypi.org/project/langdetect/>] de *Python*, y aquellas instancias donde se detectaba un idioma diferente al inglés se han borrado. Gracias a la gran cantidad de instancias que hay, la eliminación de las instancias no afecta de manera significativa.
- **Corrección de formatos incorrectos:** Algunos de los resúmenes del conjunto de datos contenían errores de formato como saltos de línea (de diferentes tipos), tabulaciones o espacios en medio de los resúmenes, así que se sustituyeron por un único espacio en blanco.
- **Eliminación de caracteres no útiles:** Se han eliminado de los resúmenes del conjunto de datos caracteres que no proporcionaban ningún tipo de información, como comillas, guiones, corchetes, paréntesis, puntos, exclamaciones, interrogaciones, etc.
- **Eliminación de términos no útiles:** Se han eliminado de los resúmenes del conjunto de datos términos que no proporcionaban ningún tipo de información útil. Para ello se ha utilizado una lista de *stopwords*, reutilizada de la segunda práctica de la asignatura, pero a la cual se han añadido términos que se han ido encontrado al realizar pruebas. Evidentemente, no se han incluido todos los términos sin información útil posibles, pero dada la gran cantidad de términos que puede haber, creemos que la cantidad de *stopwords* obtenida es suficiente.
- **Cambios requeridos por la aproximación usada:** La aproximación que se utilizada en este trabajo requiere que los resúmenes esté en minúsculas. Al contrario que ocurría en la segunda práctica, no es necesario dividir los resúmenes en términos, por lo que no se utilizara la cadena de caracteres que contiene el resumen de manera directa.
- **Eliminación de géneros no útiles:** Después de analizar un poco más a fondo los diferentes géneros que se usan, se ha observado la falta de un criterio claro para elegir que géneros usar, dando lugar a géneros de países, como *spain*, al mismo tiempo que el género *spanish literature*. También se encontraron géneros muy concretos y sin mucho sentido, como *Amazon* o *Apple*. Esto llevó a que fuera necesario revisar manualmente todos los

géneros, los cuales eran en torno a 850, y eliminar aquello que se consideraron no útiles. Al finalizar esta revisión quedaron 625 géneros.

Una vez aplicado todo esto, los datos ya están preparados para su uso. Sin embargo, queda una última cosa. La cadena de caracteres que contiene los géneros se separara según el caracter '|' en los diferentes géneros para su posterior uso. Como en este trabajo se va explorar una clasificación **multi-clase** con el primer género, y una clasificación ***multi-label*** con todos los géneros, para la primera solo se utilizará el primer género, y para la segunda se usarán todos.