Clasificador de libro en base al autor

Resumen

El siguiente trabajo consiste en entrenar un modelo que sea capaz de reconocer al autor en base a la obra. Nos enfocaremos en las obras de autores hispanohablantes, tomando autores de diferentes países. Además de la capacidad de clasificar textos en base al autor nos interesa aplicar el modelo para analizar la coautoría.

Corpus y análisis exploratorio

El formato del corpus de cada autor es un csv con el siguiente formato:
link, text_metadata, text

Donde:

- link es el link a la página web donde se encuentra el texto
- text_metadata cuenta con información como: título, género, longitud, autor
- text: es el texto del cuento.

Para la construcción de nuestro corpus, recurrimos a una base de datos de autores recopilada del repositorio GitHub (apéndice 1.1). Esta base de datos incluía una amplia gama de autores, de diferentes países latinoamericanos, cada uno con varias de sus obras asociadas. Con el objetivo de tener un conjunto de datos diverso, decidimos seleccionar autores de diferentes nacionalidades. Para ello, filtramos de la lista inicial (apéndice 1.2) y seleccionamos aquellos autores que contaban con por lo menos diez obras de las cuales aproximadamente diez serán elegidas, garantizando así uniformidad sobre los datos

En el marco de este trabajo no trabajaremos con el texto crudo sino que será particionado previamente en secuencias de oraciones. La longitud de estas secuencias de oraciones que serán el *input* de nuestro modelo representan un hiperparámetro del mismo. Buscaremos con qué valor el modelo presenta el mejor rendimiento. Este enfoque nos parece interesante ya que es un punto medio entre un enfoque bien granular que tomaría oraciones individuales y uno bien general que tomaría directamente cuentos completos.

Propuesta de análisis

Con respecto al análisis que será realizado para la segunda entrega, en primera instancia se evaluará la capacidad del modelo de deducir la autoría de textos que el modelo no ha visto previamente. Estos textos serán textos escritos por los autores con los que el modelo haya sido entrenado. Un análisis particular, como fue mencionado anteriormente es el de coautoría. Por ejemplo, dado un fragmento del libro “Cronicas de Bustos Domecq”, escrito por Borges y Bioy Casares, el modelo le asignará una probabilidad a cada autor de que este lo haya escrito, en base al entrenamiento realizado sobre textos propios de ambos autores. Otro aspecto interesante para analizar es la importancia que tienen las expresiones propias de cada región sobre la decisión tomada por el modelo, le dará más importancia a las palabras particulares o a la temática del texto.

Apéndice

1.1

[Repo Corpus]

1.2

País	Autor	Número de obras
Argentina	Julio Cortázar	55
Argentina	Adolfo Bioy Casares	14
Argentina	Jorge Luis Borges	62
Chile	Baldomero Lillo	25
Colombia	Alvaro Mutis	11
Colombia	Luis Vidales	15
Cuba	Alejo Carpentier	9
El Salvador	Salarrué	9
Guatemala	Augusto Monterroso	45
México	Juan José Arreola	45
México	Alfonso Reyes	37
México	Julio Torri	24
Nicaragua	Rubén Darío	13
Perú	Julio Ramon Ribeyro	27
Puerto Rico	Manuel A. Alonso	9
República Dominicana	Juan Bosch	8
Uruguay	Felisberto Hernández	15
Uruguay	Mario Benedetti	34