Proyecto Final de Aprendizaje de Máquinas.

Autores

Nombre y Apellidos	Grupo	Correo
Thalia Blanco Figueras	C-512	lia.blanco98@gmail.com
Eziel Ramos Piñón	C-511	e.ramos@estudiantes.matcom.uh.cu
Ariel Plasencia Díaz	C-512	arielplasencia00@gmail.com

Acerca de la Situación Problémica:

El grupo de periodismo de datos, dada a la amplia cantidad de documentos que escribió el Apóstol José Martí, ha decidido llevar a cabo diversas tareas relacionadas al aporte a la literatura cubana que proporcionó con su obra. Una de las tareas propuestas es determinar a partir de determinado texto, que tanto se parece a los que escribía Martí, en base al estilo de escritura empleado; o sea, llevar a cabo similaridad de documentos basándose en el estilo de escritura.

Acerca del Objetivo:

El objetivo que se propone este trabajo es de proveer de una herramienta automatizada que permita a partir del aprendizaje con una serie de datos asociados a la obra Martiana, se pueda predecir, para otros documentos, si adoptan el estilo de escritura de Martí.

Acerca de los Datos:

Como fuente de datos para determinar el estilo de escritura de José Martí, se emplean sus Obras Completas. Estas están conformadas por 29 documentos de un aproximado de 400 páginas cada uno y en los que se recopilan todos sus versos, cartas, artículos periodísticos, escritos literarios, discursos, borradores, traducciones y demás. El estilo de escritura de Martí se estará midiendo solamente con sus textos en español, por tanto, no se tendrán en cuenta los escritos ni en inglés ni francés que realizó. Como el estilo a evaluar es el de Martí, tampoco se tendrán en cuenta las traducciones que hizo a escritos de otros autores. Luego, si no se tienen en cuenta las categorías anteriormente mencionadas, se reduce de 29 a 23 el número de obras.

Acerca del Estilo:

El medir un estilo de escritura es una actividad compleja ya que son demasiados los factores en que se puede basar la clasificación. Por ello se lleva a cabo una extensa búsqueda, acerca de trabajos anteriores asociados a esta temática, para poder obtener las características que se tienen en cuenta durante este proceso. Los resultados de la búsqueda demuestran una serie de factores que comúnmente son tenidos en cuenta y que vienen vinculadas al análisis léxico, sintáctico y estructural de los documentos. A partir de las necesidades propias de este problema que es medir un estilo literario de un escritor del sigo XIX, se consideran que las características que se han de tener en cuenta son:

- Promedio de palabras por oraciones, normalizado por la cantidad de oraciones del documento.
- Promedio de largo de palabras, normalizado por la cantidad de palabras del documento.
- Frecuencia de palabras cortas (palabras de tamaño menor o igual a 3), normalizada por el total de palabras del documento.
- Cantidad de palabras diferentes, normalizado por el total de palabras del documento.
- Frecuencia de signos de puntuación por oraciones, normalizada por el total de palabras.
- Frecuencia de `stopwords` (palabras que no proporcionan significado a los textos), normalizada por el total de palabras.
- Frecuencia de palabras raíces, normalizada por el total de palabras.
- Frases de sustantivos (aquellas frases que contienen un sustantivo como cabecera).
- Frases verbales (incluye verbos, formas verbales, y toda expresión que lleve a cabo la función verbal).
- Frases compuestas por sustantivos continuados por adjetivos.
- Frases compuestas por adjetivos continuados por sustantivos.
- Frases compuestas por sustantivos continuados por frases verbales.

Para llevar a cabo el análisis de los documentos por los criterios anteriores, se emplea `spacy`, biblioteca que proporciona muchas ventajas para el procesamiento de textos.

Con las características anteriores, se convierten los documentos en vectores numéricos. La clase `DocumentVector` ubicada en el archivo `dataVector/docVector.py`, es la encargada de transformar estos documentos; y emplea a su vez, al archivo `dataVector/sentVector.py` para el análisis a nivel de oraciones. Esta clase ofrece a su vez un método para escribir el vector resultante de un documento, en un `csv` que pueda ser posteriormente empleado para el proceso de aprendizaje de los datos.

Acerca del Aprendizaje:

Para poder predecir la similaridad de estilos entra otros documentos con respecto a los escritos por Martí, se hace uso de algoritmos de aprendizaje que permiten a través de los datos de entrada, asociarlos o no a los escritos por el Apóstol. Para el aprendizaje se van a emplear también otros textos escritos por otros autores, para que los algoritmos puedan aprender de estilos martianos y no martianos. Para el aprendizaje de estilos no martianos, se recopilan 21 escritos de otros autores.

Para darle solución al problema se emplean dos alternativas:

- Emplear `Naive Bayes` que es un algoritmo de lenguaje supervisado basado en la teoría de probabilidades.

Al ser de tipo supervisado se basa primeramente en el entrenamiento de los datos y posteriormente en la validación, pero como conocemos una serie de documentos escritos por Martí y una serie que no, podemos emplear estos documentos como conjunto de entrenamiento, y sus clases correspondientes serían si fueron o no escritas por Martí (se emplea 1 como identificador de textos Martianos y 2 para los demás textos).

Una forma de evaluar el funcionamiento de este algoritmo puede ser entrenar con un subconjunto de este y el resto ser empleado para la evaluación, y como se conoce de estos las clases en que pertenecen, luego comparar cuántas predicciones coincidieron con las reales clases.

Para el problema en específico se entrenan con todos los documentos que se tienen, y los que se desean calificar, son transformados en vectores según sus características, y empleados como conjunto de entrenamiento. Este problema se considera más en un enfoque no supervisado ya que realmente no se espera una calificación a priori del documento, pero en caso de que existiera, puede ser empleada para evaluar el documento.

La implementación ofrecida para este algoritmo se encuentra en `algorithms/naivebayes`.

- Emplear `KMEANS` que es un algoritmo de lenguaje no supervisado que agrupa los datos en grupos (cluster) según su cercanía.

  Con los vectores asociados a los documentos con que se cuentan para el aprendizaje, se emplea `KMEANS` para agruparlos en grupos. Como por cada uno de estos vectores se conoce si lo escribió o no Martí, por cada grupo se puede conocer el porciento de documentos Martianos:

  $$\frac{documentos\ \ de\ \ Martí}{total\ \ de\ \ documentos\ \ del\ \ cluster} * 100$$

  La función de similaridad empleada para calificar un texto de entrada según su estilo, viene asociada al porciento de documentos martianos del grupo al que pertenece. El grado de aceptación establecido es de al menos un 75%.

  Luego por cada documento a evaluar, se calcula su vector, y a partir de este, el grupo al que pertenece (según su cercanía o similaridad a los datos que lo componen), y se la aplica la función de similaridad. Los resultados que se arrojen dirán si adopta o no el estilo de Martí.

  La implementación ofrecida para este algoritmo se encuentra en `algorithms/kmeans`.

Acerca de los Resultados:

Para evaluar el rendimiento de los algoritmos se emplean una serie de documentos de los cuales se conoce su autor, ubicados en la carpeta `testData`, en la carpeta `testData/marti` se encuentran escritos de Martí y en `testData/otros`, escritos de otros autores.

Al comparar los resultados que se obtienen para estos ejemplos, las predicciones que se obtienen de aplicar `KMEANS` y `Naive Bayes` sumamente parecidas y los valores no muy buenos. De aplicar `Naive Bayes` solo el 43% son evaluados correctamente y el 50% al aplicar `KMEANS`, pero ambos algoritmos evalúan en la mayoría de los casos de prueba el mismo resultado, o sea que ambos aceptan o se equivocan (excepto en un caso).

Limitantes y Recomendaciones:

Uno de los factores que influye en la baja precisión es la poca cantidad de documentos de aprendizaje con que se cuenta, ya que se tienen solamente 23 documentos, y por tanto no se están analizando sus escritos sino más bien sus colecciones de escritos, agrupadas en las Obras Completas.

También las Obras Completas al ser transformadas de `pdf` a `txt`, incluyeron muchos textos asociados a los comentarios que proporcionan sus recopiladores. Al no seguir estos un patrón común, no se pudo emplear ninguna herramienta para eliminarlos todos, por lo que se tuvieron que extraer manualmente, jugando el error humano un factor importante que puede haber obstaculizado los resultados obtenidos en la construcción de los datos de aprendizaje.

Para adquirir una mayor precisión en los algoritmos se proponen diferentes alternativas:

- En vez de emplear las obras completas, emplear cada uno de los escritos por separado, ya que al estar estos agrupados, la vectorización no se lleva a cabo a nivel de escrito sino a colección de escritos.
- Agrupar los escritos por géneros y llevar a cabo el aprendizaje a nivel de género.
- Extender la cantidad de características a evaluar para medir el estilo. Este número tampoco puede ser demasiado grande porque esto provoca a su vez el aumento del volumen del espacio haciendo que los datos se turnen dispersos.