Natural Language Processing (Procesamiento del Lenguaje Natural)
Esta serie de pasos para procesar datos de texto y aplicar un algoritmo de aprendizaje fueron los utilizados como marco de trabajo, con base a Pramod Singh y su libro Machine Learning with PySpark: With Natural Language Processing and Recommender Systems.

Reading the corpus (Lectura del cuerpo/datos)
El corpus es la colección completa de documentos de texto que se utilizan como la ingesta de datos. Por ejemplo, tenemos miles o millones de críticas de películas de cine que juntas conforman una colección que necesitamos procesar y analizar para nuestro uso.

Tokenization (Tokenización)
La tokenización es el método de dividir la colección de palabras de un documento de texto en palabras separadas/individuales (los tokens). Se caracteriza por eliminar los caracteres innecesarios como los signos puntuación.

Stopword removal (Limpieza de palabras reservadas)
Las colecciones generalmente contienen palabras muy comunes como 'esto', 'lo', 'fue', 'a', 'era', 'eso', etc. Estas palabras se conocen como palabras de parada o palabras reservadas y se caracterizan por añadir poco valor al análisis. Si se van a utilizar en el análisis, aumenta la sobrecarga del cálculo sin agregar demasiado valor o conocimiento. Por lo regular, siempre se considera una buena idea eliminar estas palabras de la colección de los tokens.

Bag of Words (Vocabulario)
Existe la necesidad de representar los datos de texto en forma numérica para que pueda ser utilizado por algoritmos de Machine Learning o cualquier otro análisis. Los datos de texto generalmente no están estructurados y varían en su longitud. La metodología de Bag of Words permite convertir la forma del texto en un vector numérico de las palabras en los documentos de texto.

Se debe establecer la lista de palabras únicas que aparecen en todos los documentos, la cual se convierte en el vocabulario. El acercamiento de la metodología no considera el orden de las palabras en el documento y el significado semántico de las mismas, solo importa la aparición de palabras válidas y se hace una representación con 1 o 0.

Count Vectorizer (Conteo de vectores)
El conteo de vectores toma en cuenta la frecuencia con la que aparece la palabra en el documento en particular. El inconveniente de usar este método de conteo es que no considera las ocurrencias de esas palabras en otros documentos, se tienen vocabularios y un conteo para cada documento, restringiendo la validación de un mayor impacto en el vector de la característica o palabra.

Latent Dirichlet Allocation (Asignación Latente de Dirichlet)
Es un modelo que permite que conjuntos de observaciones puedan ser justificados por grupos que antes no eran observados directamente y que pueden llegar a explicar porqué algunas partes de los datos son similares.

Para la revisión de observaciones en documentos, el modelo parte de la premisa que cada documento está compuesto por una mezcla de un pequeño número de categorías y la aparición de cada palabra en un documento se debe a que existe una relación con una de las categorías a las que el documento pertenece. La clave consiste en la hipótesis de que el uso de una palabra es ser parte de un tema y que comunica la misma información sin importar dónde se encuentra en el documento.

Si los documentos se comparan de forma individual, puede darse el caso de que ciertos temas no sean recogidos en la revisión, y sólo cuando toda la colección es vista se empiezan a notar ciertas categorías. Por ejemplo, algunas palabras que aparecen con menos frecuencia en los documentos únicos, pero son comunes en muchos documentos diferentes de la colección probablemente indican que existe un tema común entre los documentos.

Referencias

- Singh, P. (2019). Machine Learning with PySpark: With Natural Language Processing and Recommender Systems. Recuperado de https://doi.org/10.1007/978-1-4842-4131-8
- Latent Dirichlet Allocation (s.f.). Recuperado en noviembre 20 de 2019 de https://es.wikipedia.org/wiki/Latent_Dirichlet_Allocation