# Machine learning for tex

Carlos Andres Hernandez Gomezm

6 de mayo de 2019

# Índice general

# Capítulo 1

# Introducción

Este documento presenta los apuntes realizados del libro *Machine Learning for Text* con ISBN 978-3-319-73531-3 (eBook, 2018) de Charu C. Aggarwal un distinguido investigador de IBM; El motivo de redactar este documento es aprender sobre el procesamiento del lenguaje natural para ser aplicado en el trabajo terminal (TT) **Recolector y clasificador de noticias** con número 2019-B013, de la **Escuela superior de cómputo** (ESCOM) del **Instituto politécnico nacional** (IPN).

## 1.1. Machine learning for text

La extracción de información útil con varios tipos de algoritmos estadísticos es denominado **Extracción de datos**( *Text mining* ), **Analítica de texto** (*Text analytics*) o **Aprendizaje automático para texto** (*Machine learning for text*), en este documento se utilizará de forma indistinta. En los últimos años este campo ha incrementado por el desarrollo de la web, redes sociales, correos electrónicos, bibliotecas virtuales. Algunos ejemplos para obtener información son:

- **Bibliotecas digitales**: El uso de la información electrónica ha superado la producción de libros y publicaciones impresas, este fenómeno ha proliferado la producción de bibliotecas digitales, estas pueden ser almacenadas y ser usadas para extraer información útil.

- **Noticias electrónicas**: Existe un movimiento masivo para pasar las noticias impresas hacía la publicación electrónica, esto permite que sean almacenadas para su análisis y extracción de información sobre eventos y perspectivas importantes. Sitios como *Google news* etiquen las noticias para hacer recomendaciones al lector basado en su anterior comportamiento o intereses específicos.

- ***Web and Web-enabled applications***: La web contiene una gran cantidad de información en hipertexto, con links y otro tipo de recursos, la cual puede ser utilizada par el proceso de descubrimiento de nuevo conocimiento, al igual las *Web-enabled applications*[1] permiten obtener información que puede ser analizada.
- **Redes sociales**: Las redes sociales son un campo que está proliferando, debido a su naturaleza donde cada usuario contribuye con sus propias publicaciones.

Algunas de las aplicaciones son las siguientes:

- Etiquetar la web, permite al usuario encontrar paginas de interés
- Los proveedores de correos, utilizan la información almacenada para mostrar publicidad de interés al usuario
- Algunas páginas ordenan su contenido de acuerdo a su importancia
- El análisis de las opiniones es un campo de importancia así como el análisis de sentimientos

El orden se las palabras en un texto brindan un significado semántica el cual no puede ser inferido solo con la frecuencia de las palabras. Sin embargo, se pueden hacer varias predicciones sin contemplar la semántica. Existen 2 tipos de representaciones que son populares en las aplicaciones de **text mining**:

- ***Text as a bag-of-words***: Es la representación mas común. No se contempla el orden de las palabras el proceso. El conjunto de palabras en el documento se convierten en *Sparse multidimentional reprentation*, el cual corresponde a la dimensión en esta representación. Se utiliza para la clasificación, sistemas de recomendación.
- ***Text as a set of sequences***: En esta representación se extraen sentencias, el orden de las palabras si importa. La unidad son sentencia o párrafos. Es utilizado en aplicaciones que necesitan un fuerte uzo de la semántica, esta área se acerca mucho al modelado de lenguaje y procesamiento del lenguaje natural

## 1.2. Definiciones

### 1.2.1. Corpus

En el desarrollo de las aplicaciones se tiene un conjunto de datos los cuales forman el *corpus*, el orden de los datos se coloca en $10^5$ es decir 100,000 datos o incluso $10^6$. En la mayoría de los datos toman el valor de 0, solo un pequeño conjunto toma un valor positivo.

[1] ***Web and Web-enabled applications***: Aplicaciones de escritorio que son accedidas remotamente des de un buscador como Internet explorer.

### 1.2.2. Learning algorithm

El término *Learning algorithm* se utilizará para hacer referencia a los algoritmos que descubran patrones en el texto, o como se pueden usar los patrones para predecir valores específicos en los datos.

## 1.3. Pre-procesamiento y computación similar

El Pre-procesamiento es necesario para convertir el formato no estructurado en un formato estructurado. A menudo el texto contiene información extraña como etiquetas, *anchor text*[2], y otras características. En muchos casos las palabras son variaciones de otras (Sinónimos) por el tipo de redacción, el contexto, para eliminar redundancia. Algunas palabras simplemente tienen faltas de ortografía. El proceso de convertir una secuencia de caracteres en una secuencia de palabras(Tokens), es llamado **Tokenización**. Además por cada palabra repetida se crea un token, es decir al repetirse una palabra 3 veces, se crearán 3 tokens correspondientes. Alguno de los pasos mas comunes para el procesamiento de texto en bruto son los siguientes:

- **Extracción de texto**: En caso de recuperar información de la web, se tiene que limpiar el texto ya que contiene *anchor text*, etiquetas.Se debe buscar los bloques que brinden información útil para el análisis, ya que algunos bloques contienen publicidad o información no relacionada. Para esto se tiene que realizar un **parseo**[3] o técnicas de extracción especializadas.

- **Remover stop-words**: **Stop words**, son pronombres, *articles* y preposiciones que deben ser removidas para mejorar la compresión del texto.

- **Stemming, case-folding, punctuation**: Las palabras que derivan de la misma raíz como hundimiento, se hundió, se reducen a hundir. Una palabra puede tener diferentes significados dependiendo el contexto como la palabra Ŕosaṕuede hacer referencia a una flor o el nombre de una persona, por lo tanto se requiero la euristica del lenguaje específico para poder tomar una decisión en como debe ser interpretada. Los signos de puntuación como el guión medio deben ser tratados con mucho cuidado para realizar una buena tokenización.

- ***Frequency-based normalization***: Palabras con poca frecuencia son mas discriminatorias que las de alta frecuencia. Por lo tanto se pondera la importancia de los documentos con base al calculo de la **frecuencia inversa del documento**(*fid*) en la colección. Si $\psi_i$ es el número de documentos en el cual la palabra aparece, y $\psi$ es el número total de documentos, la *fid* se calcula como $log(\frac{\psi}{\psi_i})$. La importancia de un documento se calcula

[2]Es el texto mostrado en los enlaces o hipervínculos, **Texto de anclaje** en español.
[3]Es la acción de analizar el texto de forma especializada

multiplicando la **frecuencia de término** (*ft*) en el documento por la *fid*. Mientras la *ft* brinda la cantidad de veces que una palabra aparece en el documento la *fid* especifica la importancia en la colección; Se define como **ft-fid** o *tf-idf* (Por su sigla en ingles *Term frequency – Inverse document frequency*). La figura 1.1 muestra como en proporción a la aparición baja de la palabra en los documentos gana un valor superior (Es decir gana importancia); La figura 1.2 muestra como una palabra gana importancia al tener una frecuencia alta en el documento.

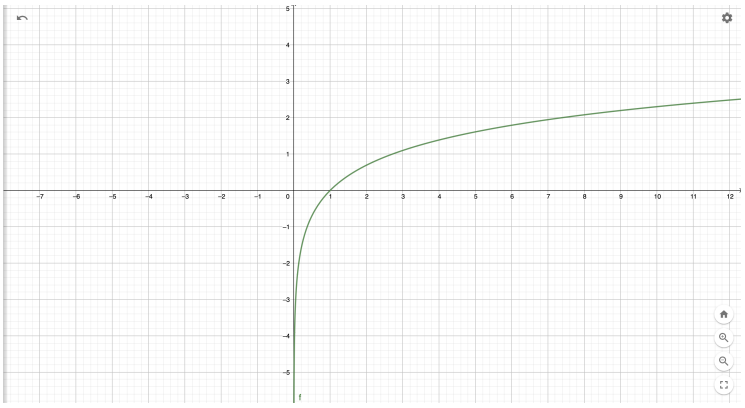

Figura 1.1: Gráfica *Inverese document frequency*

Figura 1.2: Gráfica de *Term frequency*