Reporte_Preprocesamiento

AFP

21 de enero de 2016

En este documento se justifican las decisiones tomadas a la hora de realizar el preprocesamiento

1- Eliminar urls ya que no aportan información

2- Remover #6d o #6D ya que todos los tweets lo tienen y por lo tanto no aportan información

3- Se estandariza todas las palabras a minúscula

4- Se remueven los signos de puntuación

5- Se usa "stripWhitespace" para eliminar múltiples espacios en blanco que son un básicamente un estándar en preprocesamento de datos y sirve para dar una matriz de palabras mucho más compacta

6-Se elimina palabras como artículos y pronombres (mediante el uso de stopword("spanish")) , además de eliminar la palabra "rt" que básicamente es una palabra muy repetida en twitter pero que no ofrece ningún tipo de información

7-Se quita el acento de las palabras ya que al ser twitter una red social es bastante común que se omitan los acentos y al existir la misma palabra con acento y otros con acentos afecta la verdadera frecuencia de ésta

8-Con la función removeURL se remueven todas las palabras que cumplan con la condición establecida por la Expresión Regular, que básicamente son todos aquellos URL que estén completos, es decir que tengan protocolo (http o https), seguidos de "://", un conjunto de letras, un "." y por ultimo otro conjunto de letras para la extensión.

9-Con la función removeURL2 se remueven todos aquellos URL que quedaron luego de filtrar los primeros, que son los incompletos, es decir todos que solamente estén escritos hasta los "://" y que sean "https"