Twitter como mecanismo de violencia simbólica.

Alejandro Ramirez Muñoz - 2152779
David Villabona Ardila - 2122910

Universidad Industrial de Santander

April 4, 2019

Tabla de Contenido

- Comprender la declaración del problema.
- Preprocesamiento y limpieza de los tweets.
- Extraer características de los tweets limpios.
- Construcción de modelos: Análisis de sentimiento.

Comprender la declaración del problema.

- Teniendo en cuenta el indice de proliferación global de información por medio de las redes sociales, y el impacto que genera este entorno en nosotros, ya que nadie esta inmiscuido de la burbuja generada por el uso de las mismas.

- Esa relación internet-persona es tan grande, que ya hace parte de nosotros. Expresamos nuestras opiniones con las consecuencia de mirar la de los demás, es ahí donde entramos nosotros, analizando, procesando y clasificandolos datos, para saber hasta que punto encontrar violencia simbólica, ya que metidos en un mundo con restriciones infimas, lo que se busca es controlar el uso deliberado de la ya mencionada violencia simbólica, para asi hacer que este mundo paralelo sea mas llevadero y se use con mejores fines.

Preprocesamiento y Limpieza de los Tweets.

- **Eliminando los identificadores de Twitter (@usuario):**
  Los tweets contienen identificadores de twitter como (@usuario). Eliminaremos todos estos identificadores de twitter de los datos ya que no transmiten mucha información.
- **Eliminación de puntuaciones, números y caracteres especiales:**
  Las puntuaciones, los números y los caracteres especiales tampoco nos generan las características que buscamos.
- **Eliminar palabras cortas:**
  Hemos decidido eliminar todas las palabras que tienen una longitud menor a tres caracteres. Por ejemplo, términos como "hmm", "oh", "the", son de poca utilidad.

Extraer características de los tweets limpios.

- Para analizar un dato preprocesado, es necesario convertirlo en características. Dependiendo del uso, las características del texto se pueden construir utilizando técnicas variadas: Bolsa de palabras, TF-IDF, son las que usaremos en este proyecto.

  - **Características de la bolsa de palabras.**
    Crea una matriz, donde cada columna contiene las caracterísicas mas frecuentes del dataset, y cada fila será la frecuencia con la que aparecen en cada tweet.
  - **Características de TF-IDF.**
    Este es otro método que se basa en la frecuencia, y tiene en cuenta no solo la aparición de una palabra en un solo tweet, sino en todo el dataset.

Construcción de modelos.

- Ahora hemos terminado con todas las etapas de preprocesamiento previo necesario para obtener los datos en la forma adecuada. Ahora construiremos modelos predictivos en el conjunto de datos utilizando los dos conjuntos de características: Bolsa de palabras y TF-IDF.

- **Métodos a utilizar:**
  - Logistic Regression.
  - Support Vector Machine.
  - Random Forest.

Referencias

- https://relopezbriega.github.io/blog/2017/09/23/procesamiento-del-lenguaje-natural-con-python/

E-Books

- Thank You.