Text Mining en Social Media

PAN-AP17

David Sánchez, Adrián Díaz y Joan Buigues

R Session Aborted

R encountered a fatal error.

The session was terminated.

Start New Session

Preproceso de los datos - VARIEDAD

- Modismos locales
- Creación diccionarios por variedad (país)
- Limpieza: eliminar duplicados
- Normalización de valores: minúsculas
- Mismo número de columnas, mismo nombre de variables
- Establecemos semilla

Comparativa de modelos

- SVM: 52,29%
- RF: 53,93%
- KNN: 34,50%
- NB: 42,14%
- CART: 31,79%
- NN: 43%
- C50: 51,71%

Nos quedamos con RF: 53,93%

Modelado

- Modelado con el baseline de 100 palabras contra la lista individual de Chile.
  - Accvariety: 31,21%
  - Sobreajuste
- Agregamos las listas de todas las variedades
- Modelo con el baseline de 500 contra la lista conjunta
  - Accvariety: 71,57%

- Evaluamos con ~1000 palabras RF
  - 500 de la bolsa de palabras + 416 de nuestro diccionario
  - ```
    > print(paste(accgender, accvariety, accjoint, time.taken))
    [1] "0.702857142857143 0.871428571428571 0.618571428571429 34.6219050208728"
    ```
    ```