M2.851 TIPOOGÍA Y CICLO DE VIDA DE LOS DATOS – Práctica 1

1. Título del dataset. Poned un título que sea descriptivo.
   Diag-salud

2. Subtítulo del dataset. Agregad una descripción ágil de vuestro conjunto de datos por vuestro subtítulo.
   Diagnóstico de salud y *big data*

3. Imagen. Agregad una imagen que identifique vuestro dataset visualmente

   Imagen generada por combinación de imágenes de: http://optimizeottawa.com/ways-simplify-health/
   http://bigdataanalyticsnews.com/how-to-survive-as-a-small-company-in-the-age-of-big-data/

4. Contexto. ¿Cuál es la materia del conjunto de datos?
   En un contexto amplio que corresponde al análisis de las posibilidades del big data en la ayuda al diagnóstico de enfermedades y a la mejora de la salud, la materia de los datos iniciales elegidos responde a observar las indicaciones o referencias sobre esta materia que se encuentran en la comunicación de las redes sociales, concretamente en el entorno de "microblogging" Twitter.

5. Contenido. ¿Qué campos incluye? ¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido?
   El contenido de los datos son los que devuelve la API de búsqueda de Twitter utilizada, el *Package twitteR* para R
   En la siguiente tabla figuran los campos más relevantes, de los que se han obtenido datos.

Campo	Descripción
text	Texto del tweet
created	Fecha de creación
id	Identificador
statusSource	User Agent
screenName	Autor
isRetweet	TRUE si es retweet
retweetCount	Número de retweets

   Existen otros campos adicionales que figuran en la muestra, como los atributos Latitud y Longitud para la localización, de los que no se han encontrado valores en ninguna de las consultas realizadas, de todas formas se han conservado en el archivo CSV generado.
   El periodo de tiempo de los datos es 2.017.

6. Agradecimientos. ¿Quién es propietario del conjunto de datos? Incluid citas de investigación o análisis anteriores.
   Agradecimiento a Twitter de donde se obtiene la información: http://twitter.com
   Y a Bogdan Rau por su blog para utilizar la API de consulta:
   http://bogdanrau.com/blog/collecting-tweets-using-r-and-the-twitter-search-api
   Y a los autores de los tweets, según figuran en el atributo *screenName*.
   Los tweets son información pública según consta en la política de privacidad de Twitter: "… any registered user of Twitter can send a Tweet, which is public by default…"
   https://twitter.com/en/privacy

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?
   La salud y el diagnóstico precoz es una de las aplicaciones de big data que me parecen más interesantes y desconozco en que estadio de evolución se encuentra.

8. Licencia. Seleccionad una de estas licencias y decid porqué la habéis seleccionado:
   - Released Under CC0: Public Domain License
   - Released Under CC BY-NC-SA 4.0 License
   - Released Under CC BY-SA 4.0 License
   - Database released under Open Database License, individual contents under Database Contents License
   - Other (specified above)
   - Unknown License

   Licencia *Database released under Open Database License, individual contents under Database Contents License*

   La información de partida es pública tal como se indica anteriormente, se publica como "abierta" y con la condición de que se mantenga "abierta".
   Respecto a la parte individual, en realidad, todavía no se ha generado contenido adicional, es lo que correspondería si se trabajaran los datos.

9. Código: Hay que adjuntar el código con el que habéis generado el dataset, preferiblemente con R o Python, que os ha ayudado a generar el dataset
   Se incluyen los archivos:
   - diag-salud.html: informe generado con RStudio / RMarkdown con los métodos ejecutados en R y sus comentarios
   - diag-salud.Rmd: Markdown R
   - diag-salud.R: Script R
   - diag-salud- incRetweets.csv: datos sin filtrar "retuiteados"
   - diag-salud-dat1.csv: archivo intermedio inicial de datos

10. Dataset: Dataset en formato CSV
    Se incluye el archivo: diag-salud.csv