# Detección de posicionamiento de tweets respecto a un conjunto de tópicos

<https://sites.google.com/view/mentorias2020-diplodatos/detecci%C3%B3n-de-posicionamiento-de-tweets-respecto-a-un-conjunto-de-t%C3%B3picos?authuser=0>

En este proyecto utilizaremos el dataset de la conferencia Semeval 2016 sobre detección de posicionamiento respecto a un tópico en tweets. El dataset tiene tweets de cinco tópicos distintos, con entre 600 y 700 tweets de entrenamiento y entre 200 y 300 tweets de test para cada tópico. Los tweets están clasificados en tres categorías: a favor del tópico, en contra o ninguno. Además, vamos a proveer tres datasets sobre tres de los tópicos (“legalización del aborto”, “movimiento feminista” y “el cambio climático es una problemática real”) de entre 600K y 700K tweets cada uno sin etiquetar pero con metadata e información extra como el id del usuario que twiteó, la cantidad de favs que obtuvo, la cantidad de retweets, si es un retweet de qué usuario se retwiteo, a quien sigue el usuario que twiteó y quienes lo siguen, entre otras varias cosas más.

El objetivo del proyecto es utilizar la metadata de los tweets para estructurar una gran masa de datos sin estructura aparente y ver de qué manera se puede luego usar eso para mejorar o ayudar con la tarea supervisada o incluso para desarrollar un algoritmo no supervisado o semi-supervisado de clasificación.

Existe una cantidad prácticamente ilimitada de datos en redes sociales que no tiene etiqueta pero que tiene una determinada estructura que puede ser analizada y de la que puede extraerse información. Esta información puede luego ser utilizada de diversas maneras y en distintas tareas de ML.

API Rest Twitter, 180 requerimientos cada 15 minutos.

Tweets search: 450 tweets por requerimiento.

Usuarios: 900 usuarios por requerimiento.

Timelines: 1500 tweets por requerimiento. (solo puedo sacar los ultimos 3200 tweets.)

API Streaming, obtener tweets en tiempo real.

Comandos Mongo

**Ejecutar mongo en una terminal.**

mongo

En Otra terminal posicionado en la carpeta donde queremos importar o exportar datos:

Exportar Collection

mongoexport -d=test -c=feminism --limit=100000 --out=export.json

**Ejemplo utilizado para filtrar los tweets, a diferencia del anterior, se agrega campo --fields y el output es csv en lugar de json.**

mongoexport -d=test -c=feminism --type=csv --fields='\_id,query,tweet.created\_at,tweet.id\_str,tweet.full\_text,tweet.display\_text\_range,tweet.entities,tweet.user.id\_str,tweet.user.name,tweet.user.screen\_name,tweet.user.location,tweet.user.verified,tweet.user.followers\_count,tweet.user.friends\_count,tweet.user.listed\_count,tweet.user.favorites\_count,tweet.user.statuses\_count,tweet.user.created\_at,tweet.user.following,tweet.retweet\_status,tweet.retweeted\_status.created\_at,tweet.retweeted\_status.id\_str,tweet.retweeted\_status.text,tweet.retweeted\_status.user.id\_str,tweet.retweeted\_status.user.name,tweet.retweeted\_status.user.screen\_name,tweet.retweeted\_status.user.location,tweet.retweeted\_status.user.verified,tweet.retweeted\_status.user.followers\_count,tweet.retweeted\_status.user.friends\_count,tweet.retweeted\_status.user.listed\_count,tweet.retweeted\_status.user.favorites\_count,tweet.retweeted\_status.user.statuses\_count,tweet.retweeted\_status.user.created\_at,tweet.retweeted\_status.user.id\_str,tweet.retweeted\_status.user.following,tweet.retweeted\_status.retweet\_status,tweet.retweeted\_status.extended\_tweet' --out=feminism\_tweets.csv

Importar Collection

mongorestore users.bson (posicionado en la carpeta donde se encuentra el .bson)

nota: en la terminal donde se esta ejecutando mongo, seleccionar la DB correspondiente. comando -> *use test*