

# Repositorios NoSQL y análisis básico de contenido – Taller 2

Nicolas Jimenez, Oscar Forero

MINE4102 – Análisis de información sobre Big Data

Universidad de los Andes, Bogotá, Colombia

[of.forero41@uniandes.edu.co](mailto:of.forero41@uniandes.edu.co)

[en.jimenez@uniandes.edu.co](mailto:en.jimenez@uniandes.edu.co)

Fecha de presentación: noviembre 17 de 2020

## Tabla de contenido

Introducción	1
Fuente de datos	1
Consultas	2
Análisis de coyuntura sobre Twitter	2
Aplicación WEB	4
Análisis de resultados	5
Conclusiones	5
Bibliografía	5
Anexos	6

**Enlace a aplicación:** <http://mine4102-9.virtual.uniandes.edu.co:9000/>

## Introducción

Dada la coyuntura actual de violencia social en Colombia (muertes violentas, desplazamientos, violencia intrafamiliar, violencia de género, matoneo en las redes sociales e inseguridad en las ciudades y zonas rurales) se planteó realizar un ejercicio de recolección de datos, análisis de polaridad, análisis de apoyo, contradicción o matoneo en Twitter. A continuación, se describe la fuente de datos, las consultas, se realiza un análisis de coyuntura, se habla sobre la aplicación WEB construida para este ejercicio y se presenta un análisis de resultados con las conclusiones del ejercicio.

## Fuente de datos

La fuente de datos de este taller es Twitter, para recolectar la información se construyó una aplicación .net la cual nos permite conectarnos a la API de Twitter para descargar los tweets que deseamos. Esta aplicación nos da la posibilidad de agregar cuentas y palabras que queremos buscar, de esta manera la aplicación edita la URL con estos datos en los que estamos interesados. Además, recibe como parametro el Bearer Token con el que nos autenticamos en el API de Twitter. Finalmente, tiene la opción de escoger una carpeta donde almacenar los tweets (Anexo 1). Para la recolección de los tweets se consiguió la información de 92 cuentas de twitter que se pueden agrupar en 7 grupos: Entidades gubernamentales colombianas, Medios de comunicación, Políticos, Partidos políticos, Entidades no gubernamentales, ciudades y periodistas. Dentro de estas cuentas se buscó temas como: escalada de violencia social en Colombia, muertes violentas, desplazamientos de comunidades, violencia intrafamiliar o de género, violencia o matoneo en la red social, inseguridad en las ciudades o zonas rurales, lideres sociales, secuestro, extorsión, narcotráfico y paz y postconflicto. Ver Anexo 2 para lista de cuentas. Para tener en

cuenta, los tweets fueron entregados por la API en formato JSON el cual contenia como mínimo los datos de creación, el texto, el ID, el lenguaje y el ID del autor del tweet (Anexo 3). En otros casos, si era el caso, se agregaban datos sobre los re-tweets. Finalmente, los tweets recolectados se agruparon, validaron y subieron a una colección dentro de una base de datos MongoDB. A continuación, se presenta la ficha técnica de los datos recolectados:

Colección "Tweets"	
_id	ID en MongoDB
created_at	Fecha de creación de tweet
id	ID del tweet
referenced_tweets	Referencias al tweet
text	Texto del tweet
author_id	id del autor del tweet
lang	lenguaje
polarity	Calificación de polaridad
scale	Calificación según escala de polaridad
subjectivity	Subjetividad

Tabla 1: Ficha técnica colección "Tweets".

Colección "authors"	
_id	ID del tags
id	Valor del tag
name	Nombre del dueño de la cuenta
username	Nombre de usuario de la cuenta

Tabla 2: Ficha técnica colección "authors".

Colección "Tags"	
_id	ID en MongoDB
tweet_id	Fecha de creación de tweet
author_id	ID del autor del tweet
word	Palabra
tag	Tag de la palabra (Anexo 10)

Tabla 3: Ficha técnica colección "Tags".

## Consultas

Todas las consultas que realizamos a MongoDB las hicimos a través de la operación mapReduce. La primera consulta que realizamos es la de contar la cantidad de tweets y agruparlos por la escala de polaridad. Además, se realizó la consulta que nos da información sobre la naturaleza y cantidad de los tweets, es decir, si es un tweet normal, una respuestas, un retweet o una cita. Por otro lado, se realizó la consulta que nos da información para la linea del tiempo, se consulto la fecha del tweet solo teniendo en cuenta el día, el mes y el año, con esta consulta se logro realizar el histograma. Finalmente, se realizo una consulta para recuperar la cantidad de los tweets según la subjetividad de este, con esta logramos generar el gráfico de subjetividad.

## Análisis de coyuntura sobre Twitter

Para el momento en que se realizó el analisis de coyuntura se tenían cerca de 900 tweets, por lo que se escogió un subgrupo de 99 tweets como muestra para realizar el análisis de coyuntura. Estos 99 tweets pertenecen todos al grupo de cuentas de twitter relacionadas con medios de comunicación, por ejemplo: hay tweets de cuentas como RCN Radio, El espectador, Caracol Noticias, entre otros. Al analizar el contenido de los tweets se encontró que la mayoría hacian referencia a noticias del acontecer nacional relacionadas con temas como: COVID, muertes violentas, narcotrafico, inseguridad y temas relacionados. Por lo tanto, se construyó una escala de polaridad que consta de 4 niveles que busca clasificar los tweets en neutrales (N), malos (M), muy malos (MM) y muy muy malos (MMM) (Anexo 4). Con esta escala se clasificaron manualmente los tweets y se creó un archivo .arff para que con la ayuda de la herramienta Weka se pudiera preprocesar la muestra y generar modelos de clasificación basados en esta escala.

Una vez se cargaron los tweets a Weka con ayuda del archivo .arff, se procede a preprocesar la información. Para esto, nos ayudamos de los filtros de Weka, específicamente usamos el filtro no supervisado “StringToWordVector”, este filtro nos ayuda a separar los tweets en diferentes grupos. Para este ejercicio se realizaron tres metodos diferentes de preprocesado. En el primer caso se dividió el tweet por estos caracteres especiales ““ \r\n\t.,;:\'\"()?!-¿¡+\*&#\\$%\V=<>[]\_`@".” y no se permitió la agrupación de palabras. En el segundo caso, se dividieron los tweets por estos caracteres especiales “\r\n\t.,;:\'\"()?!” y se permitió la agrupación de 2 objetos o palabras. En el tercer caso se dividieron los tweets por estos caracteres especiales “\r\n\t.,;:\'\"()?!” y se permitió la agrupación de 3 objetos o palabras (Anexo 5). Una vez aplicado el filtro sobre nuestra información se encontró que algunas palabras o argupaciopnes de palabras no tenían sentido, como los números, las direcciones web o flechas. Por lo tanto, se buscaron y eliminaron estos atributos de los datos. Finalmente, se realizó el proceso de clasificación con la ayuda de los algoritmo NaiveBayes, NaiveBayesMultinomial, SMO y RandomForest.

Instancias clasificadas correctamente (%)				
Preprocesamiento	Algoritmos			
	NaiveBaves	NaiveBayesMultinomial	RandomForest	SMO
1 Palabra	42%	47,47%	37,37%	46,46%
2 Palabras	44%	47,47%	38,38%	43,43%
3 Palabras	45,54%	41,41%	35,35%	41,41%

Tabla 4: Resultado de instancias clasificadas correctamente con detalle de preprocesamiento y algoritmo utilizado.

En la tabla 4 se muestran los resultados del porcentaje de las instancias clasificadas correctamente de estos algoritmos para cada tipo de preprocesamiento, se resaltarón las 3 mejores metricas. Es importante tener en cuenta que el metodo de testeo realizado fue Cross validation con 10-Folds. En la carpeta de documentos del github del grupo se encuentra el detalle de los resúmenes de este ejercicio. Se encontró que el modelo que mejor calificación tuvo en cuanto a instancias clasificadas y medidas de error fue el NaiveBayesMultinomial con preprocesado de una palabra, ver tabla 5. Como regla encontramos que el modelo de NaiveBayesMultinomial tiene mejor rendimiento que el modelo de NaiveBayes lo cual se puede deber a que el modelo multinomial considera la frecuencia de aparición de los terminos en los tweets. También, el modelo de RandomForest fue el que obtuvo peores metricas. En cuanto al preprocesamiento no se encontró evidencia que demuestre que la división de n palabras mejore las metricas de los modelos. Por último, es importante resaltar que estos modelos aún tienen espacio para mejorar sus metricas, por lo tanto, se recomienda realizar un mejor preprocesamiento usando tecnicas como lematización o normalización de los datos y/o mejorar los hiperparametros de los algoritmos acá propuestos para realizar la clasificación.

Algoritmo	Preprocesamiento	Total instancias	Instancias clasificadas correctamente	% Instancias clasificadas correctamente	Mean absolute error	Root mean squared error
NaivaBayesMultinomial	1 Palabra	99	47	47%	0,2783	0,4251
	2 Palabras	99	47	47%	0,272	0,4477
SMO	1 Palabra	99	46	46,46%	0,3291	0,4194

Tabla 5: Detalle de algoritmo Naive Bayes y SMO más el pre-procesamiento realizado. Valores de instancias y métricas de error.

## **Aplicación WEB**

Para la construcción de la aplicación Web se utilizó el framework Django con el lenguaje de programación Python para el front-end, el cual tiene como objetivo mostrar la información y recibir parametros de entrada. Este front se complementa con un JOB que tiene como tarea asegurarse que la información nueva que ingrese a la base de datos se procese por bloques y de esta manera los tweets obtengan su analisis de polaridad, el calculo de la subjetividad y la asignación de tags.

Es importante mencionar que este JOB trabaja con la librería TextBlob la cual realiza un preprocesamiento de los tweets en donde elimina todos los caracteres especiales y los separa por una palabra, como vimos en nuestro ejercicio de estudio de modelos este preprocesamiento es adecuado. Para el análisis de polaridad de los tweets con TextBlob se realizó una nueva escala. Teniendo en cuenta que textblob clasifica la polaridad de tweet en un indice de entre -1 y 1 se creo una escala con cinco elementos que van entre malo y excelente (Anexo 6). Además, Textblob genera otro indice de subjetividad el cual es un indice de 0 a 1, esta subjetividad también se almacena para luego ser mostrada en la aplicación. Para el procesamiento de los tweets, Textblob tiene varias opciones, en este caso utilizamos la opcion de procesar los tweets con el algoritmos de NaiveBayes, que como vimos en nuestro ejercicio con Weka es útil, aunque sería ideal poder normalizar las palabras antes de ser procesadas y de esta manera se buscaría obtener una mejor clasificación por parte del algoritmo.

Por otro lado, en el back-end tenemos una base de datos NoSQL de MongoDB la cual almacena los tweets procesados y la información relacionada a estos. En el back-end MongoDB esta configurado para recibir solicitudes de consultas del front las cuales usan la estrategia MapReduce para el procesamiento escalable, para utilizar mapReduce MongoDB cuenta con un comando especifico (Anexo 7). Las operaciones mapReduce reciben documentos de una colección como input y estan en la capacidad de realizar clasificaciones o limitaciones arbitrarias antes de realizar el map. Como output esta operación puede generar un documento o almacenar los resultados en las colecciones de la base de datos.

Para la visualización de resultados se utilizó la librería Chart.js que nos permite crear gráficos dinámicos de diferentes tipos. Se construyó una lista desplegable que tiene las cuentas de twitter que obtuvimos para nuestra aplicación, en esta lista se puede elegir la cuenta en la que estamos interesados y dar clic en el botón “Obtener Resultados” para desplegar la información de los tweets. Se construyeron dos contadores, uno de tweets y otro de palabras. Además, se tienen cinco gráficos que describen los tweets en atributos como subjetividad, tags, tipo de tweet, clasificación y fecha. Finalmente, se tienen 3 tablas que muestran los tweets con su tipo, escala y subjetividad y, además, la cuenta de los hashtags que se tienen en la cuenta.

## **Análisis de resultados**

La ideación y construcción de una aplicación web para el análisis de polaridad de tweets fue una tarea con un grado de dificultad alta. Requirió el uso de diferentes aplicaciones, lenguajes y librerías. La gran dificultad encontrada fue la integración de Weka en nuestra aplicación Web, tarea que no se pudo lograr y por lo tanto se opto por la integración con la librería TextBlob. Esta segunda librería nos permitió implementar nuevas herramientas de análisis que no teníamos en cuenta como la subjetividad y los tags de los tweets. Como gran logro encontramos el uso de la base de datos MongoDB, la cual nos permitió

utilizar la herramienta mapReduce para el procesamiento escalable. Además, trabajar MongoDB con Studio 3T el cual nos facilitó la carga de información, las consultas y la visualización de los tweets en la base de datos. Por otro lado, como posibles mejoras se encuentran varias oportunidades. En primer lugar, se propone la posibilidad de analizar más de una cuenta de twitter a la vez, viendo los temas en común, las clasificaciones de los tweets y las relaciones entre las cuentas. En segundo lugar, se propone la integración de los “trending topics” de twitter en la aplicación para lograr algún tipo de análisis con esta información. Y, por último, se propone realizar un pre-procesamiento de información más profundo con el objetivo de obtener mejores métricas. Al implementar estas propuestas en la aplicación web se espera que el análisis realizado se pueda multiplicar en la dirección correcta.

## Conclusiones

Luego de idear y construir una aplicación Web para el análisis de coyuntura sobre Twitter se llegó a las siguientes conclusiones:

- La integración de herramientas útiles como Weka puede traer más problemas que soluciones cuando no se cuenta con la experiencia necesaria.
- El pre-procesamiento de información es un paso necesario e iterativo para la aplicación de algoritmos de clasificación.
- El uso de bases de datos NoSQL como MongoDB facilita y permite el uso de herramientas como MapReduce.
- Existen diferentes herramientas en el mercado, como TextBlob, para el análisis de polaridad de tweets.

## Bibliografía

- Chart.js (2020). Chart.js documentation. URL: <https://www.chartjs.org/docs/latest/>
- Django (2020). Django documentation. URL: <https://docs.djangoproject.com/en/3.1/>
- MongoDB (2020). MongoDB Documentation. URL: <https://docs.mongodb.com/>
- Plaza Sacarrera, Lucía. (2014). Análisis de polaridad en textos escritos en inglés y español. Universidad Calors III de Madrid. URL: <https://e-archivo.uc3m.es/handle/10016/22213>
- TextBlob. (2020). TextBlob: Simplified Text Processing. URL: <https://textblob.readthedocs.io/en/dev/>
- Weka (2020). Weka documentation. URL: <https://waikato.github.io/weka-wiki/documentation/>

## Anexos

### Anexo 1: Aplicación .NET para descarga de tweets.

Get Tweets

URL API Twitter

```
https://api.twitter.com/2/tweets/search/recent?
query=from:{0} {1}
&expansions=author_id&user.fields=username&ma
x_results=100&tweet.fields=id,text,created_at,lang,
referenced_tweets
```

Count Search

1

New Word

Add Word

New Account

Add Word

Bearer Token

```
AAAAAAAAAAAAAAAAAAAAAF9HJAEAAAAAOC6tB4
FLNpav548pT9BKWytX%2BA
%3D8DKYabiHXJw8AKrqMNGNGBs0FHKiXrVFE37H
3mY39viqUI3uq
```

Query Words

muertes  
violencia  
desplazamiento  
matoneo

Query Accounts

COL\_EJERCITO  
Ejercito\_Div7  
PoliciaColombia  
EjercitoCol

## Anexo 2: Lista de cuentas de twitter y grupo de clasificación.

#	Entidades en Colombia	Grupo de clasificación
1	COL_EJERCITO	Entidades en Colombia
2	Ejercito_Div7	Entidades en Colombia
3	PoliciaColombia	Entidades en Colombia
4	EjercitoCol	Entidades en Colombia
5	Ejercito_Div6	Entidades en Colombia
6	UnidadVictimas	Entidades en Colombia
7	PosconflictoCO	Entidades en Colombia
8	CamaraColombia	Entidades en Colombia
9	ComisionVerdadC	Entidades en Colombia
10	MinInterior	Entidades en Colombia
11	infopresidencia	Entidades en Colombia
12	PGN_COL	Entidades en Colombia
13	ARNColombia	Entidades en Colombia
14	FiscaliaCol	Entidades en Colombia
15	PoliciaCali	Entidades en Colombia
16	PoliciaColombia	Entidades en Colombia
17	DIJINPolicia	Entidades en Colombia
18	DEAHQ	Entidades en Colombia
19	PoliciaAntiNar	Entidades en Colombia
20	Mindefensa	Entidades en Colombia
21	GaulaMilitares	Entidades en Colombia
22	MinSaludCol	Entidades en Colombia
23	infopresidencia	Entidades en Colombia
24	SenadoGovCo	Entidades en Colombia
25	FuerzasMilCol	Entidades en Colombia
26	FuerzaAereaCol	Entidades en Colombia
27	UDF_Medellin	Entidades en Colombia
28	UNALOOficial	Entidades en Colombia
29	PosconflictoSM	Entidades en Colombia
30	Concejo de Bogotá	Entidades en Colombia
31	PoliciaCauca	Entidades en Colombia
32	GobCauca	Entidades en Colombia
33	GobValle	Entidades en Colombia
34	GoberArauca	Entidades en Colombia
35	PoliciaArauca	Entidades en Colombia
36	Fenalco_Ant	Entidades en Colombia
37	GobAntioquia	Entidades en Colombia
38	GobiernoAnt	Entidades en Colombia
39	region6policia	Entidades en Colombia
40	Bogota	Ciudades
41	AlcaldiadeMed	Ciudades
42	AlcaldiaDeCali	Ciudades
43	fdbedout	Periodistas
44	VickyDavilaH	Periodistas
45	rcnradio	Medio de comunicación
46	NoticiasCaracol	Medio de comunicación

#	Entidades en Colombia	Grupo de clasificación
47	NoticiasUno	Medio de comunicación
48	NoticiasRCN	Medio de comunicación
49	RevistaSemana	Medio de comunicación
50	ELTIEMPO	Medio de comunicación
51	elespectador	Medio de comunicación
52	wRadioColombia	Medio de comunicación
53	CaracolRadio	Medio de comunicación
54	Caracol_Cali	Medio de comunicación
55	BluRadioCo	Medio de comunicación
56	Citytv	Medio de comunicación
57	lafm	Medio de comunicación
58	NotivisionCauca	Medio de comunicación
59	caucahoy	Medio de comunicación
60	BLUAntioquia	Medio de comunicación
61	NCAntioquia	Medio de comunicación
62	radio_armenia	Medio de comunicación
63	PalomaValenciaL	Políticos
64	AlvaroUribeVel	Políticos
65	petrogustavo	Políticos
66	IvanCepedaCast	Políticos
67	GustavoBolívar	Políticos
68	JERobledo	Políticos
69	ClaudiaLopez	Políticos
70	angelamrobledo	Políticos
71	AABenedetti	Políticos
72	IvanDuque	Políticos
73	CarlosHolmesTru	Políticos
74	Felicianp Valencia	Políticos
75	HOLLMANMORRIS	Políticos
76	RoyBarreras	Políticos
77	sergio_fajardo	Políticos
78	ColombiaHumana_	Partidos políticos
79	UP_Colombia	Partidos políticos
80	MovimientoMAIS	Partidos políticos
81	PartidoFARC	Partidos políticos
82	CeDemocratico	Partidos políticos
83	ONIC_Colombia	Partidos políticos
84	PartidoVerdeCol	Partidos políticos
85	Compromiso_Ant	Partidos políticos
86	ComunesANT	Partidos políticos
87	PConservadorAnt	Partidos políticos
88	Upantioquia	Partidos políticos
89	MisionONUCol	Entidades No Gub
90	ONUHumanRights	Entidades No Gub
91	ONU_derechos	Entidades No Gub
92	violencolombia	Entidades No Gub

### Anexo 3: Ejemplo Tweet en formato JSON.

```
{
  "created_at": "2020-11-02T22:25:53Z",
  "text": "El más reciente informe del Ministerio de Salud confirmó este lunes 7.992 pacientes recuperados en Colombia.",
  "id": "1323390902019235841",
  "lang": "es",
  "author_id": "9633802"
}
```

### Anexo 4: Escala de clasificación tweets para análisis con Weka.

Atributo	Clasificación
N	Neutrales
M	Malos
MM	Muy malos
MMM	Muy muy malos

### Anexo 5: Ejemplo pre-procesamiento de los tweets en n-grupos.

1 Palabra	2 Palabras	3 Palabras
apoyo	aumento de	dejan al menos
anunciada	autoridades intervinieron	diferentes regiones del
comunitaria	ayuda económica	esperanza de hallar
diferentes	información sobre	para su financiamiento

### Anexo 6: Escala de clasificación tweets en aplicación Web.

Escala app Web	Clasificación Numérica TextBlob
Muy bueno	De 1 a 0,5
Bueno	De 0,5 a 0
Neutral	0
Malo	De 0 a - 0,5
Muy Malo	De - 0,5 a -1



## Anexo 7: Operación MapReduce en MongoDB.

```
Collection
↓
db.orders.mapReduce(
  map    → function() { emit( this.cust_id, this.amount ); },
  reduce → function(key, values) { return Array.sum( values ) },
  {
    query → { status: "A" },
    output → "order_totals"
  }
)
```