Clasificación de Noticias

"Si torturas los datos los suficiente, acabarán confesando cualquier cosa"
- Fred Menger, profesor de química e investigador -

http://www.chemistry.emory.edu/faculty/menger/index.html

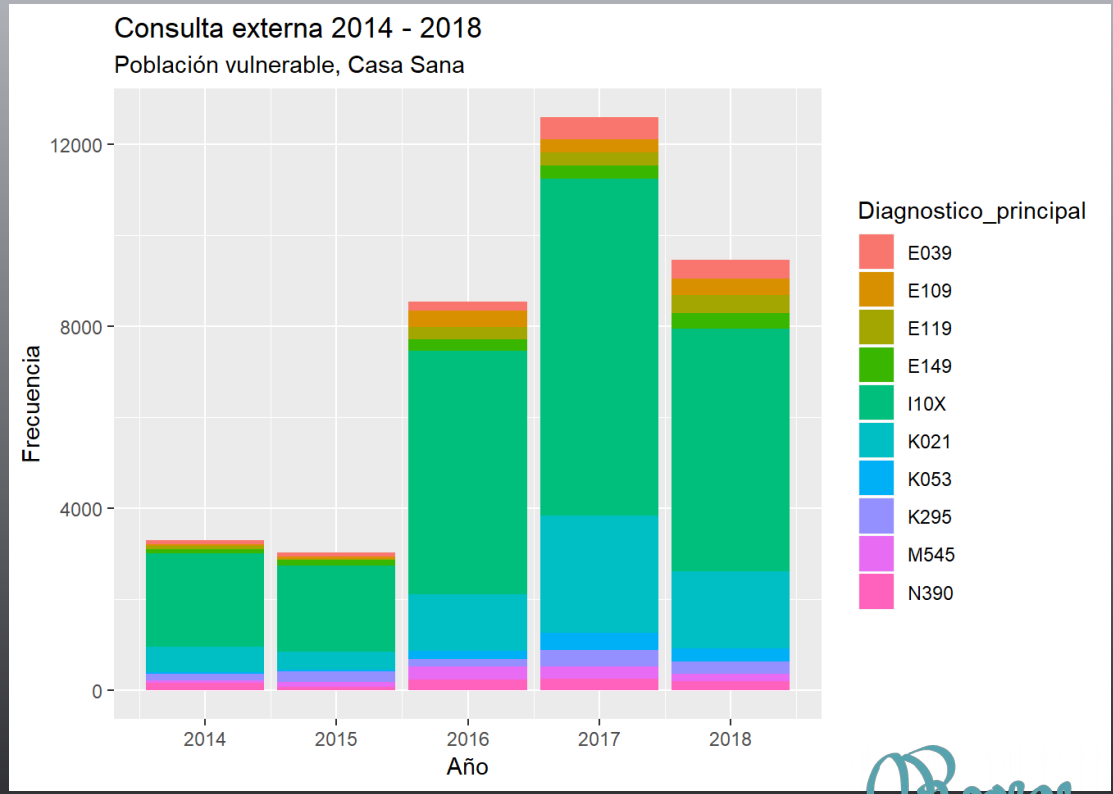
Oscar Riojas

Ingeniero mecánico de profesión, pero apasionado de las tecnologías de ML y el análisis de datos, experiencia analizando datos en el sector de salud y en arte, entusiasta maker por hobby, experto en modelado 3D e impresión 3D

Roojas

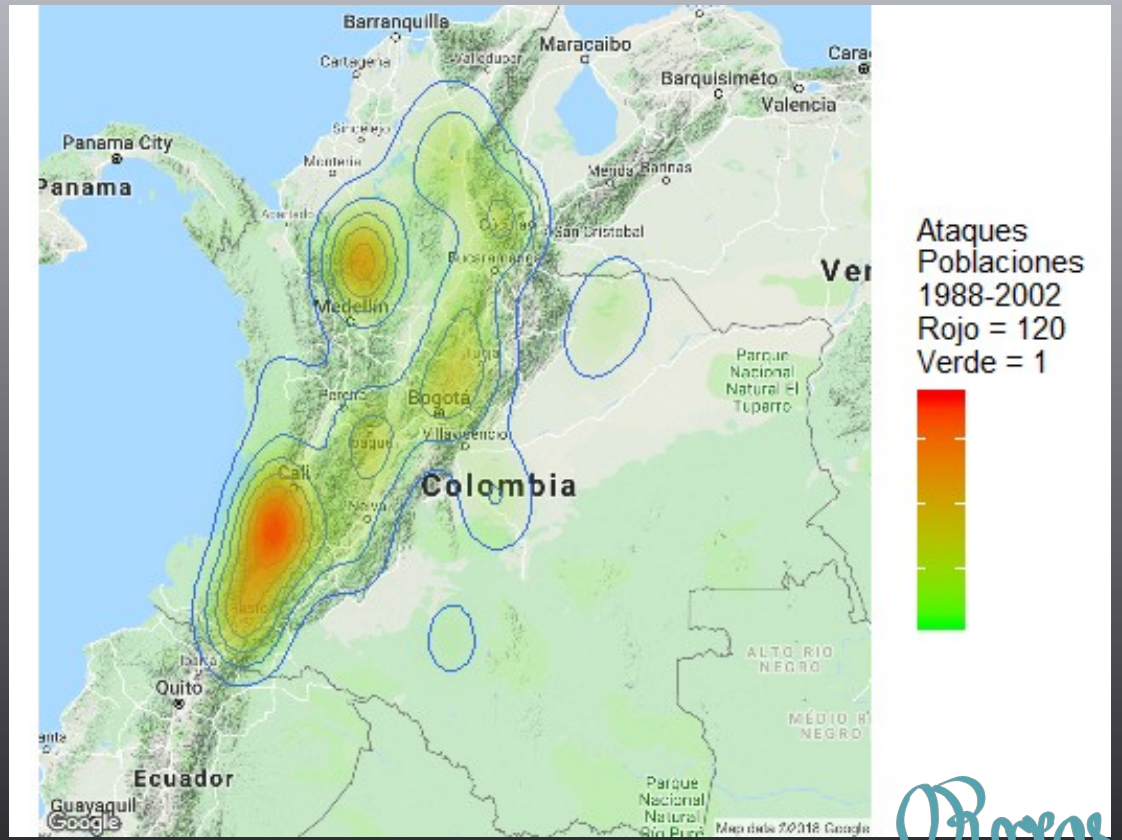
Experiencia

Análisis de información RIPS, para secretaria de salud Pereira, visualizaciones y modelo para predecir patologías con SVM con 87% de efectividad, R

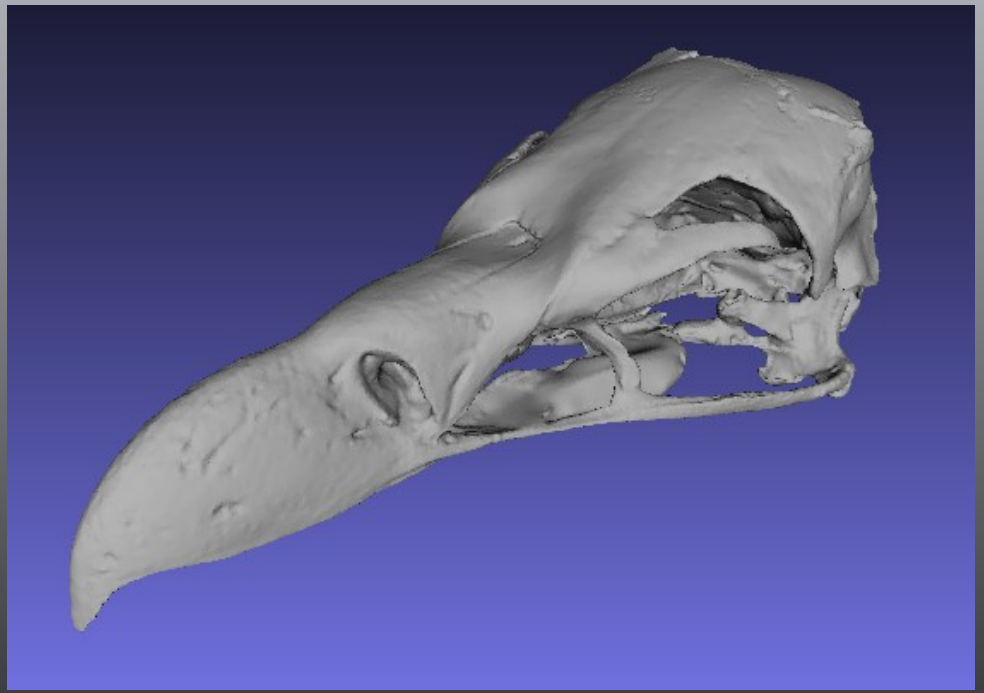


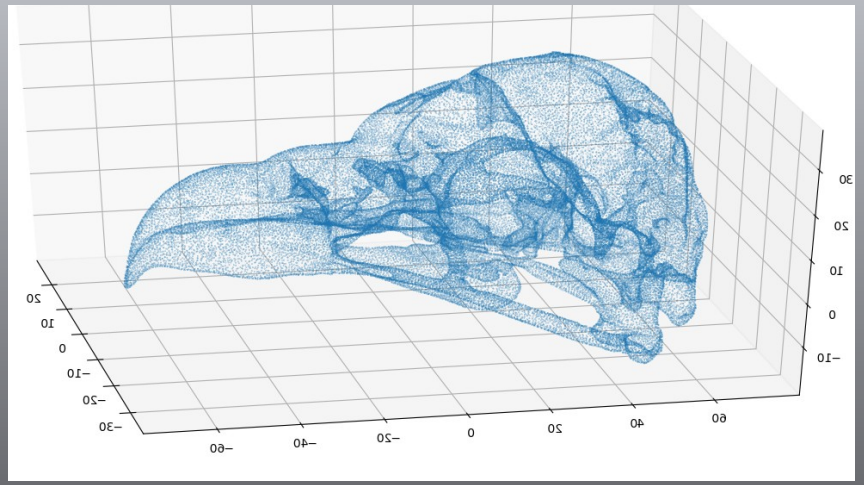
Experiencia

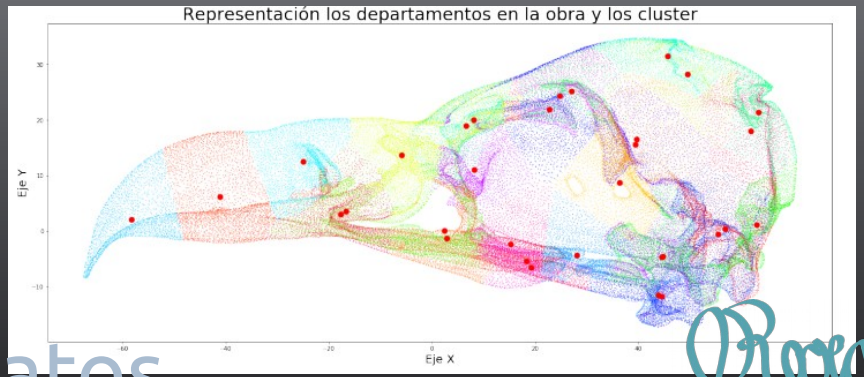
Proyecto Cóndor DNA+Art, visualización de datos del conflicto armado en Colombia con información abierta R



Experiencia



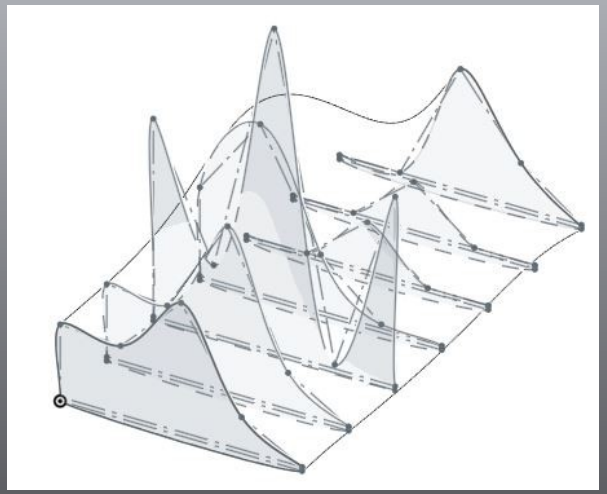


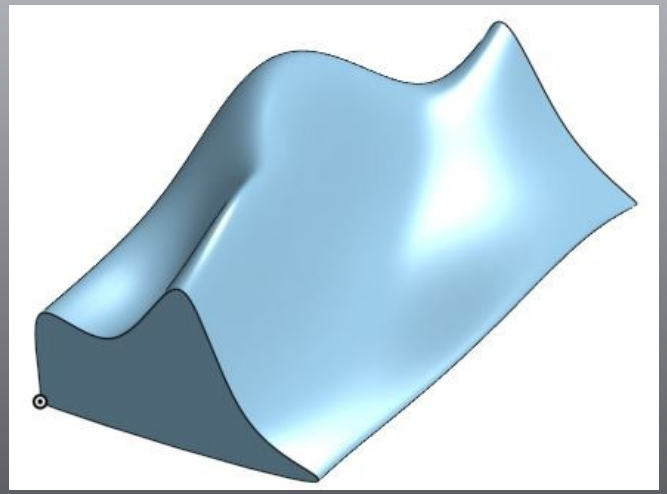


Visualización artística de datos

Rocas

Experiencia

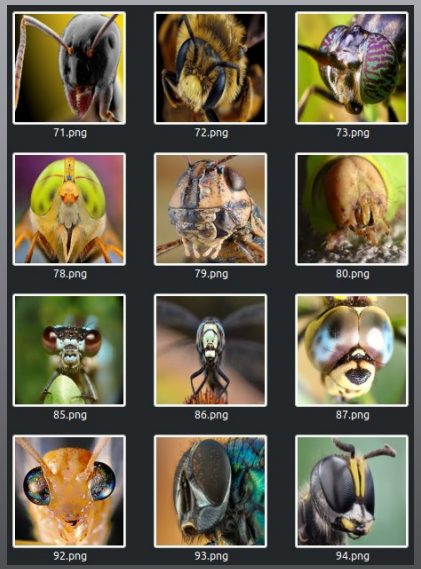


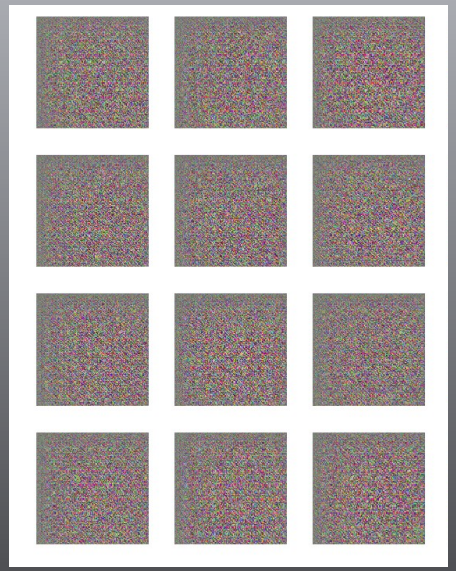


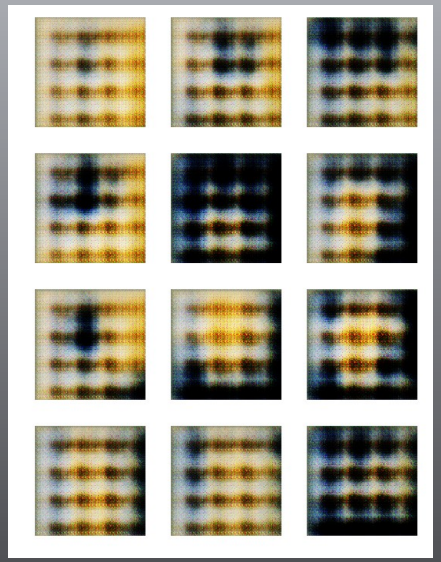
Visualización artística de datos

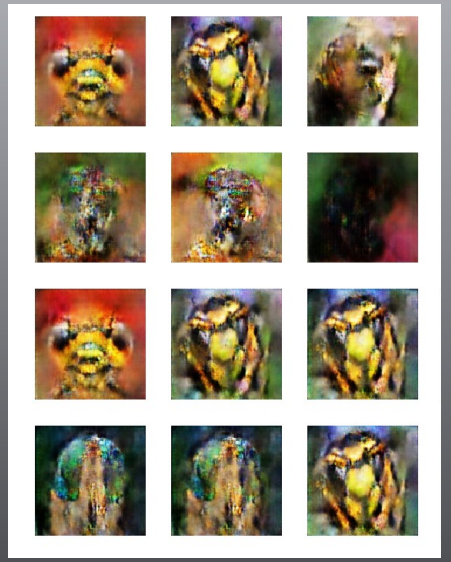
Roopas

Experiencia





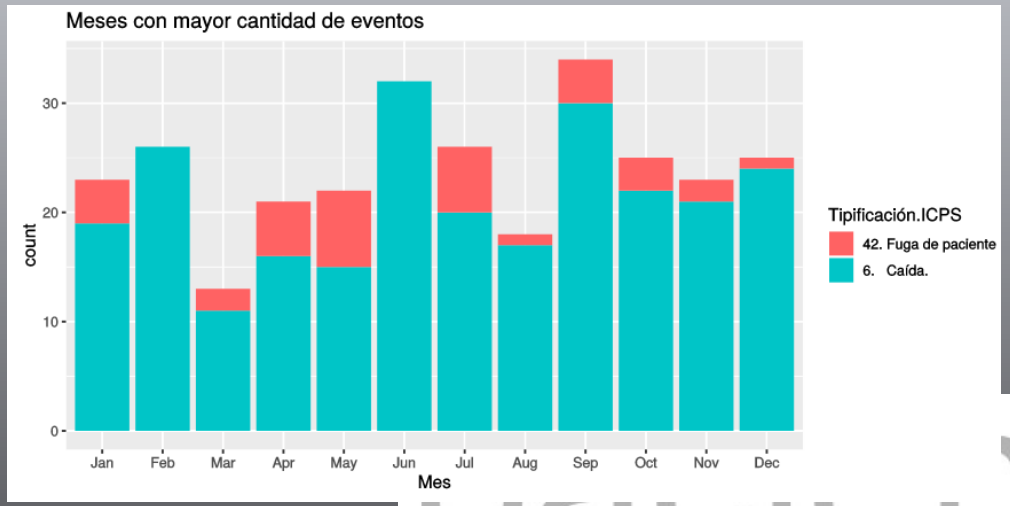




Obra, generación de insectos con la arquitectura GAN, para exposición junto Julián Anibal Henao

Rooyas

Experiencia



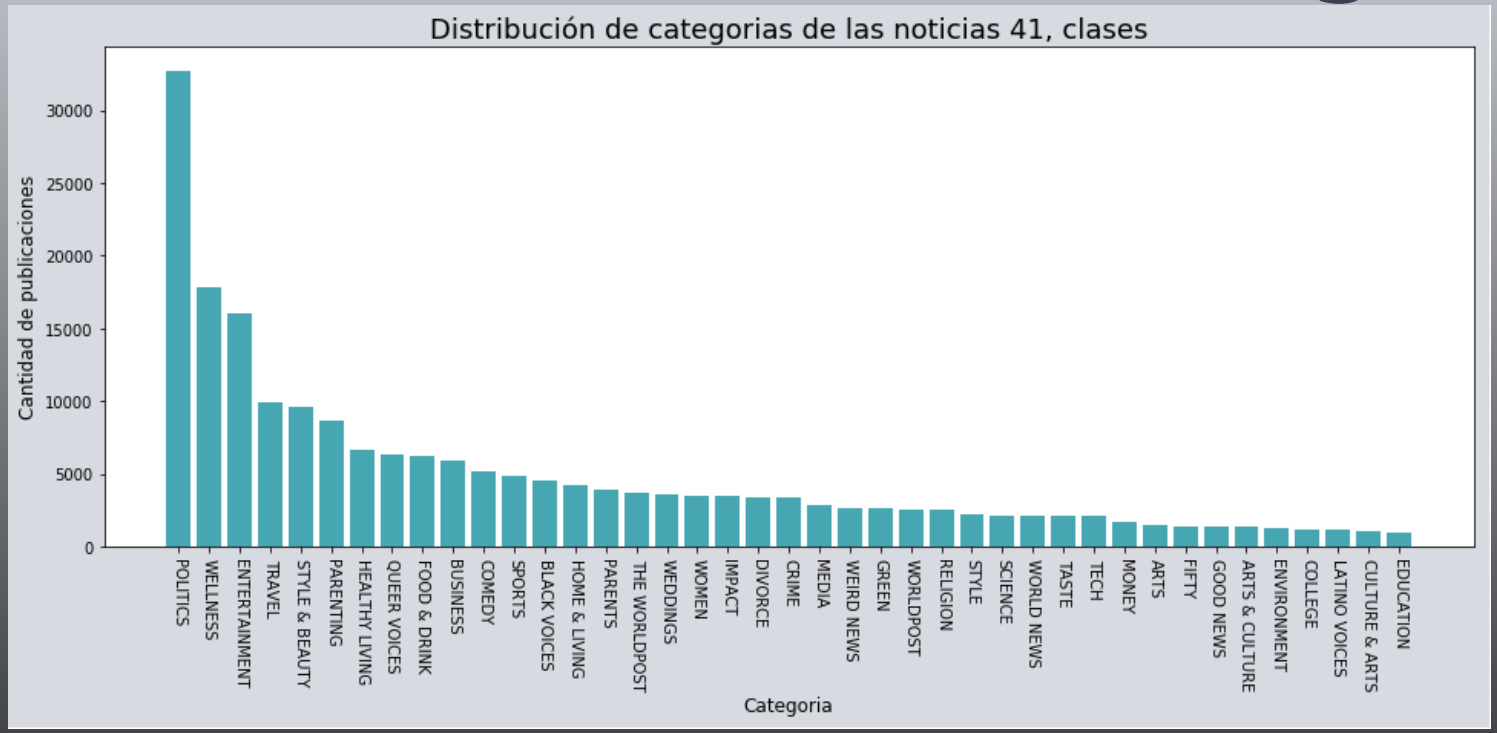
Análisis de base de datos de seguridad del paciente Clínica Nogales

Preguntas a responder...

Análisis datos de 200 mil titulares de noticias del año 2012 a 2018 obtenidos de HuffPost.

- ¿Se pueden catalogar las noticias con la descripción y los titulares?
- ¿Existen estilos de escritura asociados a cada categoría?
- ¿Qué se puede decir de los autores?
- ¿Qué información útil se puede extraer de los datos?

Distribución Categorías

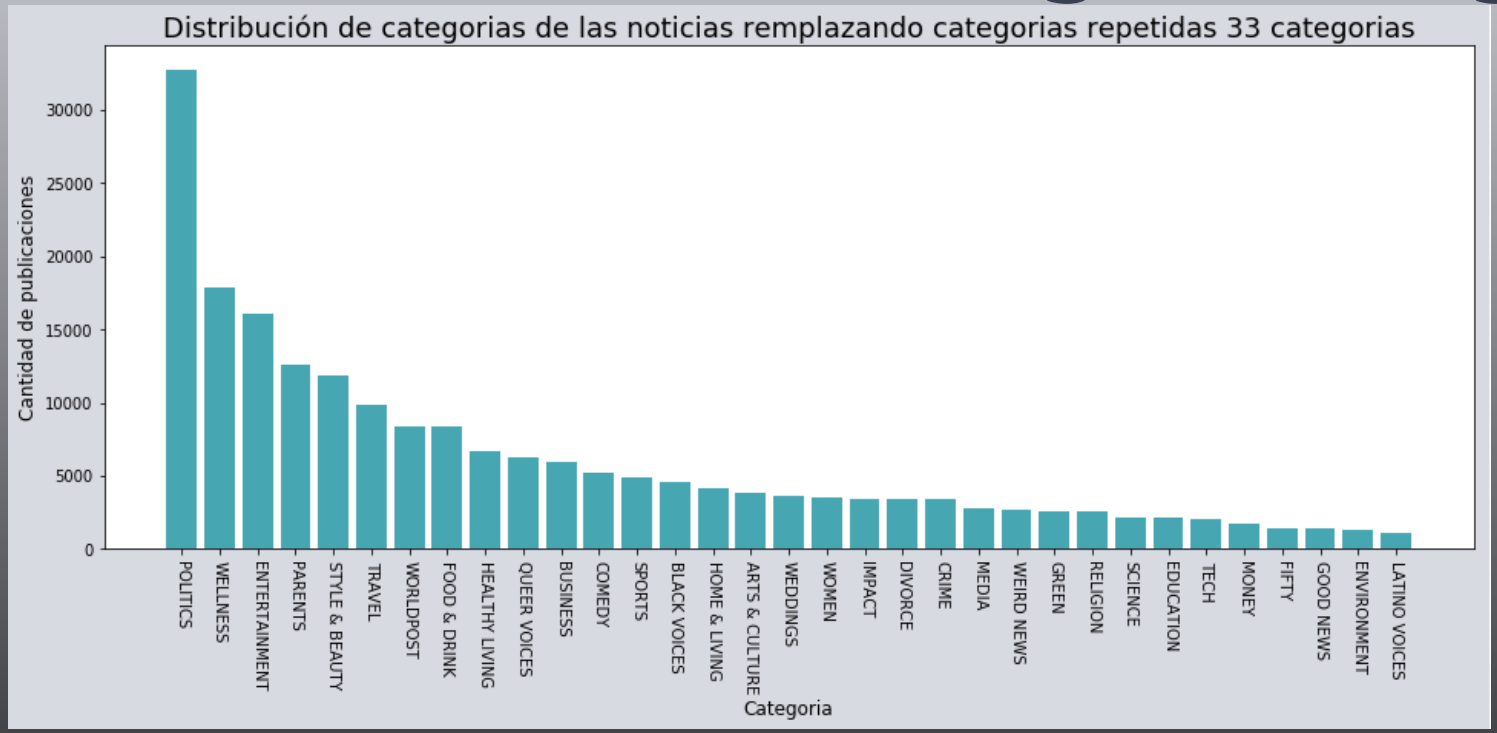


	Frecuencia	Categoria
0	32739	POLITICS
1	17827	WELLNESS
2	16058	ENTERTAINMENT
3	9887	TRAVEL
4	9649	STYLE & BEAUTY
5	8677	PARENTING
6	6694	HEALTHY LIVING
7	6314	QUEER VOICES
8	6226	FOOD & DRINK
9	5937	BUSINESS

Se tienen 41 categorías, y no tienen datos NA, 200853 observaciones, las tres categorías mas relevantes ENTERTAINMENT, POLITICS y WORLDPOST son el 25.58 % de las noticias, un total de 51376 noticias

Rocas

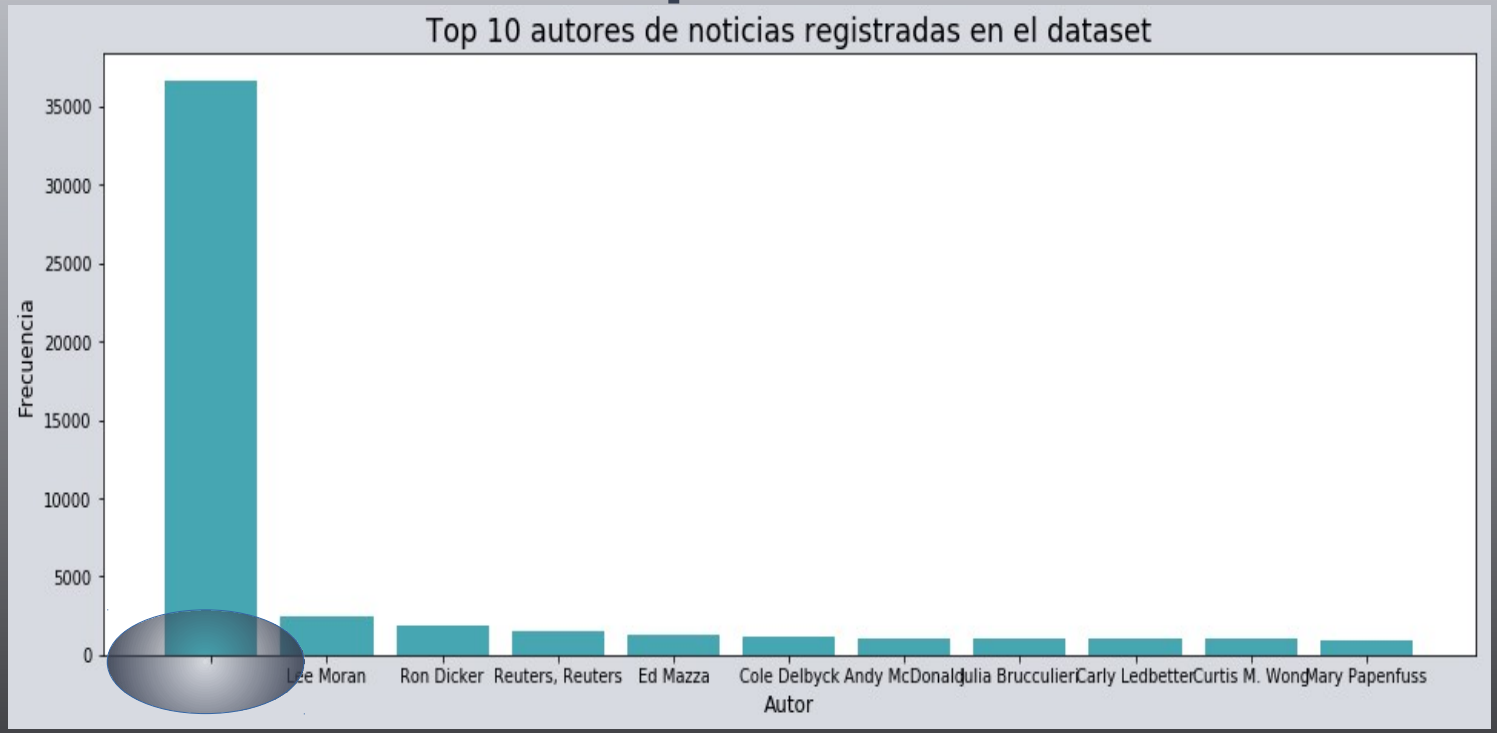
Distribución Categorías Agrupada



	Frecuencia	Categoria
0	32739	POLITICS
1	17827	WELLNESS
2	16058	ENTERTAINMENT
3	12632	PARENTS
4	11903	STYLE & BEAUTY
5	9887	TRAVEL
6	8420	WORLDPOST
7	8322	FOOD & DRINK
8	6694	HEALTHY LIVING
9	6314	QUEER VOICES

Se agruparon categorías similares para mejorar el desempeño del modelo balanceando un poco la clases de 41 a 33, las que se agruparon, por ejemplo ARTS y 'CULTURE & ARTS en ARTS & CULTURE, el top se mantiene igual

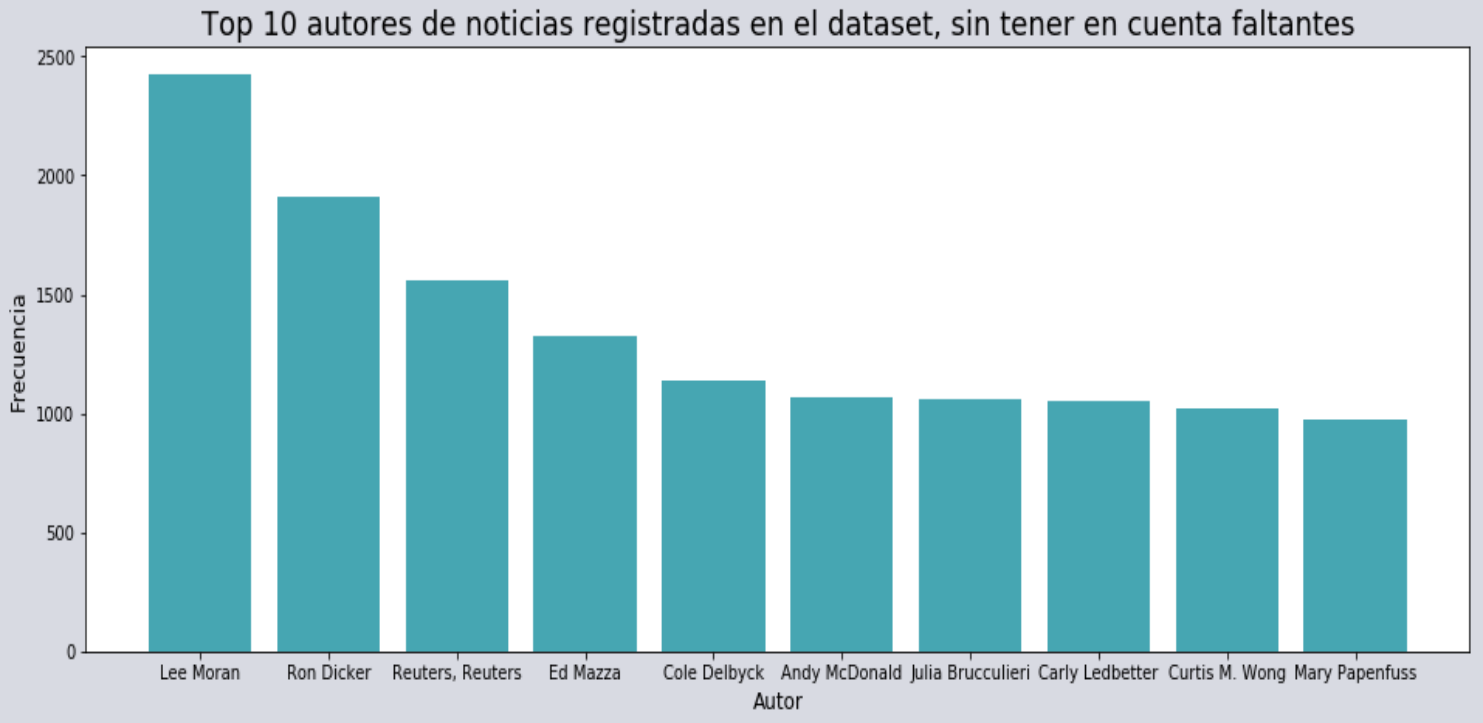
Top 10 de Autores



	Frecuencia	authors
0	36620	
1	2423	Lee Moran
2	1913	Ron Dicker
3	1562	Reuters, Reuters
4	1322	Ed Mazza
5	1140	Cole Delbyck
6	1068	Andy McDonald
7	1059	Julia Brucculieri
8	1054	Carly Ledbetter
9	1020	Curtis M. Wong
10	974	Mary Papenfuss

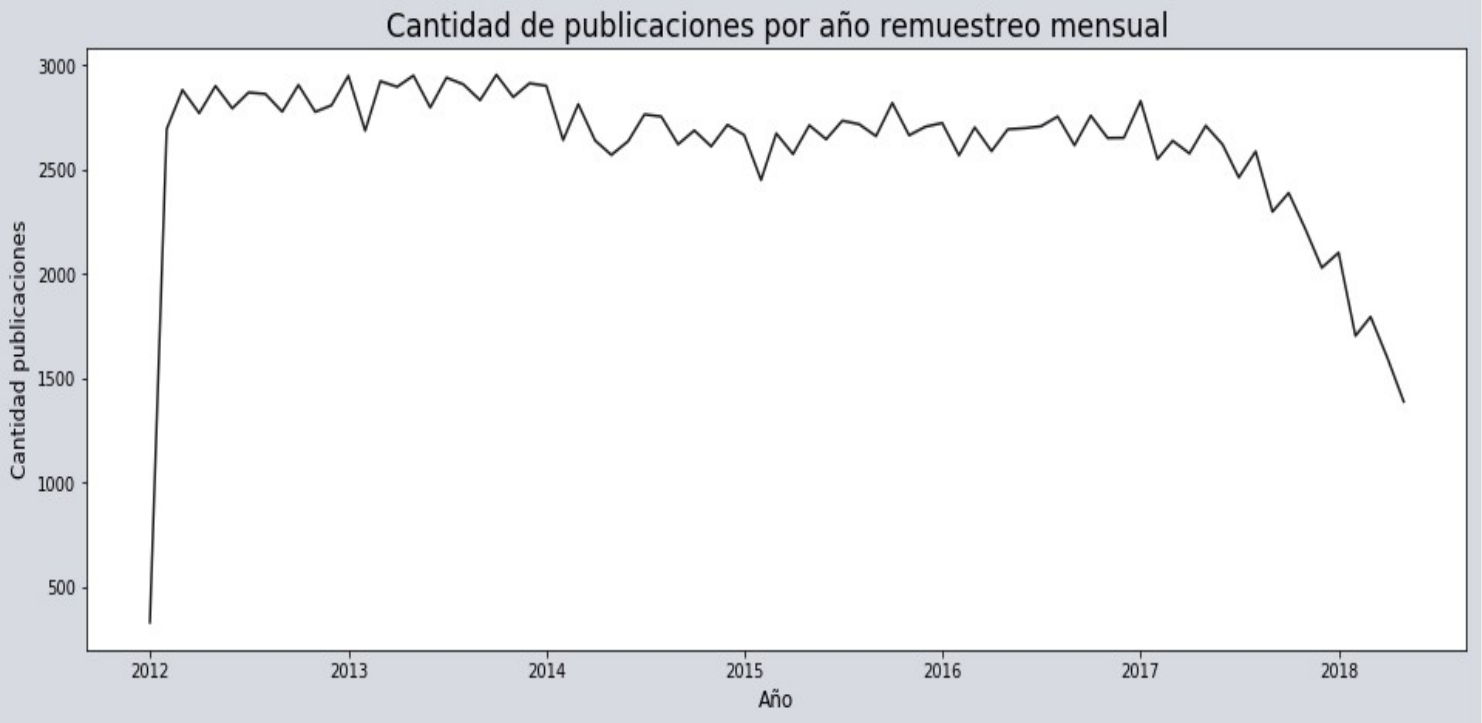
En el análisis se puede observar que se tienen 36620 registros sin el autor esto equivale al 18.23% del total de los artículos por esta razón se procede a visualizar los datos sin tener en cuenta los valores que no tienen el auto

Top 10 de Autores modificado



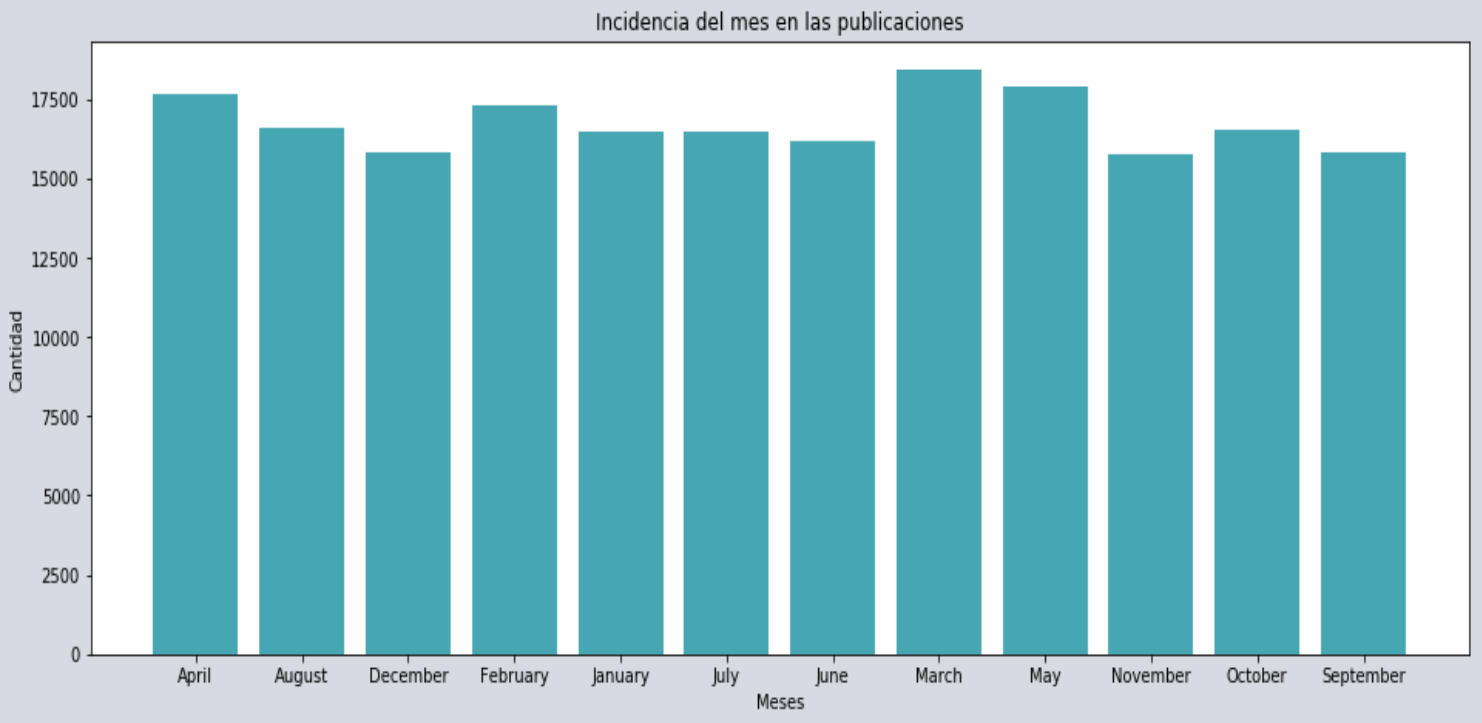
	Frecuencia	authors
1	2423	Lee Moran
2	1913	Ron Dicker
3	1562	Reuters, Reuters
4	1322	Ed Mazza
5	1140	Cole Delbyck
6	1068	Andy McDonald
7	1059	Julia Brucculieri
8	1054	Carly Ledbetter
9	1020	Curtis M. Wong
10	974	Mary Papenfuss

Fechas publicaciones



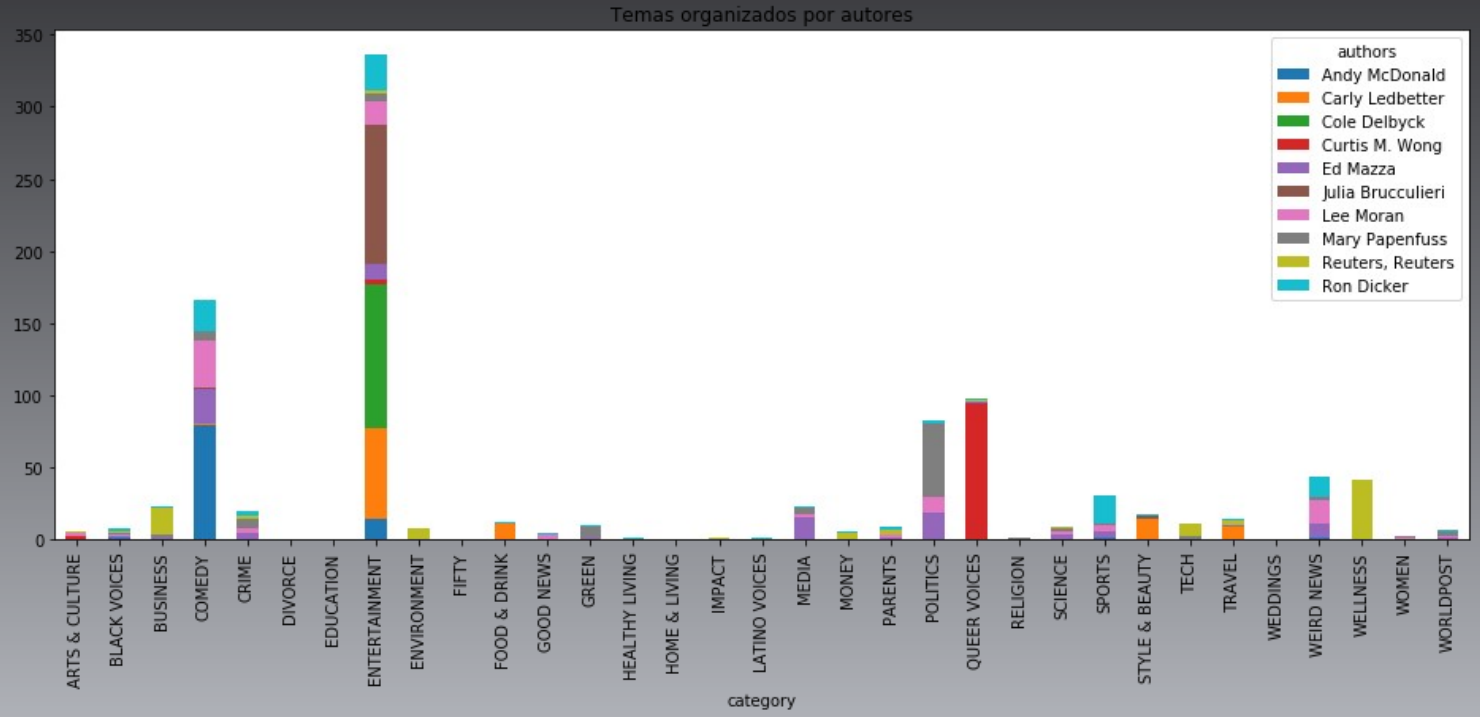
En el análisis de esta variable se puede observar como la cantidad de publicaciones tienen un tendencia a disminuir a partir del 2014

Fechas publicaciones



El mayor valor de publicaciones sumando los meses es: 18418 y el menor es : 15756, para un rango de 2662, este rango es menos al 1% del total de datos, por tanto no se puede suponer incidencia de la época del año para publicar

Autores Vs. Categoría



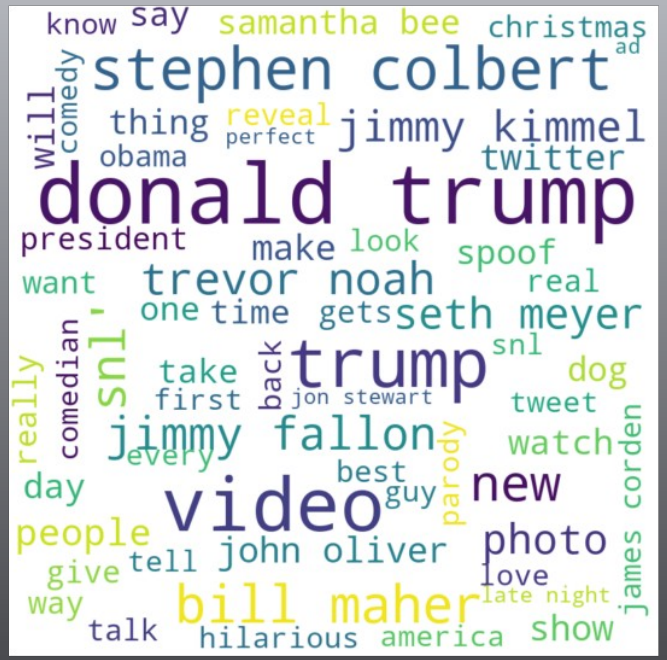
Se puede observar en los datos que los temas donde se publican mas noticias no son los mismo de los autores que mas publican, el top 10 de autores en publicar publican en la categoría ENTRERTAINMENT, esto hace mas desbalanceados los datos y al publicar un articulo en con estos datos la mayor probabilidad es que sea de política pero no de un autor del top 10

Autores, Categoría y Año

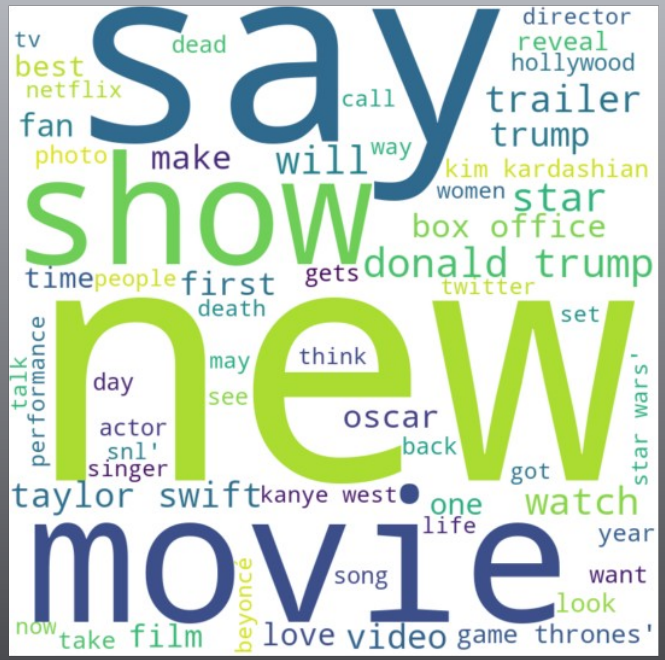
Con respecto a estas tres variables se puede observar como los autores que mas publican no son tan especializados y algunos no son tan activos en los diferentes años

Palabras predominantes

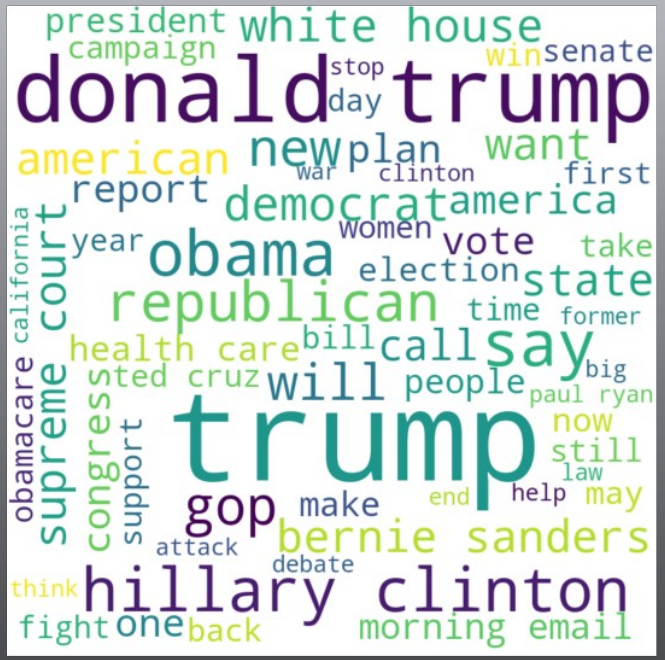
COMEDY



ENTERTAINMENT



POLITICS



Con respecto a esta variables se puede observar palabras de mucha frecuencia y se observa como la palabra “Trump” es recurrente en varias categorías de noticias

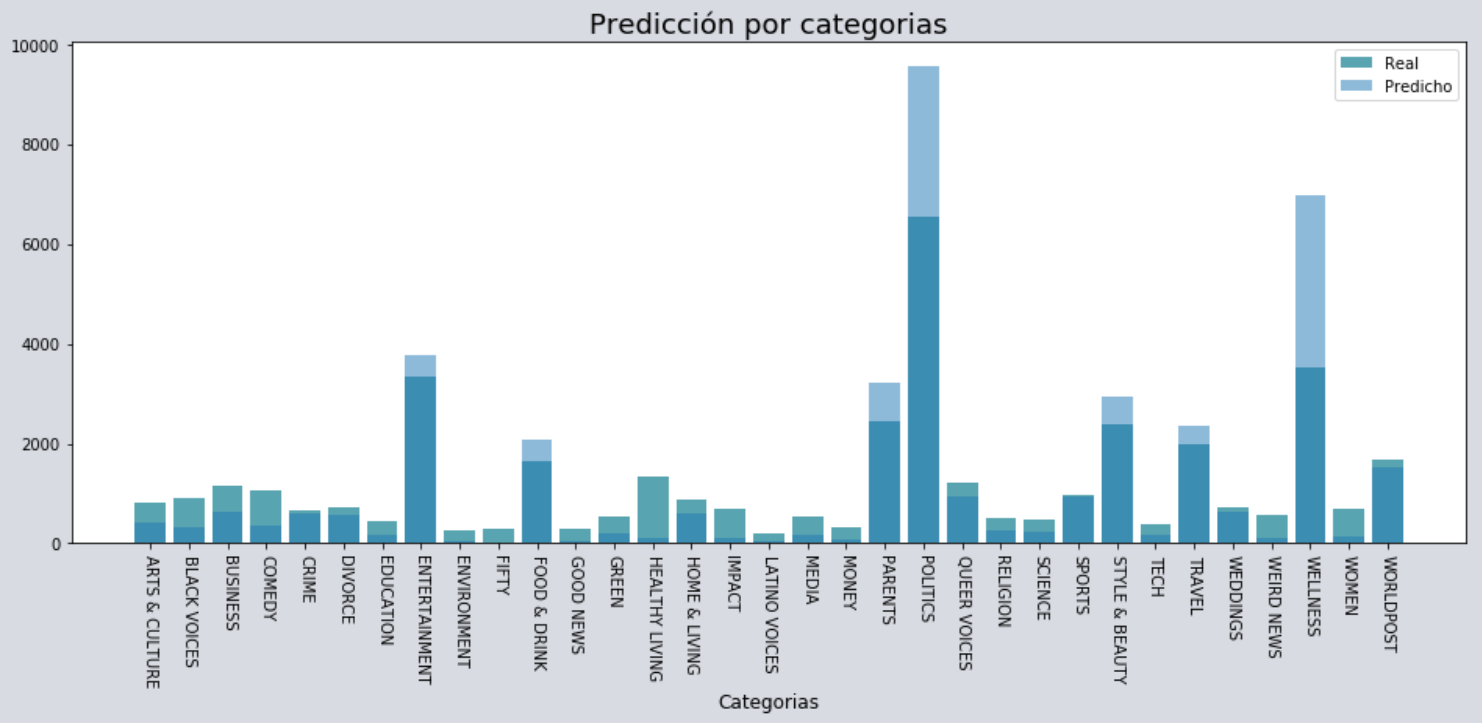
Modelos Aplicados

Modelo	Precisión	Comentario
Complement NB	61.32%	Modelo bayesiano, tiene la particularidad de funcionar bien con datos desbalanceados como en este caso
LinearSVC	underfitting	Este modelo no convergió de manera adecuada
NNT	19.28%	Para este tipo de clasificación se debe, investigar en una arquitectura adecuada

Modelos Aplicados

- Tiempo de ejecución modelo: CPU times: user 17.6 s, sys: 110 ms, total: 17.7 s Wall time: 17.7 s
- Parametros del modelo: ComplementNB(alpha=0.8, class_prior=None, fit_prior=True, norm=False)

Modelos Aplicados



Se puede observar como el modelo sobre clasifica algunas categorías y otras no, para próximos análisis es bueno en base a esta información ajustar los pesos, de la categoría del modelo

Respuestas datos

¿Se pueden catalogar las noticias con la descripción y los titulares?

Se puede catalogar la categoría de las noticias con una efectividad del 61,32% utilizando un clasificador Bayesiano, utilizando como datos de entrada la descripción y el titulo de la noticia

Respuestas datos

¿Existen estilos de escritura asociados a cada categoría?

Al analizar las palabras por las que están compuestos los artículos, se puede observar para cada categoría la relevancia de ciertas palabras clave, que definen su categoría de manera intuitiva

Respuestas datos

¿Qué se puede decir de los autores?

El top el top 10 de autores producen el 6.74% del total de los artículos, así mismo las categorías predominantes no están relacionadas en cantidad con los temas más publicados, la categoría mas publicada es politica, pero el top de autores escriben sobre entretenimiento, finalmente 36620 registros están sin autor esto equivale al 18.23%

Respuestas datos

¿Qué información útil se puede extraer de los datos?

- No se observa estacionalidad en las publicaciones no hay un mes en especial para publicar
- Las palabras “Donald Trump” aparecen predominantes en varias categorías, es decir generan mucha información en las publicaciones
- La tendencia a publicar esta bajando a partir del año 2014
- Algunas categorías estaban repetidas lo cual genera mayor des balance de los datos

Respuestas datos

¿Qué información útil se puede extraer de los datos?

- No se observa estacionalidad en las publicaciones no hay un mes en especial para publicar
- Las palabras “Donald Trump” aparecen predominantes en varias categorías, es decir generan mucha información en las publicaciones
- La tendencia a publicar esta bajando a partir del año 2014
- Algunas categorías estaban repetidas lo cual genera mayor des balance de los datos