CLASIFICADOR DE NOTICIAS

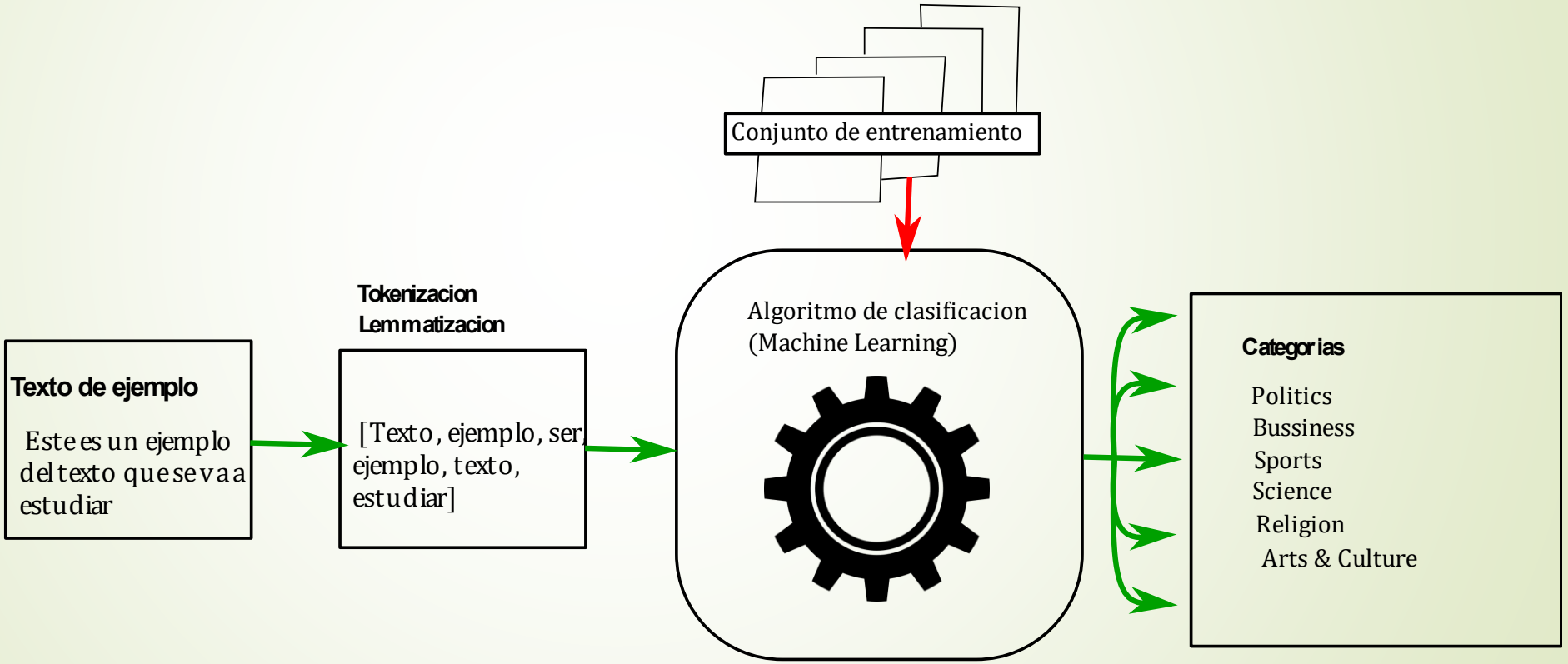
PRUEBA ML ENGINEER Whale & Jaguar

Cesar Nieto

Catalogando noticias a partir de el titulo y la descripción corta

Es posible catalogar noticias a partir de la descripción y los titulares?

Algortimo de clasificación



Desempeño del algoritmo de clasificación.

Entrenamiento con 40000 textos de prueba

Se utilizo regresión logística.

CATEGORY	precision	recall	f1-score	support
ARTS	0	0	0	4
ARTS & CULTURE	0.57	0.24	0.34	170
BLACK VOICES	0.57	0.38	0.46	331
BUSINESS	0.44	0.14	0.21	134
COLLEGE	0.33	0.08	0.12	13
COMEDY	0.69	0.45	0.54	418
CRIME	0.71	0.45	0.55	187
EDUCATION	0.45	0.22	0.29	65
ENTERTAINMENT	0.53	0.76	0.63	1244
FIFTY	0	0	0	11
GOOD NEWS	0	0	0	35
GREEN	0.52	0.16	0.25	147
HEALTHY LIVING	0.55	0.48	0.51	390
IMPACT	0.17	0.01	0.02	126
LATINO VOICES	0.8	0.1	0.18	120
MEDIA	0.61	0.23	0.34	198
PARENTS	0.62	0.58	0.6	333
POLITICS	0.63	0.93	0.75	3491
QUEER VOICES	0.82	0.65	0.73	475
RELIGION	0.42	0.12	0.19	120
SCIENCE	0.5	0.07	0.12	45
SPORTS	0.58	0.44	0.5	176
STYLE	0.5	0.26	0.34	133
TASTE	0.74	0.53	0.62	132
TECH	0.43	0.06	0.1	54
THE WORLDPOST	0.49	0.3	0.37	421
TRAVEL	0.67	0.33	0.44	88
WEIRD NEWS	0.51	0.19	0.28	181
WOMEN	0.47	0.34	0.39	289
WORLD NEWS	0.48	0.29	0.36	450
WORLDPOST	0	0	0	19
avg/total	0.59	0.6	0.56	10000

$$P = \frac{tp}{tp + fp}$$

$$R = \frac{tp}{tp + fn}$$

Estilo de escritura de cada categoria

¿Existen estilos de escritura asociados a cada categoría?

Algoritmo de estimación de palabras mas utilizadas

- Escoger el 50 artículos de cada categoría.
- Para cada categoría, tokenizar y lematizar cada articulo. Encontrar las 30 palabras mas frecuentes de cada articulo.
- Buscar las 50 palabras que mas aparecen en artículos.
- Una vez obtenidas las palabras mas comunes en cada categoría, descartar las que se repiten en 15 categorías o mas.
- Realizar el diagrama wordcloud con las palabras mas comunes y con la frecuencia que aparece en los artículos.

Palabras mas usadas por categoría

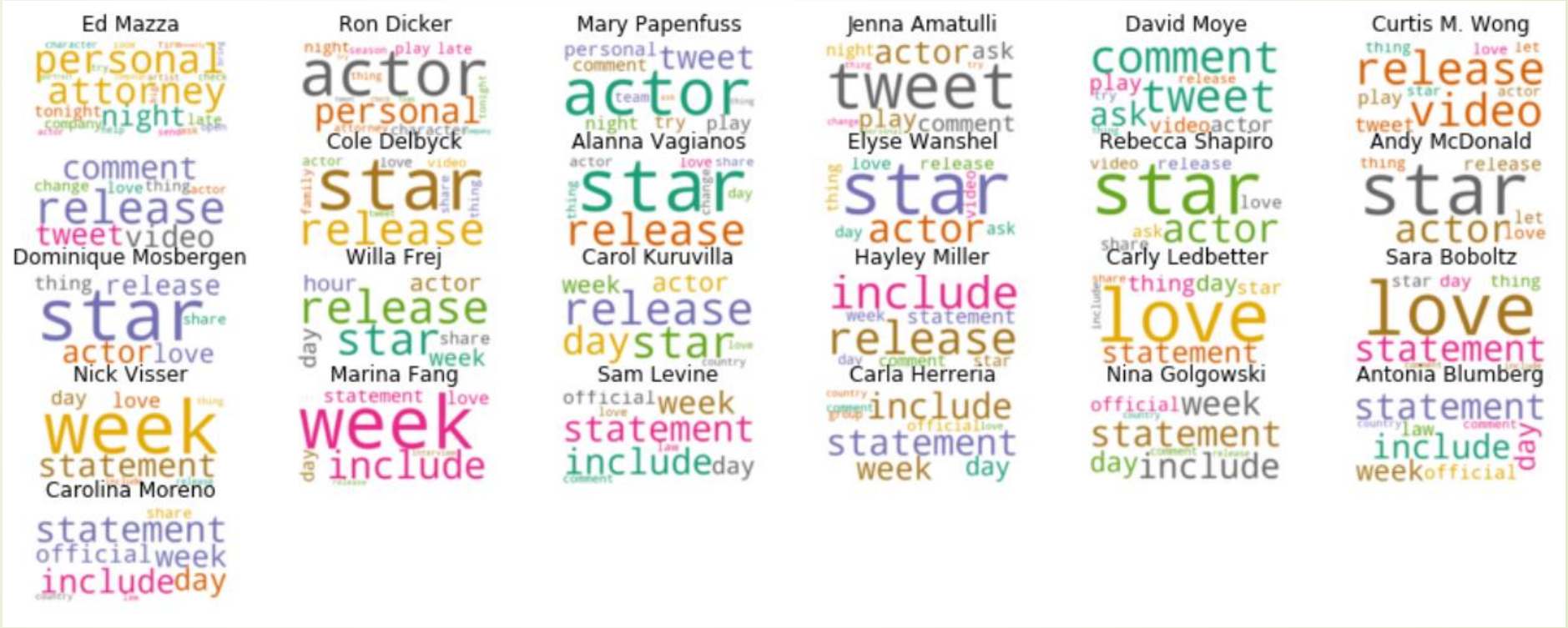
Estilo de escritura de cada autor

¿Qué se puede decir de los autores a partir de los datos?

Algoritmo de estimación de palabras mas utilizadas por autor

- Escoger el 50 artículos de cada autor.
- Para cada autor, tokenizar y lematizar cada articulo. Encontrar las 30 palabras mas frecuentes de cada articulo.
- Buscar las 50 palabras que mas aparecen en artículos considerando todos los artículos por autor.
- Una vez obtenidas las palabras mas comunes en cada categoría, descartar las que se repiten en 15 autores o mas.
- Realizar el diagrama wordcloud con las palabras mas comunes y con la frecuencia que son utilizadas por autor.

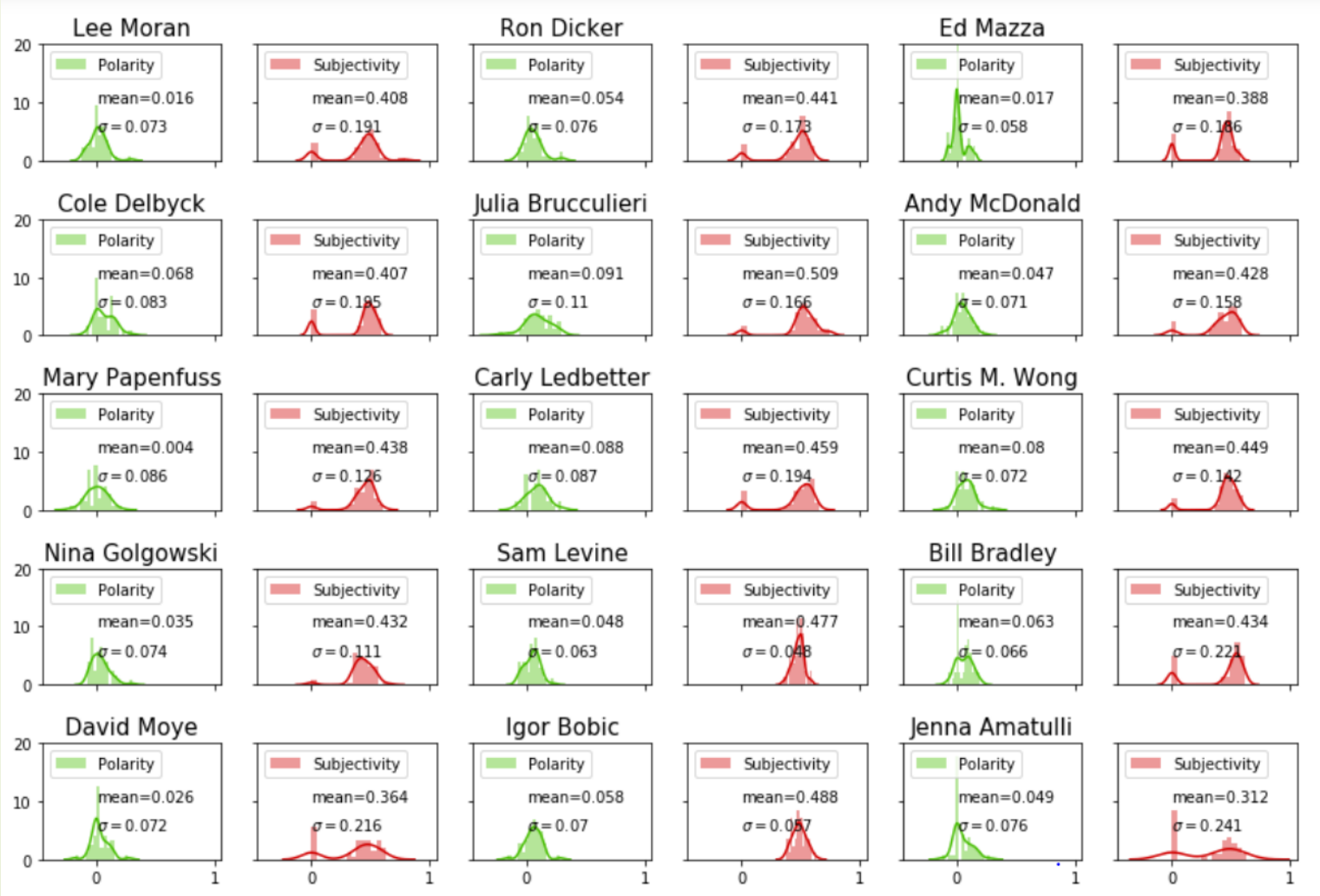
Palabras mas usada por autor



Algoritmo para el estudio de sentimientos de cada autor

- Escoger el 50 artículos de cada autor.
- Para cada autor, tokenizar y lematizar cada articulo.
- Para cada articulo, estimar los coeficientes de polarización y de subjetividad. Almacenar estos puntajes para todos los artículos.
- Realizar la grafica de distribución de los datos de polarización y subjetividad para cada autor
- Estimar el promedio y varianza de los histogramas de polarización y subjetividad.

Estudio de sentimientos por cada autor



Identificación de temas de las noticias

Algoritmo para la obtención de temas

- Escoger el 50 artículos de cada categoria.
- Para cada categoria, tokenizar y lematizar cada articulo.
- Calcular las palabras clave mas utilizadas en cada articulo
- Utilizar estas palabras mas usadas para obtener clusters de palabras utilizando el algoritmo Latent Dirichlet Allocation.
- Obtener los cinco clusters principales con 10 palabras cada uno.

Topic Modeling: Latent Dirichlet Allocation (LDA)

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
CRIME	victim, case, assault, prosecutor, sexual, man, woman, attorney, abuse, lawyer	child, bus, girl, home, kid, school, nah, leave, family, face	suspect, victim, case, investigator, crime, school, authority, site, kill, hour	assault, woman, sexual, victim, lawsuit, trial, case, jury, guilty, attorney	shooting, release, officer, footage, hour, room, survivor, hearing, sentencing, period
ENTERTAINMENT	set, hard, circumstance, offer, easy, feel, hour, scene, interview, happen	victim, hard, let, murder, suspect, rule, hour, circumstance, interview, crime	school, scene, season, violence, gun, happen, series, teen, leave, cast	woman, set, sexual, let, hard, excuse, interview, feel, offer, hour	movie, offer, row, easy, let, hard, hour, rule, watch, excuse
WORLD NEWS	plane, flight, crash, official, kill, site, passenger, cuban, far, circumstance	nuclear, summit, test, korean, meeting, site, expert, program, talk, exercise	diary, abuse, page, museum, publish, entry, family, research, write, secret	group, attack, hour, israeli, military, nuclear, kill, responsibility, hard, protest	deal, party, nuclear, sanction, plastic, site, european, country, government, deep
IMPACT	farmer, worker, rice, yield, change, increase, community, country, series, climate	job, worker, income, basic, automation, virus, idea, community, union, human	woman, farmer, ikigai, tree, world, plant, plastic, study, economy, seed	community, city, business, series, percent, content, home, big, land, change	social, anxiety, mental, symptom, disorder, feel, medium, bike, health, illness
POLITICS	suit, shooting, lawsuit, winery, yacht, false, victim, file, cruise, conspiracy	ad, child, school, disclosure, public, issue, political, information, gun, site	bot, conservation, russian, tweet, group, month, research, request, meeting, spanish	neighbor, leader, love, immigrant, federal, declaration, rule, service, citizen, political	woman, abortion, rule, campaign, information, offer, share, responsibility, hard, easy

Topic Modeling: Latent Dirichlet Allocation (LDA)

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
WEIRD NEWS	squirrel, woman, student, man, seat, post, offer, point, hard, allegedly	octopus, fawn, millennial, end, offer, hard, varie, site, rule, designate	cake, order, customer, online, profane, store, sword, graduation, write, censor	message, send, ade, zombie, site, man, allegedly, hard, text, urine	let, site, offer, responsibility, hard, easy, register, excuse, impossible, circumstance
BLACK VOICES	protest, player, anthem, black, national, understand, stand, violence, sit, patriotism	black, white, look, video, dance, person, let, hurt, easy, violence	woman, black, man, rule, designate, hard, let, hour, easy, offer	feat, black, actress, film, protest, french, woman, hit, nomination, actor	black, officer, video, white, student, woman, stop, hard, violence, man
WOMEN	woman, man, sexual, assault, let, rule, hard, period, offer, hour	woman, sexual, student, allegation, harassment, festival, comment, director, male, ask	woman, sexual, man, let, rule, excuse, hour, responsibility, offer, wear	abortion, gift, woman, rape, case, look, care, mom, feel, sign	woman, girl, young, abuse, petition, man, leader, social, event, offer
COMEDY	offer, hard, let, hour, responsibility, locality, register, designate, site, citizen	hat, red, love, hit, porn, convinced, fear, rock, confuse, star	swamp, drain, shut, immigrant, let, outside, royal, black, watch, actually	coin, watch, hour, offer, excuse, site, impossible, designate, goodbye, coronavirus	episode, prove, animal, goofy, entry, competition, password, muslim, racist, late
QUEER VOICES	employee, organization, store, boy, participant, change, announce, scout, terminate, support	gay, offer, site, citizen, hour, hard, easy, let, community, period	transgender, man, photo, easy, woman, wear, site, gender, position, character	therapy, conversion, book, child, practice, young, family, love, kid, parent	church, happen, certificate, pastor, memorial, denomination, congregation, place, character, interim

Topic Modeling: Latent Dirichlet Allocation (LDA)

	Topic, 0,	Topic, 1	Topic, 2	Topic, 3	Topic, 4
SPORTS	athlete, excuse, sport, site, cheerleader, team, varie, offer, college, locality	college, player, school, basketball, athlete, sport, huffpost, hand, play, team	team, game, player, play, hard, pardon, point, rule, let, protest	sport, anthem, win, game, performance, song, race, play, million, finish	athlete, abuse, sexual, rule, team, female, hour, woman, let, level
BUSINESS	arbitration, sexual, harassment, group, card, breach, write, woman, assault, employee	tax, percent, city, million, offer, billion, gun, hour, rule, deal	woman, man, store, employee, business, trade, simmon, datum, offer, image	housing, city, appeal, community, environmental, resident, fight, build, supportive, homelessness	woman, sexual, assault, case, harassment, file, driver, letter, worker, public
TRAVEL	flight, attendant, travel, passenger, plane, kid, book, trip, air, good	stay, town, region, city, island, couple, worth, love, place, country	city, travel, good, beach, food, hotel, summer, weather, restaurant, month	pet, animal, airline, dog, fly, cargo, death, carrier, travel, high	room, hotel, list, travel, germ, good, summer, remote, destination, affect
MEDIA	woman, sexual, allegation, story, harassment, tape, let, network, host, hour	news, union, organization, responsibility, medium, platform, site, begin, anti, newspaper	post, write, blog, sexual, woman, network, site, harassment, claim, hard	mean, coate, question, let, actually, talk, write, story, hard, lot	business, tabloid, coverage, boss, seize, financial, lawyer, allege, payment, month
TECH	user, datum, rule, information, app, site, hour, hard, offer, let	password, user, account, bug, store, internal, twitter, believe, mask, log	emoji, flamethrower, control, information, application, birth, site, offer, hard, pill	information, robot, rule, public, forget, ruling, conviction, search, remove, court	speech, lawsuit, hate, bug, housing, child, family, discrimination, disability, exclude

Caracterización de las descripciones de los artículos

Basándote en el texto de la descripción corta, caracteriza este dataset.

Distribución del numero de palabras clave de la descripción de los artículos