Project 2 -Deduplication of Spanish Articles

Group Members: Mavis Francia, Tanushri Singh, Ishan Sharma, Vyaas Shenoy

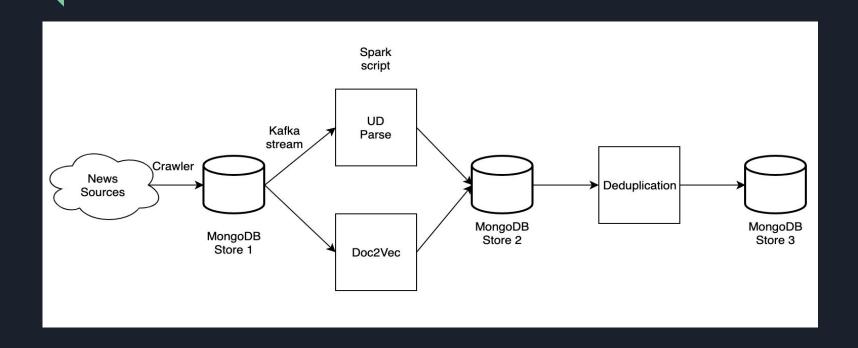
Motivation

Our project is to be a part of a pipeline for political event coding.

Our particular piece of the pipeline

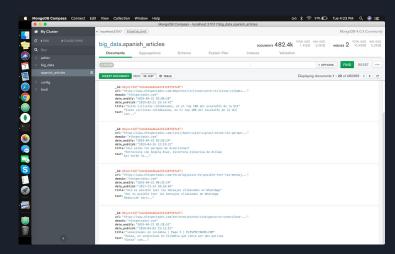
- collects Spanish news articles with a web crawler
- parses the articles into Universal Dependencies (UD)
- performs content-level deduplication
- stores the results in MongoDB

Approach - Architecture

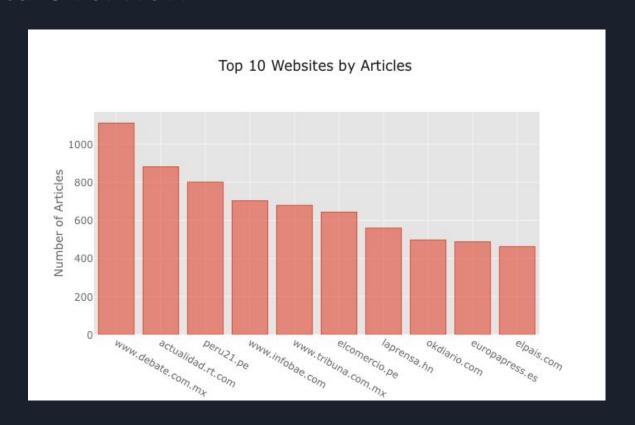


Crawler Design

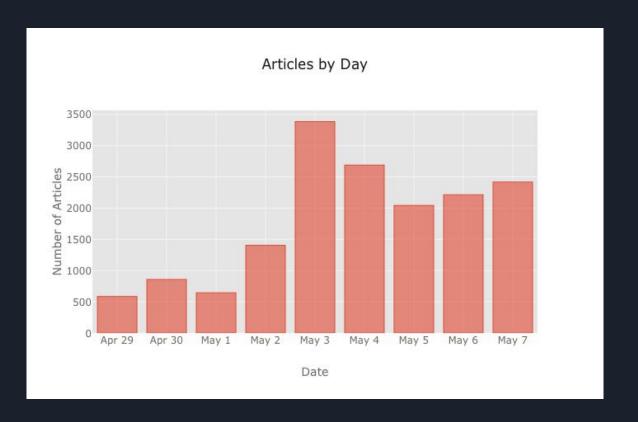
- Inconsistent data sources existed
- Crawler Options: News Please, Scrapy, RSS
- News Please: 500-600 articles per day
- Scrapy: Manual work needed per website, not scalable
- RSS: Latest articles, standard format, limited to 10-20 articles



Data Overview



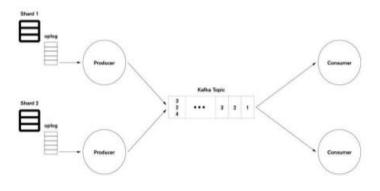
Data Overview



Interconnecting MongoDB

- Reading data in and streaming individual tables off through Kafka
- Streaming in information from Kafka to Spark in order to perform UD Parse
- Writing back to Mongo with Doc2Vec and UDparse value

MongoDB As a Kafka Producer





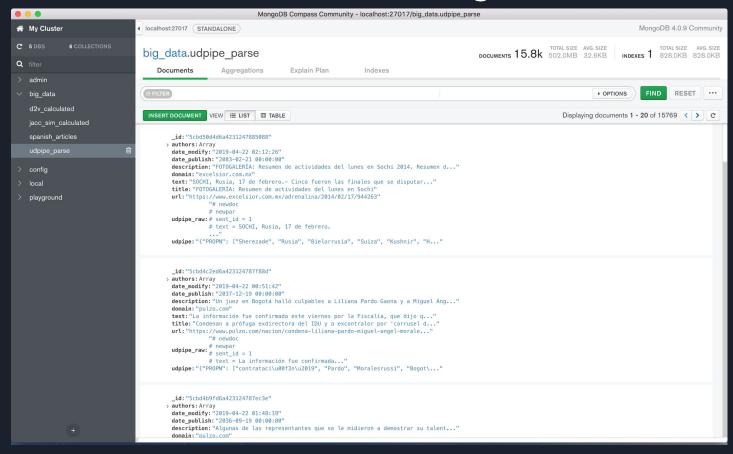
Implementation of UD Parse

- Installed ufal.udpipe package and loaded a pre-trained Spanish model (AnCora)
- Stored the articles in a Spark dataframe
- Wrote a user-defined function to parse the text for each article
- Wrote the parse to its own column in the dataframe
- Extracted important parts-of-speech from the parse and stored in a dictionary
- Wrote the POS dictionary to its own column in the dataframe
- Wrote the dataframe with UD parse columns back to MongoDB

UD Parse at Run-Time

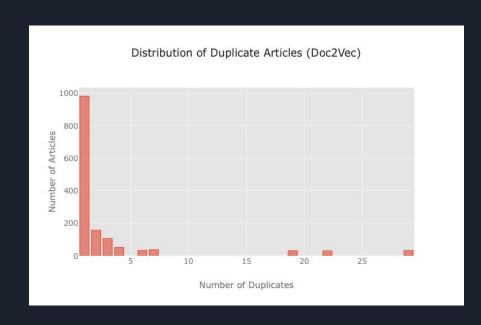
• • •			<u> </u>	treaming — java ∢ Python s	treamToSpark.py — 179×48				
t-Analytics/Streaming — ja	va • Python streamToSpark	.py java -Xm	x512M -Xms512Mconfig	/zookeeper.properties	java -Xmx1G -Xms1G -serveafka	config/server.properties	charmProjects/Big-D	ata-Management-Analytics — -b	bash +
Setting default log level to "WARN". To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel). 19/05/07 15:49:56 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform using builtin-java classes where applicable No data received from stream									
 id udpipe_raw +	authors d udpipe	ate_modify	date_publish	description	domain	text	title		
5cbd50d4d6a423124	[] 2019-04-2	2 02:12:26 208	3-02-21 00:00:00	FOTOGALERÍA: Resu	excelsior.com.mx SOCHI	, Rusia, 17 FOT	OGALERÍA: Resu	https://www.excel	# newdoc
5cbd4c2ed6a423124 # newpar { "PROPN":	[] 2019-04-2	2 00:51:42 203	7-12-19 00:00:00	Un juez en Bogotá	pulzo.com La in	formación fu Con	denan a prófug	https://www.pulzo	# newdoc
5cbd4b9fd6a423124 # newpar { "PROPN":	[] 2019-04-2	2 01:48:19 203	6-09-19 00:00:00	Algunas de las re	pulzo.com Las c	andidatas al ¿Cu	ál de estas ca	https://www.pulzo	# newdoc
5cce26a5d6a423061 # newpar { "PROPN":		4 22:39:54 203	6-03-19 00:00:00	Un vocero de la f	quien.com De ac	uerdo con el Mue	re Stephen Haw	https://www.quien	# newdoc
5cbd4bdcd6a423124		2 01:19:32 203	5-05-20 00:00:00	Los actores Lucia	pulzo.com Los i	ntérpretes d Pan	ameños enloque	https://www.pulzo	# newdoc
5cbd4be8d6a423124		2 00:37:30 203	5-05-20 00:00:00	El delantero de l	pulzo.com Carlo	s Bacca de e [Vi	deo] El golazo	https://www.pulzo	# newdoc
5cbd4baed6a423124		1 23:52:12 203	5-05-19 00:00:00	Un pistolero abri	pulzo.com Actua	lidad RT ind Un	muerto y 4 her	https://www.pulzo	# newdoc
5cbd4c32d6a423124		2 00:51:05 203	4-10-20 00:00:00	En partido en Ara	pulzo.com El bu	en jugador f Sus	tituyen a iraq	https://www.pulzo	# newdoc
5cbd4c37d6a423124 # newpar { "PROPN":	[] 2019-04-2	1 23:47:13 203	1-12-19 00:00:00	El entrenador arg	pulzo.com Millo	narios convo Rus	so no pudo evi	https://www.pulzo	# newdoc
5cbd4ba4d6a423124		2 01:41:06 202	9-04-20 00:00:00	La cantante antio	pulzo.com Boter	o comentó qu "No	soy bisexual,	https://www.pulzo	# newdoc
5cbd4c09d6a423124 # newpar { "PROPN":	[] 2019-04-2	1 23:57:40 202	8-12-20 00:00:00	Usaron atuendos y	pulzo.com Pero	las estrella [Fo	tos] Las Karda	https://www.pulzo	# newdoc
5cbd4bb9d6a423124 # newpar { "PROPN":	[] 2019-04-2	1 23:48:15 202	8-02-19 00:00:00	Así lo aseguró la	pulzo.com La co	ndición para Rue	da tendría pri	https://www.pulzo	# newdoc
5ccdbe2f9b24d49ef # newpar { "PROPN":	[] 2019-05-0	4 10:00:39 202	7-12-20 00:00:00	Un iris verde gem	glamour.es Nuevo	producto y Dar	ia Werbowy y B	https://www.glamo	# newdoc
5cbd4bcad6a423124 # newpar { "PROPN":	[] 2019-04-2	2 01:52:37 202	6-10-19 00:00:00	La exsenadora le	pulzo.com López	se refiere, Nue	vo mensaje de	https://www.pulzo	# newdoc
5cbd4c00d6a423124		2 00:37:17 202	6-05-19 00:00:00	Los 2 referentes	pulzo.com Nairo	Quintana y Nai	ro y Landa, ot	https://www.pulzo	# newdoc

UD Parse saved back to Mongo



Doc2Vec Implementation

- Model trained on 78,154 articles from 2019
- Vectors for new articles inferred from model
- Cosine similarity calculated between vectors
- Threshold: 60%



Doc2Vec Pipeline

```
Processing https://www.europapress.es/internacional/noticia-biden-presenta-candidato-sindicatos-primer-acto-politico-precampana-20190429234840.html [0/585] - 16:11:14
Processing https://www.abc.es/internacional/abci-akihito-modernizador-monarquia-mas-antiqua-mundo-201904300145 noticia.html [1/585] - 16:11:14
Processing https://www.abc.es/sociedad/abci-trafico-activa-operativo-especial-para-puente-mayo-comunidad-madrid-201904300130 noticia.html [2/585] - 16:11:15
Processing https://www.analitica.com/actualidad/actualidad-nacional/esperan-que-no-siga-el-retardo-procesal-en-caso-de-juan-requesens/ [3/585] - 16:11:16
Processing https://www.abc.es/familia/padres-hijos/abci-marca-pautas-ninos-tres-horas-ejercicio-y-mas-10-sueno-201904300121 noticia.html [4/585] - 16:11:16
Processing https://lta.reuters.com/articulo/boeing-accionistas-idLTAKCN1S51HW?symbol=BA.N [5/585] - 16:11:17
Processing https://lta.reuters.com/articulo/boeing-accionistas-idLTAKCN1S51HW [6/585] - 16:11:18
Processing https://www.laprensa.hn/sucesos/1280118-410/mexicano-victima-ultimada-vehiculo-comayagua [7/585] - 16:11:19
Processing https://www.laprensa.hn/sucesos/1280117-410/nino-baja-mangos-muere-electrocutado [8/585] - 16:11:19
Processing https://www.abc.es/motor/reportajes/abci-operacion-salida-puente-mayo-2019-mejores-horas-para-viajar-evitando-atascos-201904300110 noticia.html [9/585] - 16:11:20
Processing https://www.abc.es/deportes/futbol/abci-champions-league-tottenham-ajax-sorpresa-anunciada-para-final-201904300058 noticia.html [10/585] - 16:11:21
Processing https://www.abc.es/deportes/futbol/abci-fernando-llorente-destino-debe-champions-201904300057 noticia.html [11/585] - 16:11:21
Processing https://www.analitica.com/actualidad/actualidad-internacional/40-venezolanos-fueron-deportados-de-peru-con-antecedentes-penales/ [12/585] - 16:11:22
Processing https://www.analitica.com/actualidad/actualidad-nacional/cruz-roja-continua-entrega-de-avuda-humanitaria-a-la-espera-de-nuevo-cargamento/ [13/585] - 16:11:23
Processing https://lta.reuters.com/articulo/politica-honduras-protestas-idLTAKCN1S529J-OUSLT [14/585] - 16:11:23
Processing http://www.expansion.com/mercados/2019/04/30/5cc74a29268e3e9e798b4597.html [15/585] - 16:11:24
Processing https://www.analitica.com/actualidad/actualidad-internacional/felipe-vi-entrega-este-martes-los-premios-de-periodismo-rev-de-espana/ [16/585] - 16:11:25
Processing https://andro4all.com/2019/04/mejores-alternativas-camscanner-escanear-documentos [17/585] - 16:11:26
Processing https://lta.reuters.com/articulo/venezuela-migrantes-peru-idLTAKCN1S522Y-OUSLT [18/585] - 16:11:26
Processing https://www.proceso.com.mx/581827/astudillo-espera-alternativas-para-suplir-zee-en-la-costa-grande [19/585] - 16:11:27
Processing https://www.diariosur.es/deportes/ciclismo/muere-espanol-fernando-civera-titan-desert-20190429220805-ntrc.html [20/585] - 16:11:28
Processing https://lta.reuters.com/articulo/alphabet-resultados-idLTAKCN1S528N [21/585] - 16:11:29
```

502.107 | E ▲ | IITE 0 ▲

Jaccard Similarity Performance on 2 articles that are known to be similar

Article 1:

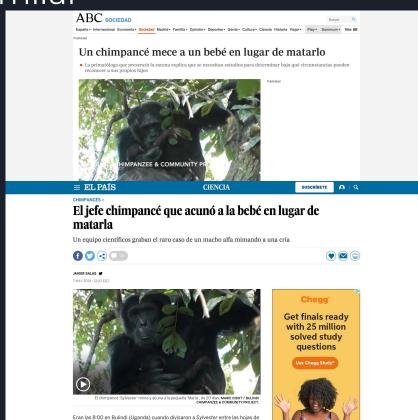
https://www.abc.es/sociedad/abci-chimpance-mece-bebe-lugar-mat arlo-201905071807 noticia.html

Article 2:

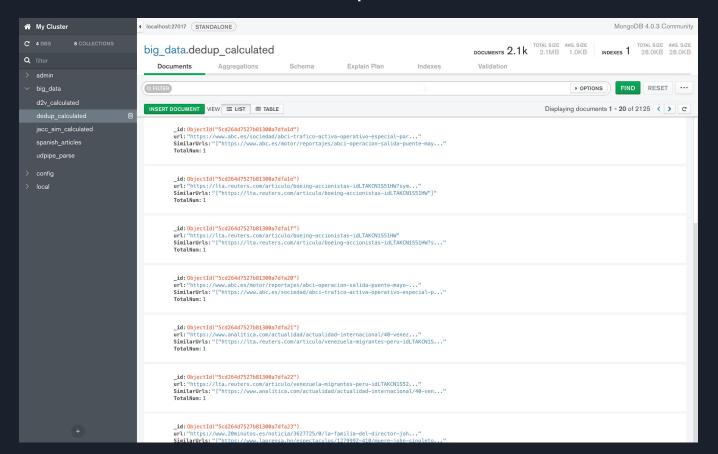
https://elpais.com/elpais/2019/05/06/ciencia/1557142204_604164.html#?ref=rss&format=simple&link=link

The Jaccard Similarity for the proper nouns of these two resulted in 35.714%

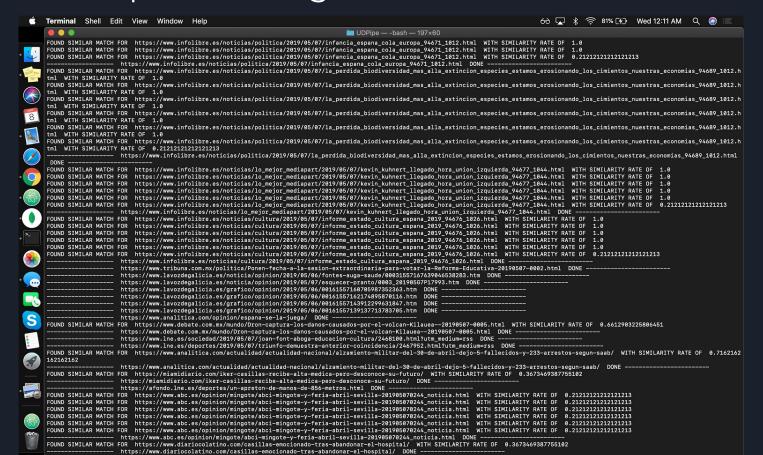
If two articles were not in the same context, had low resulting similarity ratios.



Results from De-Duplication



DeDuplication Algorithm at Runtime



Future Work

- Take article length into consideration (crawler grabs 1-2 word "articles" which are actually video/audio clips)
- Automate News Please crawler
- Train new models frequently to keep vocabulary up to date
- Integrate Doc2Vec with Spark
- Use a better measure of similarity for UDParse output

Challenges

- Crawler Design
- Implementation of UD Parse
- Interconnecting MongoDB with Kafka as well as Spark
- Measuring similarity using UD Parse output
- Similarity measures for deduplication

Challenge: Similarity measures for deduplication

- Doc2Vec model training takes 20 minutes for 1500 articles
- Doc2Vec models can't be trained with new data
- Comparing 2 articles from the training data does not require too much time after training the model
- Comparing 2 articles not in the training data takes too much time
- Articles with drastically different lengths can get similar scores
- 2 duplicate articles can have a similarity rating varying from 60% to 100%. We had to fix a
 lower bound on the similarity rating approximation to 60%.

Related Literature

- Solaimani et al. implement a distributed workflow for political event coding by using Apache Spark, MongoDB, Stanford CoreNLP, and PETRARCH [1].
- Mustafa performs deduplication by performing word embeddings and DBSCAN, then with a notion of cluster *purity* determines how much overlap two articles have in the topics they discuss [2]. If clusters are mostly pure, i.e. belonging to one article or the other, then they are likely not duplicates. If clusters are mostly impure, i.e. with points belonging to both articles, then they are likely duplicates.

References

[1] M. Solaimani, R. Gopalan, L. Khan, P. T. Brandt and B. Thuraisingham, "Spark-Based Political Event Coding," 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), Oxford, 2016, pp. 14-23.

[2] A. M. Mustafa, "Novel Class Detection and Cross-Lingual Duplicate Detection Over Online Data Stream." Order No. 10970248, The University of Texas at Dallas, Ann Arbor, 2018.

[3] models.doc2vec – Doc2vec paragraph embeddings https://radimrehurek.com/gensim/models/doc2vec.html

[4] Universal Dependency https://universaldependencies.org