

An Information Security Approach to Feature Engineering





Brian P. Murphy ReliaQuest Chief Architect

- Driving technical vision on the GreyMatter platform. Graduate of the University of Limerick in Ireland.
- Worked at Splunk, Elasticsearch, Loggly.
- ReliaQuest partners with Splunk and other Fortune 1000 customers in managing the Security Model across their entire organization.

Feature Engineering







Feature Engineering Definition

• Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work.

- "Coming up with features is difficult, time-consuming, requires expert knowledge. 'Applied machine learning' is basically feature engineering."
 - Andrew Ng, Machine Learning and AI via Brain simulations



Feature Engineering Outcome Desired

- Create encoders that can represent individual elements in a set of related values while also maintaining their relationships.
- URIs / Domains / Log messages
 - Classify and detect outlier and/or malicious activity
- Geo locations
 - Typically similar attacks will come from the same geo regions, but not necessarily the same countries.



Feature Engineering Machine Learning

• Most algorithms require columnar numerical values

• Models suffer from GIGO (Garbage In/Garbage Out)

 Memory constraints normally impossible to hold the entire data set in RAM



Feature Engineering Common Options

• Feature hashing

One hot encoding

Ordinal/Label encoding



Feature Engineering

- Feature hashing
 - Locality of information is lost
- One hot encoding
 - Can result in feature explosion
 - Dealing with new values not seen in training set
- Ordinal/Label encoding
 - Locality is not always obvious
 - Dealing with new values is hard also



Feature Engineering Proposed solution

• Min Hash Shingle

• Min Hash ngram

Geo Hash



Feature Engineering ngram/shingling

- Ngrams are contiguous sequences of characters from events.
 - Camlis -> 3gram -> ["cam","aml","mli","lis"]

- Shingles are to tokens what ngrams are to characters.
 - Camlis is really awesome -> 3shingle -> ["camlisisreally", "isreallyawesome"]



Feature Engineering Min Hashing

Jaccard Similarity

$$J(A,B)=rac{|A\cap B|}{|A\cup B|}=rac{|A\cap B|}{|A|+|B|-|A\cap B|}.$$

- Very accurate.
- But need to compare every element of a set with every other element of the set.
- Min Hash
 - Approximates Jaccard by creating 'k' hash functions and hashing each ngram or shingle, then finding the min value for each hash.
 - Events that share min hashes are similar.

Vassilvitskii, Sergey - COMS 6998-12



Feature Engineering Min Hash Benefits

- Encodes all values into a known number of columns.
- New values are handled and if they are similar to fitted values they will be similarly encoded.
- Linear time event by event processing.



Feature Engineering Geo Hash Approach

 Partitions the globe into a hierarchical NxN (32x32) grid, using a zcurve.





Feature Engineering Geo Hash Approach

- Calculate a bivariate (3D) normal distribution with the location at the peak of this distribution.
- Based on distance from the grid peak assign values across the grid.
- Results in NxN columns representing a decaying weight from the detected action.



Feature Engineering Geo Hash Benefits

- Allows the encoding of all possible geolocations into known number of columns.
- New locations are allowed for in the encoding.
- Maintains relationships between geolocations.
- Lookups and values can be precomputed to accelerate encoding.



Feature Engineering Demo



4 -> C () localhost strittlen-US/spo/Solunk ML Topikit.hearch/dupatch.	mole ratio-tildiopies.generalitype-statistics.ldiopies.geoese	rarch.mode-verbouelidisplay.pape.se 🗰 🔍 🍵	
splunk>enterprise App: Splunk Machine Learning T •	🔥 Administrator •	Messages Settings Activity	• Help • Find Q
Showcese Experiments Search Models Classic	Settings Docs 12 Video Tutoriels 12	s 🕥 -	iplunk Machine Learning Toolkit
New Search			Seve As Close
<pre>index=_internal uri=/services/server* minhashogram field by uri</pre>	uri fields + uri+ stats values(uri_minh	ash_#), values(uri_minhash_1), values(uri	Lminhash_5) All time + Q
✓ 1,718 events (before 10/26/19 2:00:57.000 PM) No Event Sample	ng =	Job = 11 = 40L	ē ⊥ ♥ Verbose Mode =
Events (1,710) Petterns Statistics (3) Visualization			
50 Per Page *			
uri t 🗸 🗸	values(uri_minhash_0) t 🖌	values(uri_minhash_1) = 🖌	values(uri_minhash_5) t 🖌
/services/server/info	71725668	257158735	314895557
/services/server/info/server-info?count=0	71725668	257158735	179533159
/services/server/info?count=0	71725668	257158735	179533159

sp	lunk>ent	terprise	App: Splunk	Machine Learn	ing T •			Administrator •	Messages	 Settings • 	Activity •	Help •	Find	q
	owcase		ts Searc	h Models	Classic •	Settings	Docs 12	Video Tutoriais LB			<u> (</u> sp	unk Machir	ve Learnin	g Toolkit
N	lew Se	arch										Si	ive As ¥	Close
index=_internal uri== minhashngram field=uri head 10000 apply urikmeans] stats count, first(uri) by cluster											Lost	24 hours •	Q.	
2	7,356 event	s (10/25/19 I	:00:00.000 A	M to 10/26/19 8	25:05:000 AM)	No Event 1	Sampling •			Job • II I		A	Verbose	Mode •
Ev	vents (7,356) Patterns Statistics (46) Visualization													
50	0 Per Page	• / For	nat Previ	ew =										
/ rster	count.													
	0	first(uri) -												
28	5	/static/1	ing/skins/def	ault/a.gif										
28	24	/services	AS/späunk-sy	stem-user/Spla	nk_ML_Toolkit	/telemetry-m	etric							
. 9	231	/services	N5/nobody/sp	lunk_instrumer	vtation/telement	try/general								
4	182	/services	A5/nobody/5g	lunk_ML_Toelki	t/properties/	nlspl/defaul	t?output_mo	de=json&count=-1						
41	159	/services	N5/nobody/5g	lunk_ML_Toolki	it/configs/con	f-mlspl?outp	ut_mode=jso	nilcount=-1						
45	36	/services	AS/admin/Spi	unk_ML_Toolkin	/deta/uL/nev/	default								
39	12	/services	N5/admin/Spi	unk_ML_Toolkit	/data/ui/mana	per?count=-1								
5	217	/services	/server/infe											
29	1	/services	/search/jobs	/export										
40	85	Annulan	Authenticat	Ion/users/admi	alastrat ander	1 ison								

	accorden-Utilapp, Splum	A_ML_Toolk Aschine Le	it/search/iq-search/k	15index%30_inte	nalk20urik30*1	Administrator •	Messages •	Settinos •	. 9. 🗶	Helo •	Pind	
Showcase Experie	nents Search	Mode	is Cinsic•	Settings	Docs 12	Video Tutoriais 12			💽 Sp	unk Mact	ine Learn	ing Toolkit
New Search											Seve As •	Close
index*_internal urt	** #inhashngra	m field≕v	ri head 1000	0 apply uni	Ameans3 se	wich cluster=4				Le	st 24 hours	• Q
√ 133 events (10/25/19	2:00:00.000 PM to	10/26/19	2:04:54.000 PM)	No Event St	mpling •			lob = III II	~ ð	*	Verbos	e Mode •
Events (133) Potto	ms Statistics	Visual	zation									
Format Timeline •	- Zoom Out	+ Zoom		× Deselect							1 ho	ur per column
									_			_
		List •	≠ Format	20 Per Page	-			CPrev 1	2 3	4 5	6 7	Next 2
K Hide Fields	III All Fields	á T	ìme	Event								
SELECTED FIELDS # bytes 5 # host 1 # source 2		> 1	0.26/19 :03:58.023 PM	127.0.0.1 - 1t7output_m bytes = 3545 sourcetype =	adetin [26/0c ode=json&coun b host=ubs splunkd_acce	t/2019:14:03:58.02 t=-1 HTTP/1.0* 200 intu source = /hon Iss	3 -8489] "GET /ser 3545 1m me/gaeltach/FEDemo	vicesNS/nobody ohsplunk/van/log	/Splunk_ML,	Toolkit/p Rd_access	log	slapl/defau
a sourcetype 2 INTERESTING FIELDS a clientip 1 If cluster 1		> 1	0.26/19 :03:56.808 PM	127.0.0.1 - 117output_m bytes = 3545 sourcetype =	adetn (25/0c ode=json&coun 5 host=ubs splunkd_acce	t/2019:14:03:56.88 t=-1 HTTP/1.0* 200 intu source = /hor Iss	8 -8480] "GET /ser 3545 1es me/gaeltadh/FEDemo	vicesNS/nobody o/splunk/van/log	/Splunk_M.,	Teelkit/p	operties/	wlspl/defau
# cluster_distance 5 # count 1 # date_hour 5 # date_mday 2		> 1	0.2679 :03:55.910 PM	127.0.0.1 = 1t7output_m bytes = 3545	admin (26/0c ode=json&coun b host=ubs	t/2019:14:03:55.910 t=-1 HTTP/1.0* 200 untu source = /hor	8 -8489] "GET /ser 1545 1es me/geeltedh/FEDemo	vicesNS/nobody o/splunk/van/log	/Splunk_ML,	Toolkit/p	operties/	alapl/defau

splunk:>enterprise App: Splunk Machine Learning T =							🔥 Administrator •	🔵 Messages 🔹	Settings •	Activity	r= He	ip• Fini	1 Q	
Showcase	Experiments	Search		Classic •		Docs E	Video Tutorists L2				Splunk N	Aachine Lea	rning Toolkit	
New Sea	arch											Save As	 Close 	
Index*_internal uri** minhashngram field*uri head 10000 opply urikmeans3 search cluster=4 stats first(uri), count by cluster_distance														
✓ 139 events (10/25/19 2:00:0	0.000 PM to 10	0/26/19 2:06:	41.000 PM)	No Event Se	mpling •		J	iob • II II	с. <i>н</i> .	ð 1	Ø Verb	ose Mode *	
Events (139)	Patterns	Statistics (5)	Visuelize	tion										
50 Per Page	✓ Format	Preview •												
cluster_d	istance t 🖌	first(uri) 0										1	count - /	ř
62268	186898824.42	/servicesNS	i/nebody/5p1	unk_ML_Toolk	st/propertie	s/mlspl/def	ault?output_mode=jso	n&count=-1					128	
1.17855431	72030556e+16	/en-US/splu	rkd/raw/s	ervicesN5/ad	min/user-pre	fs/data/use	r-prefs/general						5	
3.37410514	19322736e+16	/servicesNS	/admin/Splu	nk_ML_Toolk:	it/data/looks	p-table-fil	es/mlspl_example_h	ard_drives_Standar	dScaler_1.csv	loutput_	node-jsor		1	
3.38242621	51073484e+16	/servicesNS	Vadmin/Splu	nk_ML_Toolki	it/data/looks	p-table-fil	es/mlspl_example_h	ard_drives_Standar	dScaler_1.csv				2	
1.08549250	36232244e+16	/en-US/splu	nkd/raw/a	ervices/sear	ch/timeparse	r7output_mo	de=json&time=-14h&_=	1572092886531					1	

	iearch?sid=1572093010.160&dispatch.sample_n	atio=1&display.general.type=statis	ticsädisplay page search.mode-verbese 9, ★ 0 🔞	0 😐 💌 🖂 💭
splunk>-enterprise App: Splunk Machine Lea	ning To •	🔥 Administrator =	Messages + Settings + Activity + Help	Find Q
Showcase Experiments Search Models	Classic - Settings Docs	2 Video Tutorisis L2	🕤 Splunik Mi	achine Learning Toolkit
New Search				Seve As Close
<pre>source==port_443_external.json.out host="ubum euclid latfield=dst_geo.latitude lonfield=ds head 10000 apply geoeuclidkmeans stats count, values(dst_geo.country_name) ac</pre>	u" index="dotconf" dst_geo.latitud it_geo.longitude countries, walwes(dst_geo.city_na	er* dst_geo.longituder*, me) as cities by cluster		Al time * Q
10,000 events (before 10/26/19 8:30:10.000 AM)	No Event Sampling =		Job = II II - do L	Verbose Mode =
Events (10,000) Petterns Statistics (20)	Asualization			
50 Per Page * 🖌 Format Preview *				
cluster 8 🖌	count e 🖌 countries e		✓ cities t	1
•	964 Canada Japan United State		Boydton Lansing Port Coguitlam Tokyo	
1	2003 United State	•	Ashburn Chantilly Macon Reston	
10	123 Hong Kong United State	•	Ann Arbor Central Clifton Parsippany	
11	252 France		Austin	

<pre>source**port_443_external.json.out host**u8 dst_geo.longitude head 10000 geohas geohash_81, geohash_20, geohash_21, geo</pre>	untu" index="dotconf" h latfield=dut_geo.la hash_dp, geohash_dr	dst_geo.latitude=== titude lonfield=dst_g	fields = dst_geo.city_ eo.longitude stats va	name, dst_peo.country Lues(dst_peo.city_nam	_name, dst_geo.latitude, e), count by geohash_00,	Al time • Q
10,000 events (before 10/23/19 10:58:18.000 AM)	No Event Sampling			Job *		Verbose Mode
Events (10,000) Petterns Statistics (34)	Visualization					
50 Per Page ▼						
geohesh_00 I / geohesh_01 I /	geohash_20 0 /	geohash_21 0 🖌	geohash_dp 🔹 🖌	geohash_dr i 🖌	values(dst_geo.city_name) 0	/ count i /
0.00010355174771 0.000107179706846	0.000299148118386	0.000373948040373	1.0	0.550376015022	Ann Arbor Chicago Columbus Elk Grove Village Lansing Lombard Ouk Brook Toronto	127
0.000152924697811 0.000157295648899	0.000373548048373	0.000453648885236	0.676686238696	0.486555996988	Atlanta Roswell Suwance	11
0.000107575648362 0.00011991003346	0.000477577522548	0.000613505181179	0.559376915922	0.313701839369	Des Holmes	194
9.96456385395e-05 9.55286846536e-05	0.000181549756484	0.000221569753371	0.559376915922	1.0	Clifton New York North Bergen Parsippany Secaucus Staten Island	12

<pre> fields + dst_geo.city_name, dst_geo.country_name, dst_geo.latitude, dst_geo.longitude head 10000 geohash latfield=dst_geo.latitude lonfield=dst_geo.longitude apply geokmeans2 stats count, values(dst_geo.country_name) as countries, values(dst_geo.city_name) as cities by cluster</pre>									
10,000 events ((before 10/25/19	5.58.50.000 PM	No Event Sampling *			Job • II II	. • i	Verbose Mode	
Events (10,000)	Patterns	Statistics (19)	Visuelization						
50 Per Page *	✓ Format	Preview *							
	cluster	1 Z	count 0 🖌	countries #	/	cities 1		1	ł
		•	3381	United States					
			4125	United States		Ashburn Beltsville Boydton Chantilly Clifton New York North Bergen Parsippany Reston Secaucus Staten Island Mashington			
		18	277	United States		Austin Dallas Irving San Avtonio			

Feature Engineering



Duration (seconds)		Component	Invocations	Input count	3
1	0.00	command.addinfo	1	10,000	
	28.82	command apply	8	10,000	
1	0.00	command fields	2	60,912	
	25.31	command.geohash	8	10,000	
1	0.00	command.head	8	10,000	
1	0.00	command prehead	8	30,456	
1	0.17	command.search	8		



Thank You

https://github.com/GaelTadh/rq_feature_engineering (coming soon)

pmurphy@reliaquest.com