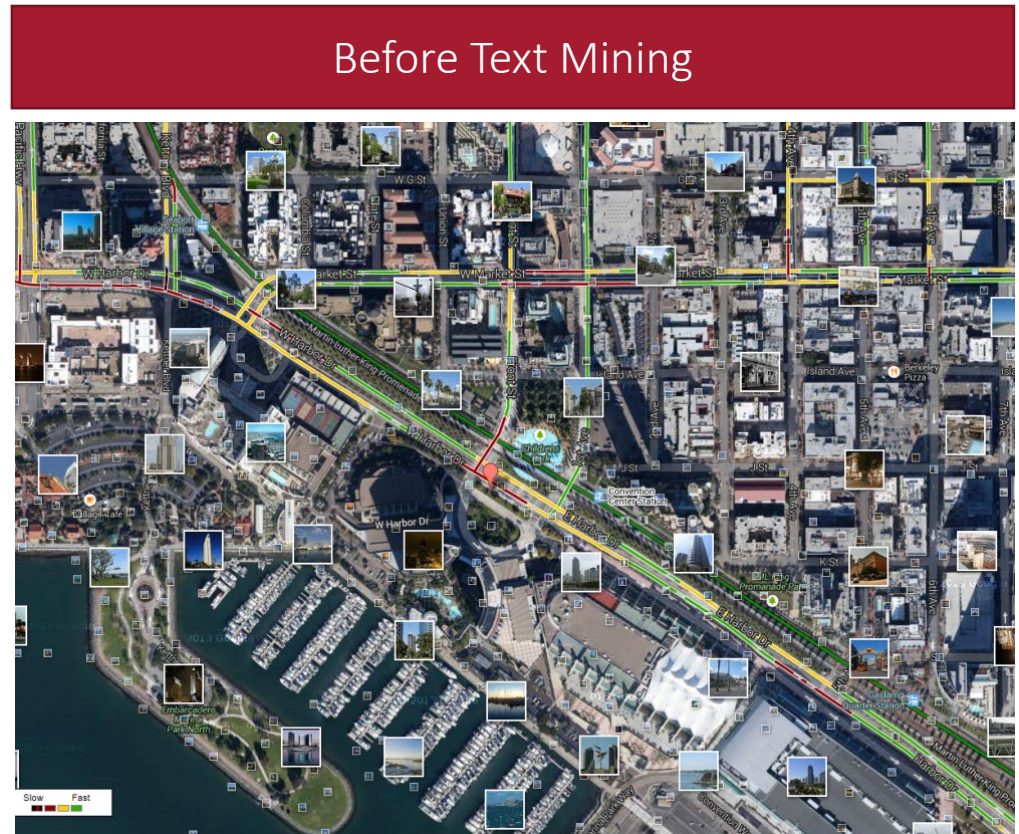

Text Mining



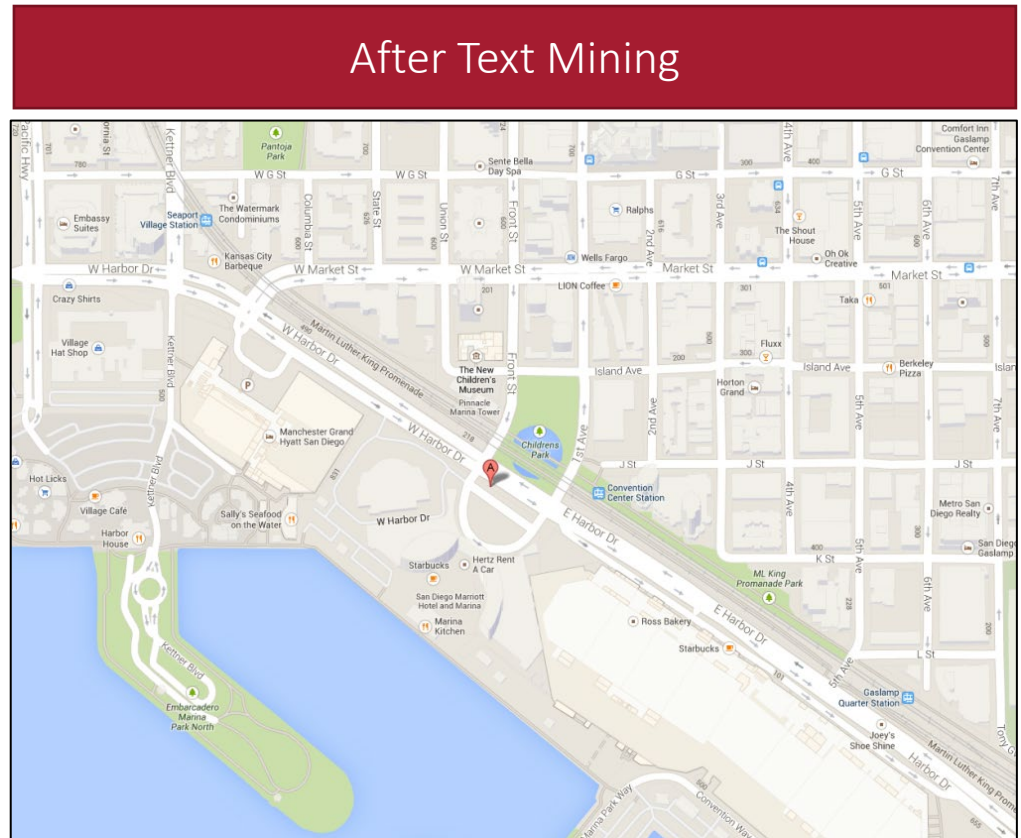
What is Text Mining?

- Extract new insights from text
- Let's you drink from a fire hose of information
- Language is hard; many unsolved problems
 - Unstructured
 - Expression is individualistic
 - Multi-language/cultural implications

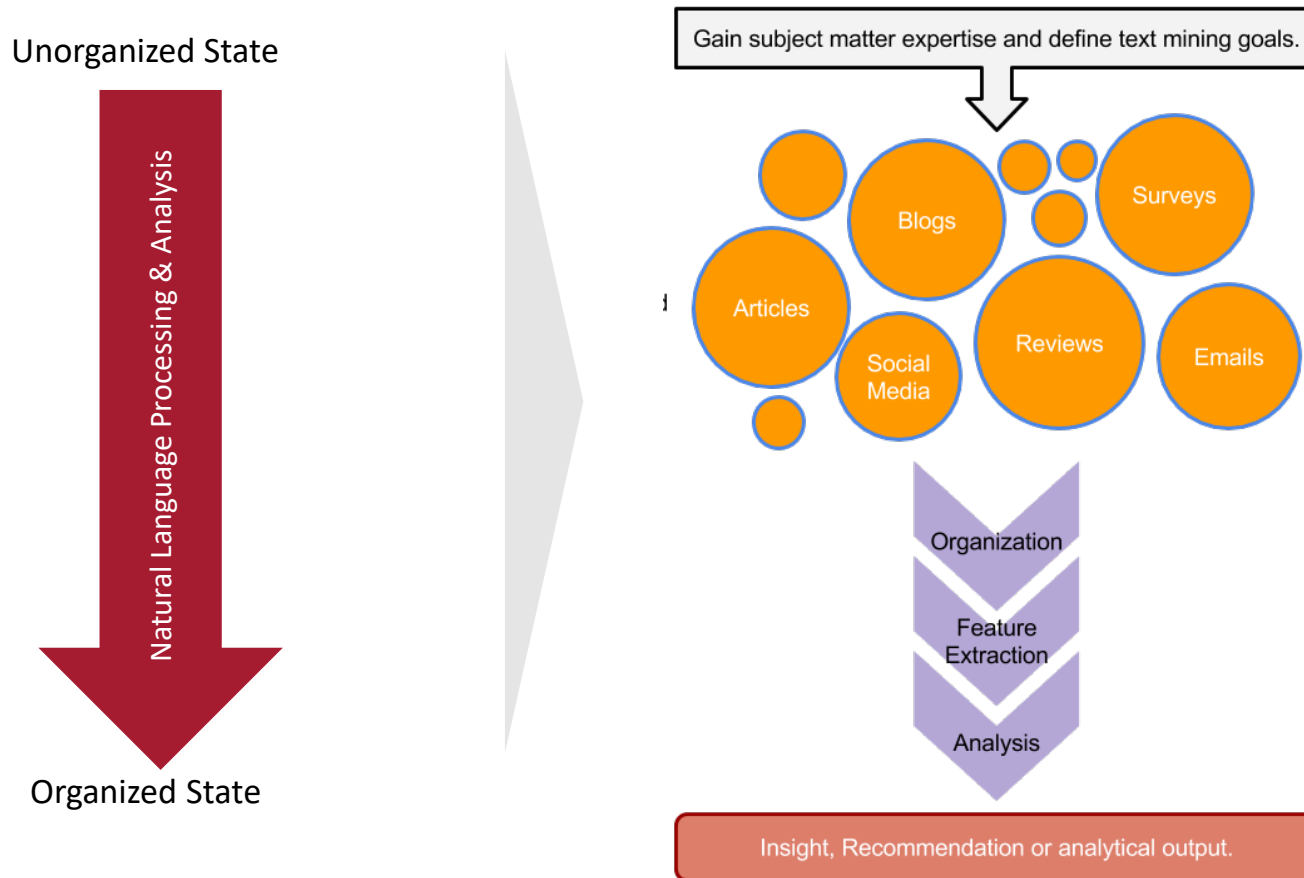


What is Text Mining?

- Extract new insights from text
- Let's you drink from a fire hose of information
- Language is hard; many unsolved problems
 - Unstructured
 - Expression is individualistic
 - Multi-language/cultural implications

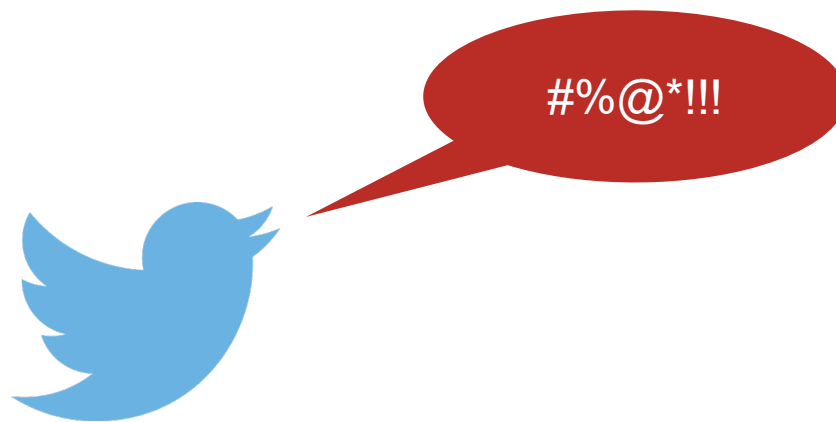


TM Project Workflow



A safe/supportive learning environment

- This text has never been read. “Keyboard Courage” is rampant which may entail some less than ideal topics.
 - Twitter is a realistic Natural Language Channel
 - It is a great place to get topics, and messy challenging data.
 - As a safe learning environment, no offense is intended, merely exposure to a real data set. If offended, please reach out and I will get you additional data sets.



Let's ~~Practice~~ Review!

A_more_string_manipulation.R

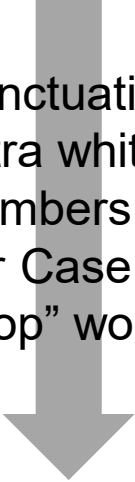
B_real_text_Search_Manipulation.R



R for Cleaning Steps

🐦 Tomorrow I'm going to have a nice glass of Chardonnay and wind down with a good book in the corner of the county :-)



- 
- 1.Remove Punctuation
 - 2.Remove extra white space
 - 3.Remove Numbers
 - 4.Make Lower Case
 - 5.Remove “stop” words

🐦 tomorrow going nice glass
chardonnay wind down good book
corner county

Meta Example

'\$doc_id'

'\$text'

META: \$favorited, \$created ...

doc_id	text	favorited	replyToSN	created	truncated	replyToSID	id	replyToUID	statusSource	screenName	retweetCount	retweeted	longitude	latitude
1	@ayyytylerb that is so true drink lots of coffee	FALSE	ayyytylerb	8/9/2013 2:43	FALSE	3.65664E+17	3.65665E+17	1637123977	<a href="http://twitter.α thejennagibson		0	FALSE	NA	NA
2	RT @bryzy_brib: Senior March tmw morning at 7:25 A.M. ii	FALSE	NA	8/9/2013 2:43	FALSE	NA	3.65665E+17	NA	<a href="http://twitter.α carolynicosia		1	FALSE	NA	NA
3	If you believe in #gunsense tomorrow would be a very go	FALSE	NA	8/9/2013 2:43	FALSE	NA	3.65665E+17	NA	web	janeCkay	0	FALSE	NA	NA
4	My cute coffee mug. http://t.co/2udvMU6XIG	FALSE	NA	8/9/2013 2:43	FALSE	NA	3.65665E+17	NA	<a href="http://twitter.α AlexandriaOOTD		0	FALSE	NA	NA
5	RT @slaredo21: I wish we had Starbucks here... Cause coff	FALSE	NA	8/9/2013 2:43	FALSE	NA	3.65665E+17	NA	<a href="http://twitter.α Rooosssaaaa		2	FALSE	NA	NA
6	Does anyone ever get a cup of coffee before a cocktail??	FALSE	NA	8/9/2013 2:43	FALSE	NA	3.65665E+17	NA	<a href="http://twitter.α E_Z_MAC		0	FALSE	NA	NA
7	"I like my coffee like I like my women...black, bitter, and	FALSE	NA	8/9/2013 2:43	FALSE	NA	3.65665E+17	NA	<a href="http://twitter.α Charlie_31191		0	FALSE	NA	NA
8	@dreamwwediva ya didn't have coffee did ya?	FALSE	dreamwwediva	8/9/2013 2:43	FALSE	3.65664E+17	3.65665E+17	1316942208	<a href="http://twitter.α JessicaSalvato5		0	FALSE	NA	NA
9	RT @iDougherty42: I just want some coffee.	FALSE	NA	8/9/2013 2:43	FALSE	NA	3.65665E+17	NA	<a href="http://twitter.α kaytiekirk		1	FALSE	NA	NA
10	RT @Dorkv76: I can't care before coffee.	FALSE	NA	8/9/2013 2:43	FALSE	NA	3.65664E+17	NA	<a href="http://tapbots.c lissteria		2	FALSE	NA	NA
11	No lie I wouldn't mind coming home smelling like coffee	FALSE	NA	8/9/2013 2:43	FALSE	NA	3.65664E+17	NA	<a href="http://twitter.α DOPECROOK		0	FALSE	NA	NA
12	RT @JonasWorldFeed: Play Ping Pong with Joe. Take a tou	FALSE	NA	8/9/2013 2:43	FALSE	NA	3.65664E+17	NA	<a href="http://www.ec TiffCaruso		6	FALSE	NA	NA
13	Have I ever told any of you that Tate Donovan bought my :	FALSE	NA	8/9/2013 2:43	FALSE	NA	3.65664E+17	NA	web	CurlysCrazyMofo	0	FALSE	NA	NA
14	RT @JonasWorldFeed: Play Ping Pong with Joe. Take a tou	FALSE	NA	8/9/2013 2:43	FALSE	NA	3.65664E+17	NA	web	JoeJonasVA	6	FALSE	NA	NA
15	@HeatherWhaley I was about 2 joke it takes 2 hands to hc	FALSE	HeatherWhaley	8/9/2013 2:42	FALSE	3.65647E+17	3.65664E+17	26035764	<a href="http://twitter.α AnnaDuleep		0	FALSE	NA	NA
16	RT @MoveTheSticks: Charlie Whitehurst looks like he sho	FALSE	NA	8/9/2013 2:42	FALSE	NA	3.65664E+17	NA	<a href="http://twitter.α mpr4437		42	FALSE	NA	NA
17	Coffee always makes everything better.	FALSE	NA	8/9/2013 2:42	FALSE	NA	3.65664E+17	NA	web	sharkshukri	0	FALSE	NA	NA
18	RT @AdelaideReview: Food For Thought: @Annabelleats	FALSE	NA	8/9/2013 2:42	FALSE	NA	3.65664E+17	NA	<a href="http://twitter.α thepaulbaker		1	FALSE	NA	NA
19	RT @LittleMells: Imfao!!!" @bryanlaca: nahhh Melanie u i	FALSE	NA	8/9/2013 2:42	FALSE	NA	3.65664E+17	NA	web	bryanlaca	1	FALSE	NA	NA
20	I wonder if Christian Colon will get a cup of coffee once th	FALSE	NA	8/9/2013 2:42	FALSE	NA	3.65664E+17	NA	<a href="http://www.my Shauncore		0	FALSE	NA	NA
21	Shouldn't have drank coffee I'm jittery as fuck.	FALSE	NA	8/9/2013 2:42	FALSE	NA	3.65664E+17	NA	<a href="http://twitter.α DylanBaur		0	FALSE	NA	NA
22	#good_morning <U+2615><ed><U+00A0><U+00BD><ed><U	FALSE	NA	8/9/2013 2:42	FALSE	NA	3.65664E+17	NA	<a href="http://instagran LadyMonyAna1		0	FALSE	NA	NA
23	@kungfupussy You might need to do a bulk shipment to N	FALSE	kungfupussy	8/9/2013 2:42	FALSE	3.65664E+17	3.65664E+17	19478601	<a href="http://janetter.i Gridlock_Coffee		0	FALSE	NA	NA
24	Gold Coast JCC Friday News features profile on new coffe	FALSE	NA	8/9/2013 2:42	FALSE	NA	3.65664E+17	NA	web	_GCJCC	0	FALSE	NA	NA
25	Sometimes I start dancing on my coffee table because I ca	FALSE	NA	8/9/2013 2:42	FALSE	NA	3.65664E+17	NA	<a href="http://twitter.α Rilevdreams		0	FALSE	NA	NA

- ID is for organization
- Text is the information we want to examine
- Meta adds context to our observations.

For Bag of Words, how is data organized?

Term Document Matrix						
	Tweet1	Tweet 2	Tweet3	Tweet4	...	Tweet_n
Term1	0	0	0	0	0	0
Term2	1	1	0	0	0	0
Term3	1	0	0	2	0	0
...	0	0	3	0	1	1
Term_n	0	0	0	1	1	0

Document Term Matrix					
	Term1	Term2	Term3	...	Term_n
Tweet1	0	1	1	0	0
Tweet2	0	1	0	0	0
Tweet3	0	0	0	3	0
...	0	0	0	1	1
Tweet_n	0	0	0	1	0

Code to Create the DTM/TDM and change to a matrix

```
txtDtm<-DocumentTermMatrix(txtCorpus)
txtTdm<-TermDocumentMatrix(txtCorpus)
txtDtmM<-as.matrix(txtDtm)
txtTdmM<-as.matrix(txtTdm)
```

Why are DTM & TDM Sparse? What do they represent?

??

The matrices are sparse (many 0's) so additional steps may be needed to extract information.

Let's ~~Practice~~ Review!

C_text_organization.R

