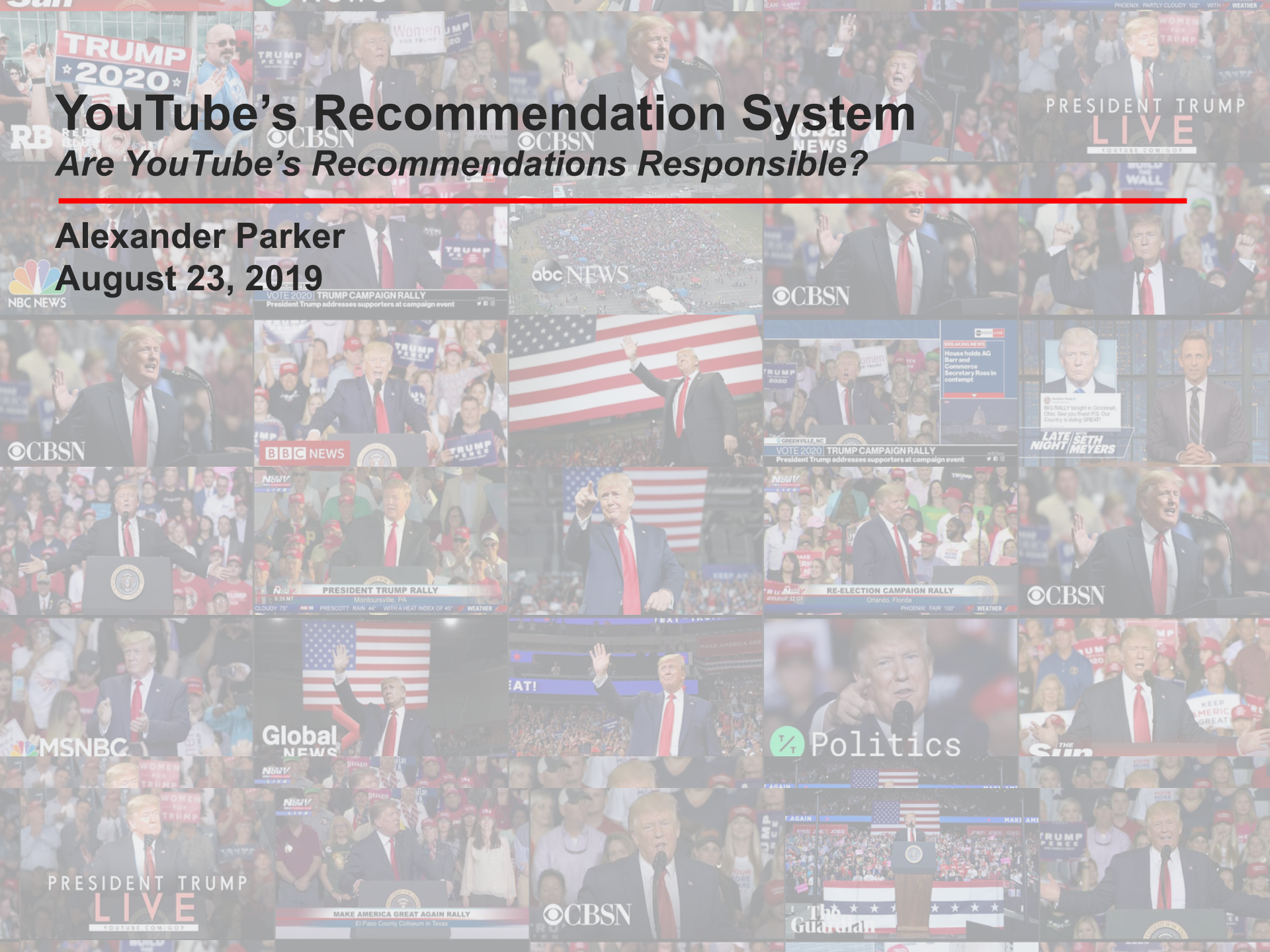
YouTube's Recommendation System

Are YouTube's Recommendations Responsible?

Alexander Parker

August 23, 2019



Criticism of YouTube's Algorithm

Concern over YouTube's tendency to recommend inaccurate, controversial videos

"The algorithm doesn't seek out extreme videos, [engineers] said, but looks for clips that data show are already drawing high traffic and keeping people on the site. Those videos often tend to be sensationalist and on the extreme fringe, the engineers said"

- Wall Street Journal (1)

"In effect, YouTube has created a restaurant that serves us increasingly sugary, fatty foods, loading up our plates as soon as we are finished with the last meal. Over time, our tastes adjust, and we seek even more sugary, fatty foods, which the restaurant dutifully provides"

- Zeynep Tufekci (2)

"We'll continue that work this year, including taking a closer look at how we can reduce the spread of content that comes close to—but doesn't quite cross the line of—violating our Community Guidelines. **To that end, we'll begin reducing recommendations of borderline content and content that could misinform users in harmful ways**—such as videos promoting a phony miracle cure for a serious illness, claiming the earth is flat, or making blatantly false claims about historic events like 9/11."

- YouTube's Official Blog (3)

1) *February 7, 2018.* https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478.
2) *March 10, 2018. New York Times opinion piece.* https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html.
3) *January 25, 2019.* https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html.

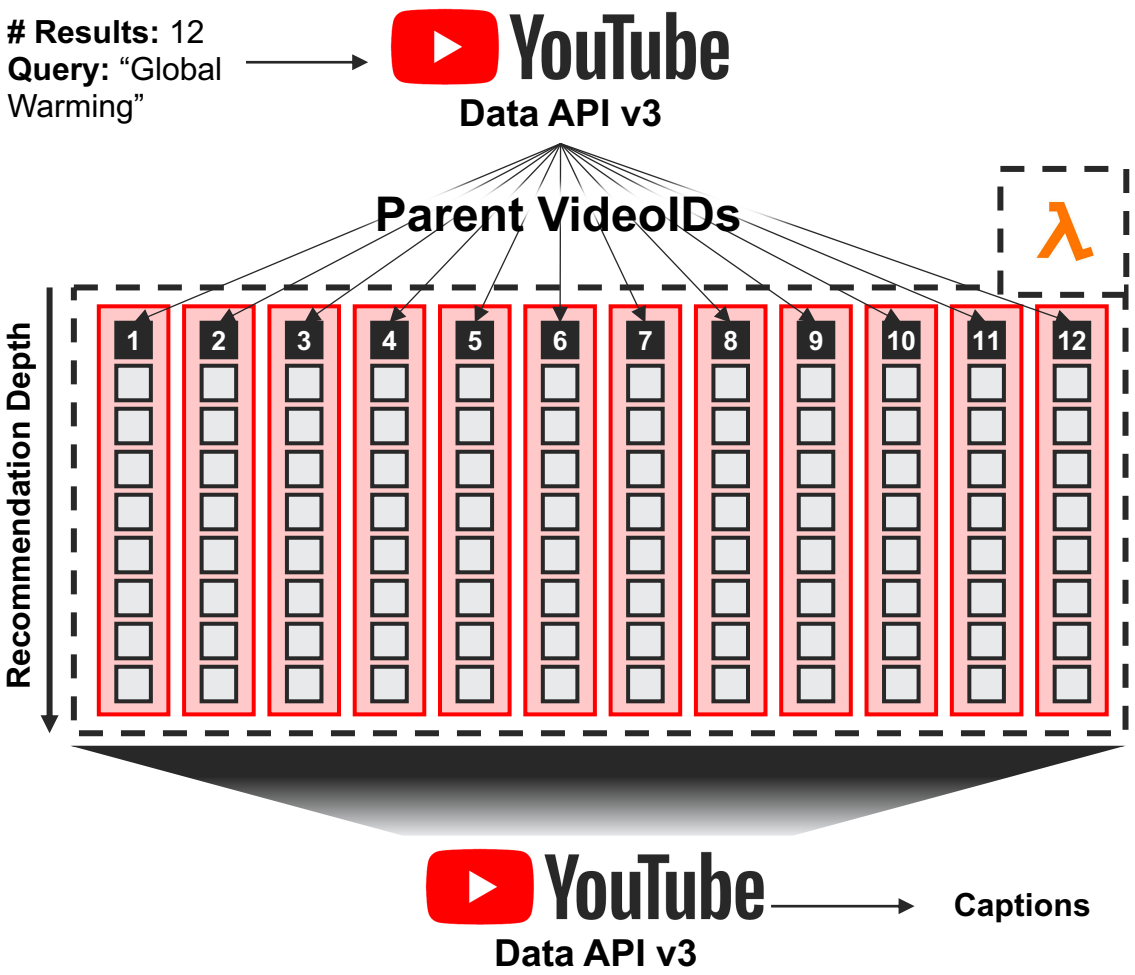
Data Collection and Project Structure

Parallelized webscraping via AWS lambda functions

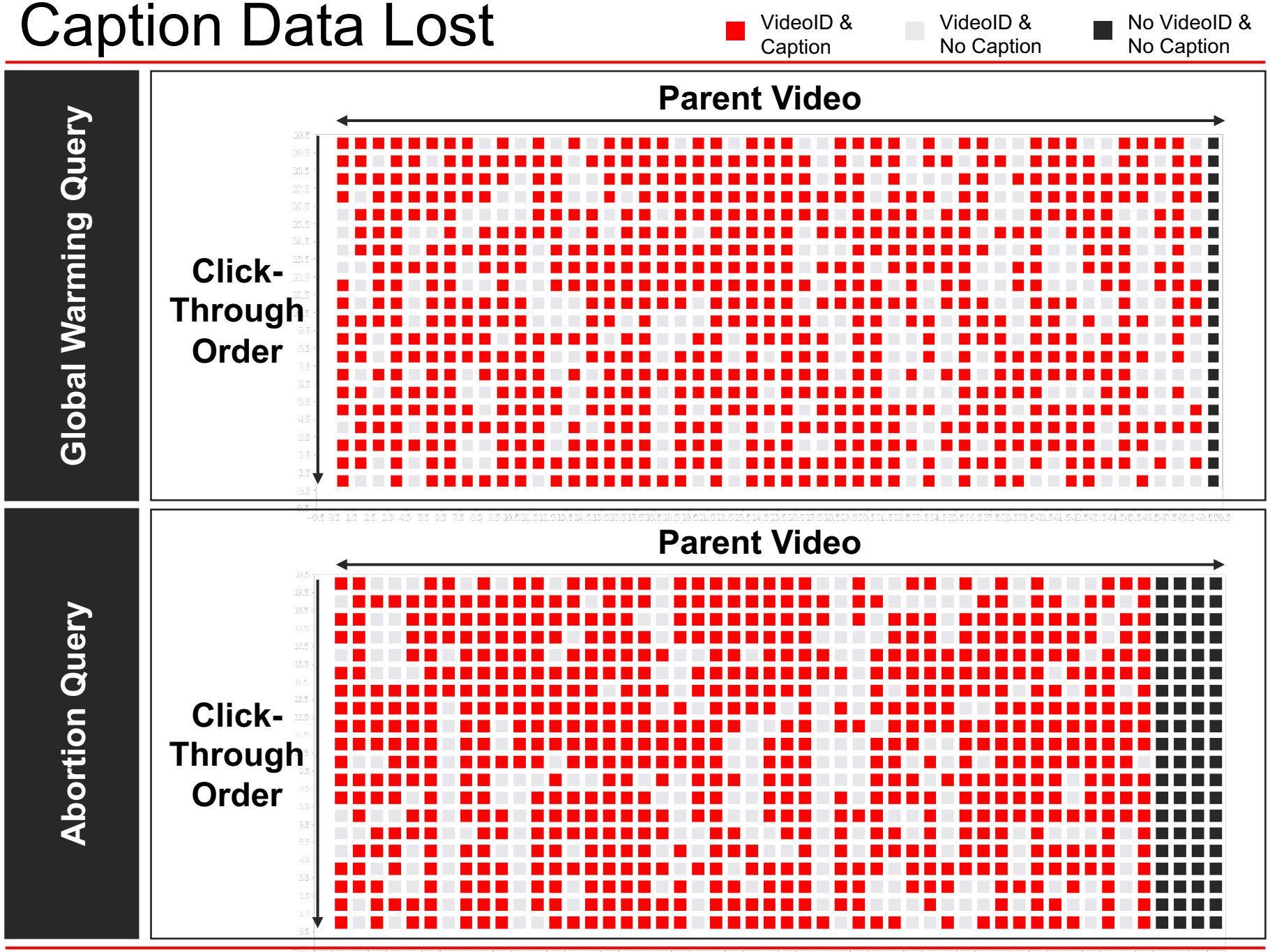
Key Points

- Recommendations are not available through API
- Data collection designed to emulate a user clicking through videos
- AWS lambda functions allowed for parallelized webscraping for each "parent" video
- **Pros:**
  - Does not reflect user viewing history
  - Roughly simulates actual an user
- **Cons:**
  - Does not reflect user viewing history
  - Does not fully "watch" the videos

Data Collection Process



Caption Data Lost



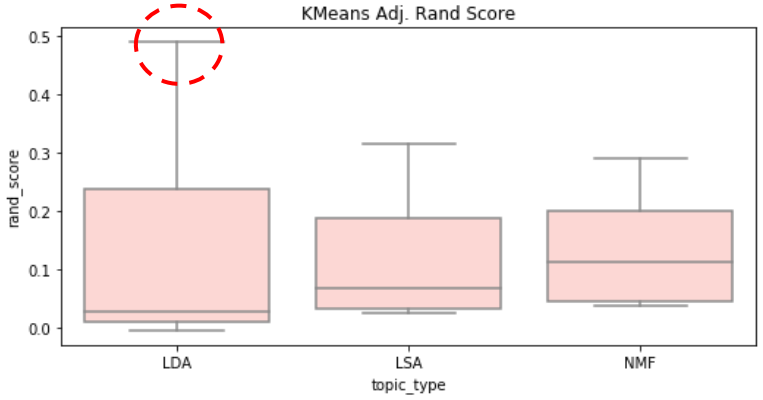
Preprocessing Methodology

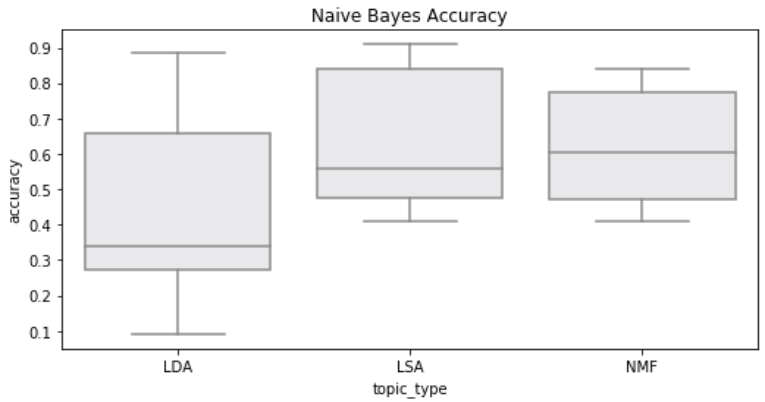
Scoring on ability to discern four political search queries

Key Points

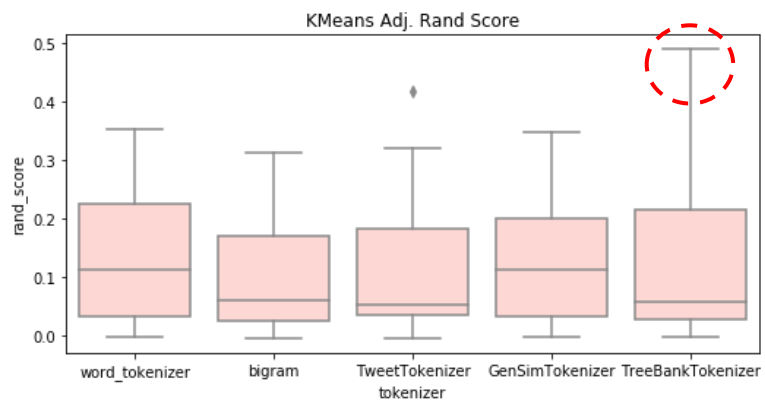
- 600 combinations of preprocessing/topic analysis were scored on clustering/classifying
- LDA topic analysis and TreeBankTokenizers had the largest impact on scores

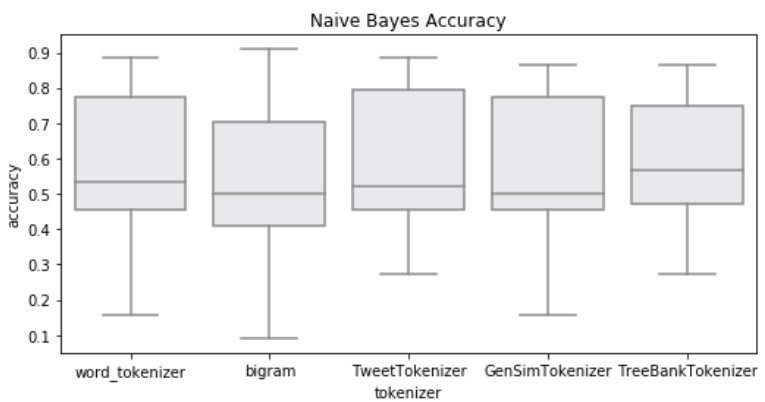
Scoring by Topic Analysis



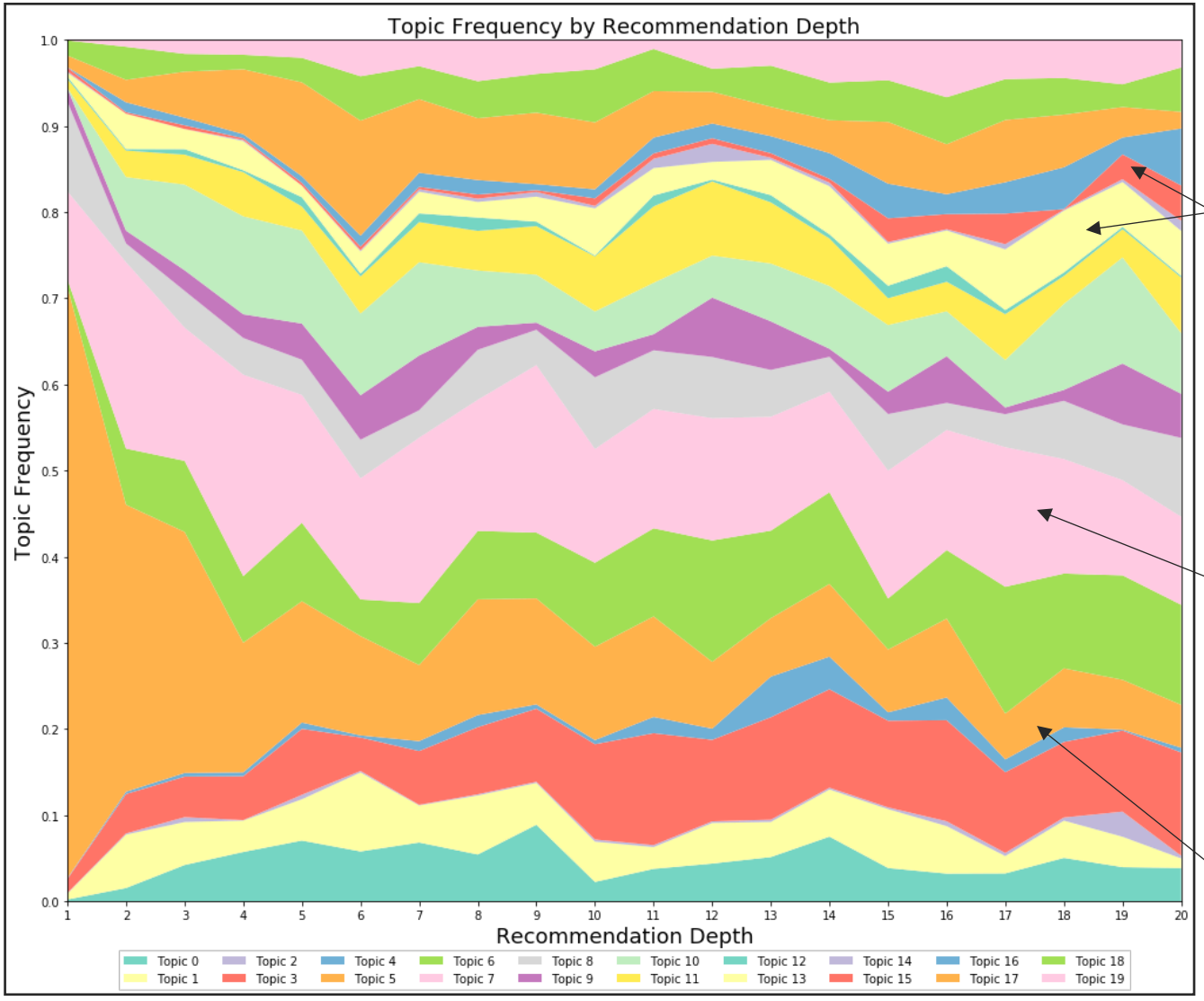


Scoring by Tokenizer





Topic Analysis: “Donald Trump Rally”



Religion

- Conspiracy
  - Church
  - Secret
  - Ancient
- Evangelizing
  - God
  - Power
  - Nation

Political Comments

- Trump
- Clinton
- Hillary
- Democrat
- Republican

Trump Talk

- Great
- American
- Thank

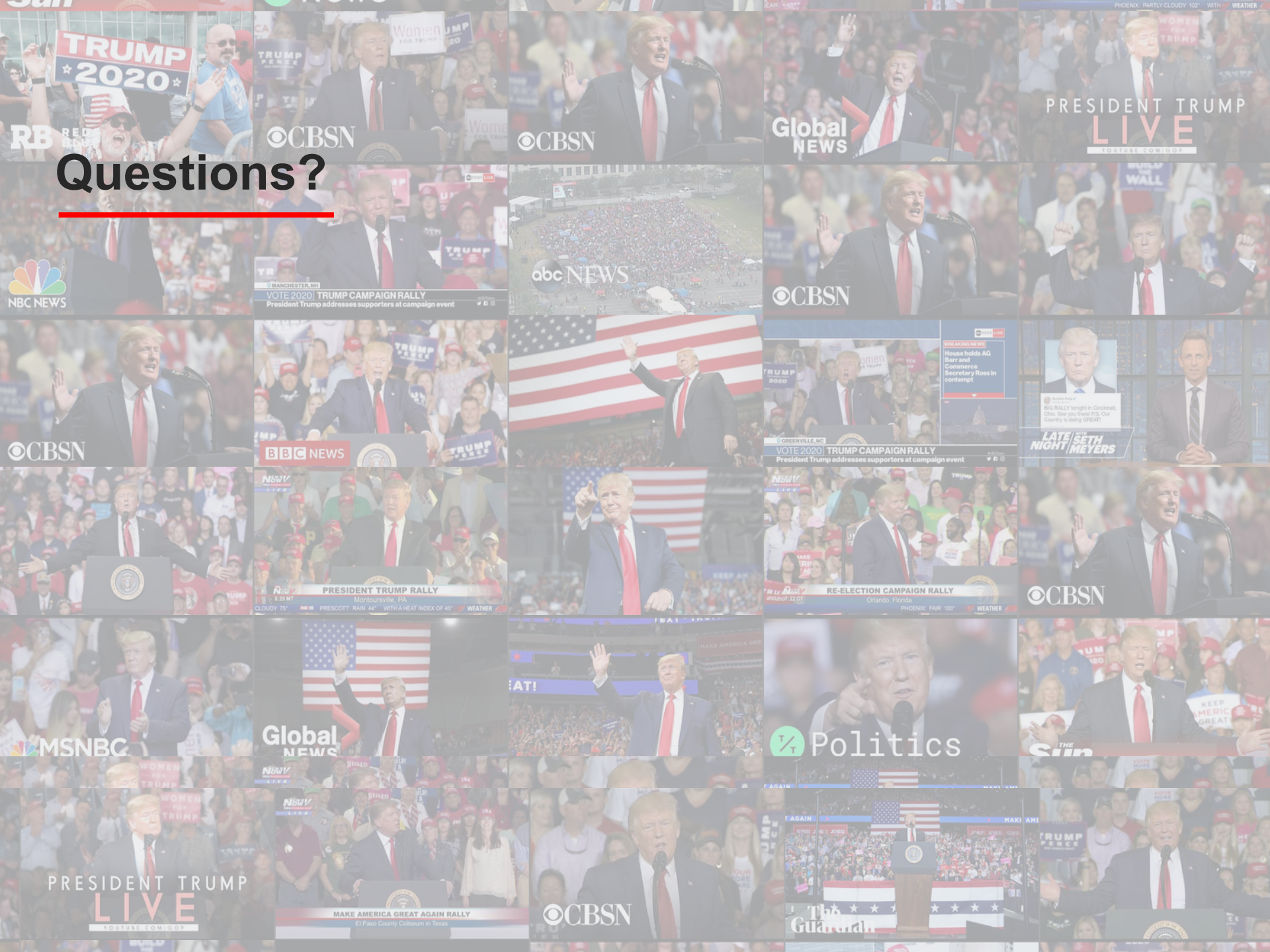
Conclusion

Conclusion

- The content of YouTube's search results does change over successive recommendations
  - Some of these changes maintain the original videos broad theme and introduce similar topics
  - A small percentage of recommendations move toward very different, controversial topics
  - Anecdotally, polarizing videos are more likely to have even more controversial recommendations

Future Work

- Data Collection:
  - Improvements to simulate user engagement like fully watching videos
  - Random walk: selecting a video at random and only looking at immediate recommendations
- Modeling:
  - Clustering with sentiment analysis and additional video meta data
  - Generalizing model to look at many topics


TRUMP 2020
CBSN
Global NEWS
PRESIDENT TRUMP LIVE
NBC NEWS
VOTE 2020 TRUMP CAMPAIGN RALLY
President Trump addresses supporters at campaign event
abc NEWS
BBC NEWS
LATE NIGHT WITH SETH MEYERS
PRESIDENT TRUMP RALLY
RE-ELECTION CAMPAIGN RALLY
MSNBC
Politics
MAKE AMERICA GREAT AGAIN RALLY
The Guardian

Questions?

Appendix: Preprocessor Scores

Top 20 by cluster score

Score Summary

topic_type	vectorizer	tokenizer	stemmer	min_df	max_df	rand_score	accuracy
LDA	CV	TreeBankTokenizer	PorterStemmer	0.1	0.9	0.49	0.82
LDA	CV	TweetTokenizer	SnowballStemmer	0.1	0.9	0.42	0.80
LDA	CV	TreeBankTokenizer	LancasterStemmer	0.2	0.8	0.41	0.66
LDA	CV	TreeBankTokenizer	WordNetLemmatizer	0.1	0.9	0.41	0.66
LDA	CV	TreeBankTokenizer	LancasterStemmer	0.1	0.9	0.40	0.70
LDA	CV	TreeBankTokenizer	None	0.2	0.8	0.40	0.75
LDA	CV	word_tokenizer	WordNetLemmatizer	0.1	0.9	0.35	0.64
LDA	CV	GenSimTokenizer	WordNetLemmatizer	0.1	0.9	0.35	0.80
LDA	CV	TreeBankTokenizer	SnowballStemmer	0.2	0.8	0.35	0.70
LDA	CV	TreeBankTokenizer	WordNetLemmatizer	0.2	0.8	0.34	0.84
LDA	CV	TreeBankTokenizer	PorterStemmer	0.2	0.8	0.34	0.70
LDA	CV	word_tokenizer	None	0.1	0.9	0.34	0.84
LDA	CV	word_tokenizer	LancasterStemmer	-	1.0	0.34	0.25
LDA	CV	word_tokenizer	LancasterStemmer	0.2	0.8	0.33	0.64
LDA	CV	GenSimTokenizer	PorterStemmer	0.1	0.9	0.32	0.70
LDA	CV	TweetTokenizer	WordNetLemmatizer	0.2	0.8	0.32	0.73
LDA	CV	GenSimTokenizer	None	0.1	0.9	0.32	0.80
LDA	CV	GenSimTokenizer	WordNetLemmatizer	0.2	0.8	0.32	0.73
LSA	TF-IDF	TreeBankTokenizer	WordNetLemmatizer	0.1	0.9	0.32	0.84
LSA	TF-IDF	TreeBankTokenizer	None	0.1	0.9	0.32	0.84

Appendix: Topic Words

Top 20 words

Topic Summary

Rank	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
1	think	report	bodi	know	compani	peopl	know	trump	oh	presid
2	market	case	mind	peopl	comput	want	say	presid	come	investig
3	know	laughter	everi	think	buy	know	right	think	show	report
4	look	law	day	say	best	said	okay	donald	see	trump
5	see	peopl	health	well	job	great	yeah	clinton	new	mr
6	realli	immigr	love	realli	product	countri	peopl	state	peopl	russian
7	year	prison	cell	way	appl	say	said	peopl	job	question
8	mean	know	happi	thing	devic	right	want	say	de	would
9	right	polic	live	right	store	think	gonna	would	work	campaign
10	rate	new	feel	someth	technolog	year	look	hillari	make	correct
11	economi	court	flow	make	billion	american	got	right	la	say
12	us	said	becom	mean	dollar	job	time	know	us	gener
13	well	also	abil	kind	audienc	got	mean	want	applaus	fbi
14	dollar	famili	environ	would	year	time	guy	democrat	worker	attorney
15	talk	legal	perfect	tri	movi	thank	would	vote	kid	special
16	much	would	spirit	see	sale	come	think	said	watch	justic
17	china	year	thank	good	busi	never	talk	talk	truck	thank
18	presid	investig	breath	could	sell	america	ye	polit	even	time
19	bank	week	take	happen	servic	would	tell	republican	human	well
20	fed	crime	audienc	time	time	lot	call	look	drive	ask

Rank	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19
1	money	god	energi	church	earth	let	actual	china	think	plane
2	bank	say	coal	water	scientist	name	black	chines	know	flight
3	dollar	think	plant	year	drug	lord	hole	countri	market	air
4	year	life	water	would	mission	spirit	see	think	year	fli
5	would	know	mine	time	life	life	realli	state	compani	ship
6	peopl	us	power	found	use	god	star	year	look	engin
7	world	said	system	name	year	everi	time	world	realli	design
8	new	want	pressur	famili	million	enemi	look	peopl	say	world
9	govern	question	test	peopl	mile	power	would	power	well	land
10	make	would	work	place	ga	come	right	war	lot	first
11	million	world	clean	day	impact	break	around	would	see	make
12	time	well	new	new	research	would	even	us	peopl	time
13	work	way	wind	secret	time	cast	univers	unit	mean	two
14	financi	see	nuclear	mani	execut	releas	planet	way	gonna	would
15	busi	believ	natur	centuri	hit	nation	littl	know	would	use
16	state	time	claim	call	death	upon	basic	govern	us	hour
17	debt	come	oper	sea	could	give	earth	see	busi	speed
18	system	look	area	use	planet	place	let	time	right	new
19	creat	peopl	fuel	ancient	object	word	mass	two	good	test
20	pay	live	sea	come	system	command	kind	polit	got	number