Topic Modelling Experiments

Patrick Martin

January 30, 2018

1 Overview

This project focuses on the question: Can a topic model distinguish between contentful and non-contentful words? That is, topic models tend to suffer from overly common words taking up a large portion of the probability space, and from uncommon words mostly taking up space. Given a topic model, can we distinguish these three classes of words? If we can, can this help us with other tasks?

- If we remove the non-contentful words and train a new topic model, does it perform better?
- Can we identify synonyms better?
- Can we cluster topics better?

Of course, these also need to be compared against a baseline. For removing words, we will just look at usage thresholds. For the others, we want to look at the default topic model.

1.1 The Data

Let's go with a couple different datasets.

- 500k reddit comments, randomly chosen from March 2017
- 50k reddit comments from the 20 most common subreddits from March 2017:
    1. AskReddit
    2. politics
    3. The_Donald
    4. worldnews
    5. nba
    6. RocketLeagueExchange
    7. pics

8. funny
9. videos
10. gaming
11. NintendoSwitch
12. leagueoflegends
13. news
14. Overwatch
15. Cricket
16. todayilearned
17. movies
18. SquaredCircle
19. pcmasterrace
20. gifs

2 Getting to know the model

This will be based on a 100-topic topic model. First, the gensim model .get_topics() function returns the probability distribution of words across each topic, that is the sum across words is 1. We might ask ourselves what happens if we sum across topics, which results in more or less a usage list:

the	3.055
and	1.156
for	0.784
this	0.755
that	0.601
with	0.462
like	0.434
about	0.425
deleted	0.423
out	0.408

First thing I'd like to point out is that for reddit data, 'deleted' should maybe be considered a "stop word", given its high usage. Anyway, ranking words based on their probability for each topic, here are the top ten words for a sampling of topics:

1:	the	was	and	that	had	could	for	but	with	one
6:	console	said	until	him	trying	physical	finally	seeing	imagine	word
14:	other	any	easily	handheld	means	the	allow	morning	midnight	with
17:	there's	run	real	we've	hit	win	details	purchase	weeks	can
20:	that's	question	hey	say	let	called	expect	unfortunately	smash	apologize
22:	the	back	before	able	and	times	update	came	toolbox	future
23:	the	system	and	that	luck	fact	further	pictures	close	from
25:	best	give	nice	days	kind	gaming	wasn't	idea	must	removing
33:	things	mode	mario	daily	life	change	story	share	the	main
43:	requests	standard	performance	resolution	nearly	base	ability	the	aware	larger
44:	should	went	thank	minutes	matter	thoughts	clean	for	bother	group
48:	the	and	then	into	while	with	down	get	for	often
53:	well	yeah	shit	guess	pay	perfect	charging	that	japan	except
55:	not	it's	the	i'm	but	that	and	for	just	sure
61:	will	case	again	personally	this	the	and	design	giveaway	terrible
70:	always	line	bit	everyone	reason	high	add	friends	party	for
75:	yet	items	within	systems	hasn't	including	happening	access	indie	despite
76:	consoles	appreciate	digital	cause	red	sweet	generally	doesnt	train	aiming
85:	way	you're	can't	live	cool	stop	okay	speed	uses	worst
87:	https	com	www	video	watch	there	youtube	usb	this	fans
88:	launch	instead	asking	due	streams	media	gameplay	lack	faq	bigger
89:	already	love	lol	thought	coming	problems	week	this	dude	behind
92:	megathread	does	hours	old	year	early	neon	kong	quick	gone
93:	game	lot	playing	the	edit	set	top	you'll	port	delivery
99:	switch	nintendo	did	use	played	the	person	joycons	literally	mind

As is pretty typical when ranking words like this, these topic keywords aren't particularly good at describing what is happening in each topic. If we instead rank words by p(topic—word) rather than p(word—topic), i.e. making columns sum to 1, this might make things better.

1:	had	could	few	stuff	wish	target	test	march	pokemon	lower
6:	console	said	until	him	trying	physical	finally	seeing	imagine	word
14:	other	easily	handheld	means	allow	morning	midnight	joke	interested	cut
17:	there's	run	real	we've	hit	win	details	purchase	weeks	ongoing
20:	that's	question	hey	say	let	called	expect	unfortunately	smash	apologize
22:	back	before	able	times	update	came	toolbox	future	room	settings
23:	system	luck	fact	further	pictures	close	points	cases	runs	certainly
25:	best	give	nice	days	kind	gaming	wasn't	idea	must	removing
33:	things	mode	mario	daily	life	change	story	share	main	sales
43:	requests	standard	performance	resolution	nearly	base	ability	aware	larger	valid
44:	should	went	minutes	matter	thoughts	clean	bother	group	skin	term
48:	often	scratches	selling	service	face	customer	scratch	fangled	anywhere	yours
53:	well	yeah	shit	guess	pay	perfect	charging	japan	except	holy
55:	i'm	sure	saying	says	stock	body	error	whether	discussing	shrine
61:	will	case	again	personally	design	giveaway	terrible	arrive	potential	law
70:	always	line	bit	everyone	reason	high	add	party	sound	list
75:	yet	items	within	systems	hasn't	including	happening	access	indie	despite
76:	consoles	appreciate	digital	cause	red	sweet	generally	doesnt	train	aiming
85:	way	you're	can't	live	cool	stop	okay	uses	worst	telling
87:	www	video	watch	youtube	usb	fans	youtu	utf8	statement	prevent
88:	launch	instead	asking	due	streams	media	gameplay	lack	faq	bigger
89:	already	love	lol	thought	coming	problems	week	dude	behind	looked
92:	megathread	hours	old	year	early	does	neon	kong	quick	gone
93:	game	lot	playing	edit	set	top	you'll	port	delivery	edition
99:	switch	nintendo	did	use	played	person	literally	mind	button	hopefully

This doesn't actually look much better. Words like 'well', 'had', 'did' are still taking up space that could go to more informative words. One concern might be that some of these topics are just bad; this can happen that the model groups words together in a way that doesn't really reflect a certain topic. In any case, we really care most about the most prevalent topics; if summing the columns when the rows were normalized to 1 gave usage ranks for the words, then maybe summing the rows when the columns are normalized to 1 will give usage ranks for the topics?

65	6441.37
24	5292.49
43	5265.69
50	5184.74
49	5154.73
67	5060.22
84	4960.22
76	4866.09
47	4486.29
⋮	
69	1784.73
5	1714.50
71	1702.69
13	1700.76
45	1272.95
99	1211.35
1	740.30
66	661.98
55	655.85
74	650.12
57	615.27

So if we use these topics as our guides, these are the keywords

65:	amp	preorder	shoutouts	repost	note	subject	net	previous	appears	core
	amp	preorder	shoutouts	repost	note	subject	net	previous	appears	core
24:	man	friend	kids	stream	who's	changed	sony	plan	720p	nvidia
	man	friend	kids	stream	who's	changed	sony	plan	720p	nvidia
43:	requests	standard	performance	resolution	nearly	base	ability	the	aware	larger
	requests	standard	performance	resolution	nearly	base	ability	aware	larger	valid
50:	joy	leave	return	zero	across	walk	ignore	limit	feet	feedback
	joy	leave	return	zero	across	walk	ignore	limit	feet	feedback
49:	removed	news	thinking	wow	fight	smart	fake	obvious	collection	refund
	news	thinking	wow	fight	smart	fake	obvious	collection	refund	awful
67:	funny	call	sorry	level	series	fit	bring	idk	gray	aside
	funny	call	sorry	level	series	fit	bring	idk	gray	aside
84:	website	similar	stick	ship	charged	they'll	current	donkey	jump	wont
	website	similar	stick	ship	charged	they'll	current	donkey	jump	wont
76:	consoles	appreciate	digital	cause	red	sweet	generally	doesnt	train	aiming
	consoles	appreciate	digital	cause	red	sweet	generally	doesnt	train	aiming
47:	super	each	friday	taken	noticed	units	showing	companies	characters	360
	super	each	friday	taken	noticed	units	showing	companies	characters	360
69:	reddit	comments	very	wiki	amazon	last	today	battery	with	posted
	reddit	comments	very	wiki	amazon	last	today	battery	posted	picked
5:	see	want	something	isn't	help	part	the	this	itself	huge
	see	want	something	isn't	help	part	itself	huge	running	front
71:	post	here	maybe	i'll	found	won't	walmart	this	sort	fast
	post	here	maybe	i'll	found	won't	walmart	sort	fast	amazing
13:	has	since	i'd	been	anyone	years	for	account	ago	took
	since	i'd	anyone	years	account	ago	took	hardware	number	rules
45:	have	i've	would	the	never	been	and	this	but	that
	i've	never	ever	haven't	seen	wanted	true	damage	words	explanation
99:	switch	nintendo	did	use	played	the	person	joycons	literally	mind
	switch	nintendo	did	use	played	person	literally	mind	button	hopefully
1:	the	was	and	that	had	could	for	but	with	one
	had	could	few	stuff	wish	target	test	march	pokemon	lower
66:	the	they	and	are	that	them	people	for	have	their
	they	their	bought	world	them	third	pass	places	powerful	designed
55:	not	it's	the	i'm	but	that	and	for	just	sure
	i'm	sure	saying	says	stock	body	error	whether	discussing	shrine
74:	your	you	please	questions	nintendoswitch	this	the	message	have	for
	please	questions	compose	2fr	message	link	subreddit	contact	review	submission
57:	you	your	can	and	have	are	that	don't	get	with
	these	hope	agree	easy	hate	inventory	123	reading	bottom	suck

So in actuality this pro-

cess didn't actually change much. One thing that worked well in the past was smoothing, that is words would get a slight boost to their score in every topic.

65:	amp	preorder	shoutouts	repost	note	subject	net	previous	appears	core
	amp	preorder	shoutouts	repost	note	subject	net	previous	appears	core
24:	man	friend	kids	stream	who's	changed	sony	plan	720p	nvidia
	man	friend	kids	stream	who's	changed	sony	plan	720p	nvidia
43:	requests	standard	performance	resolution	nearly	base	ability	the	aware	larger
	requests	standard	performance	resolution	nearly	base	ability	aware	larger	valid
50:	joy	leave	return	zero	across	walk	ignore	limit	feet	feedback
	joy	leave	return	zero	across	walk	ignore	limit	feet	feedback
49:	removed	news	thinking	wow	fight	smart	fake	obvious	collection	refund
	news	thinking	wow	fight	smart	fake	obvious	collection	refund	awful
67:	funny	call	sorry	level	series	fit	bring	idk	gray	aside
	funny	call	sorry	level	series	fit	bring	idk	gray	aside
84:	website	similar	stick	ship	charged	they'll	current	donkey	jump	wont
	website	similar	stick	ship	charged	they'll	current	donkey	jump	wont
76:	consoles	appreciate	digital	cause	red	sweet	generally	doesnt	train	aiming
	consoles	appreciate	digital	cause	red	sweet	generally	doesnt	train	aiming
47:	super	each	friday	taken	noticed	units	showing	companies	characters	360
	super	each	friday	taken	noticed	units	showing	companies	characters	360
69:	reddit	comments	very	wiki	amazon	last	today	battery	with	posted
	reddit	comments	very	wiki	amazon	last	today	battery	posted	picked
5:	see	want	something	isn't	help	part	the	this	itself	huge
	see	want	something	isn't	help	part	itself	huge	running	front
71:	post	here	maybe	i'll	found	won't	walmart	this	sort	fast
	post	here	maybe	i'll	found	won't	walmart	sort	fast	amazing
13:	has	since	i'd	been	anyone	years	for	account	ago	took
	since	i'd	anyone	years	account	ago	took	hardware	number	rules
45:	have	i've	would	the	never	been	and	this	but	that
	i've	never	ever	haven't	seen	wanted	true	damage	words	explanation
99:	switch	nintendo	did	use	played	the	person	joycons	literally	mind
	switch	nintendo	did	use	played	person	literally	mind	button	hopefully
1:	the	was	and	that	had	could	for	but	with	one
	had	could	few	stuff	couldn't	wish	target	test	march	pokemon
66:	the	they	and	are	that	them	people	for	have	their
	they	them	their	bought	world	aren't	third	pass	places	powerful
55:	not	it's	the	i'm	but	that	and	for	just	sure
	i'm	sure	saying	says	stock	body	error	whether	discussing	it's
74:	your	you	please	questions	nintendoswitch	this	the	message	have	for
	please	questions	message	compose	2fr	link	subreddit	contact	review	submission
57:	you	your	can	and	have	are	that	don't	get	with
	these	hope	agree	easy	hate	inventory	123	reading	bottom	you