FINANCIAL NEWS TOPIC MODELING AND RECOMMENDER

Jay Kwon

Metis Data Science and Engineering (flex program)

Module 5 – Unsupervised Learning and NLP

May 18, 2022

Goal:

1. Define Topics for Financial News Articles
2. Build a recommender for any given article

Motivation:

- people in finance must pay close attention to the financial news
- ability to categorize articles base on topics would be useful
- ability to pull up similar articles
- distribution of articles over topics is insightful
  - **investors** (individuals, wealth management firms, hedge funds, pension funds)
  - **financial news** (Bloomberg, CNBC, Wall Street Journal)
  - **ratings agencies** (S&P, Moody’s, Fitch)
  - **financial services**

DATA AND METHOD

Data:

- https://www.kaggle.com/datasets/jeet2016/us-financial-news-articles?datasetId=49948
- US News articles: Bloomberg.com, CNBC.com, reuters.com, wsj.com, and fortune.com
- Jan to May, 2018
- 111747 documents

Method:

- preprocessing included lemmatization and filtering to nouns
- LDA used to define 20 topics
- cosine similarity for recommender

Tools used:

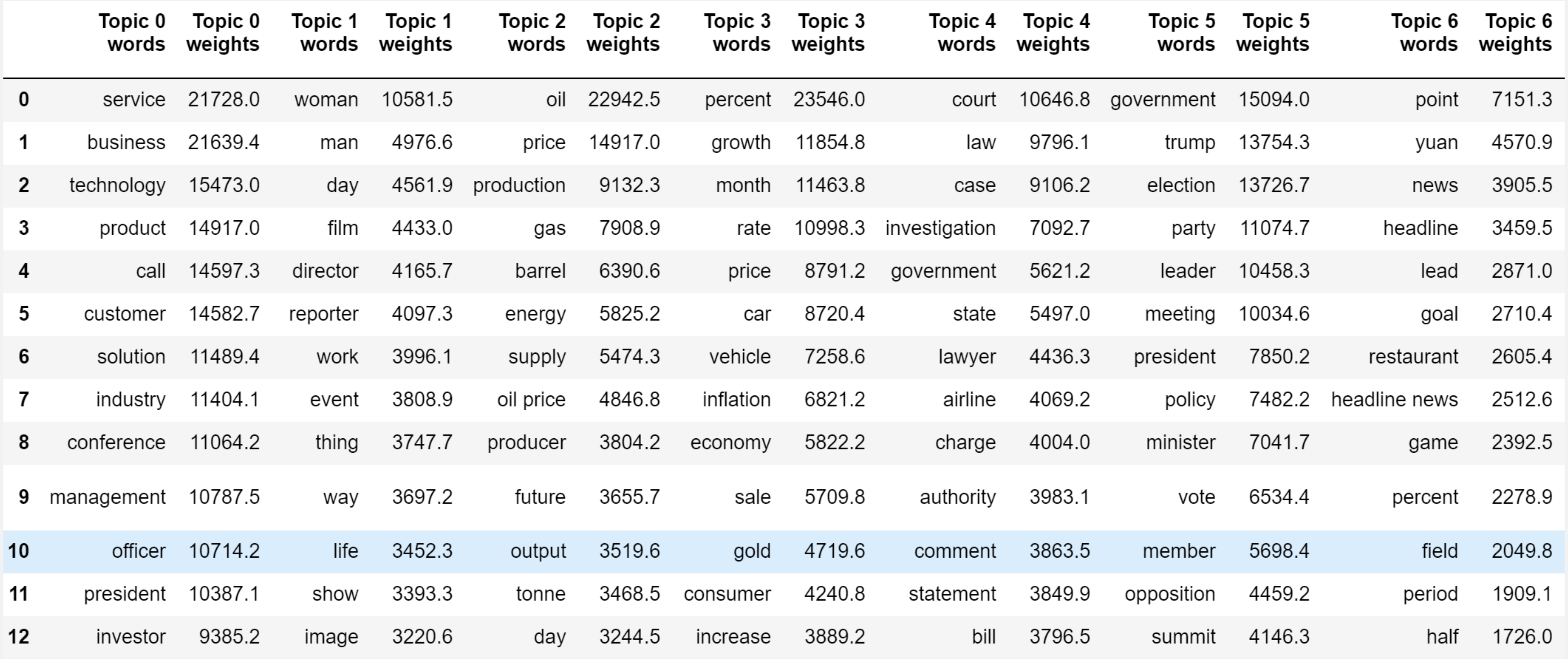
- spaCy, nltk, sklearn (LDA), Pandas, Matplotlib

Top Words for Topics 0~6

0: Management (Tech)

2: Oil

4: Lawsuits
5: US Government

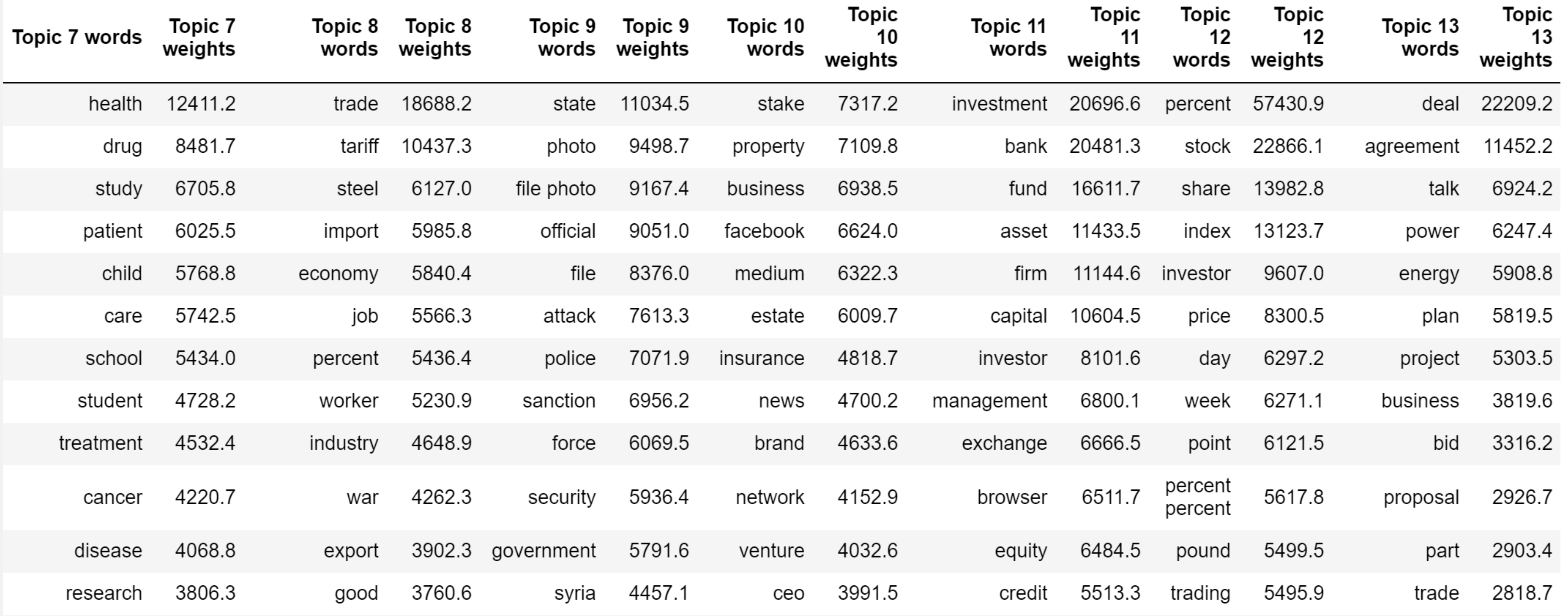


	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights	Topic 6 words	Topic 6 weights
0	service	21728.0	woman	10581.5	oil	22942.5	percent	23546.0	court	10646.8	government	15094.0	point	7151.3
1	business	21639.4	man	4976.6	price	14917.0	growth	11854.8	law	9796.1	trump	13754.3	yuan	4570.9
2	technology	15473.0	day	4561.9	production	9132.3	month	11463.8	case	9106.2	election	13726.7	news	3905.5
3	product	14917.0	film	4433.0	gas	7908.9	rate	10998.3	investigation	7092.7	party	11074.7	headline	3459.5
4	call	14597.3	director	4165.7	barrel	6390.6	price	8791.2	government	5621.2	leader	10458.3	lead	2871.0
5	customer	14582.7	reporter	4097.3	energy	5825.2	car	8720.4	state	5497.0	meeting	10034.6	goal	2710.4
6	solution	11489.4	work	3996.1	supply	5474.3	vehicle	7258.6	lawyer	4436.3	president	7850.2	restaurant	2605.4
7	industry	11404.1	event	3808.9	oil price	4846.8	inflation	6821.2	airline	4069.2	policy	7482.2	headline news	2512.6
8	conference	11064.2	thing	3747.7	producer	3804.2	economy	5822.2	charge	4004.0	minister	7041.7	game	2392.5
9	management	10787.5	way	3697.2	future	3655.7	sale	5709.8	authority	3983.1	vote	6534.4	percent	2278.9
10	officer	10714.2	life	3452.3	output	3519.6	gold	4719.6	comment	3863.5	member	5698.4	field	2049.8
11	president	10387.1	show	3393.3	tonne	3468.5	consumer	4240.8	statement	3849.9	opposition	4459.2	period	1909.1
12	investor	9385.2	image	3220.6	day	3244.5	increase	3889.2	bill	3796.5	summit	4146.3	half	1726.0

Top Words for Topics 7~13

7: Health
8: Trade (export/import)

11: Investment Management

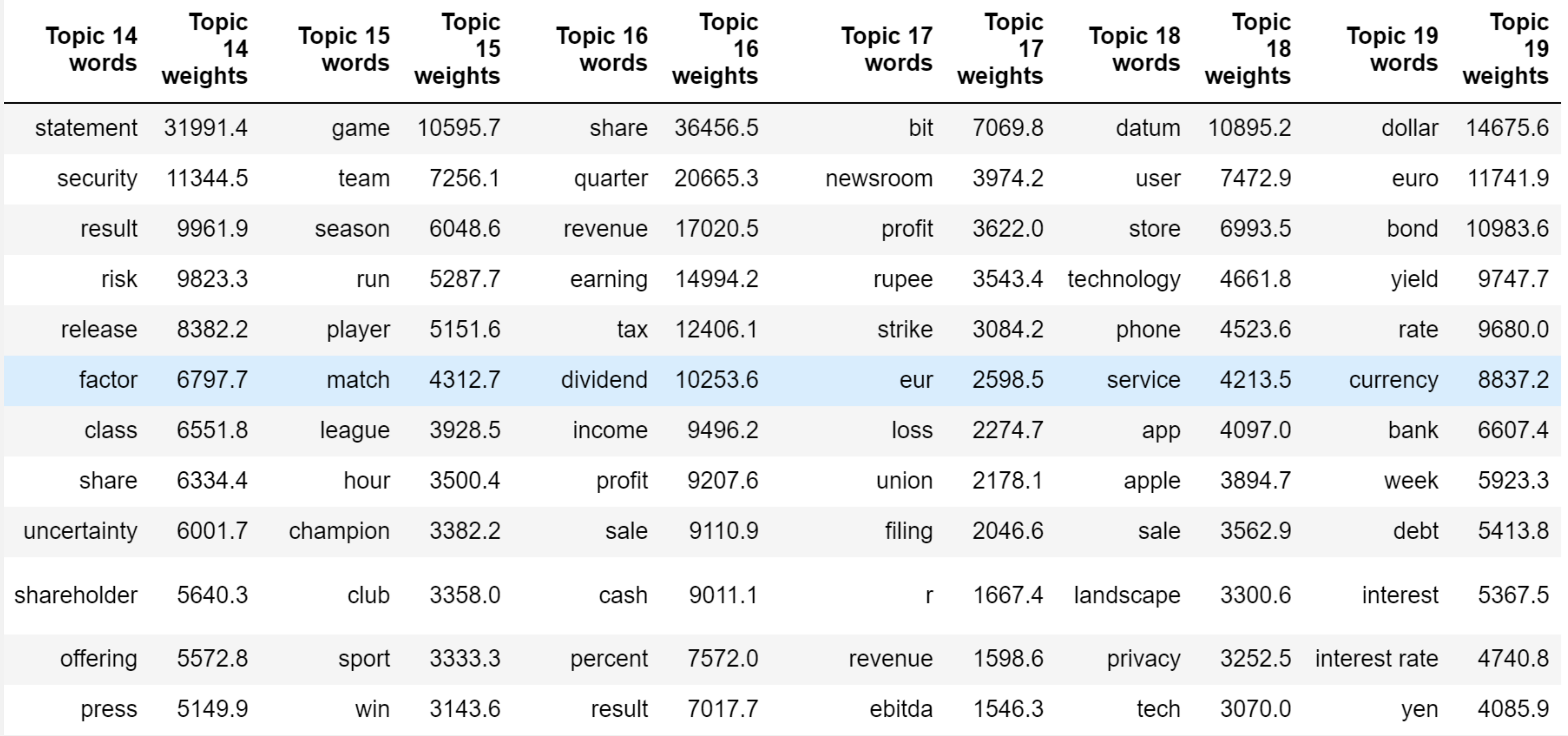


Topic 7 words	Topic 7 weights	Topic 8 words	Topic 8 weights	Topic 9 words	Topic 9 weights	Topic 10 words	Topic 10 weights	Topic 11 words	Topic 11 weights	Topic 12 words	Topic 12 weights	Topic 13 words	Topic 13 weights
health	12411.2	trade	18688.2	state	11034.5	stake	7317.2	investment	20696.6	percent	57430.9	deal	22209.2
drug	8481.7	tariff	10437.3	photo	9498.7	property	7109.8	bank	20481.3	stock	22866.1	agreement	11452.2
study	6705.8	steel	6127.0	file photo	9167.4	business	6938.5	fund	16611.7	share	13982.8	talk	6924.2
patient	6025.5	import	5985.8	official	9051.0	facebook	6624.0	asset	11433.5	index	13123.7	power	6247.4
child	5768.8	economy	5840.4	file	8376.0	medium	6322.3	firm	11144.6	investor	9607.0	energy	5908.8
care	5742.5	job	5566.3	attack	7613.3	estate	6009.7	capital	10604.5	price	8300.5	plan	5819.5
school	5434.0	percent	5436.4	police	7071.9	insurance	4818.7	investor	8101.6	day	6297.2	project	5303.5
student	4728.2	worker	5230.9	sanction	6956.2	news	4700.2	management	6800.1	week	6271.1	business	3819.6
treatment	4532.4	industry	4648.9	force	6069.5	brand	4633.6	exchange	6666.5	point	6121.5	bid	3316.2
cancer	4220.7	war	4262.3	security	5936.4	network	4152.9	browser	6511.7	percent percent	5617.8	proposal	2926.7
disease	4068.8	export	3902.3	government	5791.6	venture	4032.6	equity	6484.5	pound	5499.5	part	2903.4
research	3806.3	good	3760.6	syria	4457.1	ceo	3991.5	credit	5513.3	trading	5495.9	trade	2818.7

Top Words for Topics 14~19

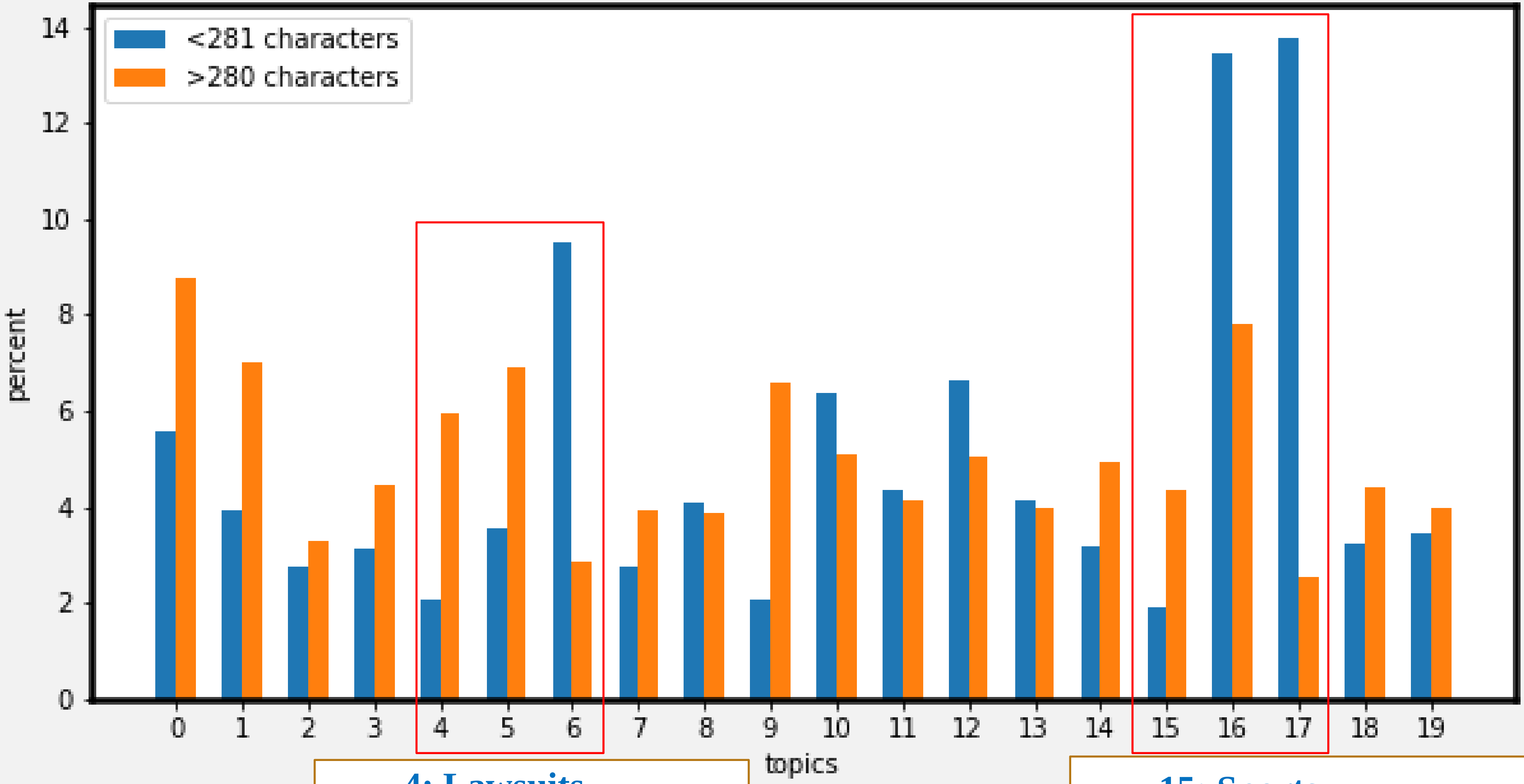
15: Sports
16: Business performance
(financial statement metrics)

Topic 18: Tech
Topic 19: Currency and Rates



Topic 14 words	Topic 14 weights	Topic 15 words	Topic 15 weights	Topic 16 words	Topic 16 weights	Topic 17 words	Topic 17 weights	Topic 18 words	Topic 18 weights	Topic 19 words	Topic 19 weights
statement	31991.4	game	10595.7	share	36456.5	bit	7069.8	datum	10895.2	dollar	14675.6
security	11344.5	team	7256.1	quarter	20665.3	newsroom	3974.2	user	7472.9	euro	11741.9
result	9961.9	season	6048.6	revenue	17020.5	profit	3622.0	store	6993.5	bond	10983.6
risk	9823.3	run	5287.7	earning	14994.2	rupee	3543.4	technology	4661.8	yield	9747.7
release	8382.2	player	5151.6	tax	12406.1	strike	3084.2	phone	4523.6	rate	9680.0
factor	6797.7	match	4312.7	dividend	10253.6	eur	2598.5	service	4213.5	currency	8837.2
class	6551.8	league	3928.5	income	9496.2	loss	2274.7	app	4097.0	bank	6607.4
share	6334.4	hour	3500.4	profit	9207.6	union	2178.1	apple	3894.7	week	5923.3
uncertainty	6001.7	champion	3382.2	sale	9110.9	filing	2046.6	sale	3562.9	debt	5413.8
shareholder	5640.3	club	3358.0	cash	9011.1	r	1667.4	landscape	3300.6	interest	5367.5
offering	5572.8	sport	3333.3	percent	7572.0	revenue	1598.6	privacy	3252.5	interest rate	4740.8
press	5149.9	win	3143.6	result	7017.7	ebitda	1546.3	tech	3070.0	yen	4085.9

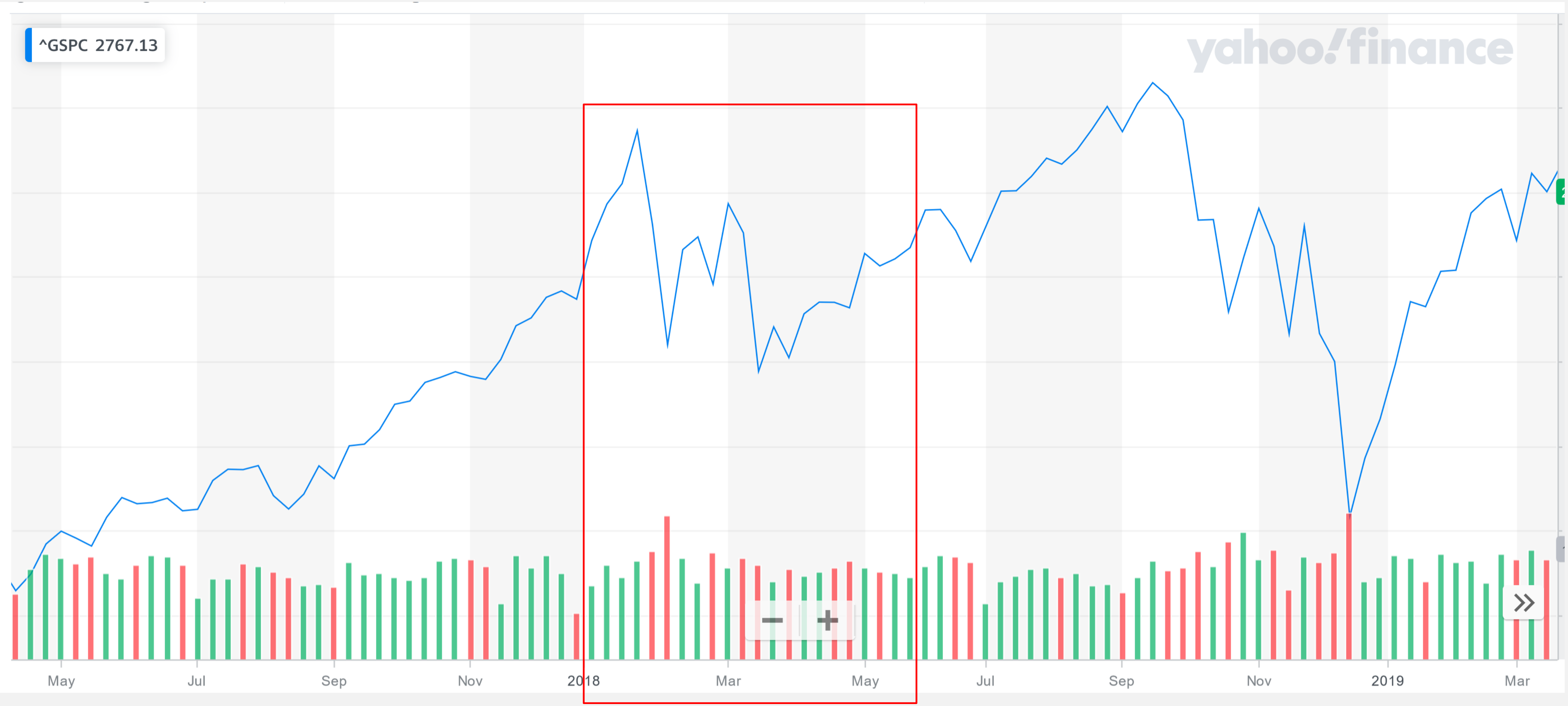
Topic Distribution by Character Count



4: Lawsuits
5: US Government
6: ?

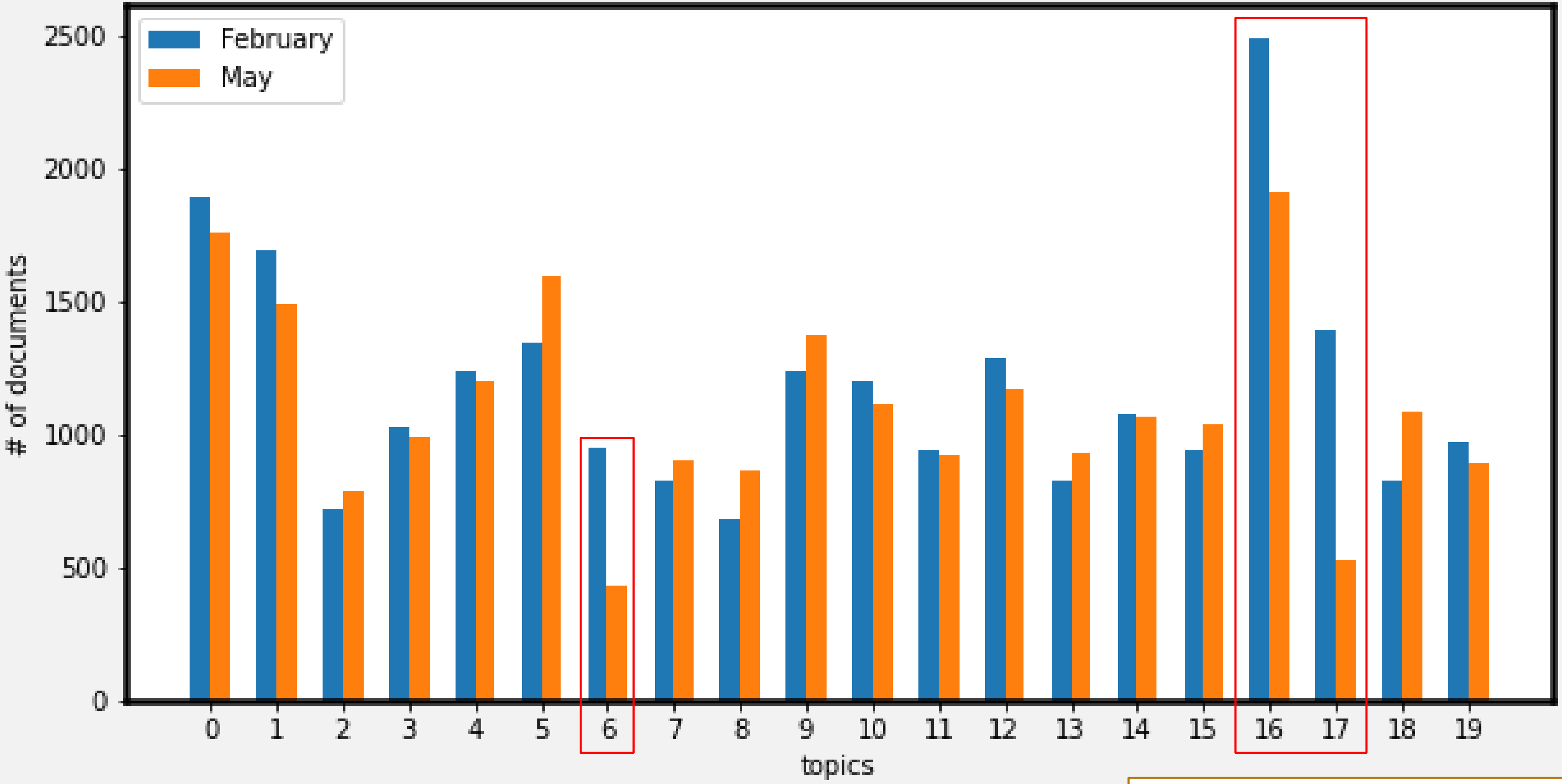
15: Sports
16: Business performance
17: ?

SP500



* sell-off in early Feb, 2018

Topic Distribution by Month: February vs May

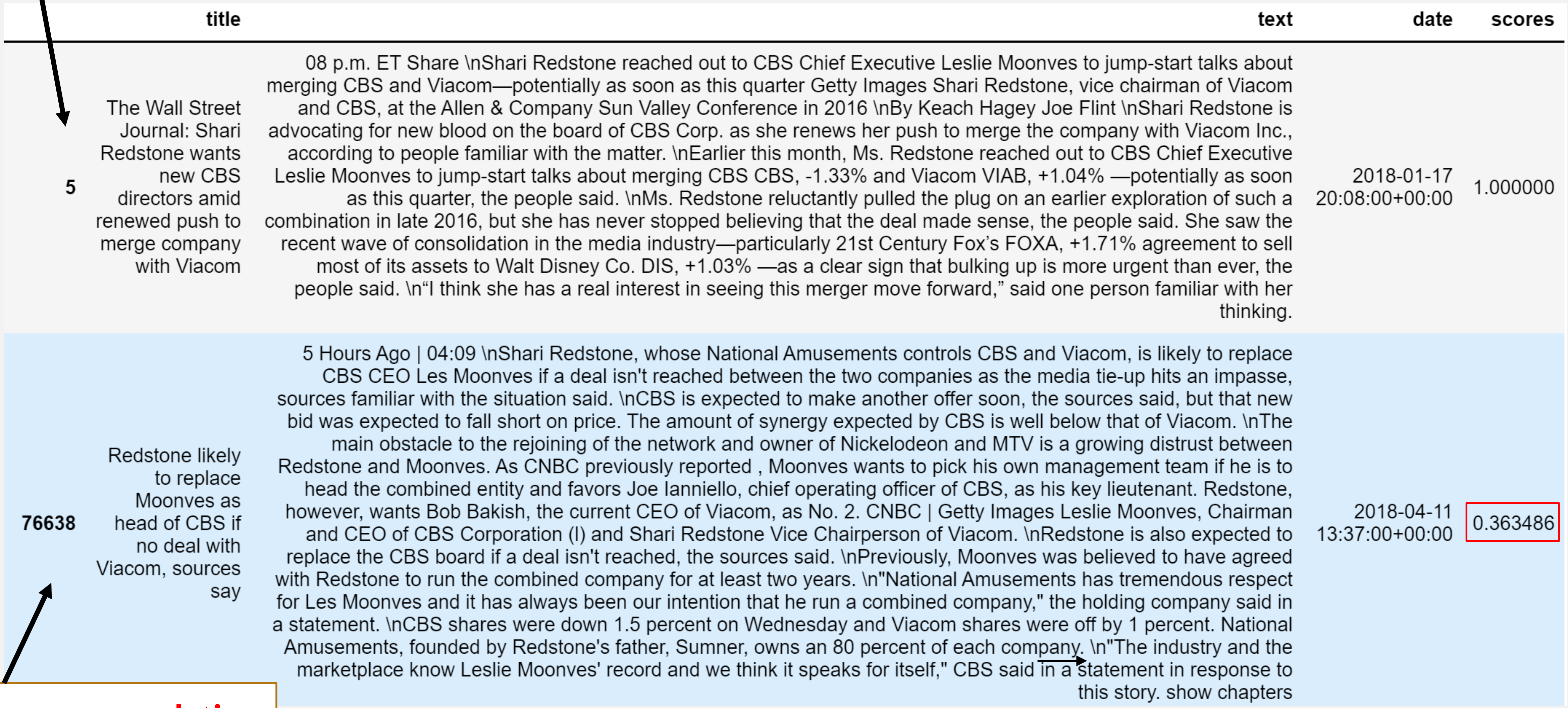


6: ?

16: Business performance
17: ?

Recommender

original article



	title	text	date	scores
5	The Wall Street Journal: Shari Redstone wants new CBS directors amid renewed push to merge company with Viacom	08 p.m. ET Share \nShari Redstone reached out to CBS Chief Executive Leslie Moonves to jump-start talks about merging CBS and Viacom—potentially as soon as this quarter Getty Images Shari Redstone, vice chairman of Viacom and CBS, at the Allen & Company Sun Valley Conference in 2016 \nBy Keach Hagey Joe Flint \nShari Redstone is advocating for new blood on the board of CBS Corp. as she renews her push to merge the company with Viacom Inc., according to people familiar with the matter. \nEarlier this month, Ms. Redstone reached out to CBS Chief Executive Leslie Moonves to jump-start talks about merging CBS CBS, -1.33% and Viacom VIAB, +1.04% —potentially as soon as this quarter, the people said. \nMs. Redstone reluctantly pulled the plug on an earlier exploration of such a combination in late 2016, but she has never stopped believing that the deal made sense, the people said. She saw the recent wave of consolidation in the media industry—particularly 21st Century Fox's FOXA, +1.71% agreement to sell most of its assets to Walt Disney Co. DIS, +1.03% —as a clear sign that bulking up is more urgent than ever, the people said. \n"I think she has a real interest in seeing this merger move forward," said one person familiar with her thinking.	2018-01-17 20:08:00+00:00	1.000000
76638	Redstone likely to replace Moonves as head of CBS if no deal with Viacom, sources say	5 Hours Ago | 04:09 \nShari Redstone, whose National Amusements controls CBS and Viacom, is likely to replace CBS CEO Les Moonves if a deal isn't reached between the two companies as the media tie-up hits an impasse, sources familiar with the situation said. \nCBS is expected to make another offer soon, the sources said, but that new bid was expected to fall short on price. The amount of synergy expected by CBS is well below that of Viacom. \nThe main obstacle to the rejoining of the network and owner of Nickelodeon and MTV is a growing distrust between Redstone and Moonves. As CNBC previously reported , Moonves wants to pick his own management team if he is to head the combined entity and favors Joe Ianniello, chief operating officer of CBS, as his key lieutenant. Redstone, however, wants Bob Bakish, the current CEO of Viacom, as No. 2. CNBC | Getty Images Leslie Moonves, Chairman and CEO of CBS Corporation (l) and Shari Redstone Vice Chairperson of Viacom. \nRedstone is also expected to replace the CBS board if a deal isn't reached, the sources said. \nPreviously, Moonves was believed to have agreed with Redstone to run the combined company for at least two years. \n"National Amusements has tremendous respect for Les Moonves and it has always been our intention that he run a combined company," the holding company said in a statement. \nCBS shares were down 1.5 percent on Wednesday and Viacom shares were off by 1 percent. National Amusements, founded by Redstone's father, Sumner, owns an 80 percent of each company. \n"The industry and the marketplace know Leslie Moonves' record and we think it speaks for itself," CBS said in a statement in response to this story. show chapters	2018-04-11 13:37:00+00:00	0.363486

top recommendation

Conclusions and Future Projects

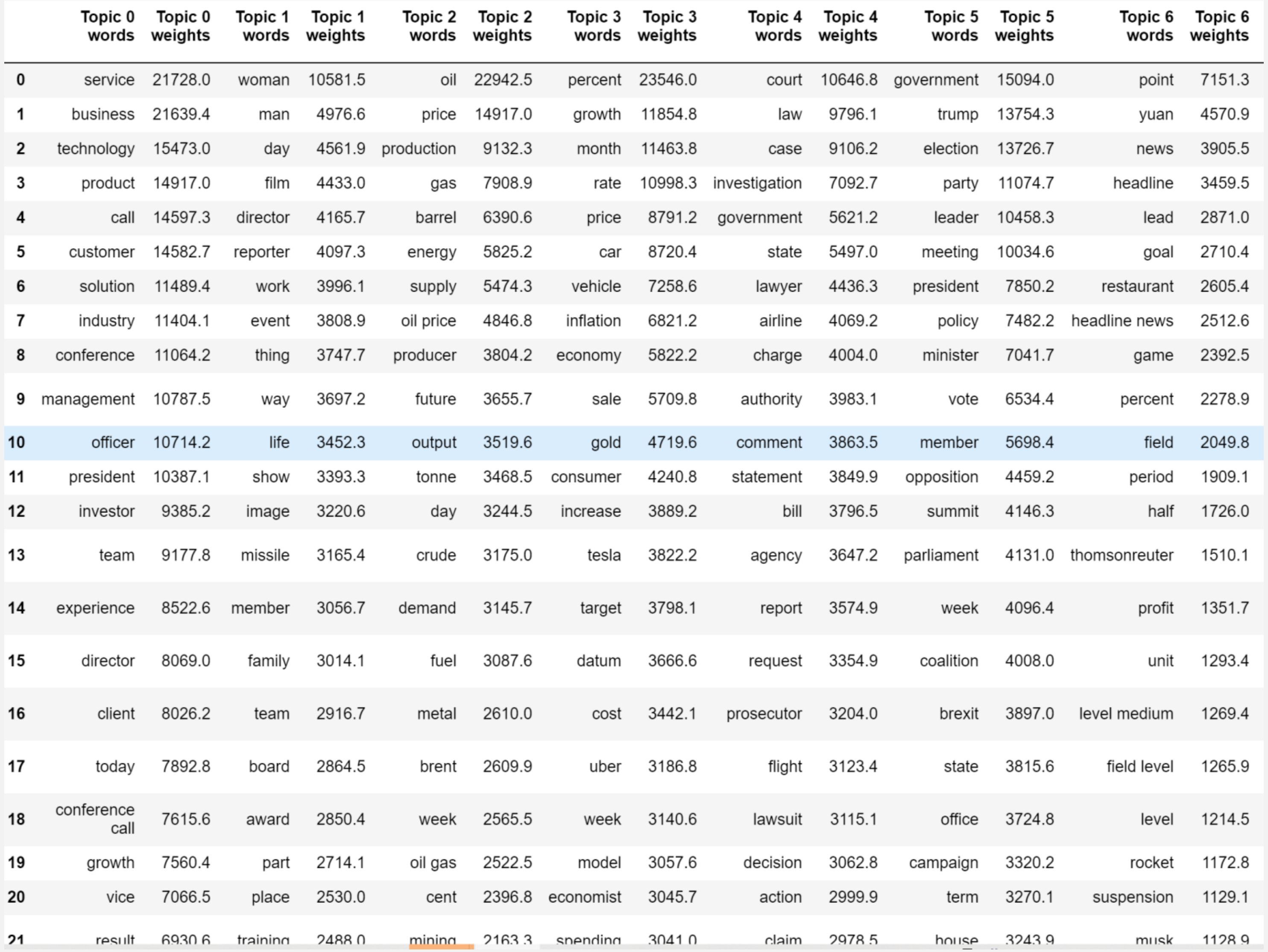
Conclusions:

- LDA modeling resulted in some clearly defined topics
- Readers can pull up articles based on topics
- Readers can use recommender to display similar articles
- some insights gained based on segmentation of data

Future Projects:

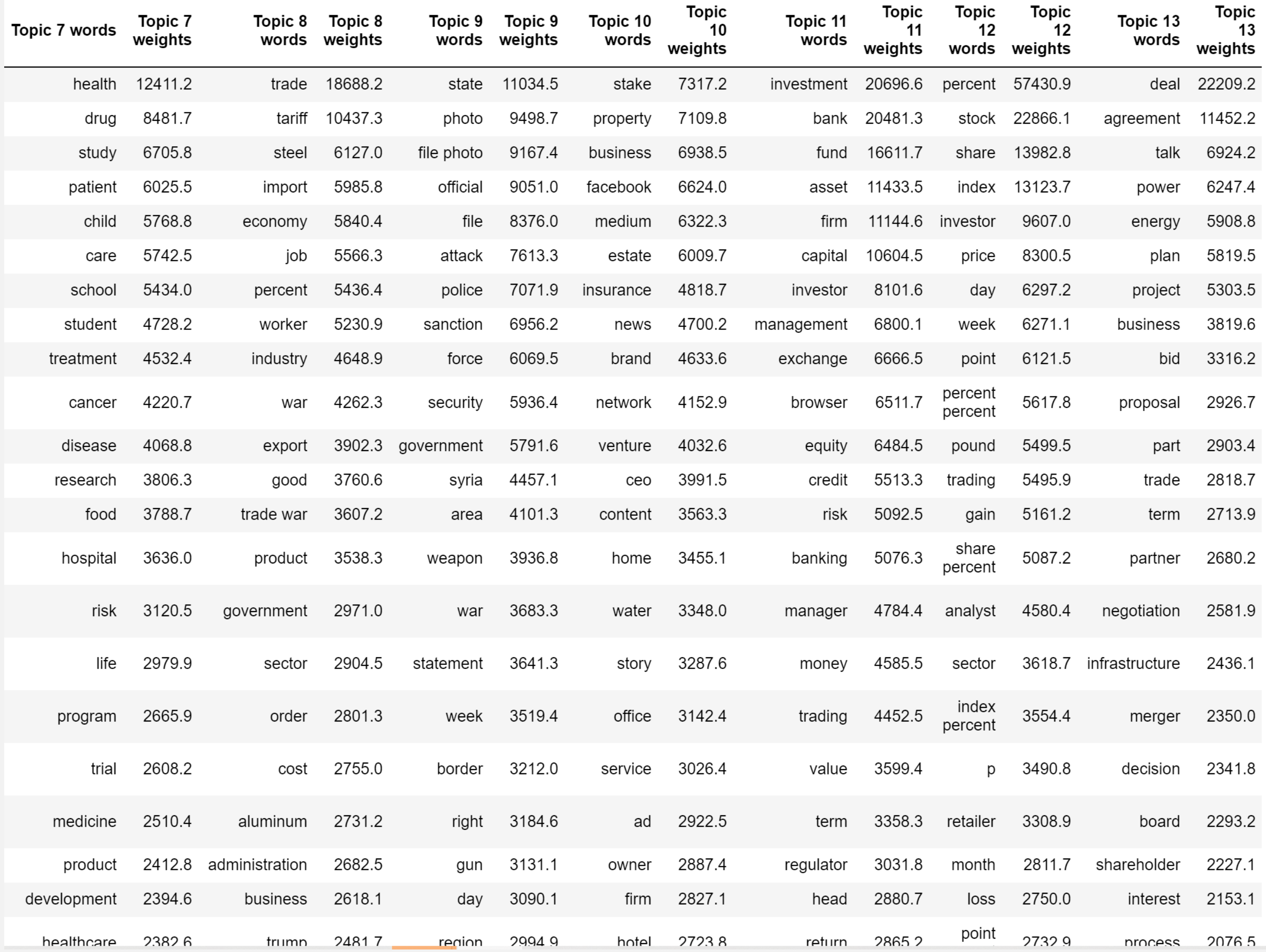
- Try increasing number of topics for better resolution
- Sentiment analysis

Supplemental



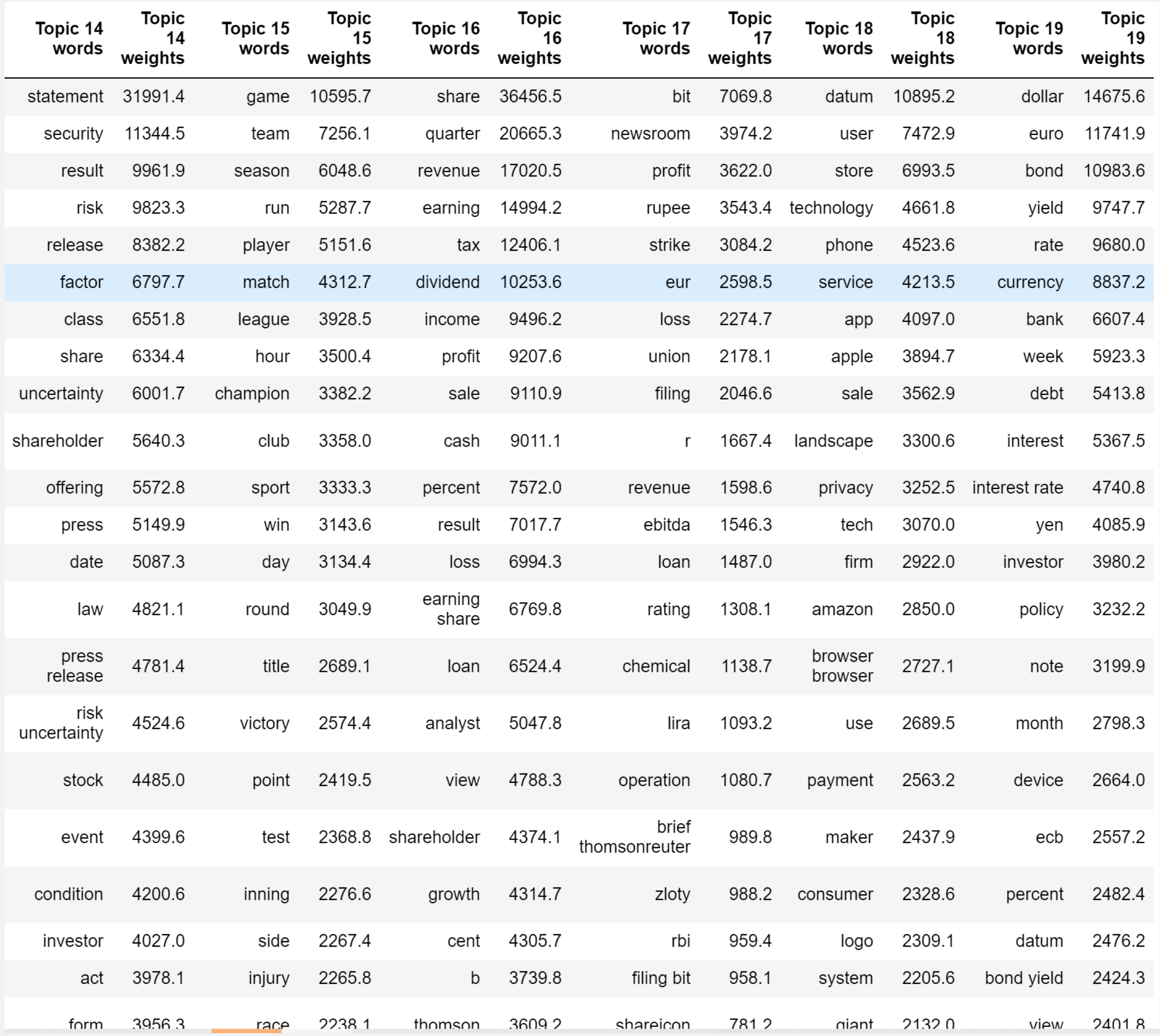
	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights	Topic 6 words	Topic 6 weights
0	service	21728.0	woman	10581.5	oil	22942.5	percent	23546.0	court	10646.8	government	15094.0	point	7151.3
1	business	21639.4	man	4976.6	price	14917.0	growth	11854.8	law	9796.1	trump	13754.3	yuan	4570.9
2	technology	15473.0	day	4561.9	production	9132.3	month	11463.8	case	9106.2	election	13726.7	news	3905.5
3	product	14917.0	film	4433.0	gas	7908.9	rate	10998.3	investigation	7092.7	party	11074.7	headline	3459.5
4	call	14597.3	director	4165.7	barrel	6390.6	price	8791.2	government	5621.2	leader	10458.3	lead	2871.0
5	customer	14582.7	reporter	4097.3	energy	5825.2	car	8720.4	state	5497.0	meeting	10034.6	goal	2710.4
6	solution	11489.4	work	3996.1	supply	5474.3	vehicle	7258.6	lawyer	4436.3	president	7850.2	restaurant	2605.4
7	industry	11404.1	event	3808.9	oil price	4846.8	inflation	6821.2	airline	4069.2	policy	7482.2	headline news	2512.6
8	conference	11064.2	thing	3747.7	producer	3804.2	economy	5822.2	charge	4004.0	minister	7041.7	game	2392.5
9	management	10787.5	way	3697.2	future	3655.7	sale	5709.8	authority	3983.1	vote	6534.4	percent	2278.9
10	officer	10714.2	life	3452.3	output	3519.6	gold	4719.6	comment	3863.5	member	5698.4	field	2049.8
11	president	10387.1	show	3393.3	tonne	3468.5	consumer	4240.8	statement	3849.9	opposition	4459.2	period	1909.1
12	investor	9385.2	image	3220.6	day	3244.5	increase	3889.2	bill	3796.5	summit	4146.3	half	1726.0
13	team	9177.8	missile	3165.4	crude	3175.0	tesla	3822.2	agency	3647.2	parliament	4131.0	thomsonreuter	1510.1
14	experience	8522.6	member	3056.7	demand	3145.7	target	3798.1	report	3574.9	week	4096.4	profit	1351.7
15	director	8069.0	family	3014.1	fuel	3087.6	datum	3666.6	request	3354.9	coalition	4008.0	unit	1293.4
16	client	8026.2	team	2916.7	metal	2610.0	cost	3442.1	prosecutor	3204.0	brexit	3897.0	level medium	1269.4
17	today	7892.8	board	2864.5	brent	2609.9	uber	3186.8	flight	3123.4	state	3815.6	field level	1265.9
18	conference call	7615.6	award	2850.4	week	2565.5	week	3140.6	lawsuit	3115.1	office	3724.8	level	1214.5
19	growth	7560.4	part	2714.1	oil gas	2522.5	model	3057.6	decision	3062.8	campaign	3320.2	rocket	1172.8
20	vice	7066.5	place	2530.0	cent	2396.8	economist	3045.7	action	2999.9	term	3270.1	suspension	1129.1
21	result	6930.6	training	2488.0	mining	2163.3	spending	3041.0	claim	2978.5	house	3243.9	musk	1128.9

Supplemental



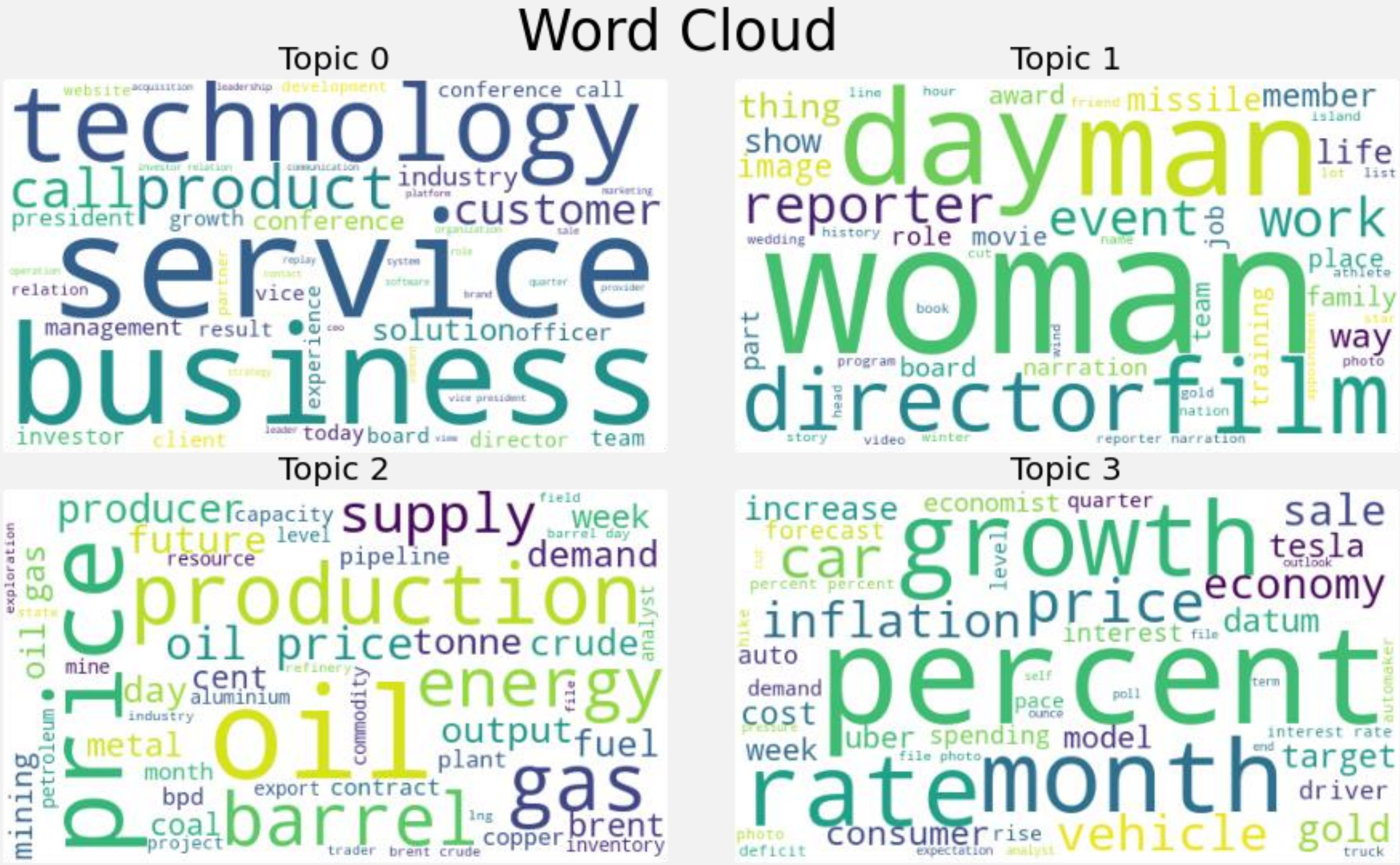
Topic 7 words	Topic 7 weights	Topic 8 words	Topic 8 weights	Topic 9 words	Topic 9 weights	Topic 10 words	Topic 10 weights	Topic 11 words	Topic 11 weights	Topic 12 words	Topic 12 weights	Topic 13 words	Topic 13 weights
health	12411.2	trade	18688.2	state	11034.5	stake	7317.2	investment	20696.6	percent	57430.9	deal	22209.2
drug	8481.7	tariff	10437.3	photo	9498.7	property	7109.8	bank	20481.3	stock	22866.1	agreement	11452.2
study	6705.8	steel	6127.0	file photo	9167.4	business	6938.5	fund	16611.7	share	13982.8	talk	6924.2
patient	6025.5	import	5985.8	official	9051.0	facebook	6624.0	asset	11433.5	index	13123.7	power	6247.4
child	5768.8	economy	5840.4	file	8376.0	medium	6322.3	firm	11144.6	investor	9607.0	energy	5908.8
care	5742.5	job	5566.3	attack	7613.3	estate	6009.7	capital	10604.5	price	8300.5	plan	5819.5
school	5434.0	percent	5436.4	police	7071.9	insurance	4818.7	investor	8101.6	day	6297.2	project	5303.5
student	4728.2	worker	5230.9	sanction	6956.2	news	4700.2	management	6800.1	week	6271.1	business	3819.6
treatment	4532.4	industry	4648.9	force	6069.5	brand	4633.6	exchange	6666.5	point	6121.5	bid	3316.2
cancer	4220.7	war	4262.3	security	5936.4	network	4152.9	browser	6511.7	percent percent	5617.8	proposal	2926.7
disease	4068.8	export	3902.3	government	5791.6	venture	4032.6	equity	6484.5	pound	5499.5	part	2903.4
research	3806.3	good	3760.6	syria	4457.1	ceo	3991.5	credit	5513.3	trading	5495.9	trade	2818.7
food	3788.7	trade war	3607.2	area	4101.3	content	3563.3	risk	5092.5	gain	5161.2	term	2713.9
hospital	3636.0	product	3538.3	weapon	3936.8	home	3455.1	banking	5076.3	share percent	5087.2	partner	2680.2
risk	3120.5	government	2971.0	war	3683.3	water	3348.0	manager	4784.4	analyst	4580.4	negotiation	2581.9
life	2979.9	sector	2904.5	statement	3641.3	story	3287.6	money	4585.5	sector	3618.7	infrastructure	2436.1
program	2665.9	order	2801.3	week	3519.4	office	3142.4	trading	4452.5	index percent	3554.4	merger	2350.0
trial	2608.2	cost	2755.0	border	3212.0	service	3026.4	value	3599.4	p	3490.8	decision	2341.8
medicine	2510.4	aluminum	2731.2	right	3184.6	ad	2922.5	term	3358.3	retailer	3308.9	board	2293.2
product	2412.8	administration	2682.5	gun	3131.1	owner	2887.4	regulator	3031.8	month	2811.7	shareholder	2227.1
development	2394.6	business	2618.1	day	3090.1	firm	2827.1	head	2880.7	loss	2750.0	interest	2153.1
healthcare	2382.6	trump	2481.7	region	2994.9	hotel	2723.8	return	2865.2	point	2732.9	process	2076.5

Supplemental



Topic 14 words	Topic 14 weights	Topic 15 words	Topic 15 weights	Topic 16 words	Topic 16 weights	Topic 17 words	Topic 17 weights	Topic 18 words	Topic 18 weights	Topic 19 words	Topic 19 weights
statement	31991.4	game	10595.7	share	36456.5	bit	7069.8	datum	10895.2	dollar	14675.6
security	11344.5	team	7256.1	quarter	20665.3	newsroom	3974.2	user	7472.9	euro	11741.9
result	9961.9	season	6048.6	revenue	17020.5	profit	3622.0	store	6993.5	bond	10983.6
risk	9823.3	run	5287.7	earning	14994.2	rupee	3543.4	technology	4661.8	yield	9747.7
release	8382.2	player	5151.6	tax	12406.1	strike	3084.2	phone	4523.6	rate	9680.0
factor	6797.7	match	4312.7	dividend	10253.6	eur	2598.5	service	4213.5	currency	8837.2
class	6551.8	league	3928.5	income	9496.2	loss	2274.7	app	4097.0	bank	6607.4
share	6334.4	hour	3500.4	profit	9207.6	union	2178.1	apple	3894.7	week	5923.3
uncertainty	6001.7	champion	3382.2	sale	9110.9	filing	2046.6	sale	3562.9	debt	5413.8
shareholder	5640.3	club	3358.0	cash	9011.1	r	1667.4	landscape	3300.6	interest	5367.5
offering	5572.8	sport	3333.3	percent	7572.0	revenue	1598.6	privacy	3252.5	interest rate	4740.8
press	5149.9	win	3143.6	result	7017.7	ebitda	1546.3	tech	3070.0	yen	4085.9
date	5087.3	day	3134.4	loss	6994.3	loan	1487.0	firm	2922.0	investor	3980.2
law	4821.1	round	3049.9	earning share	6769.8	rating	1308.1	amazon	2850.0	policy	3232.2
press release	4781.4	title	2689.1	loan	6524.4	chemical	1138.7	browser browser	2727.1	note	3199.9
risk uncertainty	4524.6	victory	2574.4	analyst	5047.8	lira	1093.2	use	2689.5	month	2798.3
stock	4485.0	point	2419.5	view	4788.3	operation	1080.7	payment	2563.2	device	2664.0
event	4399.6	test	2368.8	shareholder	4374.1	brief thomsonreuter	989.8	maker	2437.9	ecb	2557.2
condition	4200.6	inning	2276.6	growth	4314.7	zloty	988.2	consumer	2328.6	percent	2482.4
investor	4027.0	side	2267.4	cent	4305.7	rbi	959.4	logo	2309.1	datum	2476.2
act	3978.1	injury	2265.8	b	3739.8	filing bit	958.1	system	2205.6	bond yield	2424.3
form	3956.3	race	2238.1	thomson	3609.2	shareicon	781.2	giant	2132.0	view	2401.8

Supplemental


Word Cloud
Topic 0
technology
service
business
call
product
customer
industry
conference call
conference
president
growth
solution
officer
management
result
vice
relation
investor
client
today
board
director
team
experience
partner
website
development
acquisition
leadership
Topic 1
day
man
woman
director
film
reporter
event
work
thing
show
image
missile
member
award
life
role
movie
job
place
family
way
team
training
narration
board
part
book
program
history
wedding
athlete
photo
gold
nation
story
video
winter
reporter narration
Topic 2
price
production
oil
oil price
supply
energy
gas
barrel
producer
capacity
future
level
week
demand
pipeline
resource
tonne
crude
cent
aluminium
day
metal
month
output
fuel
plant
coal
bpd
export contract
brent
copper
inventory
mining
petroleum
commodity
exploration
analyst
refinery
project
trader
brent crude
mine
industry
lng
Topic 3
growth
percent
rate
month
price
inflation
car
economy
sale
tesla
increase
economist
quarter
forecast
datum
interest
auto
demand
cost
week
uber
spending
model
target
driver
consumer
rise
vehicle
gold
pace
interest rate
automaker
file photo
percent percent
level
deficit
photo
expectation
truck