Algorithmic Media Capture: Evidence from News Aggregators in Russia

Research Symposium, Spring 2021
Intensive Applied-Research Course
Advisor: Prof. Andrey Simonov
Savannah Wang

Background:

More diverse consumption sets of outlets

-> Theoretically, decrease the opportunities of media capture by the governments

-> Provide chances of allowing the government to capture the media at upstream and to censor new aggregators

Objective: Russia Online News Market

The government still allows independent news outlets to operate

Russian government does not restrict the work of news aggregators, including the market leader, Yandex News.

Anecdotal reports: Yandex omits some of the government-sensitive news such as protests or corruption

Research question:

Does Yandex censor government-sensitive news?

Overview of Data

25M+ documents with online news from Russia → 21697904 rows x 18 columns SFrame:

- Row: a news article
- Column: ['id', 'date', 'name', 'name_short', 'url', 'misses_URL', 'title_raw', 'text_raw', 'text_html', 'is_russian', 'title_tokenized', 'text_tokenized', 'propers', 'proper_counts', 'md5hash', 'is_good', duplicate_group', 'duplicate_count']

id	date	name	name_short
0	2014-11-04 13:06:00	Ридус (ridus.ru)	ridus.ru
1	2014-11-04 13:40:00	ТАСС	tass.ru
2	2014-11-04 14:10:00	ИА Regnum	regnum.ru
3	2014-11-04 12:47:00	Ридус (ridus.ru)	ridus.ru
4	2014-11-04 14:00:00	ТАСС	tass.ru
5	2014-11-04 12:47:00	Ридус (ridus.ru)	ridus.ru
6	2014-11-04 12:30:00	Радио Свобода (svoboda.org)	svoboda.org
7	2014-11-04 13:52:00	Собеседник (sobesednik.ru)	sobesednik.ru
8	2014-11-04 13:00:00	Сегодня.ua	Сегодня.ua
9	2014-11-04 14:00:39	РБК ТВ # Главные новости	РБК ТВ

title_tokenized	text_tokenized
[43.0, 1.0, 1.0, 12.0, 0.0...	[3, 1, 1, 11, 0, 16, 1, 0,...
[163, 1, 1, 12, 5, ,, 1, 0...	[119, 1, 1, 14, 0, 114, 1,...
[193.0, 1.0, 1.0, 12.0, 0....	[194, 1, 1, 1, 0, 165, 1, ...
[239.0, 0.0, 1.0, 0.0, 0.0...	[242, 0, 0, 0, 0, ,, 1, 0,...
[382, 1, 1, 12, 4, ,, 1, 0...	[,, 1, 0, 15, 0, 25, 1, 1,...
[520, 1, 1, 14, 0, 508, 1,...	[493, 1, 1, 12, 0, 79, 1, ...
[170, 1, 1, 12, 0, 307, 1,...	[538, 1, 1, 12, 0, 79, 1, ...
[#, 1, 1, 11, 0, 92, 1, 1,...	[98, 1, 0, 12, 0, 20, 1, 0...
[644.0, 1.0, 1.0, 12.0, 2....	[612, 1, 1, 12, 0, 618, 1,...
[171.0, 1.0, 1.0, 12.0, 0....	[#, 1, 1, 12, 5, ,, 1, 0, ...

url	misses_URL	title_raw
http://www.ridus.ru/news/1...	0	Австралиец прокатился на т...
http://tass.ru/ekonomika/1...	0	СМИ: парламент Венгрии одо...
http://www.regnum.ru/news/...	0	Нацбанк уже не считает кры...
http://www.ridus.ru/news/1...	0	Skype открыл регистрацию н...
http://tass.ru/mezhdunarod...	0	Порошенко: кабинет министр...
http://www.ridus.ru/news/1...	0	Закончились съемки фильма ...
http://www.svoboda.org/con...	0	Время русских маршей. Эфир...
http://sobesednik.ru/obshc...	0	В Москве на "Русский марш"...
http://www.segodnya.ua/cul...	0	Остап Ступка прокомментиро...
nan	0	День народного единства

propers	proper_counts
хэррисон уильямс оригинал_...	[1, 1, 1]
Bloomberg центральный_евро...	[1, 1, 1, 1, 1, 7, 1, 1, 1...
украина национальный_банк_...	[7, 3, 2, 5]
Skype_Translator оригинал_...	[3, 1, 1, 1, 1, 1, 1, 5, 1, 1]
львов_андрей бойко самопом...	[1, 1, 2, 1, 1, 1, 1, 1, 2...
дина абрамс джей звездный_...	[1, 1, 1, 4, 1, 1, 1, 1, 1...
кирилл_родионов путин укра...	[2, 1, 1, 1, 1, 3, 1, 1, 1...
путин день_народный_единст...	[1, 1, 1, 1, 1, 2, 1, 2, 1...
ирина ТСН богдан_ступка да...	[1, 1, 1, 2, 1, 2, 1]
москва день_народный_единс...	[1, 2, 1, 1, 1, 1, 1, 2]

text_raw	text_html	is_russian
Без компании он не остался...	Main mainpage 1b3efc7d29 ...	1
Принятые поправки позволяю...	***	1
Национальный банк Украина ...		1
Skype открыл регис...	Main mainpage 577753bb37 ...	1
"Он должен формироваться и...	***	1
Режиссер и продюсеры благо...	Main mainpage 9b4ec9be1a ...	1
Национализм и украинская в...	Владимир Рыжков Михаил Со...	1
4 ноября на территории Мос...	"Русский марш" *** Потехи...	1
Актер внес некоторые точно...		1
В: В Москве в эти минуты ...	В: В Москве в эти минуты с...	1

md5hash	is_good	duplicate_group	duplicate_count
0a4d37976e1c0b0c7f4ba376ee...	1	0	1
62d9e4fc65da6c13fe74a4b279...	1	0	1
9e9a40e4dbd1ab6298c51b9012...	1	0	1
23a6dfb9b89618ba1b0384d39c...	1	0	1
2690bdc2fe6e8524ec7d5c72d4...	1	0	1
0b0e1eefbfbea5103e4d1816d5...	1	0	1
d6cdcc8b3533da4e51fae1335e...	1	0	1
fc8901ce157ba3e5573d82b1ad...	1	0	1
32128dd5d6b5ca1bc93b0dfc79...	1	0	1
cc97caa7e1c8117f2fa07d9da2...	1	0	1

Data-Analytic Approach

Data Cleaning:

Proper_count (A sparse matrix with rows: a news article; columns: counts of proper nouns)

1) Subset of the full dataset with only columns, “propers”, “proper_counts”
2) Multiply “proper” by “proper_counts” and generating the sparse matrix

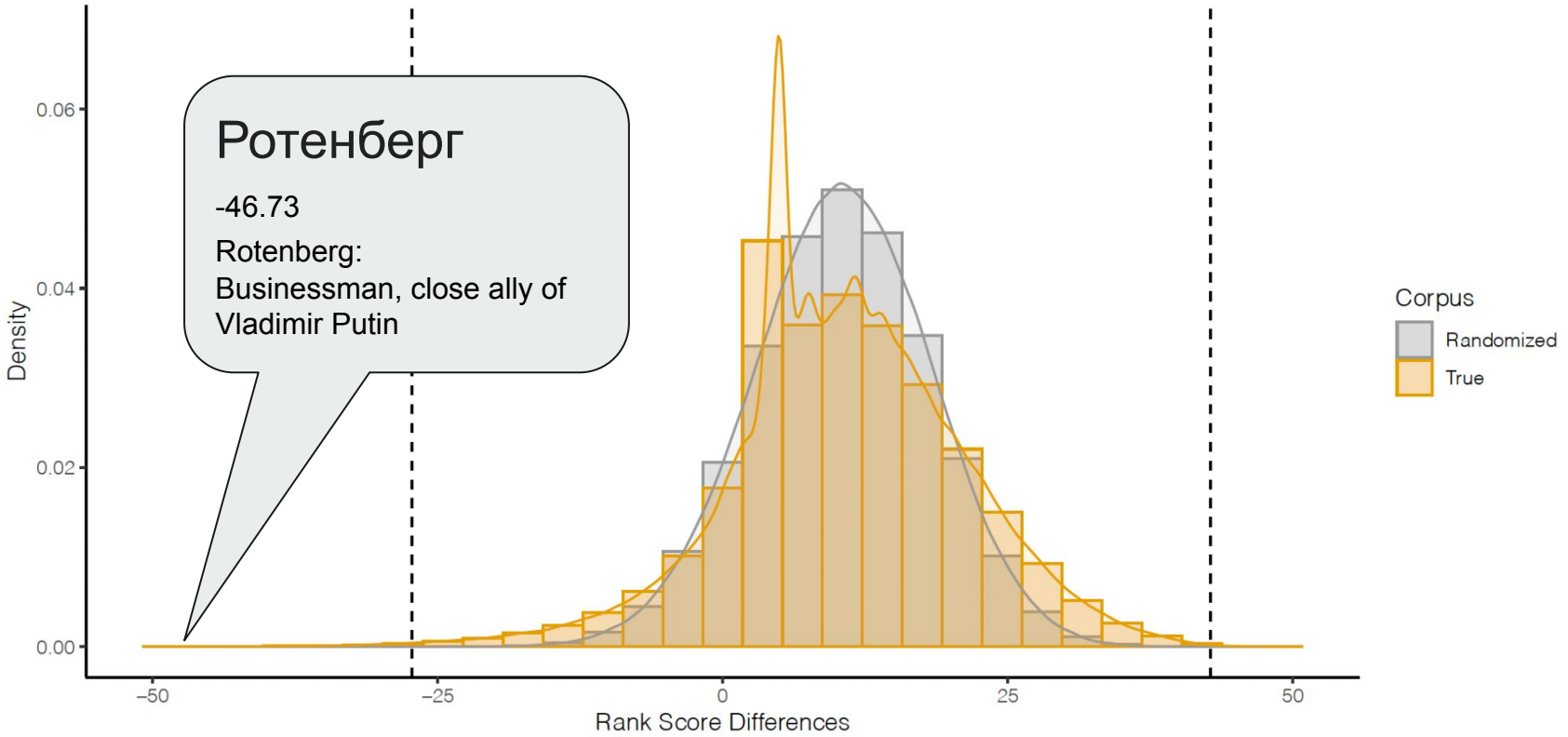
Outlet_count (A sparse matrix with rows: a news outlet; columns: counts of proper nouns)

1) Dictionary with key: outlet; and value: list of rows(articles) associating with this outlet
2) Building a sparse matrix in which each row is the sum of all rows with same outlets

Text Analysis:

Sensitivity detection among Government-Controlled outlets (GC) and Independent outlets(INP)

1) Compute share, S_{vj}, of each proper noun, v, by a outlet, j
2) Rank each v with increasing order among all outlets, j’s
3) Compute average rank of v for different outlet types, and the difference of GC and INP, $Rank_v(GC\text{-}INP)$
4) Permute word counts within the outlets and iterate step 1-4 1000 times to compute $RandRank_v(GC\text{-}INP)$


Ротенберг
-46.73
Rotenberg:
Businessman, close ally of
Vladimir Putin
Density
0.06
0.04
0.02
0.00
−50
−25
0
25
50
Rank Score Differences
Corpus
Randomized
True

Future Directions

- Detection of sensitivity for each document depending on if its title includes any sensitive words.
- Robust sensitivity analysis for each year.