Научный семинар по исследованиям цифровой экономики «Цифровизация и демография»

3 февраля 2021 года

(по материалам НИР ЭФ МГУ «ВОСПРОИЗВОДСТВО НАСЕЛЕНИЯ В ЦИФРОВОМ ОБЩЕСТВЕ»

Калабихина И.Е. (ЭФ МГУ, рук.)

Архангельский В.Н. (ЭФ МГУ)

Николаева У.Г. (ЭФ МГУ)

Колотуша А.В. (ЭФ МГУ, аспирант кафедры народонаселения)

Абдуселимова И.А. (ЭФ МГУ, магистр ЭП)

Банин Е.П. (МГТУ им. Н.Э. Баумана, НИЦ «Курчатовский институт», аспирант)

Клименко Г.А. (ЭФ МГУ, магистр ЭП)

Шамсутдинова В.Ш. (ЭФ МГУ, аспирантка кафедры народонаселения)

https://demography.econ.msu.ru/about/science/infographics/ – проект 2 (инфографика на сайте кафедры)

Использование данных социальных сетей и поисковых систем для оценки «демографической температуры» и демографических прогнозов

И.Е.Калабихина[1], И.А.Абдуселимова[1], В.Н.Архангельский[1], Е.П. Банин[2,3], Г.А. Клименко[1], А.В. Колотуша[1], У.Г. Николаева[1], В.Ш. Шамсутдинова[1]

[1]МГУ им. М.В. Ломоносова, экономический факультет, Москва, Россия

[2]МГТУ им. Н.Э. Баумана, Москва, Россия

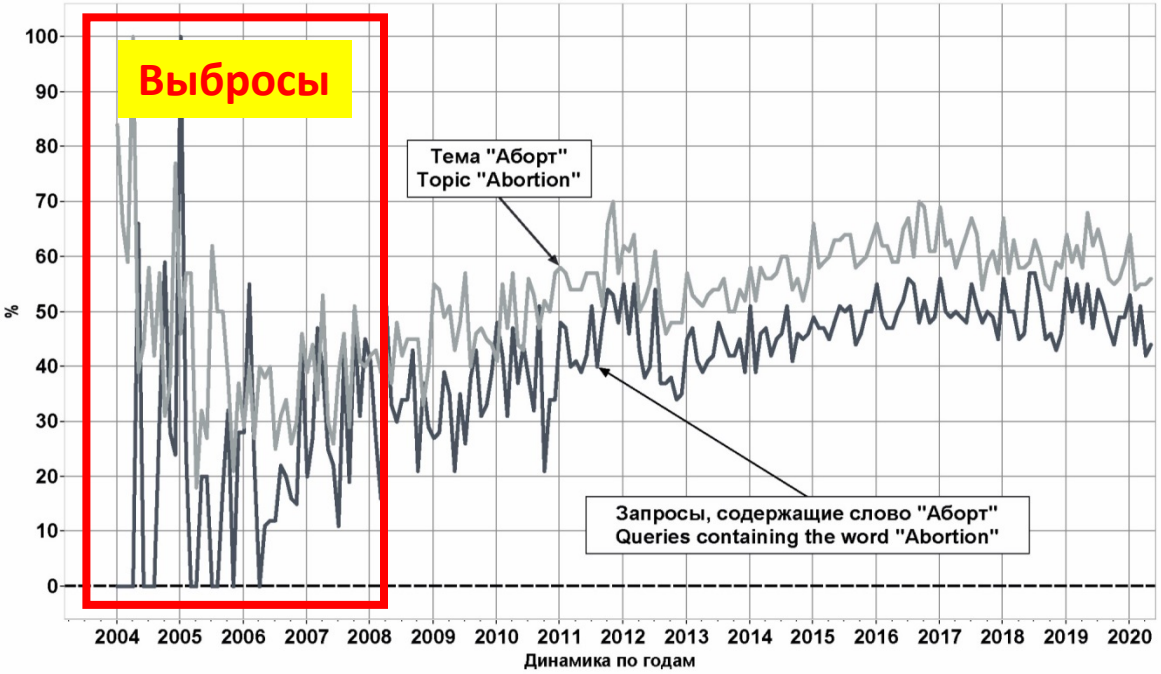
[3]НИЦ «Курчатовский институт», Москва, Россия

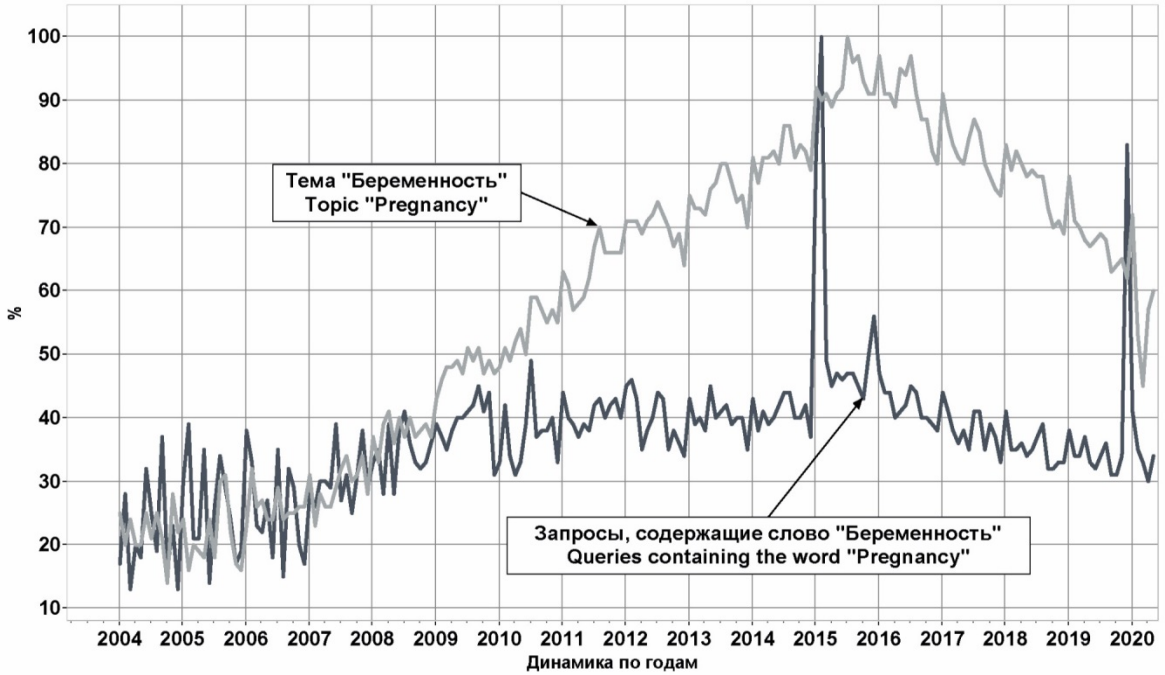
Калабихина И.Е., Банин Е.П., Абдуселимова И.А., Архангельский В.Н., Клименко Г.А., Колотуша А.В., Николаева У.Г., Шамсутдинова В.Ш. Краткосрочное прогнозирование демографических тенденций на основе данных Google trends // *Прикладная информатика*. 2020. Т. 15, № 6. С. 91-118.

Сюжет 1:

«Неклассическое» прогнозирование данных Росстата

Анализ текстовых запросов к Google Trends, 2009 - 2020





Разработка инструментария для оценки реакции населения на демографические инициативы

(поиск прокси-переменных из цифровых следов для прогнозирования)

«Беременность», «Аборт», «ЗАГС», «Роддом», «Обручальное кольцо», «Документы на развод», «Развод»

Индекс популярности

$$g_{i,r,t} = \frac{q_{i,r,t}}{Q_{r,t}} c_i$$

i – запрос

r – регион

t – временной период

c – масштаб

Выборка данных из Google Trends за период анализ 2009 - 2020)

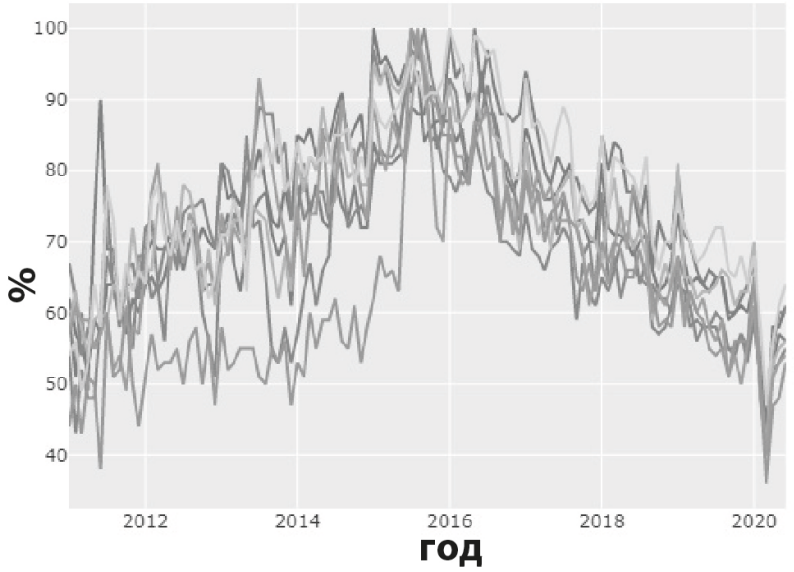
- Москва
- Московская область
- Краснодарский край
- Санкт-Петербург
- Ростовская область
- Свердловская область
- Республика Татарстан
- Республика Башкортостан
- Россия в целом

Запросы* (сокращение):

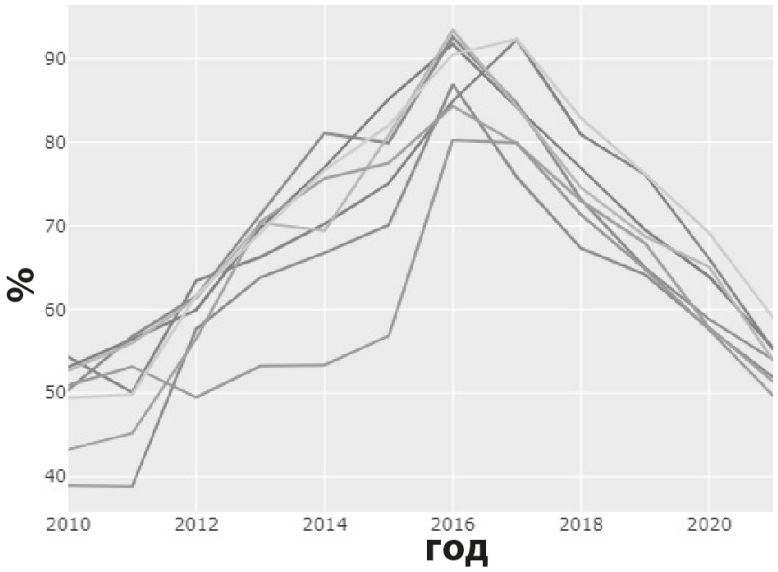
- «Беременность»
- «Аборт»
- «ЗАГС»
- «Роддом»
- «Обручальное кольцо»
- «Документы на развод»

* Параметр лежит на отрезке [0, 1]

Данные по месяцам



Данные по годам



Москва

Дагестан

Дополнительные части (анализ текстовых запросов к Google Trends, 2009 - 2020)

По годам:

- суммарный коэффициент рождаемости (**tfr**)
- аборты на 100 рождений (**A100b**)
- аборты на 1000 женщин (**A1000w**),
- количество браков на 1000 населения (**Mar**)
- количество разводов на 1000 населения (**Div**)

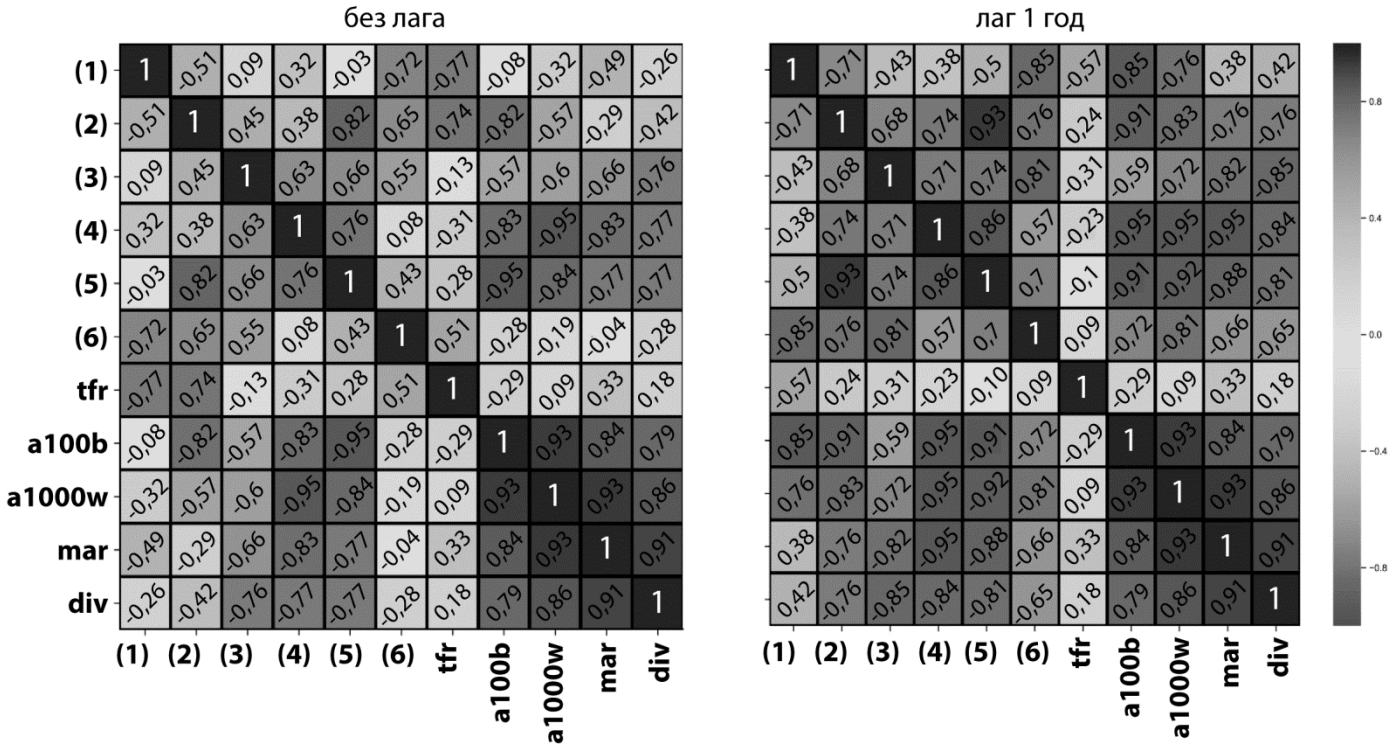
По месяцам:

- число родившихся (**birth_cnt**),
- число браков (**mar_cnt**),
- число разводов (**div_cnt**)

* Параметр лежит на отрезке [0, 1]

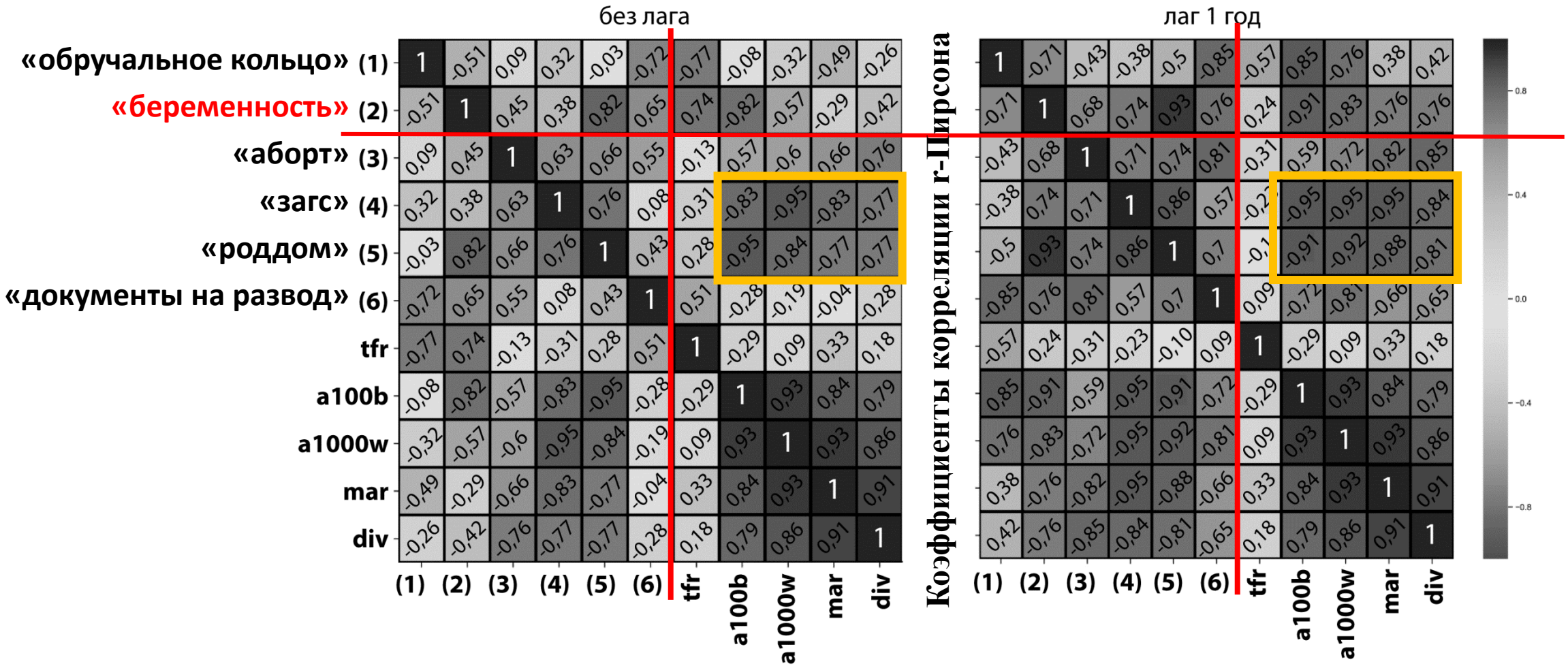
(1) – «обручальное кольцо», (2) – «беременность», (3) – «аборт», (4) – «загс», (5) – «роддом», (6) – «документы на развод»

Коэффициенты корреляции r-Пирсона



Дополнительные части (анализ текстовых запросов к Google Trends, 2009 - 2020)

(1) – «обручальное кольцо», (2) – «беременность», (3) – «аборт», (4) – «загс», (5) – «роддом», (6) – «документы на развод»



модель

ARIMA (p, d, q)(P, D, Q, s)

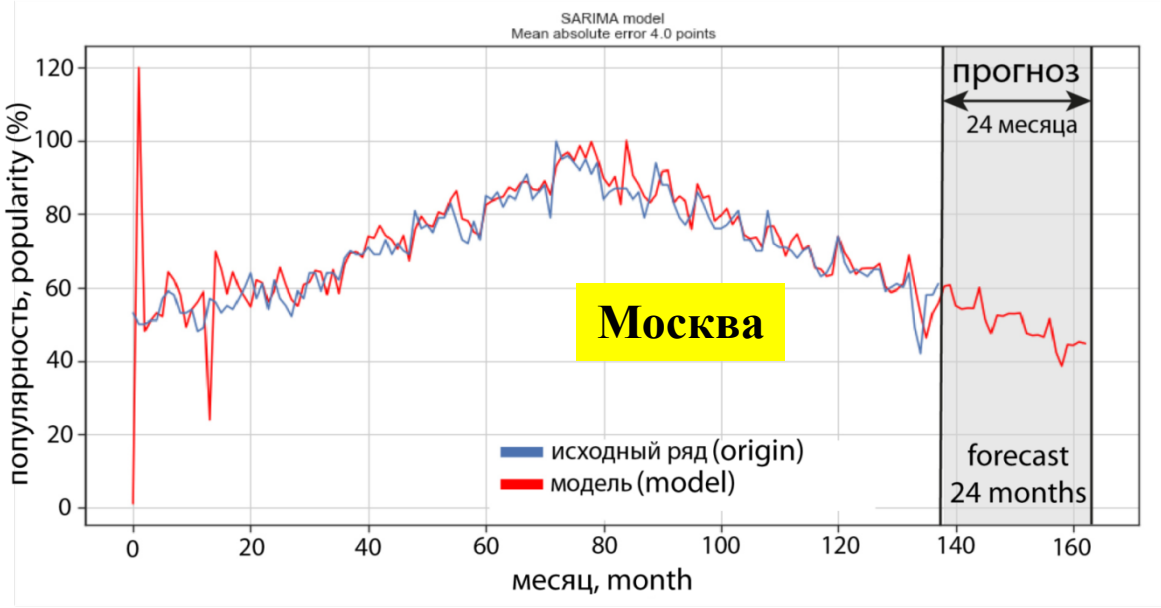
p – порядок модели AR(p)
d – порядок интегрирования исх. данных
q – порядок модели MA(q)
P – порядок сезонной составляющей SAR(P)
D – порядок интегрирования сезонной составляющей
Q – порядок сезонной составляющей SMA(Q)
s – размерность сезонности (месяцы)

ARIMA (0, 2, 2)(2, 1, 1, 12) или ARIMA (0, 2, 2)(2, 0, 1, 12)

Регион	Население (2019 г.)	r-Pearson	
		без лага	Лаг в 1 год
Москва	12646679	0,22	0,62
Москвоская обл.	7645255	0,94	0,59
Краснодарский край	5661848	0,91	0,73
Санкт-Петербург	5390977	0,89	0,85
Свердловская обл.	4315699	0,82	0,4
Ростовская обл.	4202320	0,94	0,53
Респ. Башкортостан	4044578	0,17	-0,41
Респ. Татарстан	3900758	0,6	0,078

→ Тема зашумлена

→ Мало внимания



Результаты по первому сюжету

1. Разработан алгоритм прогнозирования краткосрочных данных статистики (коэффициент рождаемости, брачность, количество абортов) на основе запросов к Google для регионов и России в целом.
2. Продемонстрирован алгоритм поиска прокси-переменных к данным статистики, которые могут быть использованы в моделях множественной регрессии, для заполнения «пропусков» в демографических данных и т.п.
3. Модель работает тем точнее, чем больший объем запросов по ключевому слову (**даже текущая активность населения регионов в некоторых случаях недостаточно хорошо отражает демографические процессы**).
4. Модель учитывает сезонные тенденции в данных запросов к Google (**но не для всех регионов она характерна**).
5. Ошибка прогноза коэффициента рождаемости на уровне страны 1,14-2,11% **(на уровне региона может быть более 5 %)**

Ищем прокси к демографическим данным По запросам к Google	Анализируем корреляции запросов с данными Росстата	Прогнозируем динамику запросов (уже по месяцам)	Оперативный прогноз данных Росстата

Сюжет 2:

«Демографическая температура» пронаталистских и интинаталистских сообществ ВКонтакте

Сбор и обработка данных

Пронаталистские группы

Групп – 314
Пользователей – 9 млн.
Кол-во комментариев – 112 900

Антинаталистские группы (Childfree-сообщества)

Групп – 9
Пользователей – ок. 100 тысяч
Кол-во комментариев – 670 000

	Ссылка на автора	Пол автора	Ссылка на комментарий	Дата и время	Текст комментария	Число лайков к комментарию
71545	https://vk.com/club86333616	NaN	https://vk.com/wall-86333616_6944?reply=6945	2015-09-06 21:25:00	Привет! я ждала твоего звонка. Очень надеялась...	2268.0
13882	https://vk.com/id214886699	Ж	https://vk.com/wall-52388302_289756?reply=289763	2019-06-26 22:23:00	таких учителей на пушечный выстрел нельзя подп...	1939.0
97763	https://vk.com/id96334626	Ж	https://vk.com/wall-170234932_468228?reply=468243	2020-04-23 09:34:00	Ни один футболист и ни одна певица не смогли б...	1077.0
107103	https://vk.com/id166418625	Ж	https://vk.com/wall-52388302_606195?reply=606213	2020-09-13 15:34:00	Не знаю, что происходит с миром, с людьми. Ка...	959.0
13904	https://vk.com/id273463315	Ж	https://vk.com/wall-52388302_289756?reply=289778	2019-06-26 22:50:00	Ну что молодец Паша!!!с малых лет был и остался...	812.0

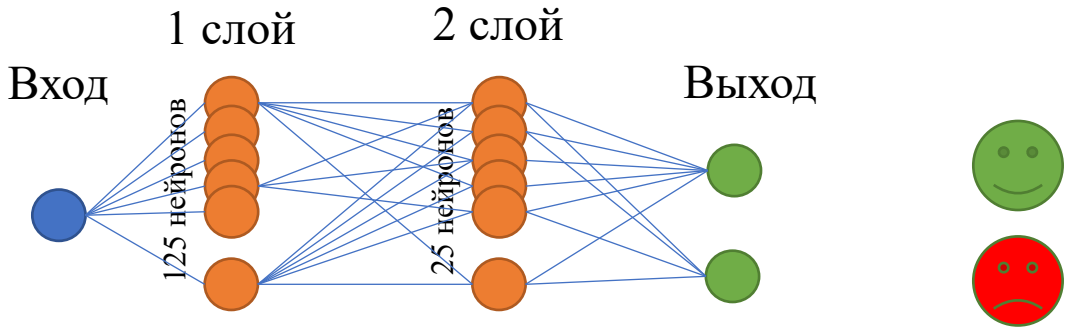
После обработки

	Текст комментария	Число лайков к комментарию	preprocessed	text_prep	text_stem	text_sw
71545	Привет! я ждала твоего звонка. Очень надеялась...	2268.0	Привет я ждала твоего звонка Очень надеялась у...	привет я ждала твоего звонка очень надеялась у...	привет ждал тво звонк очен надея услыша тво го...	привет ждала твоего звонка очень надеялась усл...
13882	таких учителей на пушечный выстрел нельзя подп...	1939.0	таких учителей на пушечный выстрел нельзя подп...	таких учителей на пушечный выстрел нельзя подп...	так учител пушечн выстрел подпуска дет	таких учителей пушечный выстрел подпускать детям
97763	Ни один футболист и ни одна певица не смогли б...	1077.0	Ни один футболист и ни одна певица не смогли б...	ни один футболист и ни одна певица не смогли б...	футболист одн певиц смогл ход так костюм спаси...	футболист одна певица смогли ходить таких кост...
107103	Не знаю, что происходит с миром, с людьми. Ка...	959.0	Не знаю что происходит с миром с людьми Как не...	не знаю что происходит с миром с людьми как не...	зна происход мир людьм помоч ребенк расплака р...	знаю происходит миром людьми помочь ребенку ра...
13904	Ну что молодец Паша!!!с малых лет был и остался...	812.0	Ну что молодец Паша с малых лет был и остался ...	ну что молодец паша с малых лет был и остался ...	молодец паш мал лет оста человек	молодец паша малых лет остался человеком

Удаление пунктуации | **Нижний регистр** | **Стеммизация** | **Стоп-слова**

Определение тональности высказываний

```
net = tflearn.input_data([None, VOCAB_SIZE])
net = tflearn.fully_connected(net, 125, activation='ReLU')
net = tflearn.fully_connected(net, 25, activation='ReLU')
net = tflearn.fully_connected(net, 2, activation='softmax')
```



Словарь* – 5000 слов

NLTK - RussianStemmer

Векторизация – TweetTokenizer

Тестовая выборка – 30 % от набора

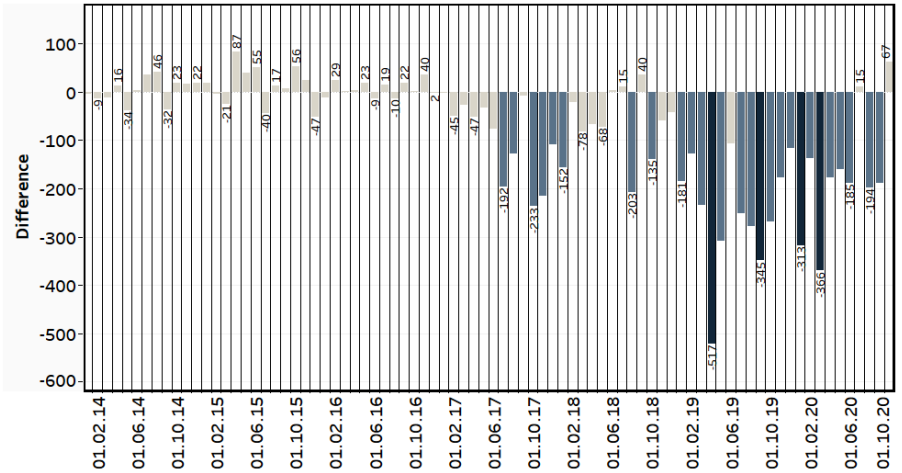
Точность на тестовой выборке
71 %

Негативные	Позитивные
Скоро увижу твои зеленые глаза в последний раз(((	Предвкушение потрясающего уикенда - Праздник света! Гулять ночи напролет и любоваться огнями
думаю, на моем веку почта России не поднимется со дна	я раааааааад, ведь сегодня впереди вся ночь, сериалы, фильмы, книга и кофе:) идеальная ночка
Угадайте чья школа учится сегодня? Причем единственная в области	Не иду завтра в школу,первый раз мама так настойчиво не отпускала меня,спасиибки)

*Обучение на готовом наборе твиттов от Вячеслава Ковалевского https://github.com/b0noI/ml-lessons/tree/master/sentiments_rus

Разность между "позитивными" и "негативными" комментариями по месяцам

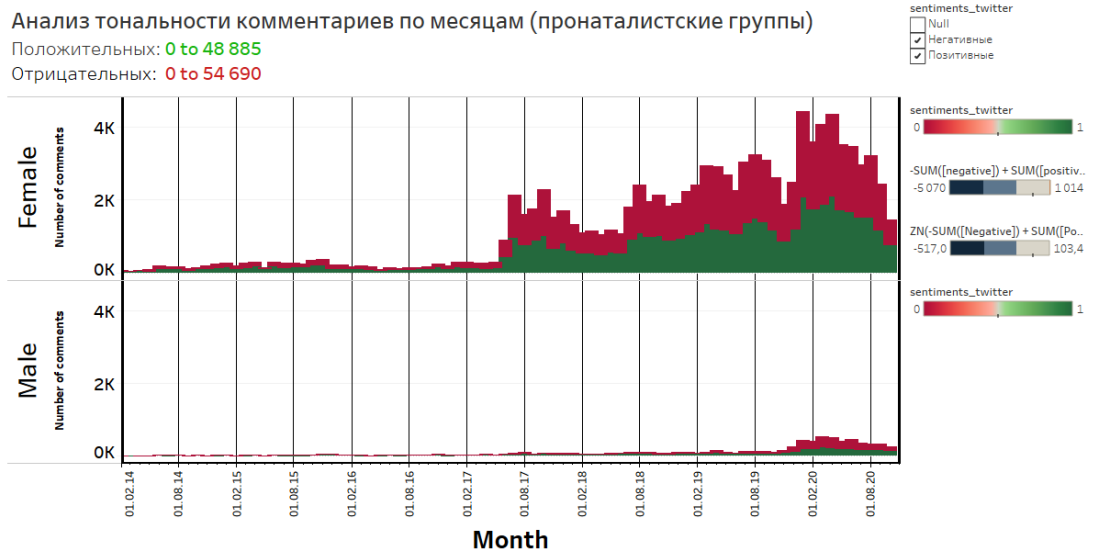
Источник: vk_posts_stem_lemm (Пронаталистские группы)



Анализ тональности комментариев по месяцам (пронаталистские группы)

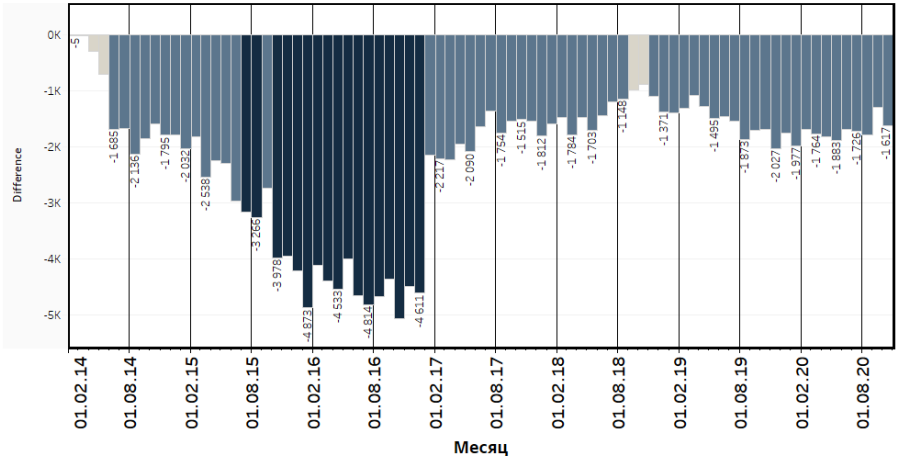
Положительных: 0 to 48 885

Отрицательных: 0 to 54 690



Анализ тональности комментариев по месяцам (антинаталистские группы, полная выборка)

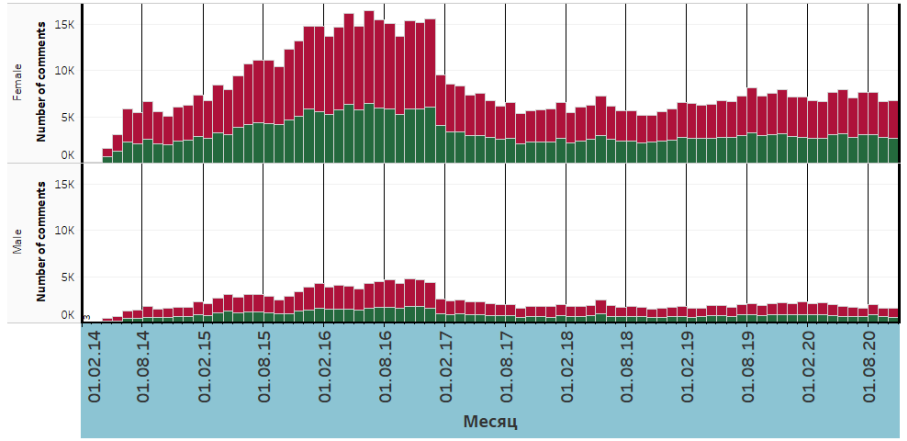
'-' - преобладание отрицательных тональностей



Анализ тональности комментариев по месяцам (антинаталистские группы, полная выборка)

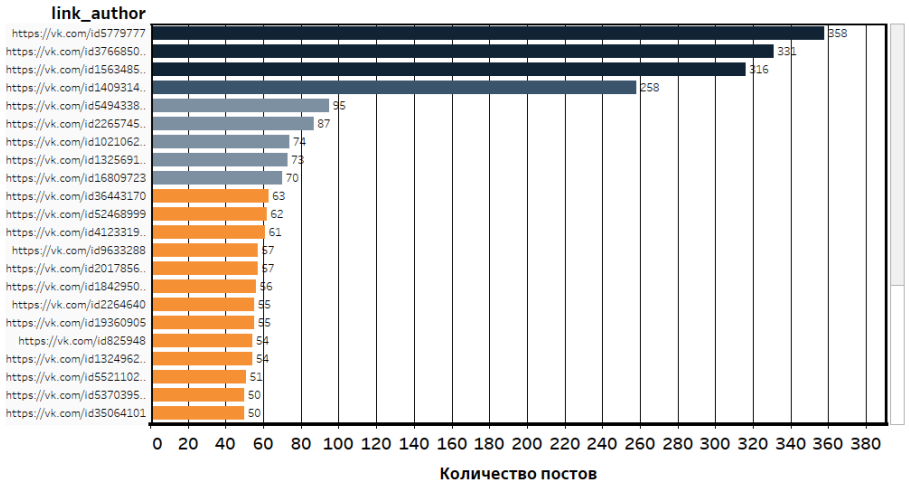
Положительных: 0 to 267 002

Отрицательных: 0 to 403 938



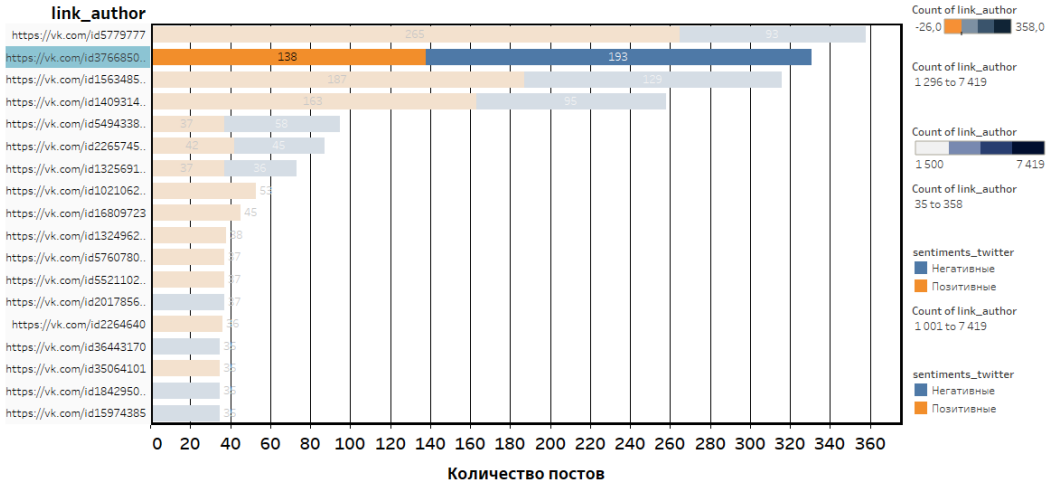
Наиболее активные участники (пронаталистские группы, полная выборка)

Источник: vk_posts_stem_lemm (Пронаталистские группы)



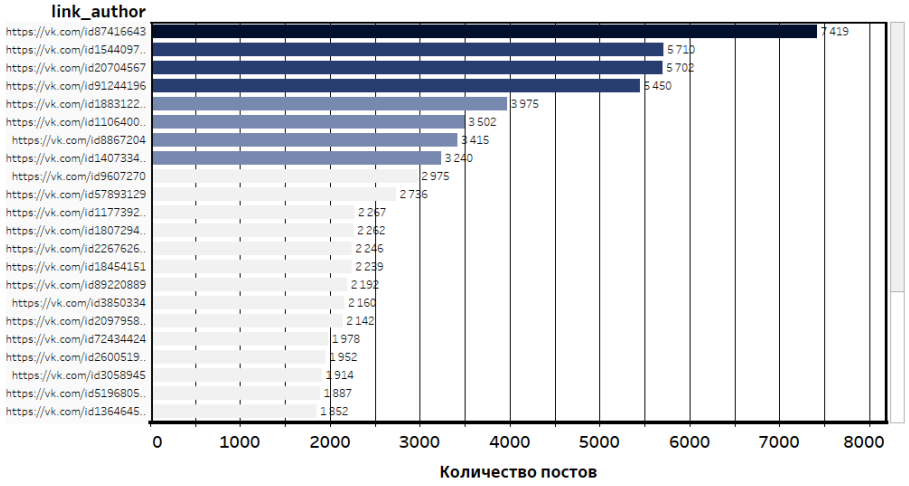
Наиболее активные участники (пронаталистские группы, полная выборка)

Источник: vk_posts_stem_lemm (Пронаталистские группы)



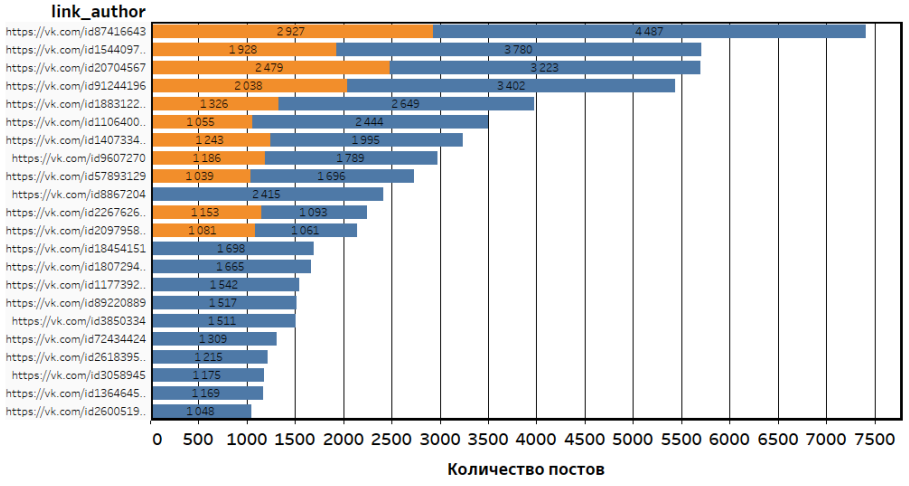
Наиболее активные участники (антинаталистские группы, полная выборка)

Источник: Antinata_vk_sentiments_preparing (Антинаталистские группы)



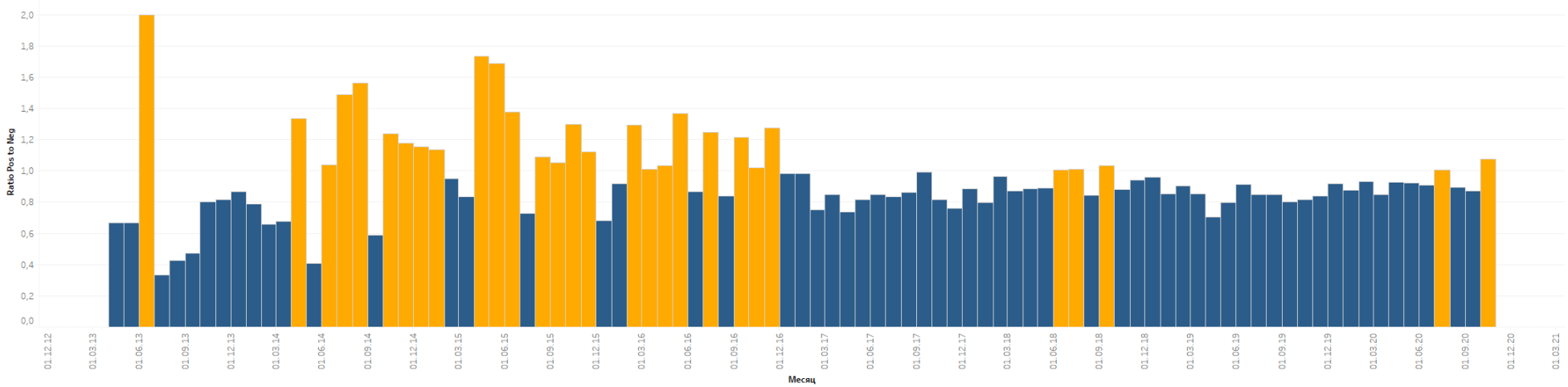
Наиболее активные участники (антинаталистские группы, полная выборка)

Источник: Antinata_vk_sentiments_preparing (Антинаталистские группы)



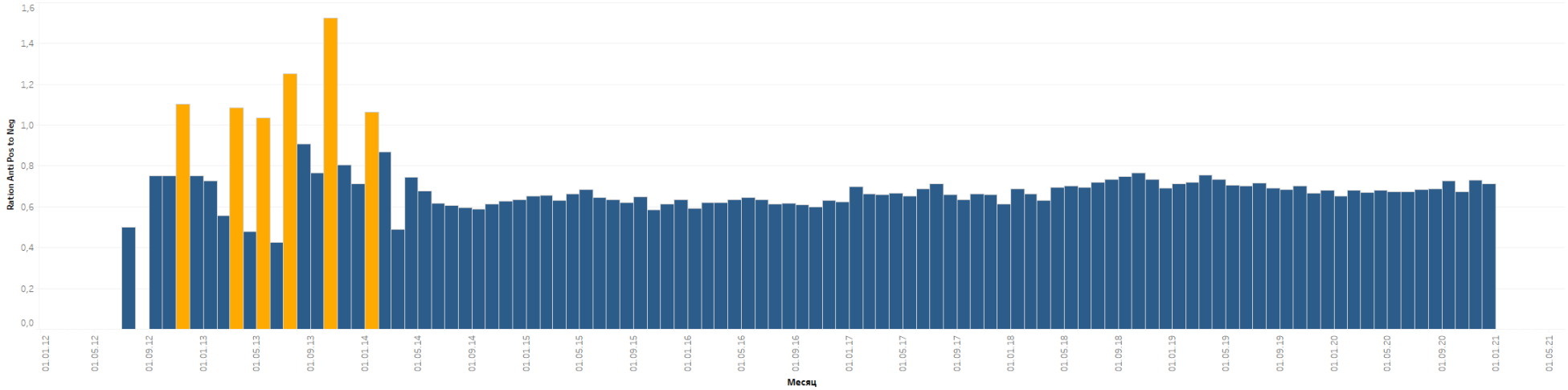
Отношение положительных комментариев к отрицательным (пронаталисты)

по месяцам



Отношение положительных комментариев к отрицательным (антиналисты)

по месяцам



Результаты по второму сюжету

1. Разработан алгоритм оценки «демографической температуры», на примере социальной сети ВКонтакте;
2. Проведены оценки различий между двумя группами пользователей (про- и антинаталистские группы);
3. Многочисленные пронаталистские группы в среднем малоактивны по сравнению с пользователями пронаталистских групп (314 групп и 112 000 комментариев против 9 групп с 670 000 комментарий,);
4. 68 356 уникальных авторов (пронаталисты) и 46 528 уникальных авторов (антинаталисты)

	Пронаталисты	Антинаталисты (Child free)
	Отношение +/- 0,894	Отношение +/- 0,655
	Уникальные авторы 68 353	Уникальные авторы 46 528
	+ на 1000 авторов 466,9	+ на 1000 авторов 395,3
	- на 1000 авторов 522,2	- на 1000 авторов 603,0

Использование данных социальных сетей и поисковых систем для оценки «демографической температуры» и демографических прогнозов

И.Е.Калабихина[1], И.А.Абдуселимова[1], В.Н.Архангельский[1], Е.П. Банин[2], Г.А. Клименко[1], А.В. Колотуша[1], У.Г. Николаева[1], В.Ш. Шамсутдинова[1]*

[1]МГУ им. М.В. Ломоносова, экономический факультет, Москва, Россия

[2]МГТУ им. Н.Э. Баумана, Москва, Россия, evg.banin@gmail.com

MSU Demography

Tableau

Zenodo.org

MSU Demography

GitHub