Реакция на демографическую политику пользователей социальной сети «Вконтакте»

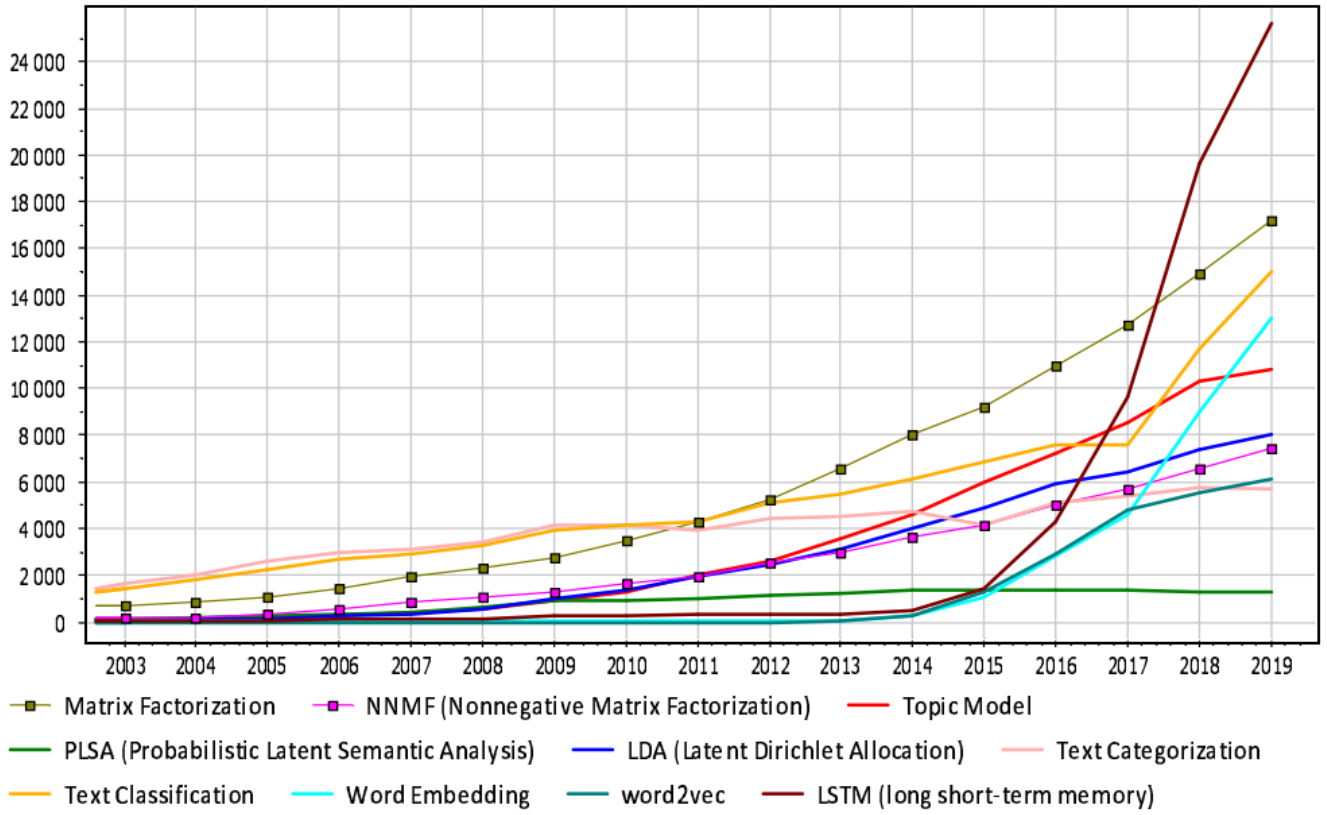
***И.Е.Калабихина*[1], *И.А.Абдуселимова*[1], *В.Н.Архангельский*[1], *Е.П. Банин*[2*], *Г.А. Клименко*[1], *А.В. Колотуша*[1], *У.Г. Николаева*[1], *В.Ш. Шамсутдинова*[1]**

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

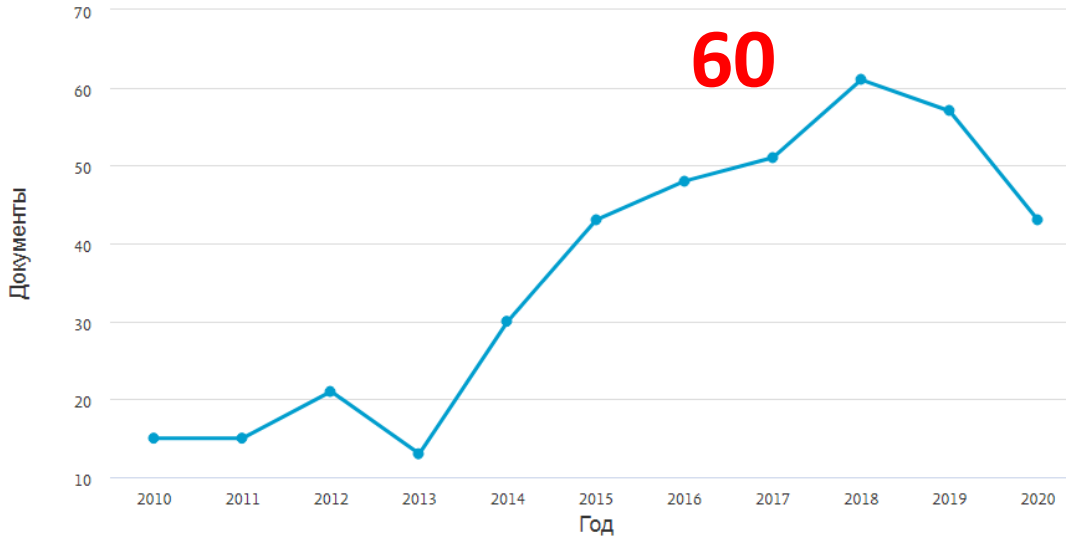
[2]МГТУ им. Н.Э. Баумана, Москва, Россия

Big data и демография

Динамика цитирования Google Scholar*



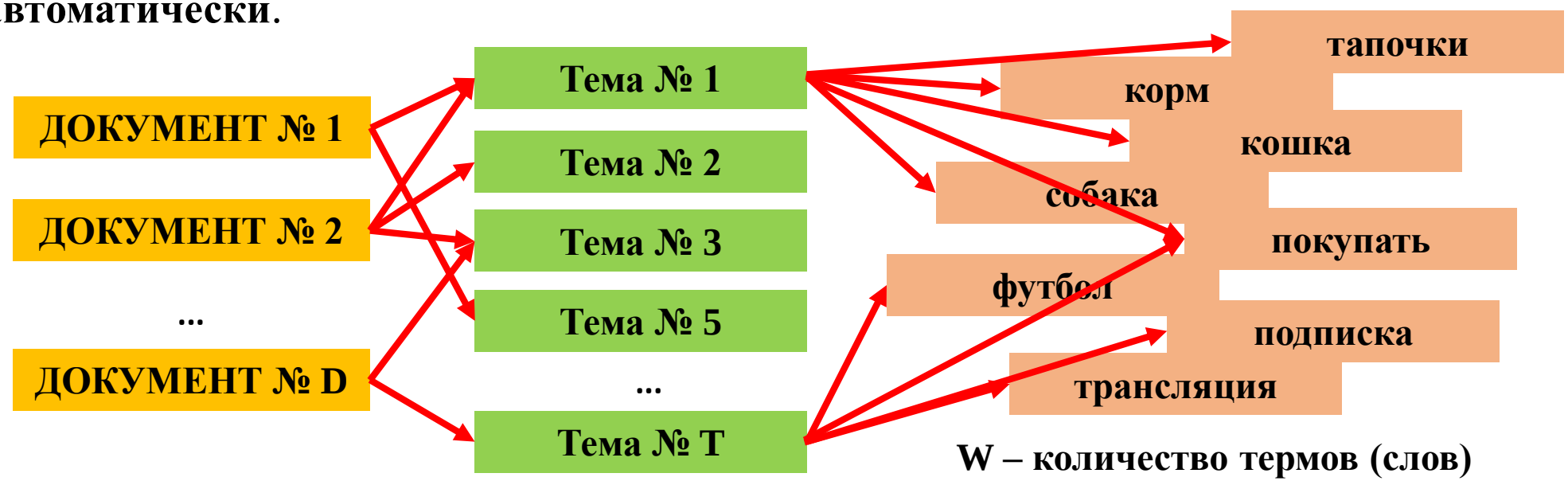
Динамика публикаций по теме «big data» в демографии



** Иллюстрация из лекции «Вероятностные тематические модели» К.В. Воронцова, МФТИ*
http://www.machinelearning.ru/wiki/images/6/67/Voron20ptm-intro.pdf

Цель: определение эмоциональной оценки пользователей соцсетей по темам, связанным с рождаемостью родительством, брачностью и т.п.
Тематическая модель – модель коллекции тестовых документов, которая определяет, к каким темам относится каждый документ

На вход алгоритма тематического моделирования передаётся **коллекция текстовых документов**. На выходе – **числовой вектор, составленный из оценок степени принадлежности данного документа каждой из тем**. Размерность выходного вектора соответствующая числу тем, может быть **задана** или **вычислена автоматически**.



$(d_i, w_i, t_i)_{i=1}^{n} \sim p(d, w, t)$

$d_i, w_i \rightarrow$ наблюдаемые
$t_i \rightarrow$ скрытые

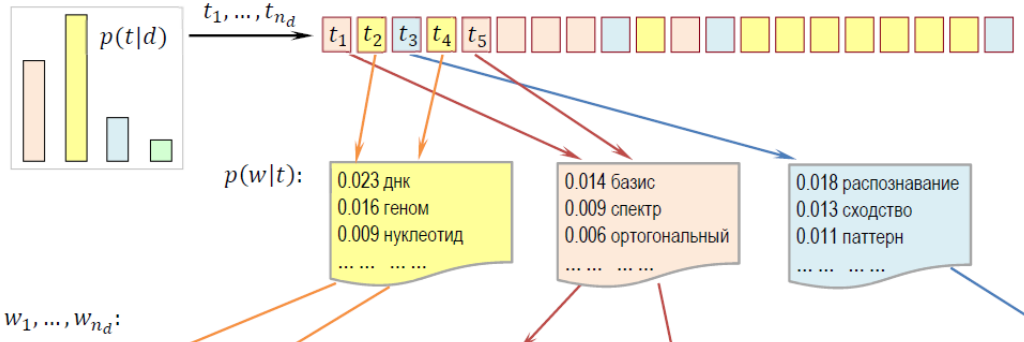
Гипотеза условной независимости

$p(w \mid d, t) = p(w \mid t)$

Порядок слов в документе **не важен** и порядок документов **не важен**

Прямая и обратная задачи тематического моделирования

$$p(w \mid t) = \sum_{t \in T} p(w \mid t) p(t \mid d)$$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Иллюстрация из лекции «Вероятностные тематические модели» К.В. Воронцова, МФТИ
http://www.machinelearning.ru/wiki/images/6/67/Voron20ptm-intro.pdf

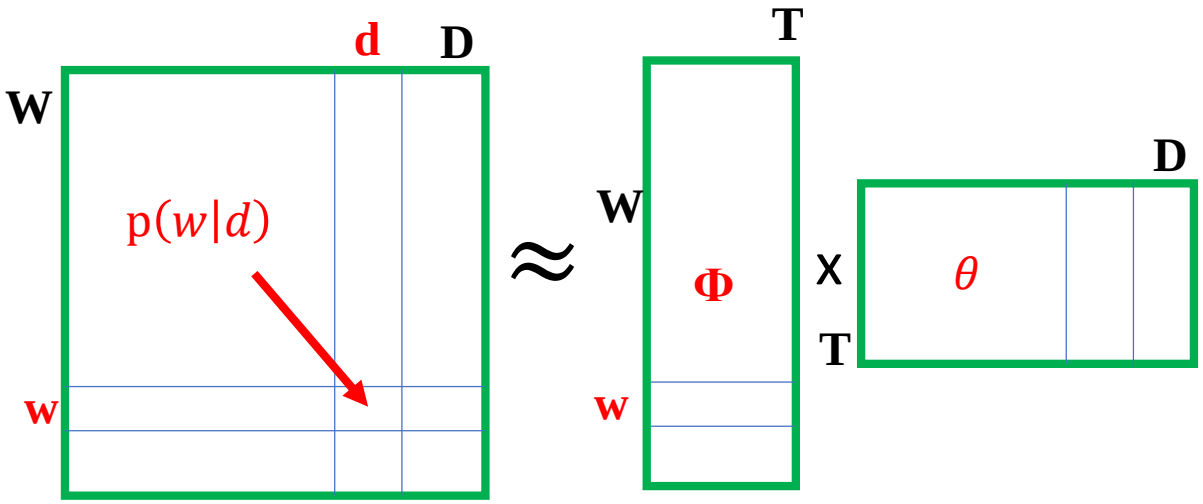
Дано: коллекция текстовых документов

n_{dw}- частоты термов в документах $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

Найти: параметры тематической модели

$$\mathrm{p}(w|d) = \sum_{t \in T} \varphi_{wt} \theta_{td}$$

$\varphi_{wt} = p(w|t)$- вероятность термов w в каждой t
$\theta_{td} = p(t|d)$- вероятность темы t в каждом d



Сбор и обработка данных

URL	Название	Подписчики
https://vk.com/club52388302	ХОРОШИЕ РОДИТЕЛИ	1482303
https://vk.com/club34677924	Беременность	1339737
https://vk.com/club170234932	Женское Здоровье	1053617
https://vk.com/club20199180	РАЗВИВАЙКА РОДИТЕЛИ И ДЕТИ В ИНТЕРНЕТЕ	794730
https://vk.com/club14395935	Pampers: Мамочки ВКонтакте	428464
https://vk.com/club69716165	МАМА: Развитие, Семья, Дети	213562
https://vk.com/club29746763	Семья Родители Дети СПб	208624
https://vk.com/club78865067	МАМА Дети Семья	202603
https://vk.com/club61700163	Полезная страничка! Здоровье | Красота | Спорт	178790
https://vk.com/club20622108	Я - МАМА: беременность, дети, семья, материнство	147412

Выборка

До обработки:
Всего групп – 813
Пользователей – 22 млн.

После:
Всего групп – 314
Пользователей – 9 млн.
Кол-во комментариев – 112 900

До обработки

	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 осталс...	812.0

После обработки

	Текст комментария	Число лайков к комментарию	preproccessed	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 осталс...	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**

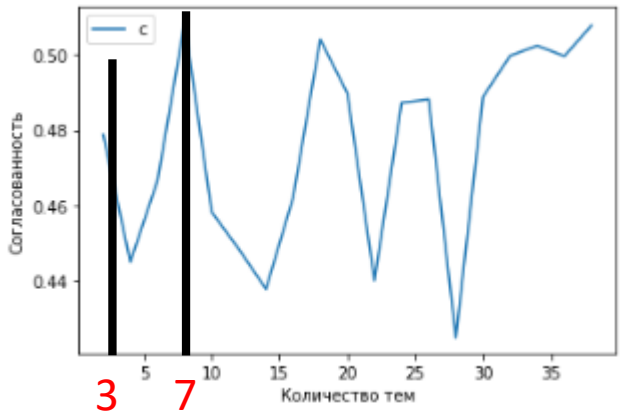
Результаты (3 темы, комментарии)

[(0, '0.011*"ребенк" + 0.011*"муж" + 0.010*"мам" + 0.010*"родител" + 0.008*"сын" + 0.008*"женщин" + 0.008*"мужчин" + 0.007*"жизн" + 0.007*"бог" + 0.007*"да_бог"'),

(1, '0.012*"мам" + 0.008*"женщин" + 0.008*"котор" + 0.007*"говор" + 0.007*"родител" + 0.007*"люд" + 0.006*"прост" + 0.006*"люб" + 0.006*"все" + 0.006*"очен"'),

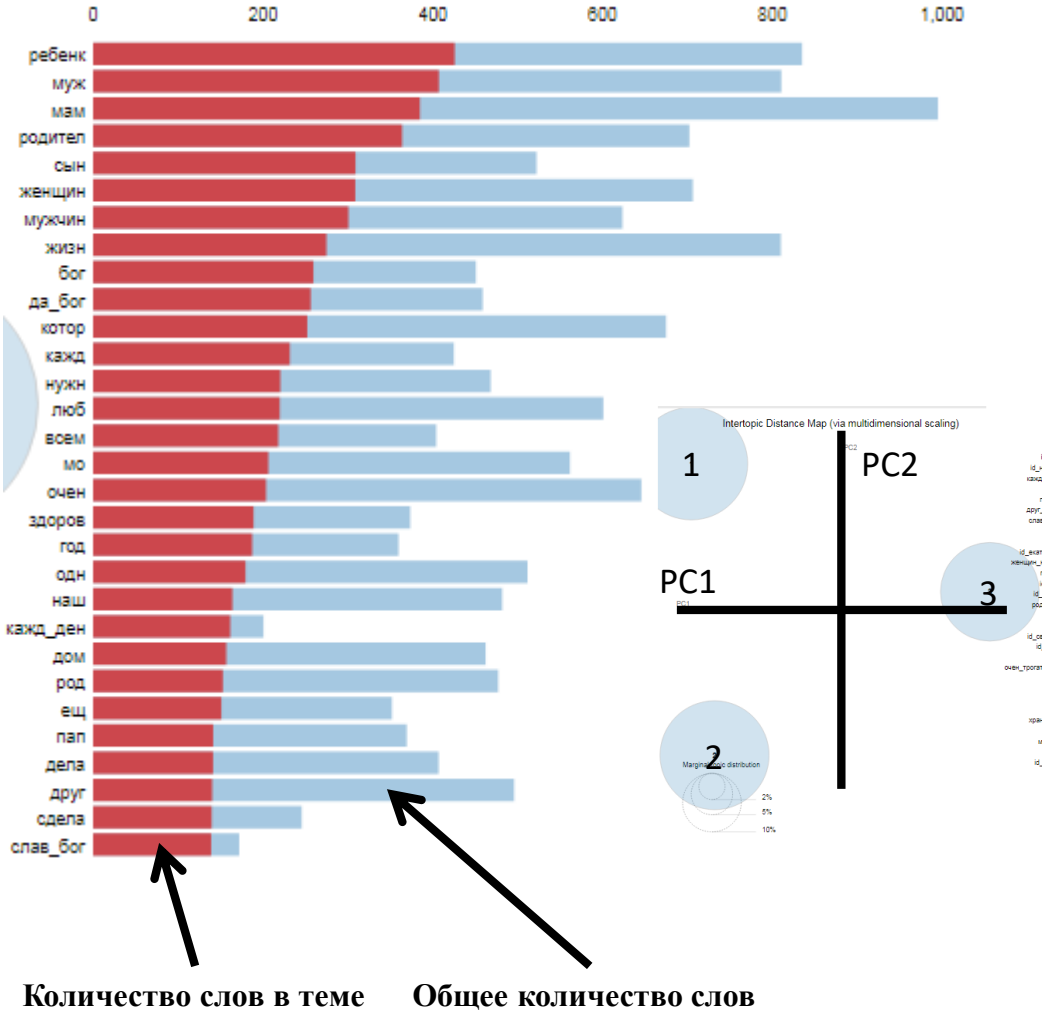
(2, '0.011*"жизн" + 0.011*"прост" + 0.009*"муж" + 0.008*"сем" + 0.008*"очен" + 0.007*"друг" + 0.007*"мам" + 0.007*"ребенк" + 0.006*"друг_друг" + 0.006*"дом"')]

Метрика согласованности (coherence)



Оптимальная разбивка на темы

Количество слов



Результаты (8 тем, комментарии)

[(0, '0.010*"мам" + 0.010*"очен" + 0.009*"одн" + 0.008*"родител" + 0.008*"мужчин" + 0.008*"люд" + 0.007*"жизн" + 0.007*"ребенк" + 0.007*"дела" + 0.006*"женщин"'),

(1, '0.010*"женщин" + 0.009*"муж" + 0.009*"ребенк" + 0.009*"id_светла" + 0.008*"мужчин" + 0.008*"люб" + 0.007*"очен" + 0.007*"мам" + 0.007*"дом" + 0.007*"сем"'),

(2, '0.014*"жизн" + 0.013*"муж" + 0.010*"мо" + 0.010*"мам" + 0.008*"прост" + 0.007*"род" + 0.007*"люб" + 0.006*"ребенк" + 0.006*"все" + 0.006*"сем"'),

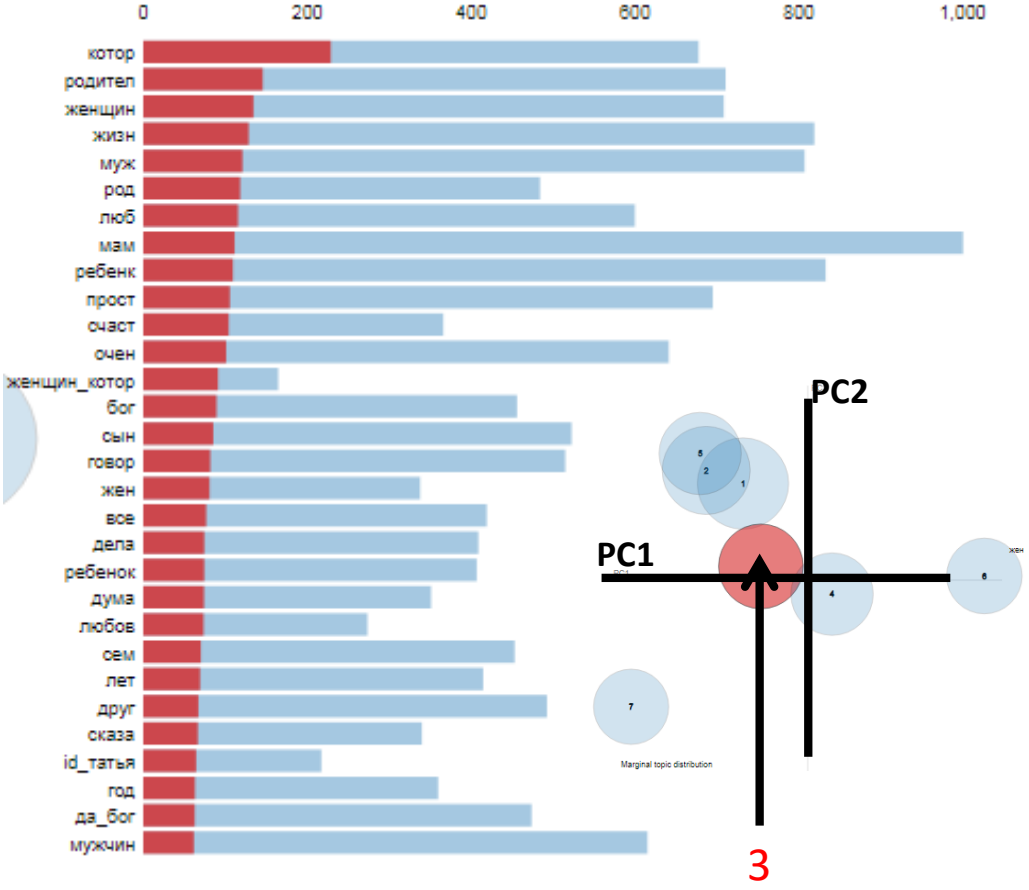
(3, '0.013*"мам" + 0.011*"наш" + 0.010*"прост" + 0.010*"жизн" + 0.009*"муж" + 0.009*"родител" + 0.009*"пап" + 0.008*"ребенк" + 0.008*"id_ел" + 0.007*"сын"'),

(4, '0.013*"ребенк" + 0.013*"женщин" + 0.011*"мужчин" + 0.008*"мам" + 0.008*"очен" + 0.008*"друг" + 0.007*"друг_друг" + 0.006*"человек" + 0.006*"всем" + 0.006*"нужн"'),

(5, '0.020*"да_бог" + 0.012*"мам" + 0.012*"говор" + 0.011*"бог" + 0.008*"одн" + 0.008*"дом" 0.008*"здоров" + 0.008*"родител" + 0.007*"да" + 0.007*"род"'),

(6, '0.015*"котор" + 0.010*"родител" + 0.009*"женщин" + 0.008*"жизн" + 0.008*"муж" + 0.008*"род" + 0.008*"люб" + 0.007*"мам" + 0.007*"ребенк" + 0.007*"прост"')]

Количество слов



Результаты (4 темы, посты)

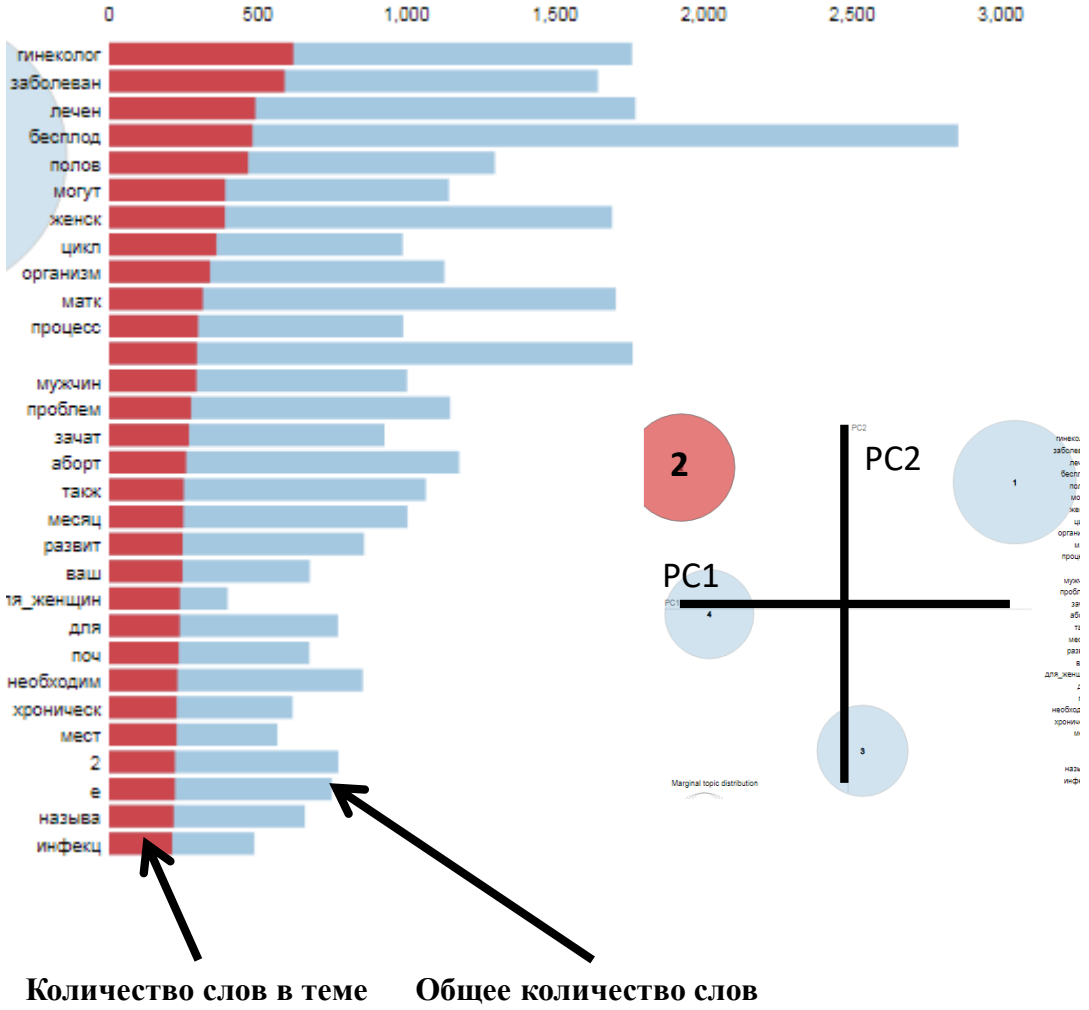
[(0, '0.008*"бесплод" + 0.003*"лечен" + 0.003*"матк" + 0.003*"заболеван" + 0.003*"женск" + 0.002*"гинеколог" + 0.002*"яичник" + 0.002*"полов" + 0.002*"явля" + 0.002*"проблем"'),

(1, '0.003*"гинеколог" + 0.003*"заболеван" + 0.003*"лечен" + 0.003*"бесплод" + 0.003*"полов" + 0.002*"могут" + 0.002*"женск" + 0.002*"цикл" + 0.002*"организм" + 0.002*"матк"'),

(2, '0.003*"бесплод" + 0.002*"лечен" + 0.002*"женск" + 0.002*" " + 0.002*"плачет" + 0.002*"матк" + 0.002*"заболеван" + 0.002*"гинеколог" + 0.002*"такж" + 0.002*"организм"'),

(3, '0.006*" " + 0.004*"бесплод" + 0.003*"матк" + 0.003*"женск" + 0.003*"гинеколог" + 0.002*"аборт" + 0.002*"месяц" + 0.002*"текст" + 0.002*"лечен" + 0.002*"проблем"')]

Количество слов



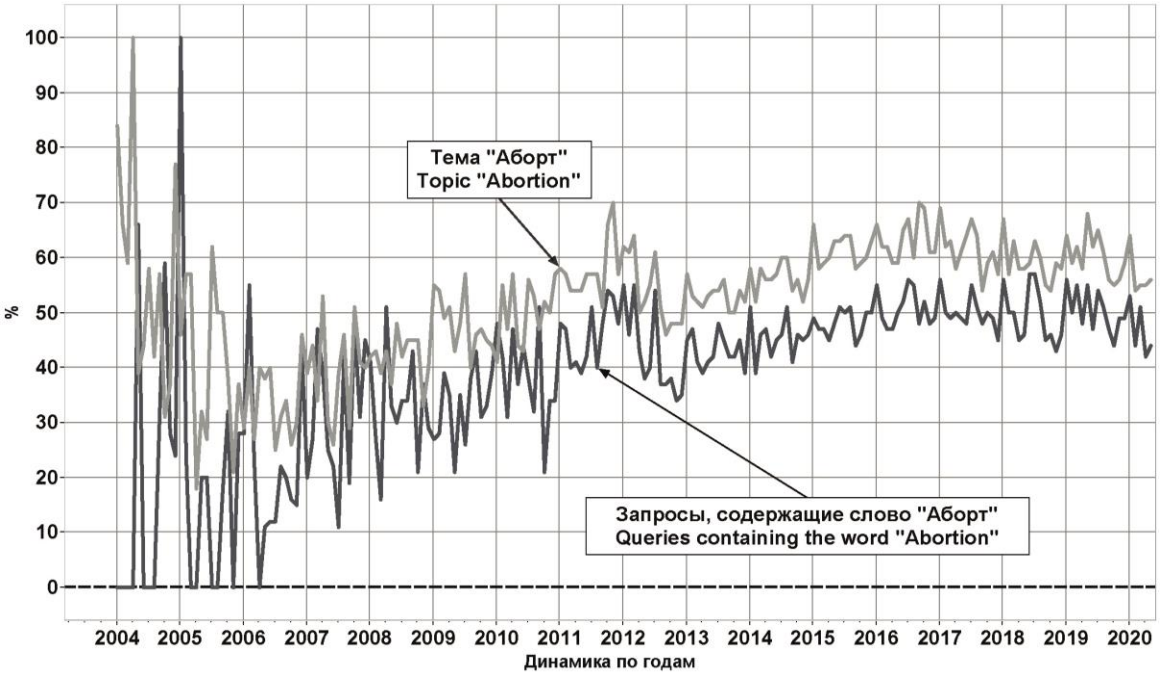
Результаты

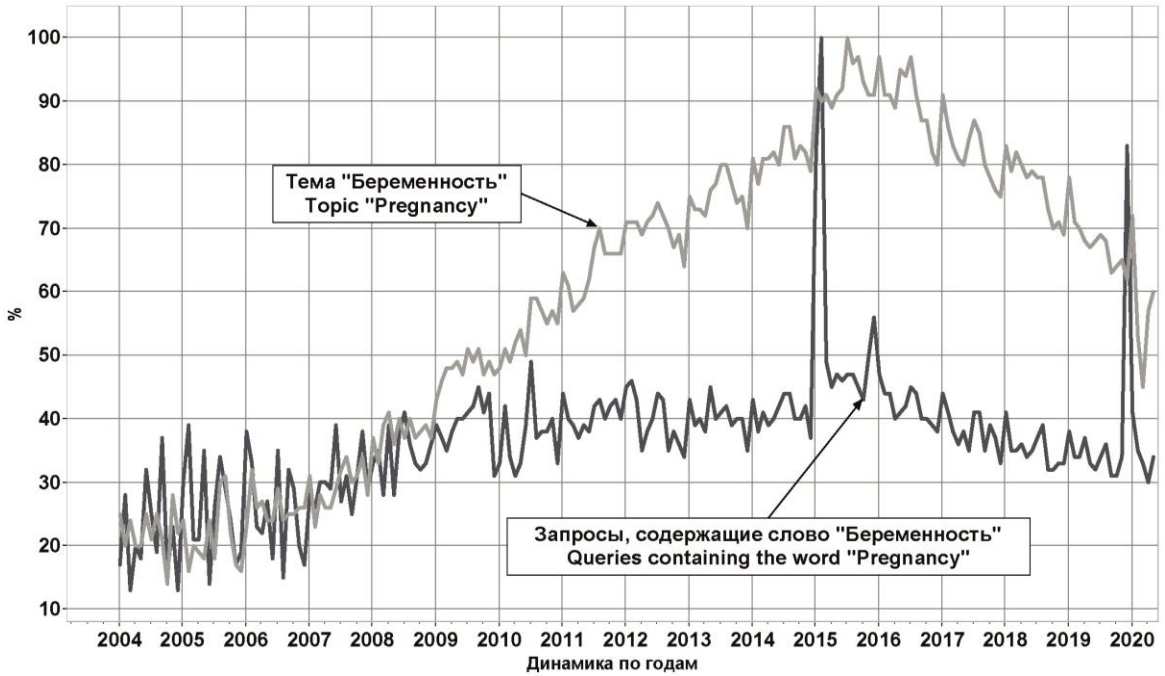
1. **Проведен тематический анализ выборки из групп Вконтакте на основе данных постов и комментариев к постам и разработан алгоритм идентификации тем, связанных с рождаемостью, брачностью и т.п.**

2. **Выделены блоки тем для дальнейшего анализа («ипотека», «беременность», «аборт»).**

3. **В дальнейшем на основании выделенных тем и временных меток будет проведён анализ тональности высказываний по выделенным темам.**

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





«Беременность» (бере), «Аборт» (абор), «ЗАГС», «Роддом» (роди), «Обручальное кольцо» (обру), «Документы на развод» (доку), «Развод» (разв)

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

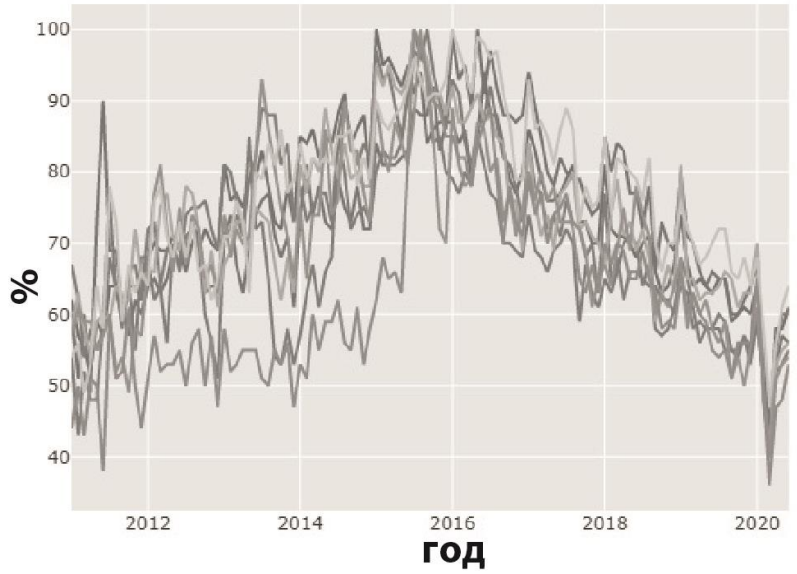
- Москва
- Московская область
- Краснодарский край
- Санкт-Петербург
- Ростовская область
- Свердловская область
- Республика Татарстан
- Республика Башкортостан
- Россия в целом

Запросы* (сокращение):

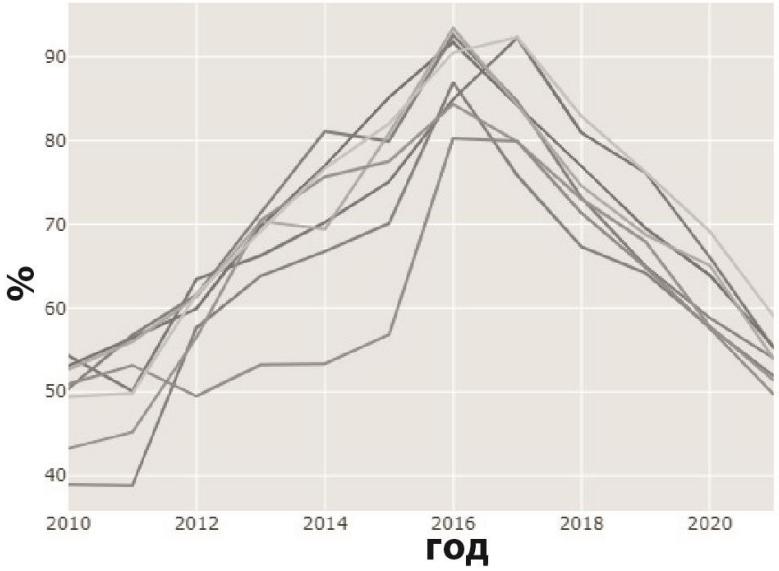
- «Беременность» (бере)
- «Аборт» (абор)
- «ЗАГС»
- «Роддом» (роди)
- «Обручальное кольцо» (обру)
- «Документы на развод» (доку)

* Параметр лежит на отрезке [0, 1]

Данные по месяцам



Данные по годам



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

По годам:

- 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)

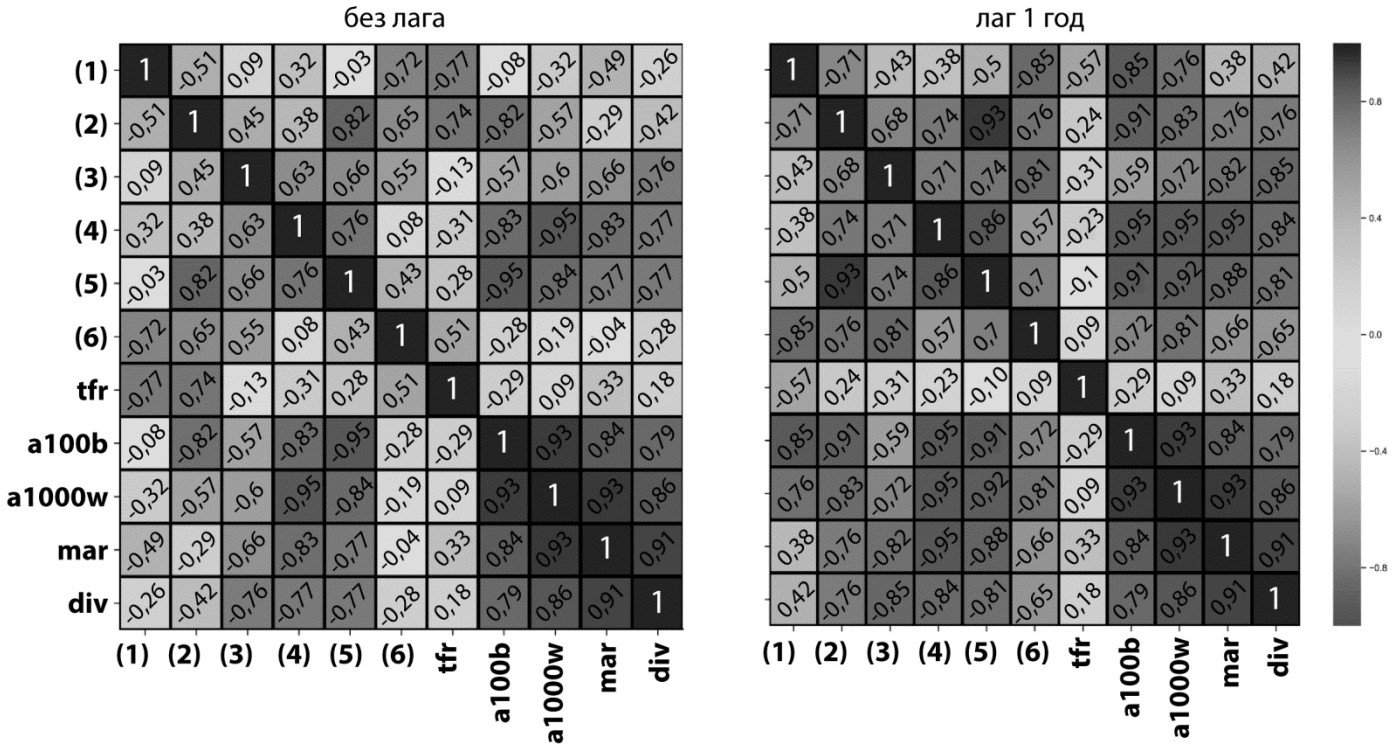
По месяцам:

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

* Параметр лежит на отрезке [0, 1]

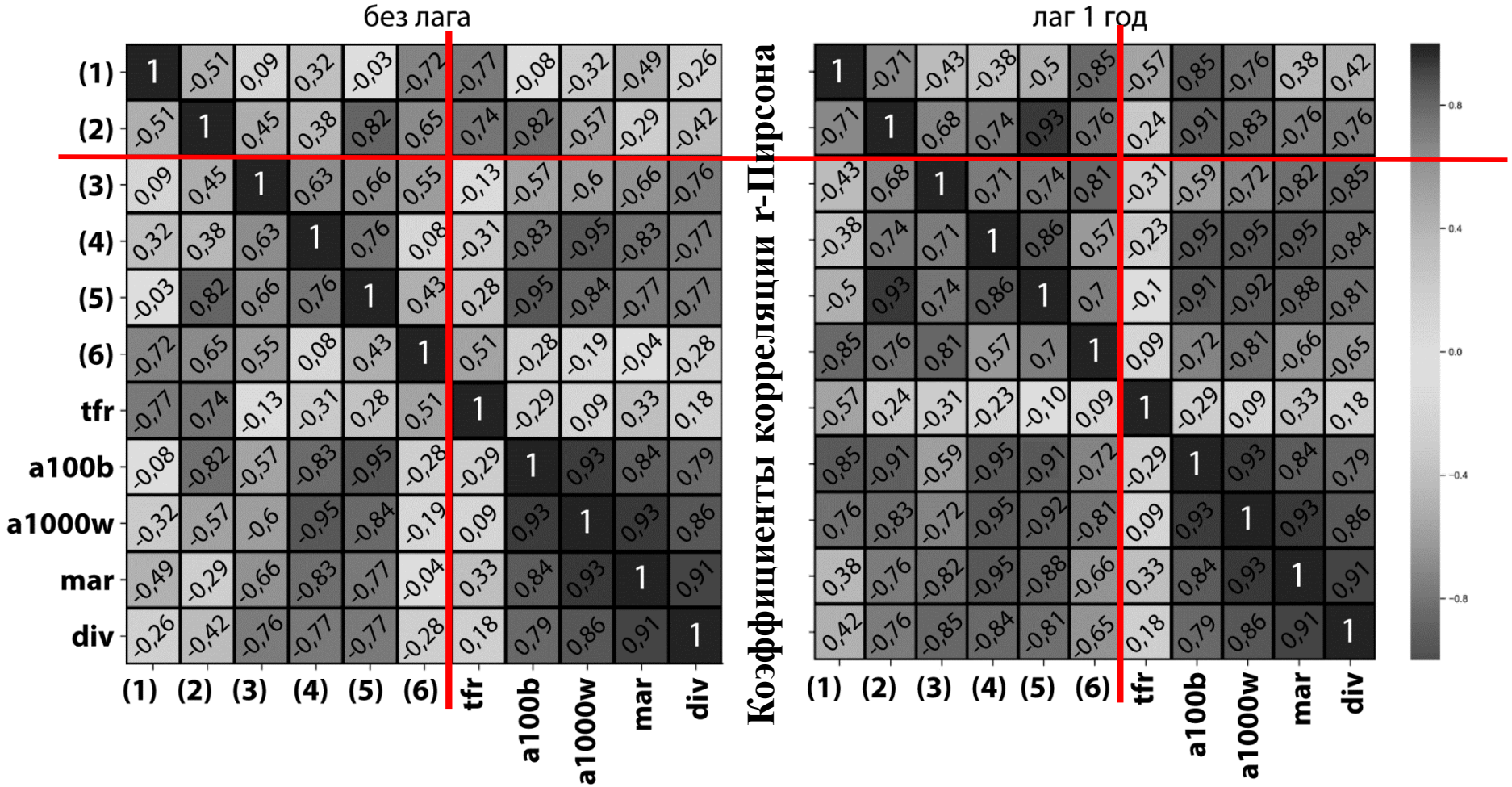
(1) – «обручальное кольцо», (2) – «беременность», (3) – «аборт», (4) – «загс», (5) – «роддом», (6) – «документы на развод»

Коэффициенты корреляции r-Пирсона



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

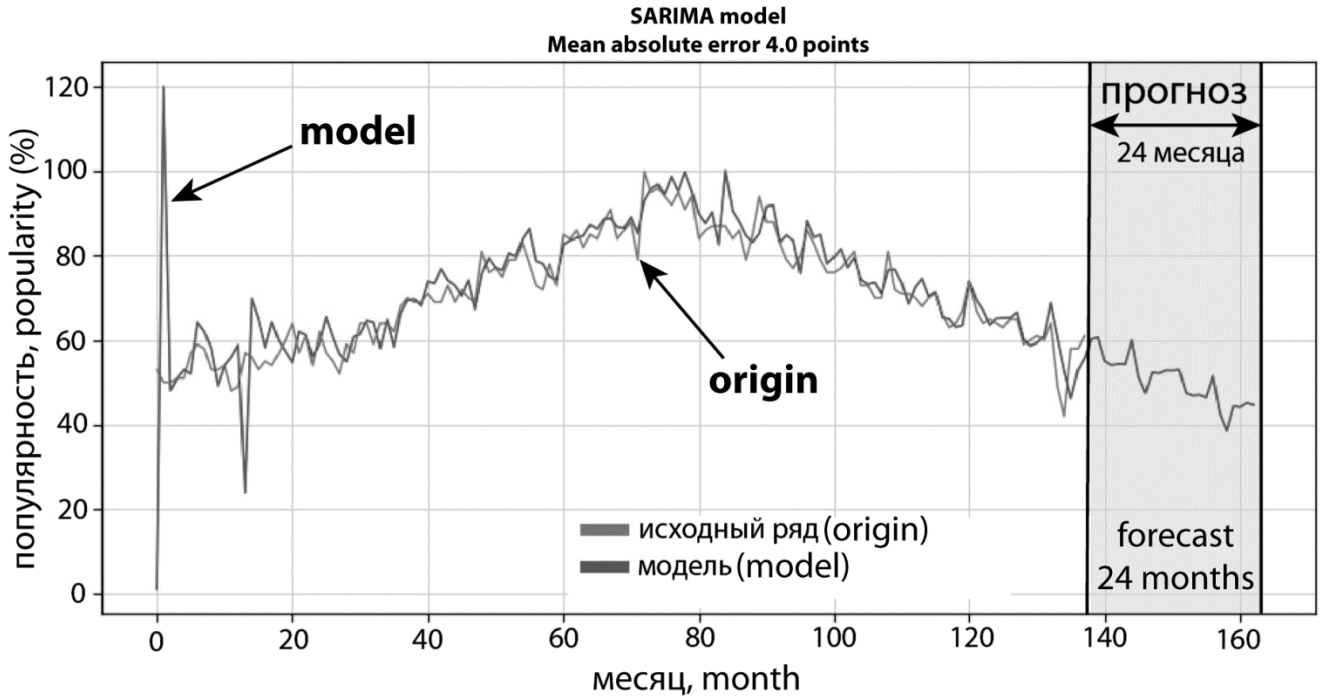
(1) – «обручальное кольцо», (2) – «беременность», (3) – «аборт», (4) – «загс», (5) – «роддом», (6) – «документы на развод»



Результаты

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.**



Реакция на демографическую политику пользователей социальной сети «Вконтакте»

***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