ОТКРЫТОЕ МНЕНИЕ

НЕЗАВИСИМЫЙ СОЦИОЛОГИЧЕСКИЙ ПРОЕКТ

Образ социологии в СМИ

Результаты инициативного аналитического исследования

Первое заседание «Клуба журналистов и социологов»

17 января 2020

О проекте

Цель исследования –- определение образа социологии в России на основе анализа текстовой информации из СМИ и экспертных публикаций за 2014 – 2019 гг.

Исследование реализовано в рамках работы **содружества «Открытое мнение»** – независимой группы профессиональных социологов, деятельность которой направлена на получение достоверной, надежной и общедоступной исследовательской информации о состоянии общественного мнения в России.

Участники проекта:

Руководитель:

Игорь Задорин (Группа ЦИРКОН)

Составление базы публикаций, подготовка массивов текстовых данных:

Анна Хомякова (Группа ЦИРКОН), Дарья Щербакова, Арюна Раднаева (НИУ ВШЭ)

Подготовка и анализ данных:

Дарья Мальцева, Екатерина Булычёва (МЛ прикладного сетевого анализа НИУ ВШЭ), Дарья Рудь (Открытое мнение), Валерия Мошенко, Илья Фомин (НИУ ВШЭ, ЦИРКОН)

Подготовка презентации:

Дарья Мальцева, Дарья Рудь

По согласованию, использованы некоторые стилистические элементы презентации компании Aventica.

Массив данных

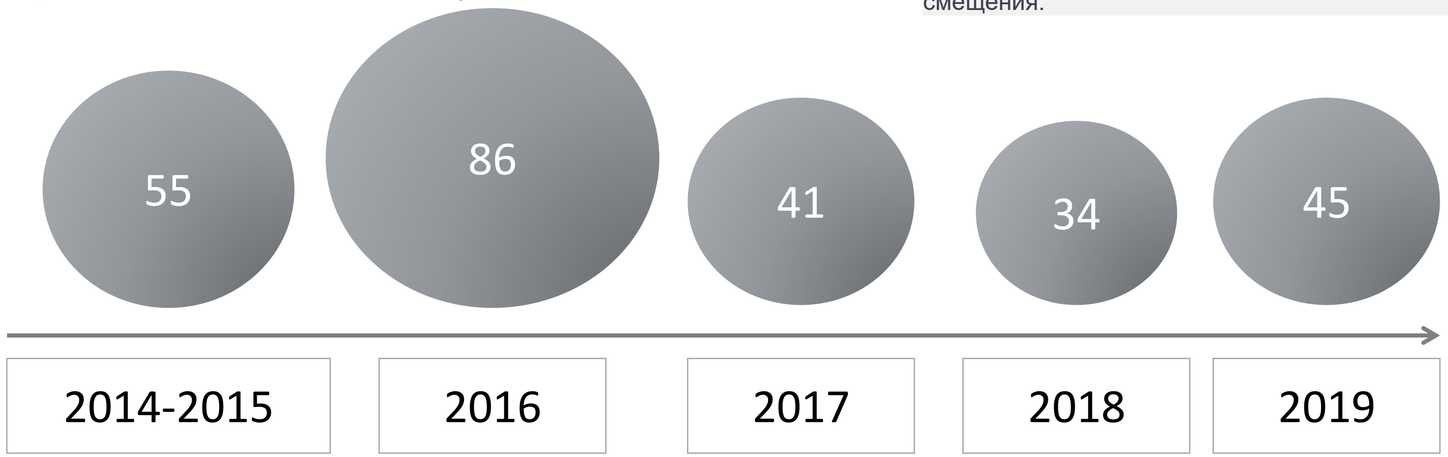
1 В массив данных вошли публикации по теме социологии и социологических данных, опубликованные в СМИ и медиа-ресурсах (блогах), отобранные экспертно.

2 Для анализа данных были взяты публикации в **текстовом** формате (аудио, видео, рисунки, инфографика не включены в финальный массив).

3 Массив был разделен **по годам** на 5 подмассивов.

4 Для отдельного анализа были выделены подмассивы из 87 публикаций (за все годы) **социологов** и 168 публикаций **журналистов**.

Примечание:
Мы хотим подчеркнуть, что **не говорим** о полноте и репрезентации собранного массива публикаций – отобранное количество публикаций может быть вызвано ограничениями информационного поиска. Из-за этого в массиве могут присутствовать некоторые смещения.



Текстовый анализ данных: динамика 2014 – 2019 гг.

Текстовый анализ данных

Исходный текст –

Токены –

Нормалиация –

Удаление стоп-слов –

Лексемы как «словарь» для дальнейшего анализа

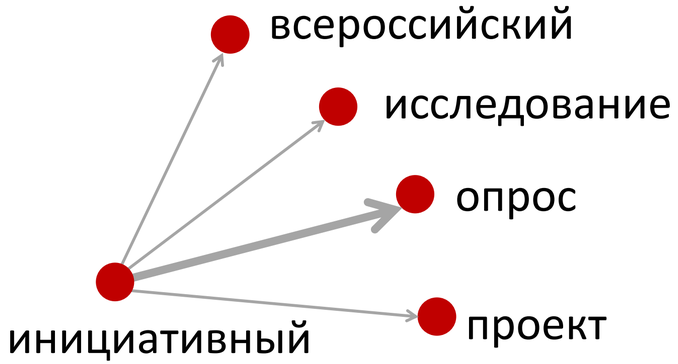
Облака слов
из наиболее часто встречающихся лексем

Индекс **TF (term frequency)**
– соотношение количества упоминания лексемы ко всему количеству лексем в массиве («важность» каждой лексемы)

Сети связей лексем,
основанные на близости по словосочетаниям с весом = количеству совместных упоминаний

Исходный текст:
«Генеральному директору ВЦИОМ Валерию Федорову в ответ …»
Токены:
['генеральному', 'директору', 'вциом', 'валерию', 'федорову', 'в', 'ответ‘,…]
Лексемы:
['генеральный', 'директор', 'вциома', 'валерий', 'фёдоров', 'ответ', …]

Биграмма:
инициативный {'всероссийский': 1, 'исследование': 1, 'опрос': 2, 'проект': 1}



Важно! В результате нормализации некоторые слова были трансформированы программой в начальную форму не совсем корректно, например ВЦИОМ – «вциома», «ФОМ» - «фома», Лев – «левый». Эта особенность может быть поправлена при следующей итерации анализа, но ее необходимо учитывать в настоящем отчете при просмотре результатов и сетевых визуализаций. Приносим свои извинения интересантам ☺

TF индекс – топ 30 лексем

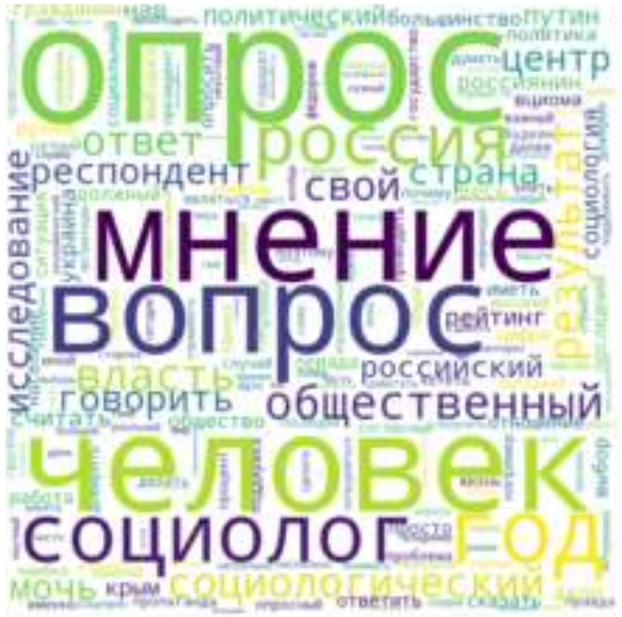
	2014-2015	2016	2017	2018	2019
1	**опрос**	**опрос**	**опрос**	**опрос**	**опрос**
2	который	человек	**мнение**	год	вопрос
3	человек	который	который	человек	который
4	**мнение**	год	**общественный**	который	человек
5	вопрос	вопрос	человек	**мнение**	**путин**
6	год	центр	год	результат	**мнение**
7	**социолог**	**россия**	результат	**россиянин**	**вциом(а)**
8	**россия**	свой	**исследование**	**россия**	**президент**
9	**общественный**	**мнение**	центр	**вциом(а)**	**доверие**
10	результат	**исследование**	вопрос	**общественный**	**рейтинг**
11	**власть**	**левада**	**голосование**	**респондент**	**доверять**
12	**социологический**	результат	проведение	**исследование**	**респондент**
13	ответ	говорить	**левада**	**доверие**	**социолог**
14	свой	**власть**	**вциом(а)**	считать	**общественный**
15	дать	дать	**россия**	**президент**	год
16	**исследование**	самый	**выбор**	**власть**	**храм**
17	центр	очень	время	дать	власть
18	**респондент**	выбор	день	**социологический**	**интервьюер**
19	говорить	**общественный**	свой	**опросить**	результат
20	**страна**	сказать	**политический**	общество	центр
21	мочь	**социолог**	дать	наш	говорить
22	**российский**	мочь	**власть**	**российский**	дать
23	**политический**	**политический**	число	свой	сказать
24	**путин**	весь	**страна**	вопрос	**екатеринбург**
25	самый	наш	самый	**социолог**	**исследование**
26	**социология**	**страна**	**социологический**	**страна**	ответ
27	**рейтинг**	**путин**	являться	гражданин	**строительство**
28	считать	просто	организация	самый	**россиянин**
29	наш	время	**обнародование**	**интервьюер**	**владимир**
30	**россиянин**	ответ	**россиянин**	время	провести

Опрос – самая часто встречающаяся лексема по всем годам

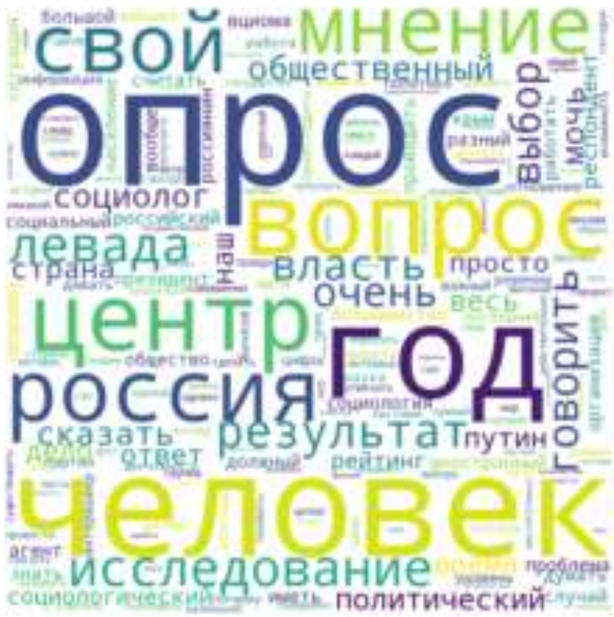
Облака слов

Примечание:
При построении облаков слов были дополнительно удалены наиболее частотные лексемы, не несущие фактической информации

2014-2015

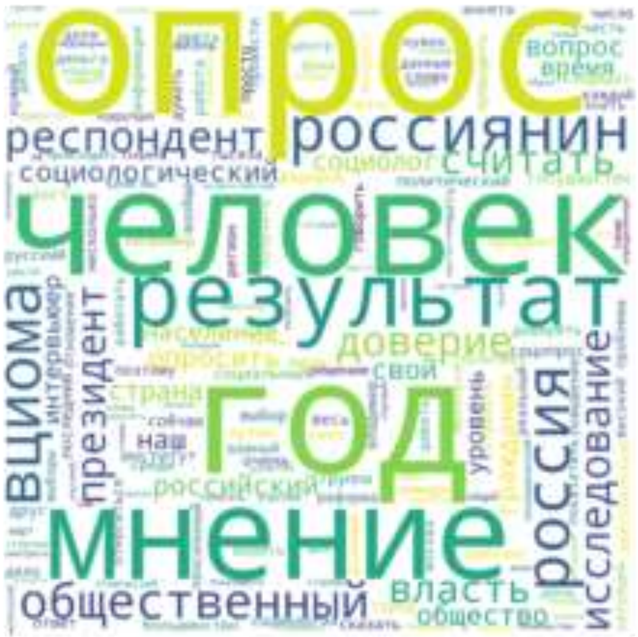


2016



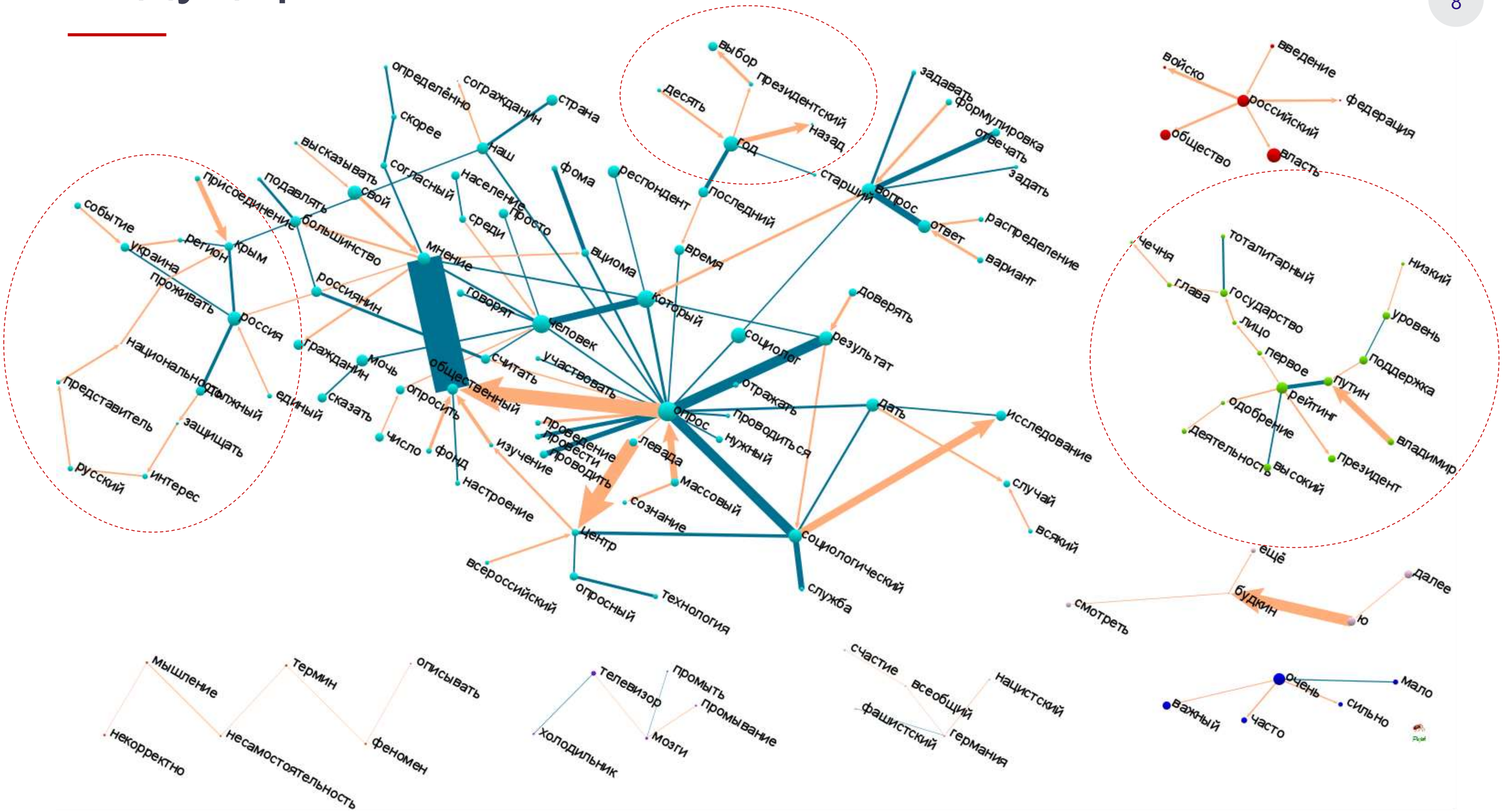
2017

2018

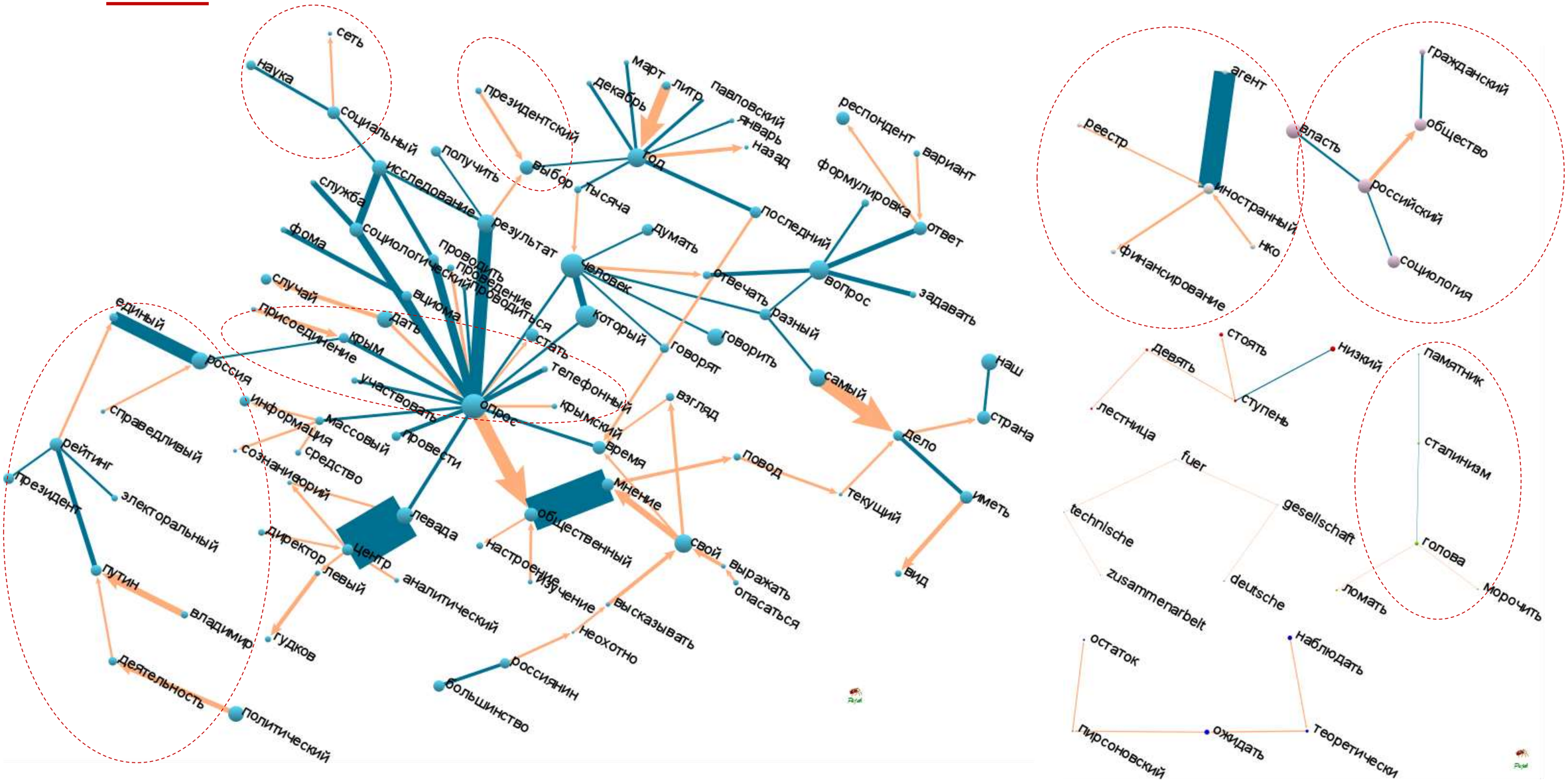


2019

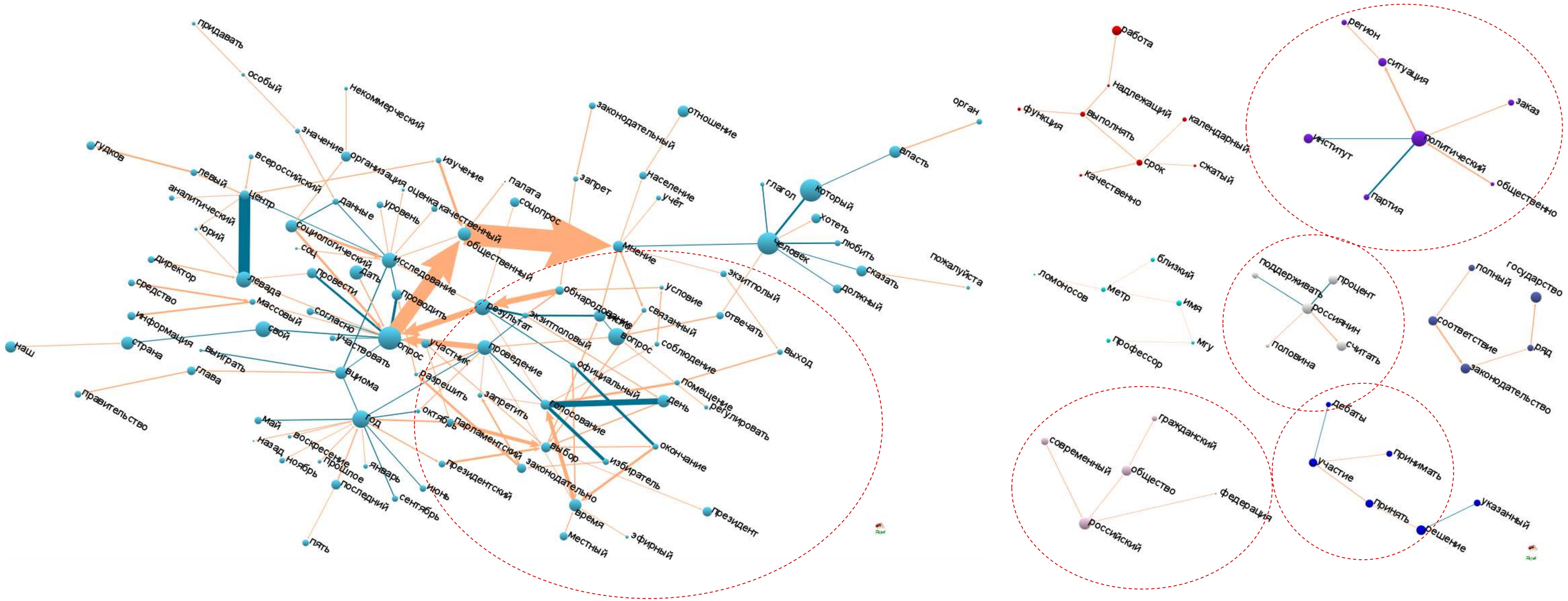
Редуцированная сеть 2014-2015



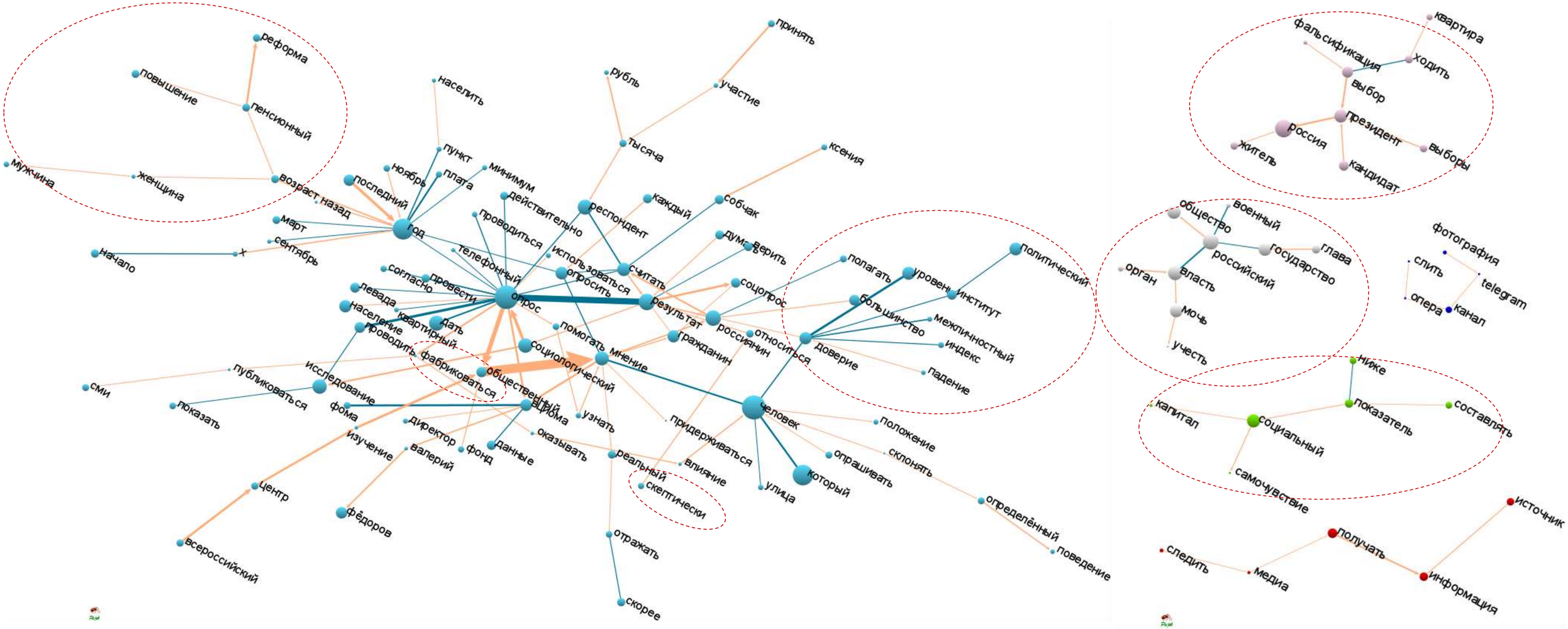
Редуцированная сеть 2016



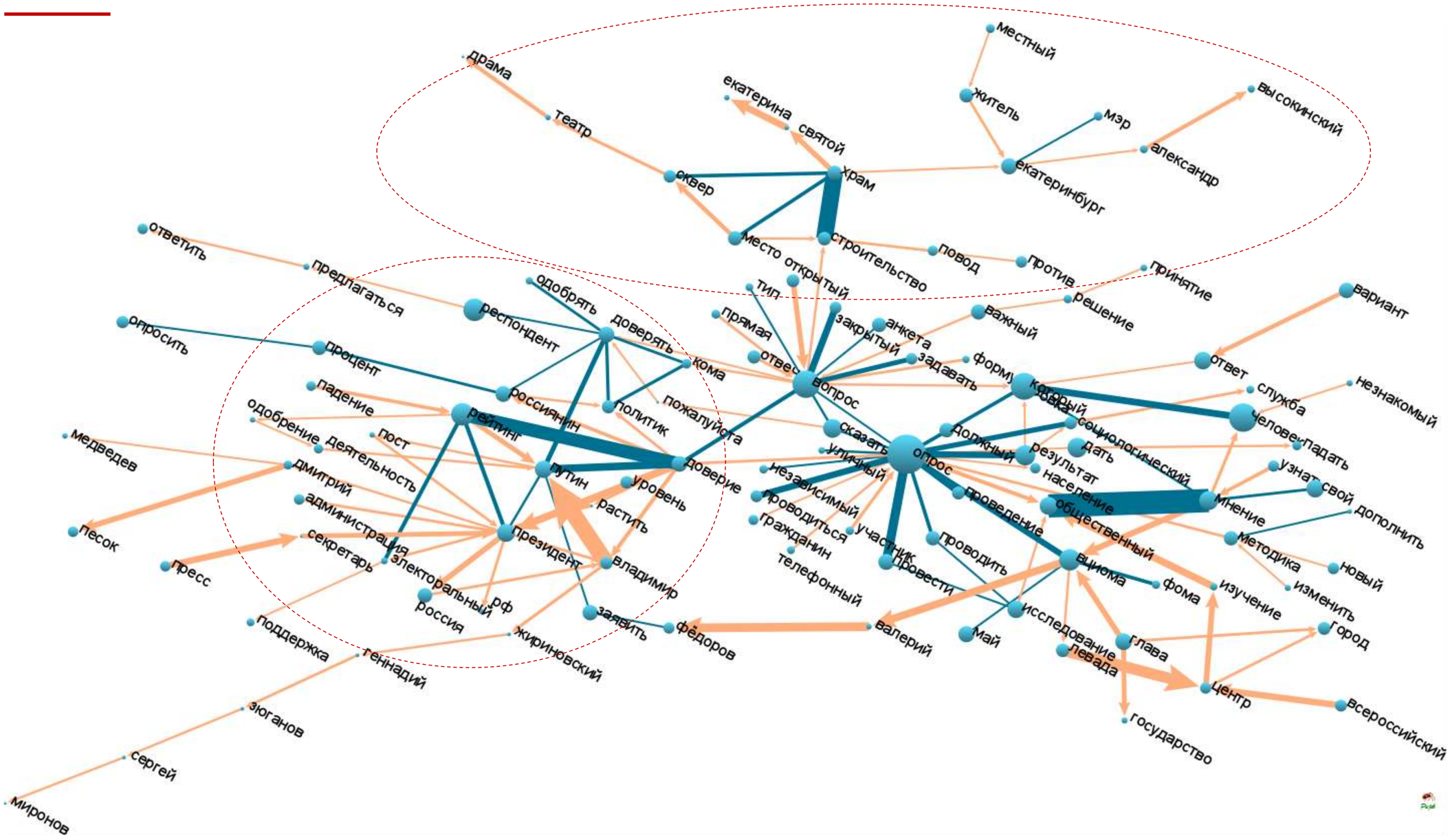
Редуцированная сеть 2017



Редуцированная сеть 2018



Редуцированная сеть 2019



Выводы и комментарии

1 Результаты анализа публикаций показывают, что «социология» представлена в массовом пространстве (СМИ и экспертных публикациях в медиа) по большей части в связи с **массовыми опросами** (о чем говорят наиболее часто встречающиеся лексемы «опрос», «респондент», «мнение», и др.).

2 При этом тематически социология в СМИ и экспертных публикациях в наибольшей степени употребляется в контексте событий из **политической сферы жизни общества (а не социальной)** – одна из постоянных тематик, возникающих из анализа текстов в разные годы, касается выборов и электоральных рейтингов («большие» выборы 2017-2018 гг.), а также присоединение Крыма (2014-2015), закон об иностранных агентах и связанные с этим события в Левада-центре (2016).

3 Вместе с тем, анализ массива публикаций позволяет зафиксировать и некоторые **событийные контексты**, относящиеся к ситуациям в социальной сфере: пенсионная реформа (2018), ситуация вокруг строительства храма в Екатеринбурге (2019).

Текстовый анализ данных: социологи vs. журналисты

Лексемы (1)

Социологи

Журналисты

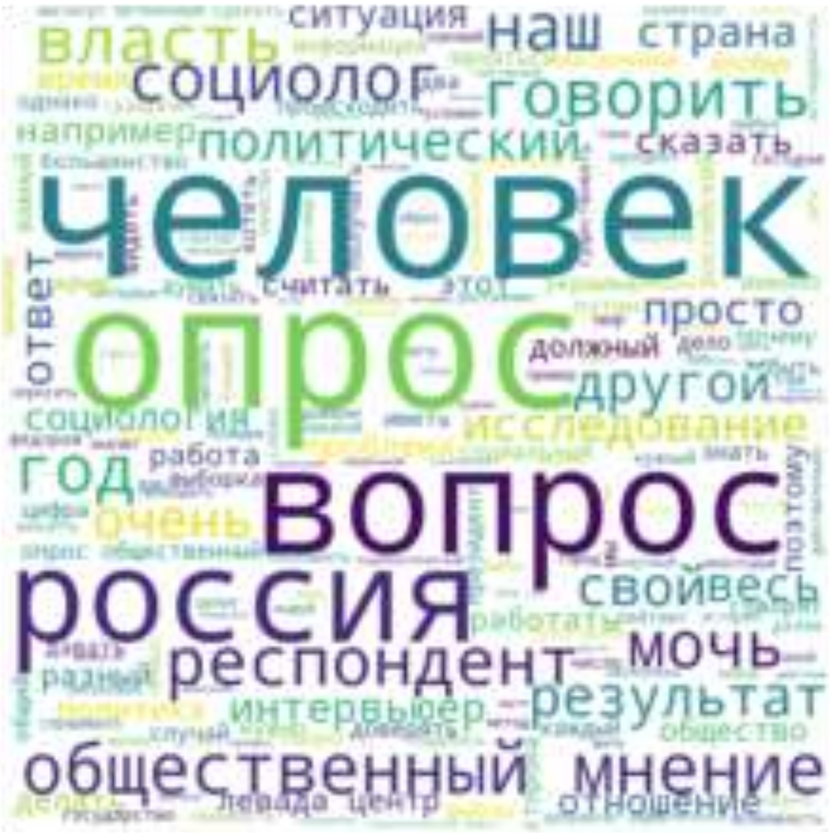
TF индекс – топ 30 лексем

№	Лексема	№	Лексема
1	**опрос**	16	ответ
2	который	17	говорить
3	человек	18	**власть**
4	**вопрос**	19	**респондент**
5	год	20	**социолог**
6	**мнение**	21	очень
7	россия	22	другой
8	**общественный**	23	мочь
9	такой	24	наш
10	тот	25	центр
11	один	26	самый
12	свой	27	страна
13	результат	28	**политический**
14	дать	29	весь
15	**исследование**	30	**крым**

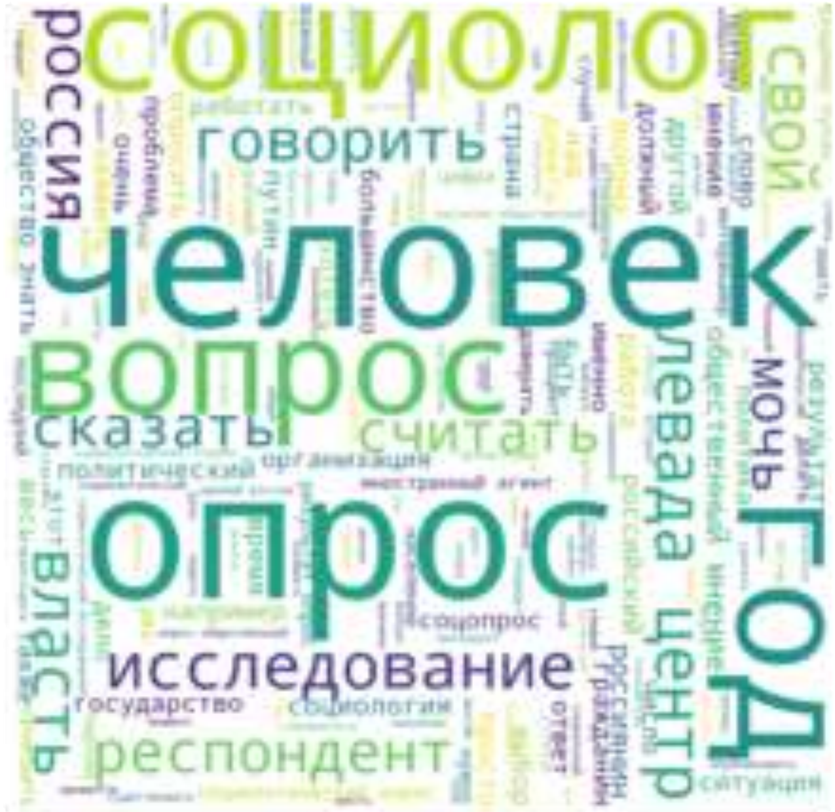
№	Лексема	№	Лексема
1	**опрос**	16	**социологический**
2	который	17	свой
3	человек	18	россиянин
4	год	19	тот
5	**мнение**	20	**путин**
6	центр	21	**общественный**
7	**исследование**	22	один
8	вопрос	23	дать
9	**социолог**	24	считать
10	**левада**	25	**президент**
11	результат	26	говорить
12	россия	27	**респондент**
13	**вциом(а)**	28	**рейтинг**
14	такой	29	**политический**
15	власть	30	самый

Облака слов

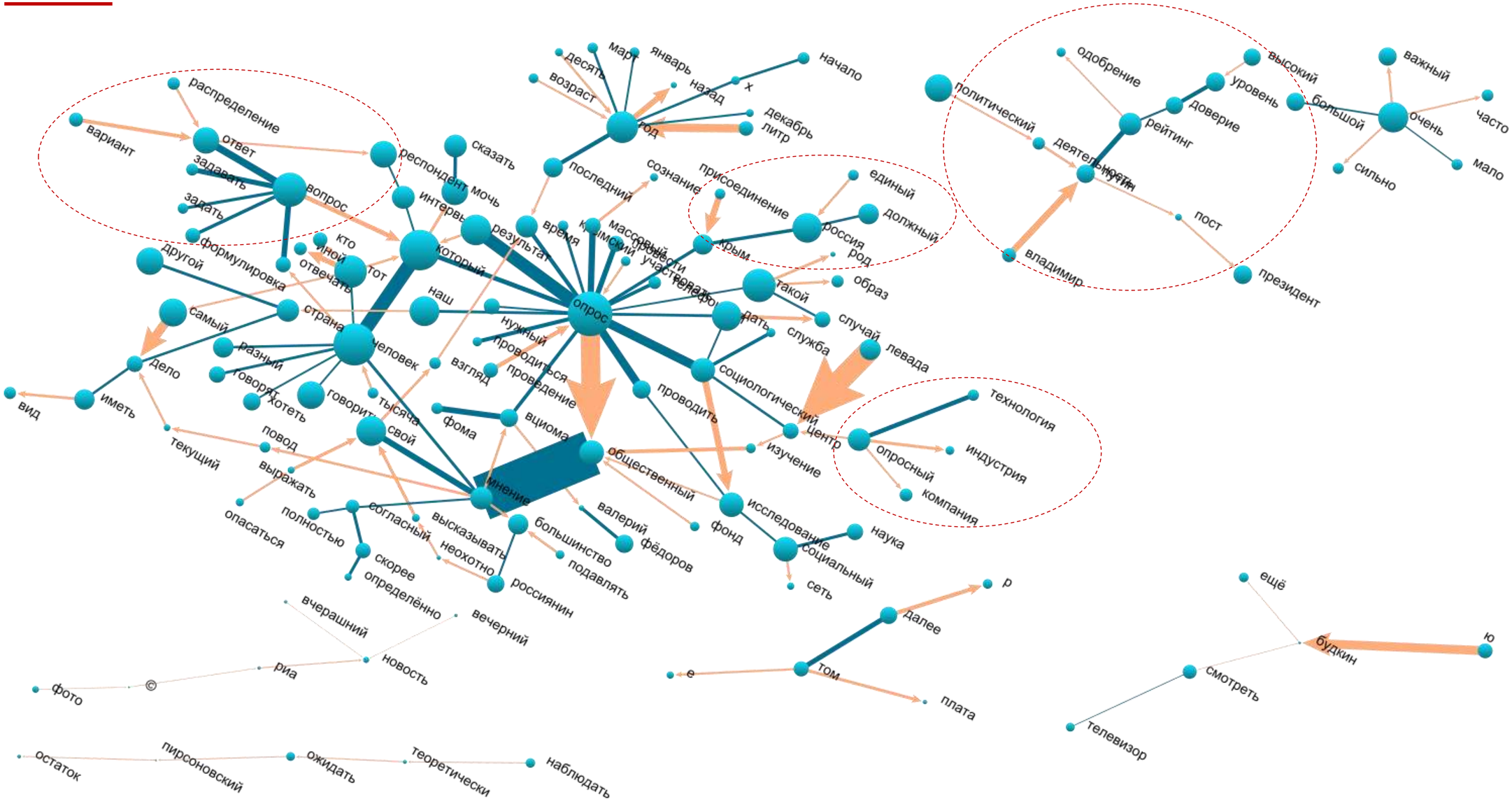
Социологи



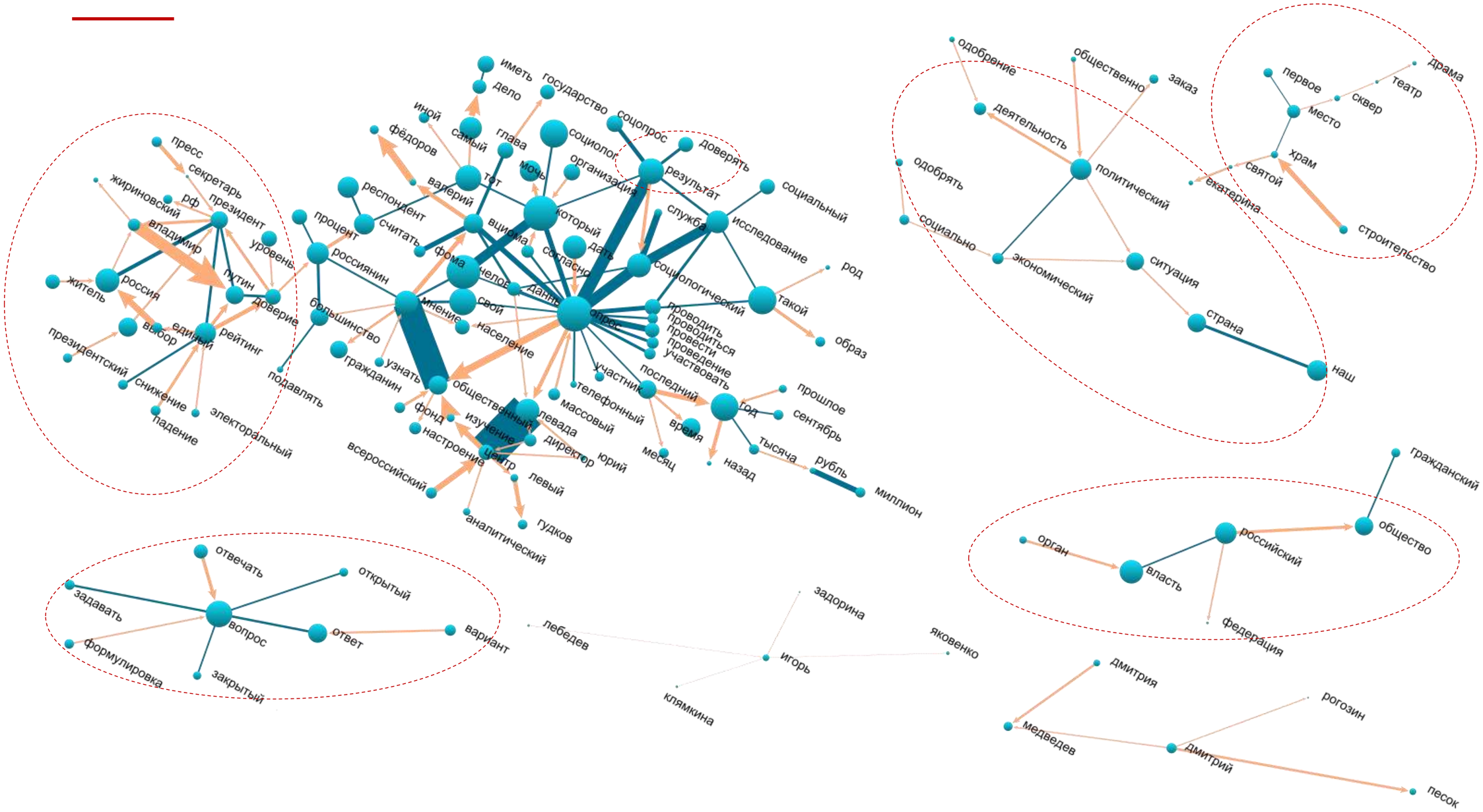
Журналисты



Редуцированная сеть - социологи



Редуцированная сеть - журналисты



Выводы и комментарии

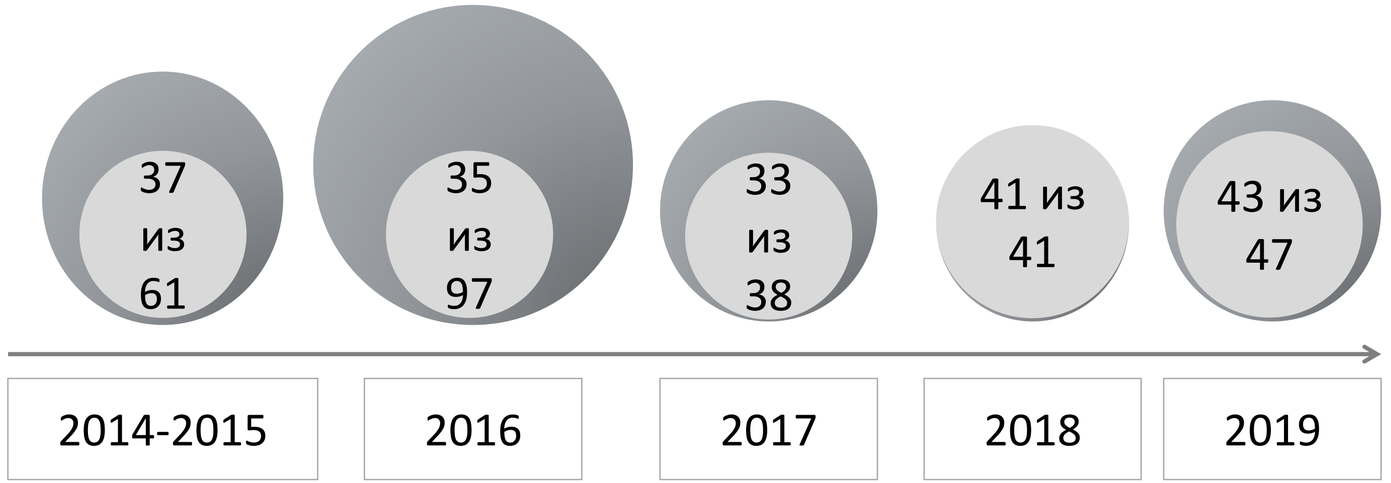
1

Разделение текстов на экспертные тексты социологов и журналистов не приводит к значительным изменениям результатов: упор в текстах все равно делается на **массовые опросы, политическую сферу жизни общества и событийные контексты**. В экспертных текстах профессиональных социологов проявляется тематика присоединения Крыма, в текстах журналистов - доверия социологическим данным, строительства храма к Екатеринбурге.

Анализ данных на основе первичного кодирования

Массив данных для кодирования

Процедуре первичного кодирования подвергнуто **30-40 статей каждого периода, всего 189 публикаций**



Процедура кодирования и анализ

1 Открытое кодирование: «от материала»

2 Осевое кодирование: связи между кодами

Коды присваиваются тексту публикации целиком
Преимущества: публикация рассматривается в качестве единого высказывания
Недостатки: несколько различных тезисов в одном тексте становятся аналитически неразличимыми
ПО: Atlas.ti 8

4 **Результат**:

- Частотные распределения кодов
- Анализ соответствий

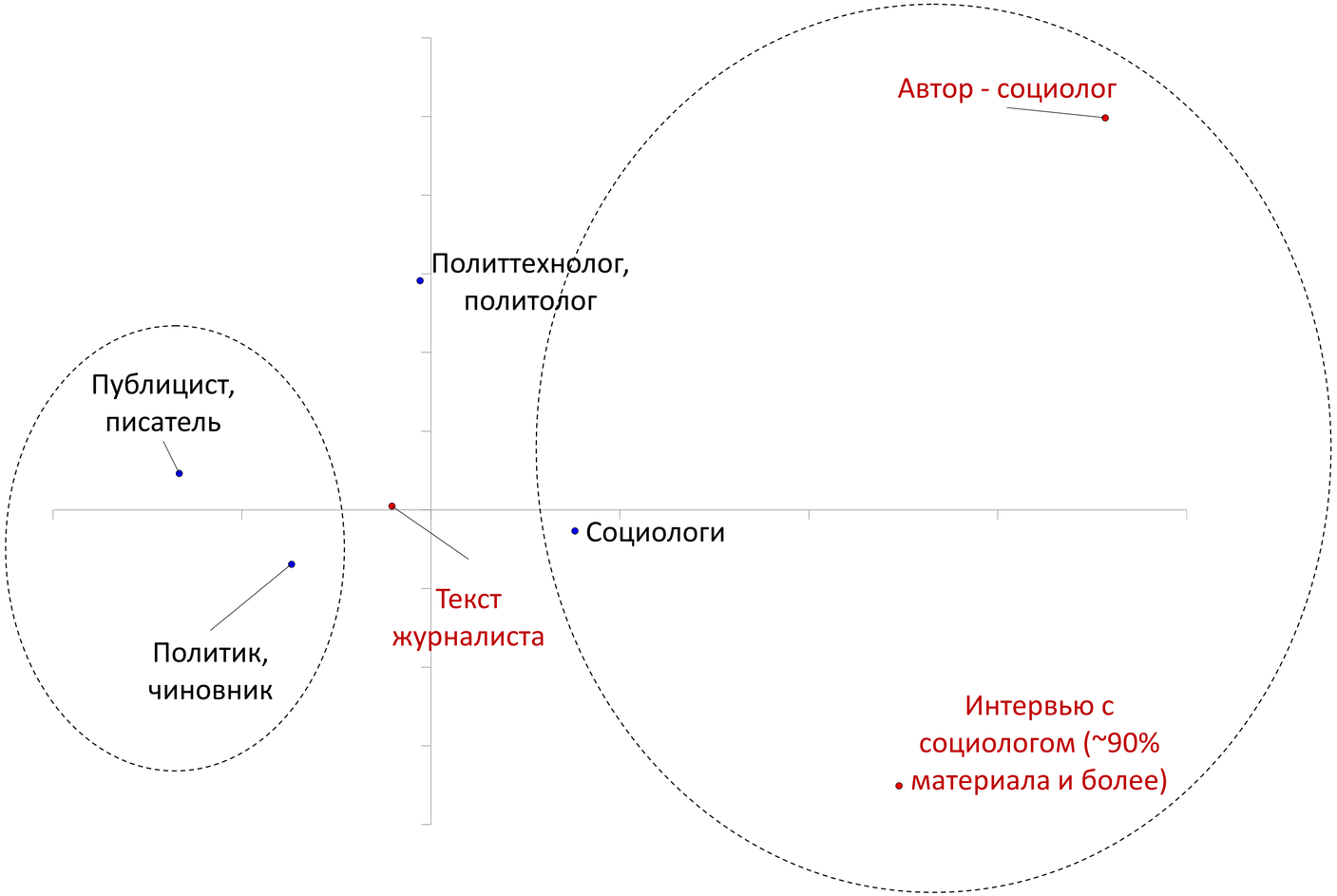
3 Кодификатор включает «тэги» для суждений **по темам**:

1. **Данные и их интерпретация**
2. **Критика работы социологов и ее основания**
3. **Субъекты критики**
4. **Поддержка социологов и ее основания**
5. **Предложения и пожелания социологам**
6. **Вопросы социологии (выборка, инструментарий, респондент, стоимость работ и др.)**
7. **Комментатор/действующее лицо**
8. **Крупные инфоповоды**
9. **Политические темы (оппозиция, Кремль, давление и др.)**
10. **Характеристика текста (от журналиста/от социолога/интервью с социологом)**

Итого: 49 кодов

Этос текстов

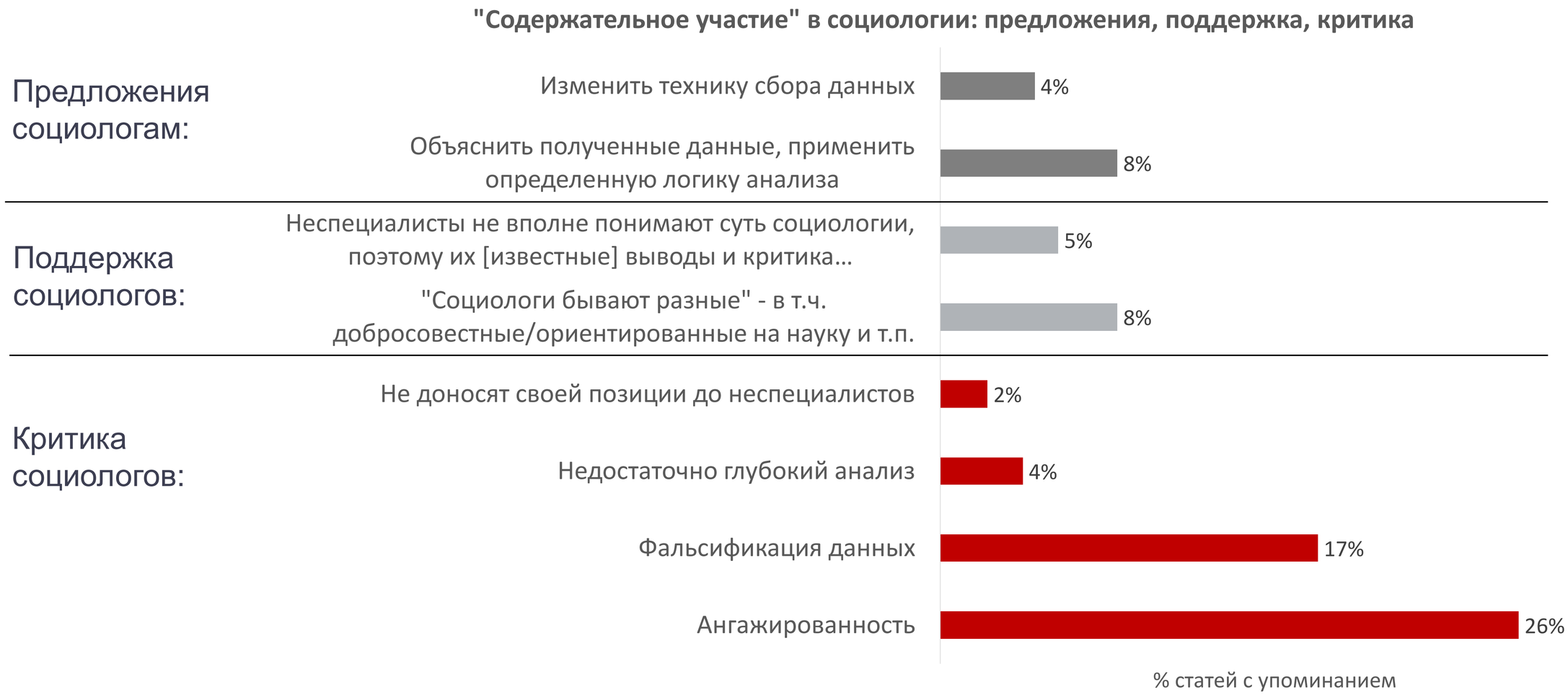
В текстах от социологов несколько чаще упоминаются другие социологи. На другом конце оси – «публицистическо-официальное» восприятие социологии.



Анализ соответствий

Активная позиция в отношении социологии

Оценки, отличающиеся от обвинений в ангажированности и фальсификациях, встречаются нечасто.



Предложения социологам: цитаты

Почему Россия отстала? Можем ли мы включиться в догоняющую модернизацию, или это отставание навсегда? Преодолимо ли наше имперское наследие? Могут ли случиться новые революции, и не опасно ли в этой связи «раскачивать лодку»? Зависит ли развитие от культуры, от институтов, от исторического пути или только от воли реформаторов? (…) почти нет профессионального исследования сформулированных выше проблем. Точнее, они иногда становятся побочным продуктом научных работ историков, экономистов, социологов и политологов, но специалисты углубляются в эти «вечные» вопросы как-то стыдливо.

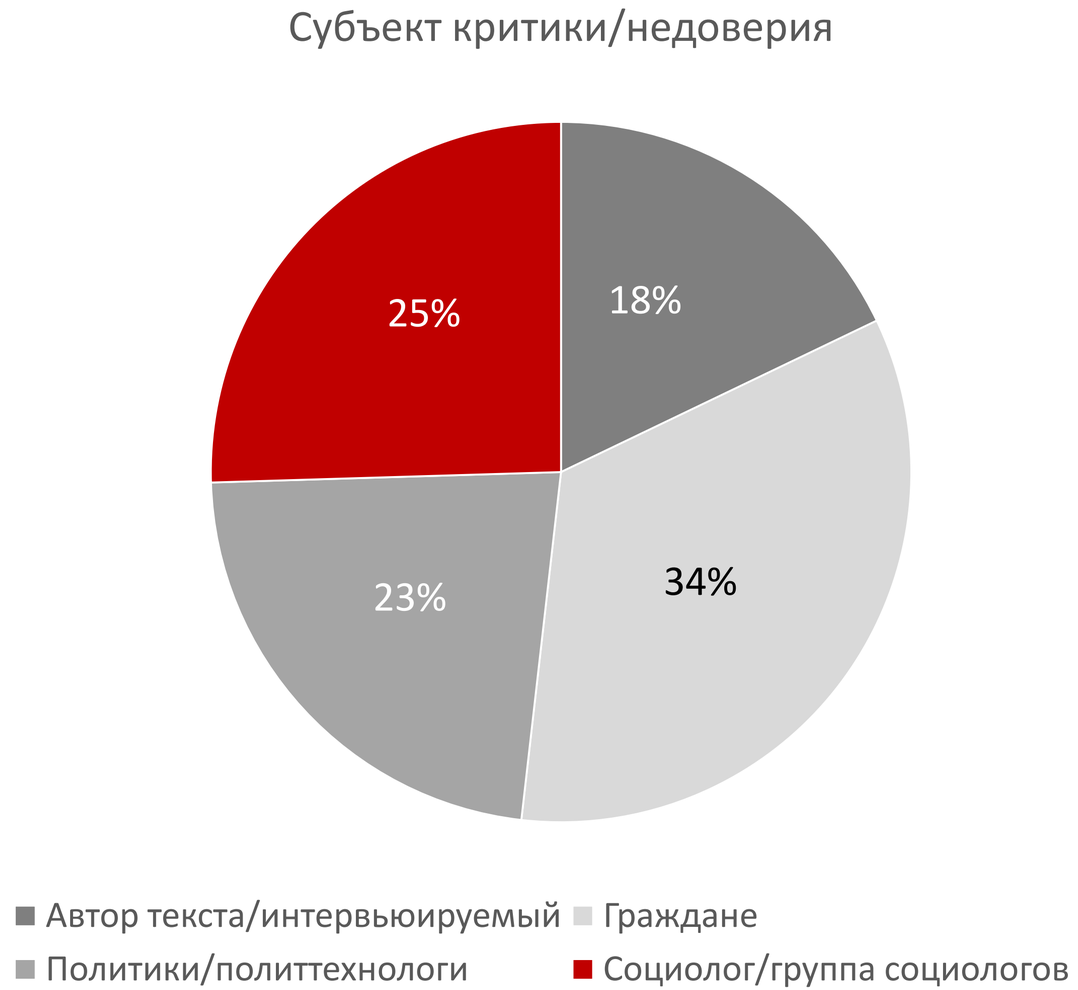
Депутат Государственной думы Александр Шерин предложил исследовательским организациям воздержаться от телефонных опросов россиян.

"Мы ждем какого-то анализа наших уважаемых специалистов, как коррелируются эти данные: как может падать уровень доверия, но расти электоральный рейтинг. Это сложный анализ, мы надеемся, что со временем он появится", - подчеркнул пресс-секретарь

Занудно задавал один и тот же набор вопросов: «Почему не регистрируете отказы? Почему не анализируете ситуацию прохождения маршрутов? Почему не накапливаете данные о реализованной выборке?».

Описание субъектов критики

Критика и недоверие в адрес социологии выражены от лица различных субъектов примерно с равной частотой.



Активная позиция и принадлежность текста

Поддержка исходит от социологов, критика и предложения – от всех остальных.

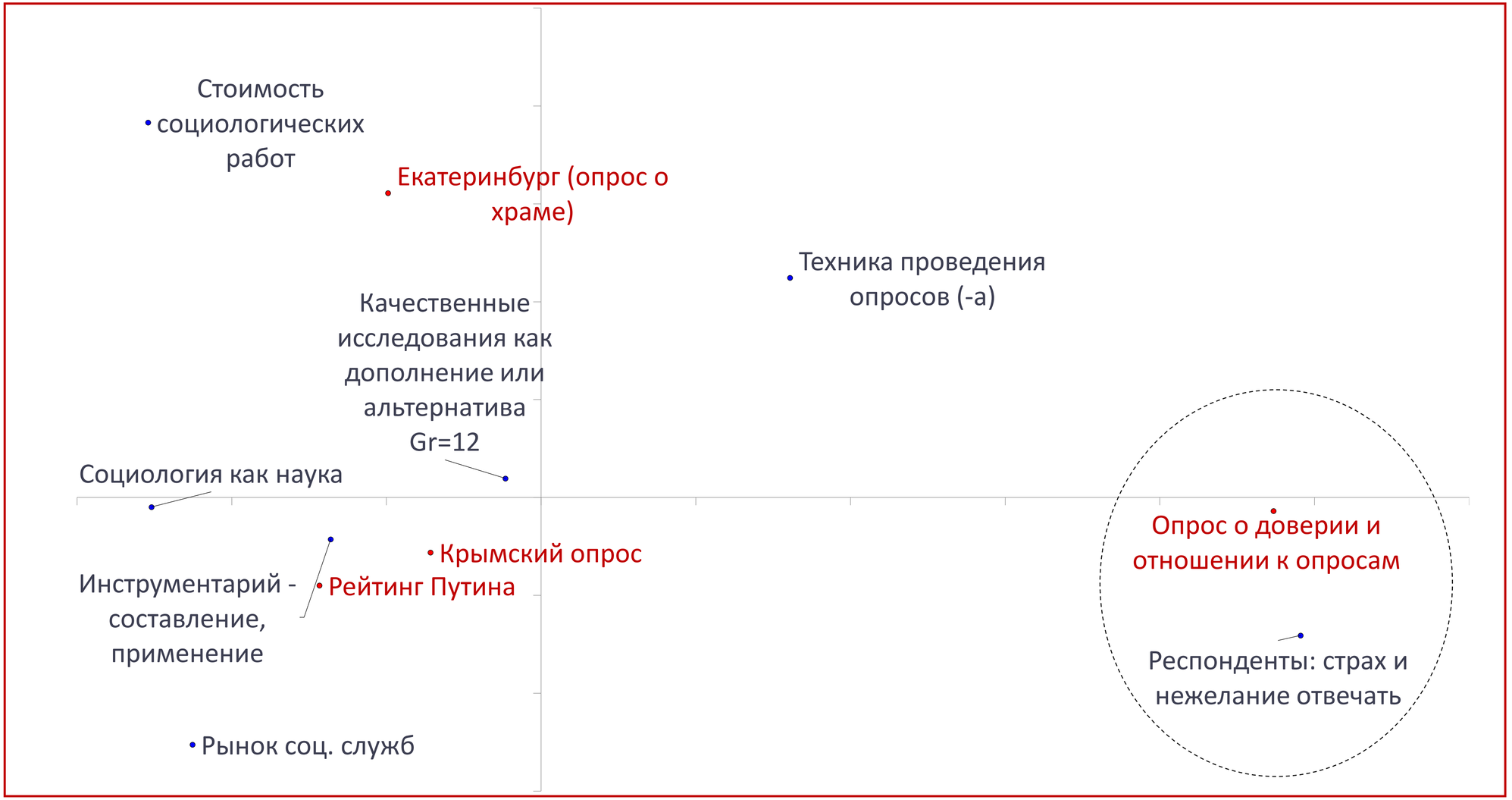
«Вопросы социологии», поставленные в текстах

Ремесло и полевая работа в массовых опросах – основной предмет описания

Более общая, стратегическая постановка вопросов обсуждается редко

Вопросы социологии и крупные инфоповоды

Различные инфоповоды порождают обсуждение различных аспектов. Философия опроса воспринимается отдельно от новостей.



Выводы и комментарии

1. Мир мнений социологов оказывается замкнутым и мало связанным с миром мнений политиков и публицистов.

2. Критика преобладает над предложениями и поддержкой.

3. Все типы авторов публикаций склонны критиковать социологов примерно одинаково (в равных долях), однако поддержка чаще исходит из социологического цеха, а предложения – от журналистов.

4. Полевая работа и конкретные узкие «технические» вопросы – основной объект описания проблем. Позиционирование социологии в целом, ее рынок и т.п. эксплицитно обсуждаются редко.

5. Различные инфоповоды генерируют общественное внимание к различным узким вопросам.

ОТКРЫТОЕ МНЕНИЕ

НЕЗАВИСИМЫЙ СОЦИОЛОГИЧЕСКИЙ ПРОЕКТ

Контакты

Игорь Задорин
zadorin@zircon.ru

Дарья Мальцева
d_malceva@mail.ru

Дарья Рудь
daria.rud.moscow@gmail.com

ОТКРЫТОЕ МНЕНИЕ

НЕЗАВИСИМЫЙ СОЦИОЛОГИЧЕСКИЙ ПРОЕКТ

Дополнительные слайды

Количественный анализ данных

1 Анализ текстовых данных выполнен в программе Python (пакеты для анализа текстовых данных и обработки естественного языка nltk, pymorphy2, wordcloud).

2 Внутри каждого подмассива (по году) каждое предложение разбито на отдельные слова (токены), произведен **подсчет общих метрик**: частота встречаемости токенов, общее количество используемых слов и различных словоформ, средняя длина слов и предложений в тексте (общие метрики).

3 Проведена нормализация (приведение токенов к начальной форме), удалены стоп-слова для русского языка (предлоги, союзы, местоимения и т.п.) – набор лексем **(«словарь») для дальнейшего анализа.** Подсчитано количество уникальных лексем в тексте, определена их часть речи, выделены наиболее часто встречающиеся лексемы.

4 Из наиболее часто встречающихся лексем построены **облака слов**. При этом дополнительно удалены наиболее частотные лексемы, не несущие фактической информации: ['это', 'хотя', 'кого', 'кроме', 'таких', 'который','какой','тот','стать','дать','один','такой','самый'].

5 На чистом массиве подсчитан индекс **TF (term frequency)** – соотношение количества упоминания лексемы ко всему количеству лексем в массиве (определена «важность» каждой лексемы).

Простой текст:
«Генеральному директору ВЦИОМ Валерию Федорову в ответ …»
Токены:
['генеральному', 'директору', 'вциом', 'валерию', 'федорову', 'в', 'ответ‘,…]
Лексемы:
['генеральный', 'директор', 'вциома', 'валерий', 'фёдоров', 'ответ', …]

Важно! В результате нормализации некоторые слова были трансформированы программой в начальную форму не совсем корректно, например ВЦИОМ – «вциома», «ФОМ» - «форма», Лев – «левый». Эта особенность может быть поправлена при следующей итерации анализа, но ее необходимо учитывать в настоящем отчете при просмотре результатов и сетевых визуализаций. Приносим свои извинения интересантам ☺

Анализ данных – сети

1 Для построения сетей была использована **близость лексем по словосочетаниям**: связи выстраивались между лексемами, стоящими рядом друг с другом. Такие пары слов называются **биграмы**.

2 Биграммы строились по правилу: к каждой лексеме («ключу») из числа лексем («словаря») были добавлены все стоящие после нее лексемы (одна лексема могло встречаться несколько раз). Связи между лексемами являются **направленными** (в порядке встречаемости лексем - i и i+1).

3 Произведен подсчет количества раз, когда пары лексем встречались вместе – получен **вес**, **соответствующий каждой связи** (количество раз, когда лексемы упоминались совместно).

4 На основе биграм построены **сети совместной встречаемости лексем** (формат .net). Для сетевого анализа использована программа Pajek).

Пример биграмы:
инициативный {'всероссийский': 1, 'исследование': 1, 'опрос': 2, 'проект': 1}

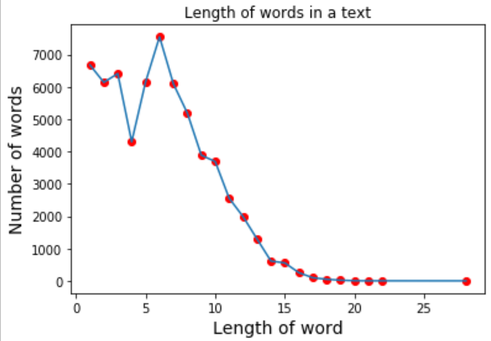
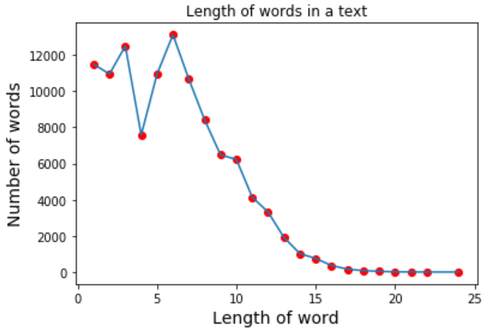
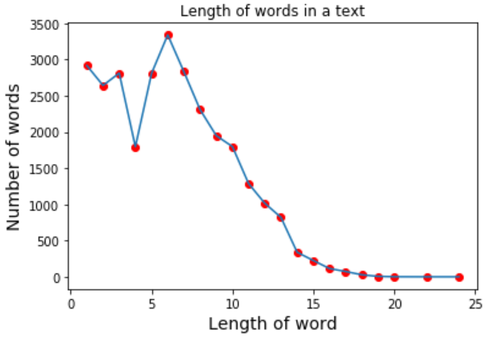
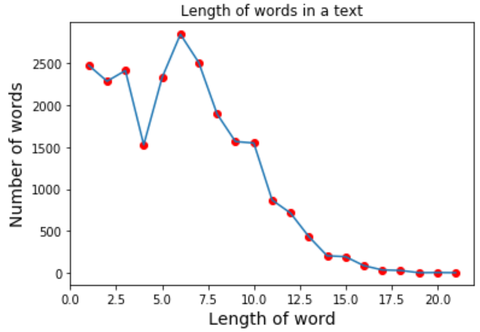
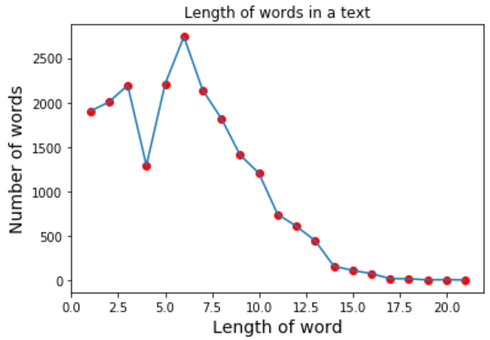
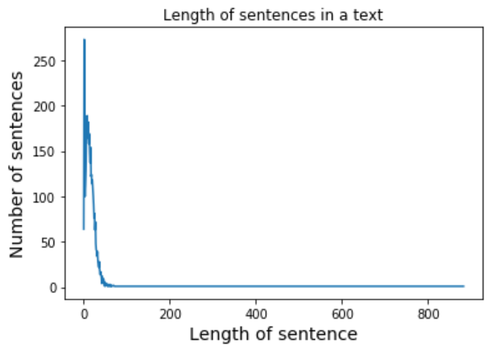
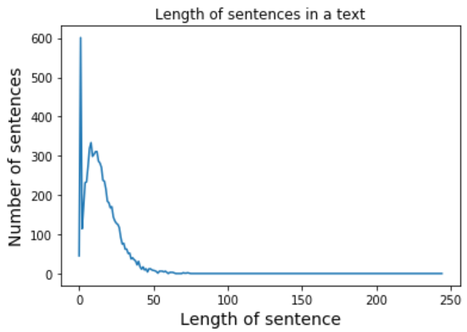
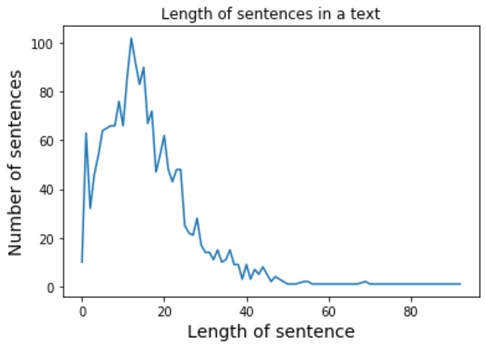
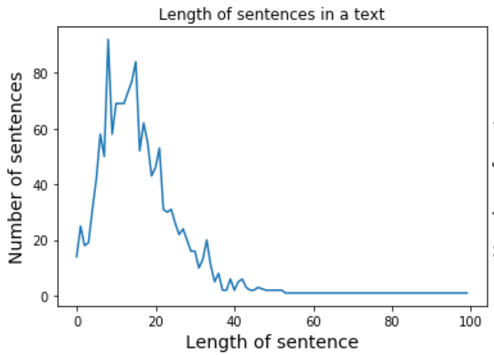
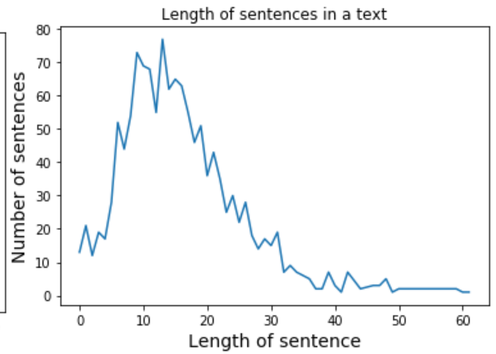
5 В случаях, когда связи между узлами (лексемами) в сети были взаимными, направленные связи были **заменены на ненаправленные** (вес связей при этом был суммирован).

6 В полученных сетях произведены незначительные трансформации и чистка (удалены т.н. «петли» - связи узла с самим собой). Получена **базовая информация** по каждой сети (число узлов и связей, узлы с наибольшим количество связей).

7 К трансформированным сетям применены подходы обрезания по весу связи (**"line cut"**) и поиска островов (**"Islands approach"**), которые позволяют выделять наиболее тесно связанные друг с другом подгруппы в сети.

8 Для визуализации сети использовался алгоритм Kamada-Kawai. Размер узла соответствует его степени входящей центральности (метрика indegree), а размер связи – ее весу.

Общие метрики по текстам

2014-2015	2016	2017	2018	2019
Токены, средняя длина				
5.99	5.85	6.21	5.99	6.0
				
Массив токенов, средняя длина предложения				
15.33	15.03	15.82	16.02	15.96
				

Представленные распределения показывают, что массивы по разным годам являются схожими.

Общие метрики по сетям

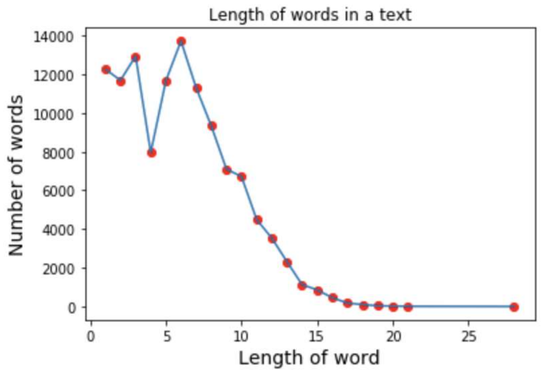
Количество	2014-2015	2016	2017	2018	2019
Узлы (лексемы)	7000	9272	4214	3933	3139
Связи направленные	31266	51375	14077	12249	10300
Связи ненаправленные	955	1793	293	273	280
Среднее количество связей узла	9.2	11.47	6.82	6.37	6.74
Плотность сети	0.0007	0.0006	0.0008	0.0008	0.0011
Размах силы связей	[1-190]	[1-243]	[1-164]	[1-61]	[2 – 116]
Line cut – редукция сети через удаление связей ниже определенного порогового значения					
Выбранное пороговое значение	10	10	7	7	10
Узлы в сети	114	185	117	78	144
Islands approach – редукция сети через выделение наиболее плотно связанных друг с другом компонентов					
Острова в размахе от [5 - 100]	8	7	9	6	1
Узлы в сети	133	119	148	125	77

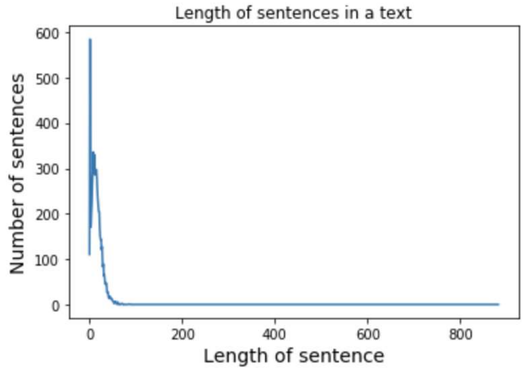
Общие метрики по текстам

Социологи

Единица	Количество
Токены, всего	117 801
Токены, разные словоформы	20 572
Лексемы, всего	9 542
Лексемы-существительные	4 208
Лексемы-прилагательные	1 889
Лексемы-глаголы	2 207

Токены, средняя длина	Массив токенов, средняя длина предложения
5.91	15.42

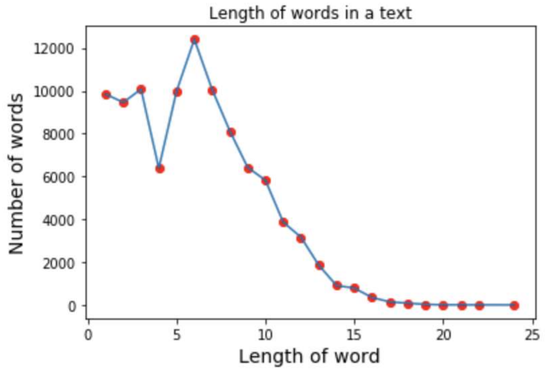


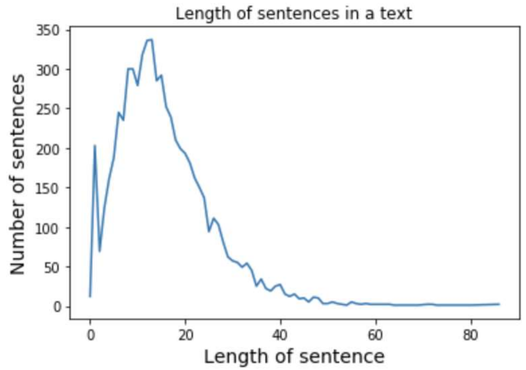


Журналисты

Единица	Количество
Токены, всего	99 759
Токены, разные словоформы	18 036
Лексемы, всего	8 607
Лексемы-существительные	3 853
Лексемы-прилагательные	1 603
Лексемы-глаголы	2 002

Токены, средняя длина	Массив токенов, средняя длина предложения
6.01	15.58





Общие метрики по сетям

Количество	Социологи	Журналисты
Узлы (лексемы)	9543	8608
Связи направленные	55793	47142
Связи ненаправленные	2159	1603
Среднее количество связей узла	12.1	11.3
Плотность сети	0.0007	0.0007
Размах силы связей	[1-301]	[1-271]
Line cut – редукция сети через удаление связей ниже определенного порогового значения		
Выбранное пороговое значение	12	12
Узлы в сети	186	158
Islands approach – редукция сети через выделение наиболее плотно связанных друг с другом компонентов		
Острова в размахе от [5 - 100]	7	9
Узлы в сети	136	146