МЕТОДЫ СОЦИОЛОГИИ

Д.И. Юдина, В.И. Дудина

СЕМАНТИЧЕСКАЯ СЕТЬ НА БИГРАММАХ КАК МЕТОД ВАЛИДИЗАЦИИ РЕЗУЛЬТАТОВ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ В СОЦИОЛОГИЧЕСКОМ ИССЛЕДОВАНИИ

Доступ к Большим Данным, представленным, в частности, текстами в социальных медиа, дает социологам новые возможности для исследований. Однако анализ этих текстовых данных осложнен их большими объемами и неструктурированностью. Значительную помощь в работе с такой информацией предоставляют статистические методы, а именно тематическое моделирование. Проблема состоит в том, что получаемая в результате такого анализа тематическая структура не гарантирует отсутствия ошибок, порождаемых как интерпретативной работой исследователя, так и свойствами самих моделей. В статье рассмотрен способ валидизации результатов тематического моделирования при помощи сравнения с результатами другого метода — построения сетевой модели. В качестве эмпирического материала в исследовании были использованы тексты с ресурса Youtube, представляющие собой комментарии к фильму «Чайка» Фонда Борьбы с Коррупцией. В ходе исследования были построены две тематические модели — «базовая» и «расширенная», в результате анализа которых была получена тематическая структура дискуссии. Методом, использованным для валидизации полученных тем, стала семантическая сеть на биграммах. Данный метод показал свою эффективность, как в качестве инструмента валидизации, так и как способ расширить множество обнаруженных тем. Одним из преимуществ метода стала возможность визуализации тематической структуры. Представленная работа показывает, каким образом можно существенно облегчить «ручную» работу социолога при работе с большим объемом неструктурированных текстовых данных при помощи математических и статистических методов.

Юдина Дарья Игоревна — Санкт-Петербургский государственный университет, Центр социологических и Интернет-исследований, социолог (dartisimus@gmail.com)

Daria Iudina — Researcher, Resource Center «Center for Sociological and Internet Research», Saint Petersburg State University (dartisimus@gmail.com)

Дудина Виктория Ивановна — кандидат социологических наук, Санкт-Петербургский государственный университет, факультет социологии, доцент кафедры прикладной и отраслевой социологии (viktoria_dudina@mail.ru)

Victoria Dudina — Associate Professor, Faculty of Sociology, Saint Petersburg State University (viktoria_dudina@mail.ru)

Ключевые слова: *большие данные, тематическое моделирование, сетевой анализ, семантическая сеть, анализ текста.*

Благодаря аккумулированию различного рода коммуникаций в сети Интернет и относительной открытости этих данных, социологам стали доступны большие объемы информации, которую они могут получить достаточно быстро и без использования опросных методов. Доступ к Большим Данным, представленным, в частности, текстами в социальных медиа, предоставил социологам новые возможности для исследований. Однако анализ этих текстовых данных осложнен их большими объемами и неструктурированностью. Видимая привлекательность использования Больших Данных уже была подвергнута вполне конструктивной критике, например, Бойд и Кроуфорд (Boyd, Crawford 2012) указали на то, что далеко не всегда исследователь имеет доступ ко всем интересующим его данным из-за политики ресурса, который представляет доступ к своим данным или вследствие того, что для анализа этой информации социологу необходимо обладать достаточно продвинутым знанием статистики и программирования или делегировать часть своей работы специалистам из этих областей. Тем не менее, Большие Данные вызывают все больший интерес со стороны социологов и заставляют искать такие способы работы с ними, которые бы позволили получать значимые социологические результаты. Текстовые данные, которые продуцируют пользователи Интернета, лишены привычной для социолога структуры, обычно задаваемой гайдом интервью или вопросами анкеты. Таким образом, неструктурированность и большие объемы данных заставляют искать дополнительные методы анализа текста, помимо качественных методов, предполагающих по большей части «ручную» интерпретативную работу исследователя.

Для решения этой задачи в зависимости от конкретной цели исследования могут применяться разные виды аналитических инструментов. Например, это могут быть автоматизированные лингвистические анализаторы (парсеры), такие, как МальтПарсер (MaltParser) (Nivre et al. 2007) или МиниПар (MiniPar) (Lin 2003), которые позволяют вычленять слова из предложения в соответствии с определенной структурой, например, «субъект-действие-объект» или какой-либо более специфичной. Но такие инструменты редко применяются в русскоязычных социологических исследованиях в силу отсутствия доступных (бесплатно) и удобных для работы аналогов этих лингвистических анализаторов для русского языка.

Почти не зависимые от естественного языка, статистические методы эксплораторного анализа текста получили бо́льшее распространение. К ним можно отнести модели тематического моделирования, наиболее популярной среди которых стала модель на основе латентного разложения Дирихле, разработанная Блеем, Энгом и Джорданом (Blei, Ng, Jordan 2003). Кроме этих моделей, в социологических исследованиях также стало популярным использование сетевого анализа через представление текста или текстов в виде семантической сети и измерение различных характеристик (преимущественно, центральности) этого графа (Freeman 1978; Nerghes et al. 2014; Light 2014; Басов, Василькова 2014). Оба метода имеют как недостатки, так и достоинства. Например, при

использовании обоих методов всегда остается вероятность, что полученные темы будут неверно интерпретированы или пропущены. Кроме того, вследствие низкой частоты встречаемости часть тем может остаться «незамеченной» тематической моделью. Среди недостатков сетевого анализа можно выделить сложность в интерпретации полученной модели из-за ее перегруженности связями между словами.

В представленном исследовании была поставлена задача — выяснить, в какой мере недостатки тематического моделирования могут быть компенсированы использованием сетевого анализа при определенном выборе параметров сетевой модели. В процессе решения этой методологической задачи были проработаны следующие компоненты исследования:

1) Получена выборка текстов из ресурса Youtube, которая оказалась приемлемой для работы, как с тематическим моделированием, так и с сетевым анализом.
2) Выбраны этапы препроцессинга данных, позволившие получить приемлемые модели данных для обоих методов анализа.
3) Представлен вариант эффективного использования сетевого анализа, позволяющий проводить валидизацию тематической модели, а также, при необходимости, допускающий возможность самостоятельного использования.

Выбор корпуса текстов

Основным ограничением при выборе текстов для анализа, наряду с языковым и содержательным, стала целесообразность использования тематического моделирования. Например, тематическое моделирование нет смысла использовать в случае, если почти все документы корпуса состоят не более чем из 10 слов, как, например, статусы в социальных сетях. Для определения тематической структуры в таком наборе текстов будет достаточно простого частотного анализа, а тематическая модель может оказаться слишком ненадежной: ее результаты будут зависеть от малого числа слов. Для того чтобы соблюсти допущения по применимости тематического моделирования, нужно ориентироваться на тексты хотя бы длиной в несколько предложений.

Наряду с этими ограничениями, остается также проблема доступности данных. Безусловно, возможности для их получения очень широки: из Интернет-источников тексты можно скачать, как при помощи API, так и другими способами, например, при помощи различных библиотек парсинга сайтов. Но проблема доступа все равно остается, поскольку администрация ресурсов может так или иначе ограничить закачку данных либо сделать ее максимально неудобной.

При отборе нужных текстов были рассмотрены несколько корпусов текстов — в основном связанных с обсуждениями кинофильмов. Анализ подобного рода дискуссий представляется интересным по нескольким причинам: пользователи сети Интернет активно обсуждают наиболее резонансные фильмы; обсуждение, как правило, вращается вокруг ограниченного спектра тематик, объединенных как сюжетом фильма, так и теми коннотациями, которые фильм вызывает; подобные обсуждения являются косвенным отражением структуры

общественного мнения, т. к. пользователи выборочно реагируют на некоторые сюжеты и интерпретируют показанное в фильме в соответствии со своими установками, опытом, предпочтениями.

В данном исследовании выбор был сделан в пользу дискуссии, развернувшейся в комментариях к фильму Фонда борьбы с коррупцией (ФБК) «Чайка», размещенному на ресурсе YouTube. Тексты из этого обсуждения удовлетворяли почти всем методологическим требованиям, за исключением полной доступности — пришлось ограничиться выборкой. С содержательной точки зрения, основную роль в выборе этих текстов сыграла их высокая общественная актуальность. В фильме «Чайка» представлены результаты расследования ФБК о бизнесе родственников генерального прокурора РФ, Юрия Чайки. Фильм был выложен в сети Интернет 1 декабря 2015 г. и за первые три дня набрал более 1 миллиона просмотров (Газета.ru 2015). Расследование получило значительное внимание как интернет-аудитории, так и различных СМИ (Новостной портал «Медуза» 2015; Телеканал «Дождь» 2015).

Так как сам фильм был размещен фондом на Youtube (Фонд Борьбы с Коррупцией 2015), то пользователи этого популярного видеохостинга имели возможность оставить под видео свои комментарии. К моменту закачки (28.12.2015), этих постов набралось около 19 тысяч. К сожалению, API Youtube не позволяет закачивать более 100 комментариев (YouTube Data API 2015), поэтому пришлось воспользоваться специальным скриптом, который скачивает комментарии путем автоматического повторения действий, который надо было бы предпринимать для ручной закачки (Bouman 2015). Этой мини-программе не удалось скачать все комментарии, видимо, потому что сервер Youtube настроен прерывать подобные «подозрительные» запросы. Тем не менее, с ее помощью удалось получить самые популярные, т. е. получившие наибольшее число «лайков», посты и комментарии общим числом 2907. Данная выборка текстов составила приблизительно 15% от всего объема совокупности постов и комментариев, размещенных с 1.12.2015 по 28.12.2015. Поскольку в нее вошли наиболее популярные посты и комментарии, то вполне можно допустить, что полученные данные содержат наиболее значимые мнения, аргументы и другую информацию. С позиции методологии качественного исследования, подобную выборку можно было бы назвать целевой, с той поправкой, что «целенаправленность» отбора в данном случае была реализована с помощью программных средств.

Препроцессинг данных

В машинном обучении этап работы с данными, включающий очистку «сырых данных» и преобразование их в подходящий формат для работы конкретных алгоритмов, принято называть «препроцессингом данных» (Kotsiantis 2006). Для применения математических методов анализа текста имеющийся корпус текстов был трансформирован в соответствующую матрицу. Матрицы для тематического моделирования и сетевого анализа отличаются, но часть этапов препроцессинга данных для их получения в данном исследовании совпадают.

Общей частью обработки корпуса текстов стали следующие шаги:

1) Из каждого поста были удалены все нетекстовые знаки (кроме дефиса, который может быть частью слова в русском языке), в том числе и цифры, а так-

же текстовые знаки длиной менее трех. Цифры было решено удалить, т. к. они могут быть частью каких-либо метаданных или использоваться в качестве знаков препинания. «Короткие» слова в постах зачастую являются опечатками, хотя в русском языке существуют слова длиной в один или два символа, но ради лучшей очистки от неинформативных знаков было решено ими пожертвовать. Тем более что чаще всего такими словами являются местоимения, которые трудно интерпретировать, не видя их связи с обозначаемыми словами.

2) Из оставшихся слов также были удалены так называемые «стоп-слова» из словаря библиотеки "NLTK" (Loper 2002) для русского языка. В данный словарь входят, в основном, предлоги и служебные слова.

3) При помощи библиотеки "Pymorphy2" (Korobov 2015) все слова в каждом тексте были нормализованы через лемматизацию. Данный анализатор не идеален (например, в нем нет слова «кремль», которое часто встречалось в анализируемом тексте), но в отсутствие других более адекватных и доступных инструментов для русского языка, можно сказать, что со своей задачей он справился.

Этапы препроцессинга, специфичные для тематического моделирования:

1) Удалены низкочастотные слова, то есть те, которые крайне редко встречаются в корпусе текстов, поэтому их присутствие в модели окажется незаметным, но они увеличат будущую матрицу, что в свою очередь увеличит время на ее анализ. В нашем случае было решено удалять слова с частотой менее пяти.

2) Создана матрица «термин-документ».

Ключевым моментом в подготовке матрицы для сетевого анализа стал выбор того, что в дальнейшем будет представлять множество вершин графа. Обычно в качестве таковых выбираются одиночные слова (unigrams), что приводит к сложно интерпретируемой семантической сети. Поэтому было решено выбрать не одиночные слова, а наиболее содержательные устойчивые выражения (collocations). Вследствие этого следующими этапами препроцессинга данных для сетевого анализа стали следующие:

1) Дополнительно к словарю стоп-слов из словаря библиотеки "NLTK" (Loper 2002) были удалены слова и символы, которые оказались высокочастотными, но, вместе с этим, малоинформативными, например: это, который, -, весь, свой, человек, мочь, кто-то, что-то, поэтому, просто, еще, хотя.

2) В качестве вершин графа был выбраны биграммы — выражения из двух последовательно встречаемых слов — с окном (расстоянием в предложении между первым и последним словами включительно) равным 4. Такой размер окна был выбран потому, что устойчивые выражения могут быть разделены в тексте вводными словами, но количество этих слов редко превосходит два. Что касается выбора именно биграмм, а не триграмм или более длинных устойчивых выражений, то, как показал дальнейших анализ, из двух биграмм не сложно понять триграмму и так далее, а длинных устойчивых выражений может быть слишком мало.

3) Для ранжирования биграмм с целью выбора наиболее информативного подмножества был использован метод отношений правдоподобия, который упорядочил биграммы таким образом, что на вершине списка оказались наиболее интерпретируемые результаты по сравнению с теми, которые выбрали другие методы (хи-квадрат, критерий взаимной информации) (Manning, Schütze 1999).

4) Отбор первых n-выражений из полученного упорядоченного множества был проведен итеративным способом. Было протестировано несколько вариантов по количеству устойчивых выражений от 500 до 3000 с шагом в 500. В результате были отобраны первые 3000, поскольку уже при данной выборке биграмм после отфильтровки тех сочетаний, которые встречаются меньше 3 раз в корпусе, их число снизилось до 896 уникальных сочетаний.

Сформирована матрица «биграмма-документ».

Тематическое моделирование

Для работы с тематическим моделированием была выбрана библиотека "topicmodels" (Hornik, Grün 2011) языка R. Выбор был обусловлен удобным интерфейсом, предоставляемым самой библиотекой, а также наличием понятного описания работы с ней. Анализ результатов тематического моделирования проводился последовательно с двумя моделями, получившими название «базовая» и «расширенная». Такое обозначение моделей было продиктовано количеством заданных тем, которые являются одним из параметров тематической модели, определяемыми самим исследователем. Поскольку в тематическом моделировании чаще употребляется слово топик (topic) вместо слова тема (theme), то в дальнейшем оба этих термина будут использоваться как синонимы.

Для построения базовой тематической модели был выбран метод на основе обычного размещения Дирихле с автоматически высчитываемыми начальными значениями гиперпараметров α и β и выборкой Гиббса в качестве алгоритма оптимизации параметров модели. Для нахождения оптимального числа тем использовался показатель внутренней валидности — среднее гармоническое правдоподобие. Так как общее число анализируемых постов относительно невелико, то оптимальное количество тем в модели будет также небольшим, поэтому для сравнения средних количество топиков бралось на интервале от 5 до 50 с шагом 1. Как видно из рис. 1, между 20-ым и 35-ым топиками значения среднего гармонического правдоподобия примерно похожи, но максимальное его значение приходится на 25 тем.

Для получения более детальной тематической структуры дискуссии, а также в целях валидизации тем, полученных от базовой модели, была также построена и проанализирована расширенная модель. Дело в том, что количество тем в базовой модели определялось на основании показателя максимального правдоподобия, который для обычной модели на основе ЛРД выше в случае, когда сходство между топиками минимально. Если же надо получить более детальные топики, то следует увеличить число тем, пробуя разные варианты вручную. В данном случае были выбраны 35 топиков, которые дали достаточно неплохую детализацию некоторым темам из базовой модели.

Для того чтобы определить, какие топики расширенной модели соответствуют топикам из базовой модели, нужно измерить близость (similarity), или похожесть, между отдельным распределением слов по каждой избранной теме из базовой модели и каждым распределением слов из расширенной. В качестве меры расстояния между этими двумя векторами был использован косинус. Эта мера была выбрана по двум причинам: во-первых, она подходит для определе-

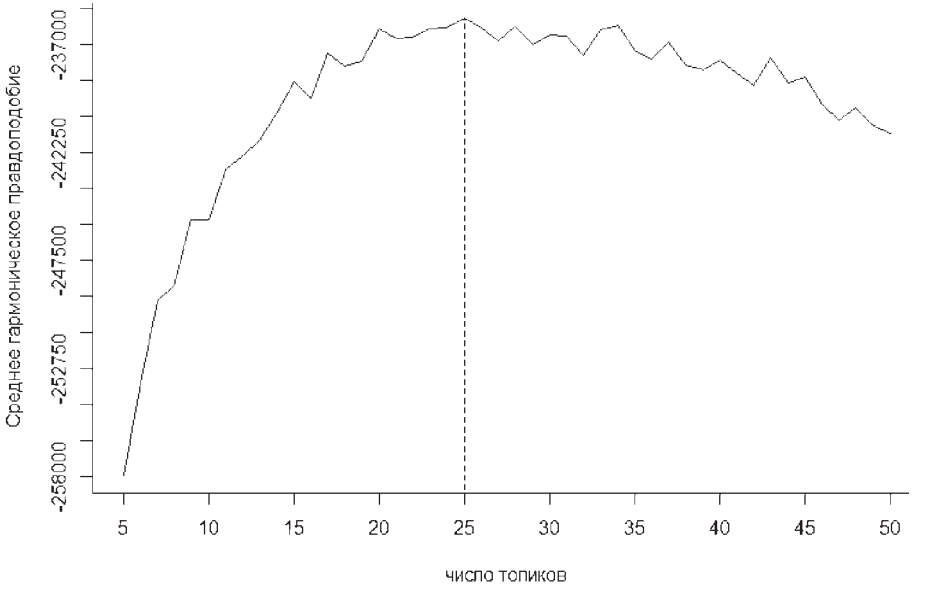

Рис. 1. График зависимости значений среднего гармонического правдоподобия от числа топиков модели Латентного Размещения Дирихле

ния расстояния между случайными векторами, во-вторых, показала свою эффективность в похожих задачах в информационном поиске (Singhal 2001).

В результате работы с тематическим моделированием были получены шесть ключевых тем дискуссии к фильму. Их названия и основные понятия представлены в табл. 1.

Таблица 1

Результаты тематического моделирования

N	Тема	Основные понятия
1	Обсуждение основных фактов и доказательств, представленных в фильме	Бизнес сына генпрокурора, связь с преступлениями, заказчик Браудер, ЕГРЮЛ, доказательства, Цапок, директор пароходства,
2	Последствия расследования для генпрокурора и авторов фильма	Суд, доказательства, ФБК, закон, уголовное дело, прокуратура, чиновники
3	Роль и образ Навального	Навальный, негодяй, патриот, чиновники, оппозиционеры
4	Взаимодействие власти и оппозиции с народом	Государство, народ, деньги, быдло, власть, оппозиция
5	Влияние Путина и нефтяных цен на жизнь в России	Путин, олигархи, Сердюков, нефть, цена, жизнь,
6	Влияние Западных стран и Украины на жизнь в России	Страна, США, Запад, Европа, Россия, война, Украина, русские

Полученные темы можно разбить на два класса: темы с 1 по 3 так или иначе связаны с самим содержанием фильма или его создателями; темы с 4 по 6 скорее относятся к социально-политическому контексту, который ассоциируется у участников дискуссии с фактами, представленными в фильме.

Семантическая модель

После проведения препроцессинга на основе матрицы «биграмма-документ» при помощи библиотеки «igraph» (Csardi, Nepusz 2006) языка R был сформирован и проанализирован взвешенный ненаправленный граф, вершины которого представляли множество биграмм, а ребра — частоту совместного употребления биграмм в постах (если частота равна 0, то ребро отсутствует). Этот граф оказался разбит на 405 связных компоненты, гигантская компонента включает в себя 302 вершины (или биграммы). Кроме нее, получилось еще 5 компонент, содержащих от 12 до 42 вершин, которые включали незначительные аспекты обсуждения, поэтому было решено их подробно не анализировать. Основная работа была проведена с анализом гигантской компоненты, поскольку она представляет собой «основное тело» анализируемой дискуссии.

Для того чтобы получить темы, на которые разбивается главная компонента, был использован метод иерархической кластеризации, в результате которого получилось 8 кластеров с модулярностью 0.25. Каждый кластер было достаточно легко проинтерпретировать, в том числе используя ключевые слова, выделенные путем ранжирования по критерию центральности по посредничеству (betweenness centrality) (Freeman 1977). Стоит отметить также значительную пользу от визуализации главной компоненты и каждого ее кластера, проведенной в Gephi (Bastian 2009) через укладку Force Atlas 2. Благодаря ей оказалась видна структура основного обсуждения.

Анализ гигантской компоненты позволил выделить восемь тем, поскольку каждый кластер удалось проинтерпретировать. Благодаря визуализации топологии графа (рис. 2) удалось выделить центральную тему дискуссии вокруг фильма «Чайка». Что не удивительно, ею стало обсуждение основных фактов из фильма. Как видно из изображения подграфа, образованного этой темой (рис. 3), отдельный аспект составляют комментарии вокруг связи Браудера с созданием «Чайки». Примечательно также то, что все темы, соответствующие контексту из расширенной модели, оказались на периферии структуры графа, а посвященные обсуждению фактов из фильма — ближе к центру.

Все полученные темы представлены в табл. 2 вместе с основными понятиями, в роли которых здесь были использованы биграммы, занявшие наиболее центральное положение в своем кластере по показателю посредничества (betweenness). На визуализации всей гигантской компоненты (рис. 2) видно, что темы 2, 5, 6 сильно друг с другом связаны, что говорит о том, что они составляют одну тему и пару ее аспектов. По биграммам каждого кластера из этой тройки можно понять, что тему здесь задает шестой кластер, а остальные два представляют собой две группы комментариев к ней: сомнения и спор по поводу связи генпрокурора с бандой Цапка, а также призывы проверить доказательства относительно этой связи.

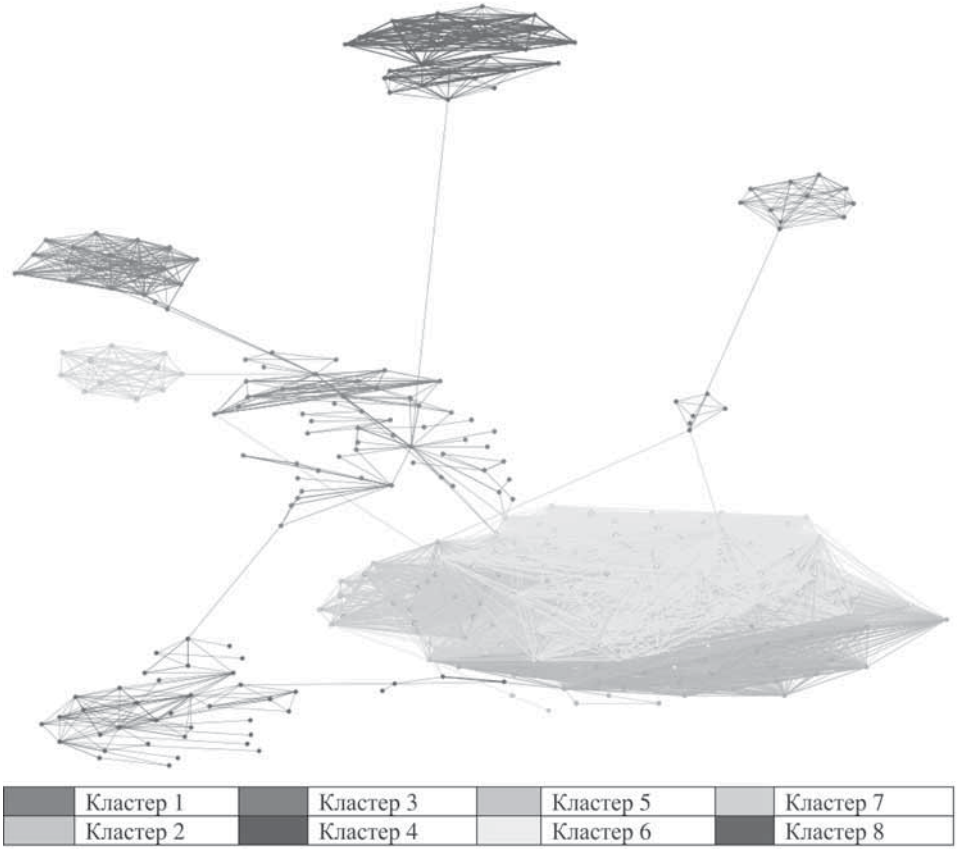

Рис. 2. Гигантская компонента графа биграмм из постов к фильму «Чайка» Фонда Борьбы с Коррупцией

Таблица 2

Темы сетевой модели

№	Тема	Центральные биграммы
1	Обсуждение фактов о фильме	(‘сын’, ‘чайка’), (‘прокурор’, ‘чайка’), (‘уголовный’, ‘дело’), (‘информация’, ‘дать’), (‘использование’, ‘положение’)
2	Спор между участниками дискуссии	(‘чайка’, ‘отмазать’), (‘сайт’, ‘егрюла’), (‘привет’, ‘тролль’), (‘ольга’, ‘лобанов’), (‘жанр’, ‘аргумент’)
3	Возможные действия оппозиции против власти	(‘против’, ‘власть’), (‘пять’, ‘колонна’), (‘юридический’, ‘лицо’), (‘медиа’, ‘продукт’), (‘жечь’, ‘покрышка’)
4	Влияние цены на нефть на экономику России	(‘путин’, ‘чайка’), (‘цена’, ‘нефть’), (‘нефть’, ‘подешеветь‘), (‘падение’, ‘нефть’), (‘рубль’, ‘нефть’)

Окончание таблицы 2

№	Тема	Центральные биграммы
5	Призыв к проверке доказательств, представленных в фильме	(‘факт’, ‘фильм’), (‘показать’, ‘фильм’), (‘доказательство’, ‘перепроверить’), (‘находиться’, ‘лично’), (‘документ’, ‘подтверждать’)
6	Связь генпрокурора с бандой Цапка	(‘жена’, ‘чайка’), (‘чайка’, ‘лопатин’), (‘связь’, ‘чайка’), (‘бандит’, ‘кущёвка’), (‘бандит’, ‘цапка’)
7	Негативные последствия показа фильма для Навального	(‘навальный’, ‘пришить’), (‘воровство’, ‘пришить’), (‘воровство’, ‘убийство’), (‘навальный’, ‘повезти’), (‘повезти’, ‘воровство’)
8	Призыв к отставке генпрокурора	(‘ген’, ‘прокурор’), (‘текст’, ‘петиция’), (‘отставка’, ‘подписать’), (‘отставка’, ‘прокурор’)

Если сравнить темы, выделенные при помощи сетевого анализа (табл. 2) и в результате тематического моделирования (табл. 1), то получится, что в достаточной степени совпали все, кроме обсуждения роли Навального и между-

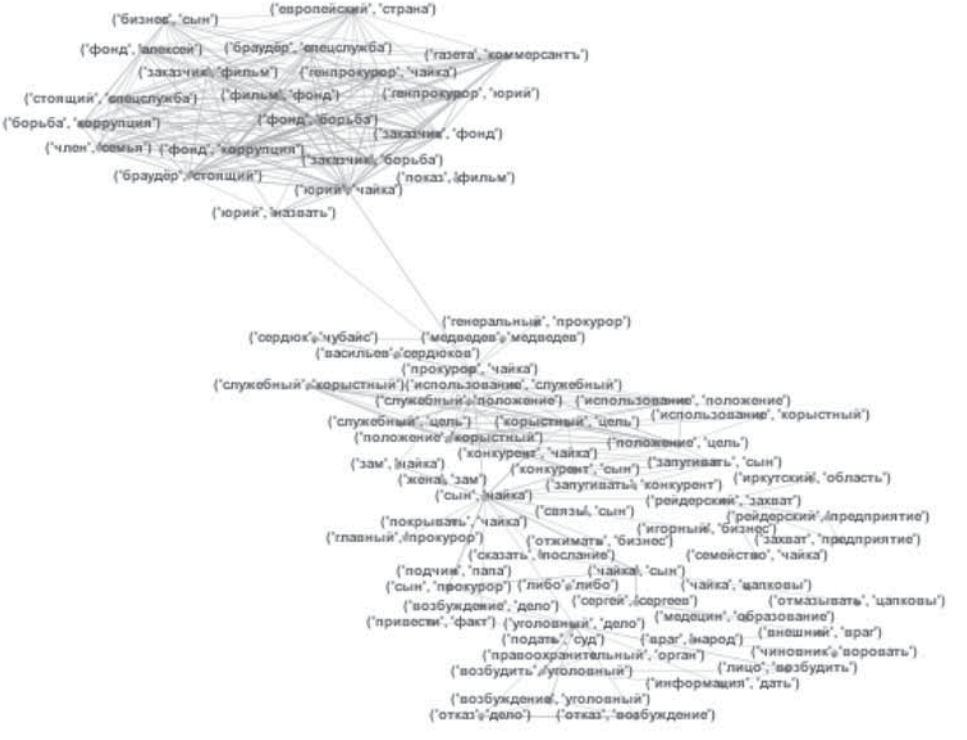

Рис. 3. Кластер 1 гигантской компоненты графа биграмм комментариев к фильму

народных отношений из тематической модели и последствий для Навального из сетевой модели. Эти темы не были выделены в отдельные кластеры, хотя некоторые биграммы могут быть к ним отнесены. Объяснением подобному несовпадению, скорее всего, является то, что обсуждение Навального и образа Запада шло по очень разным аспектам и с разной оценкой, поэтому биграммы из этих постов оказались с очень малой частотой и не были включены в исследуемую выборку.

Тема «последствия для Навального», обнаруженная сетевым анализом, не нашла прямого соответствия во множестве тем в результатах тематического моделирования, но по смыслу ее можно отнести к теме последствий для авторов фильма, хотя топы ключевых слов и биграмм в этих темах не совпадают. Вероятно, обсуждение негативных последствий для Навального появлялось в небольшом количестве постов, поэтому тематические модели ее пропустили. Этот результат говорит в пользу того, что использование сетевого анализа на устойчивых выражениях позволяет компенсировать такой недостаток тематического моделирования, как пропуск редких тем.

Выводы

Представленная работа показывает, каким образом можно существенно облегчить «ручную» работу социолога при работе с большим объемом неструктурированных текстовых данных при помощи математических и статистических методов в отсутствие мощных лингвистических анализаторов. Среди таких методов тематическое моделирование является, пожалуй, наиболее эффективным инструментом, благодаря предоставлению достаточно качественных результатов даже при скромном по числу этапов препроцессинге данных.

Что касается результативности сетевого анализа на множестве биграмм, то стоит отметить его некоторые преимущества перед тематическим моделированием: возможность получить информативную визуализацию структуры дискуссии и более быструю и простую интерпретацию самих тем благодаря устойчивым выражениям. Конечно, тематические модели имеют свои сильные стороны в виде обнаружения тем, состоящих из множества мелких аспектов, и предоставлении ключевых текстов для более содержательной интерпретации. Можно сказать, что их совместное использование позволяет компенсировать некоторые недостатки друг друга и является достаточно эффективным относительно потребления времени и сил исследователя. Кроме того, полученная картина дискуссии к фильму позволяет рассматривать семантическую сеть из биграмм как отдельный метод. Можно с уверенностью предположить, что сетевой анализ будет эффективной альтернативой тематическому моделированию на текстовых данных, представляющих собой множество коротких текстов.

В целом, реализованный подход к проведению поискового анализа на множестве неструктурированных текстов стоит рассматривать как приемлемый вариант при отсутствии продвинутых лингвистических анализаторов или невозможности тратить много времени на создание тренировочной выборки для использования моделей классификации «с учителем».

Литература и источники

Басов Н.В., Василькова В.В. Семантические сети социологического знания // *Журнал социологии и социальной антропологии*, 2014, 1: 112–138.

Газета.ru. Фильм ФБК о сыновьях генпрокурора Чайки набрал более 1 млн просмотров [http://www.gazeta.ru/tech/news/2015/12/03/n_7964111.shtml]. Дата доступа: 08.06.2016.

Новостной портал «Медуза». Расследование ФБК про семью гепрокурора. [https://meduza.io/feature/2015/12/01/rassledovanie-fbk-pro-semyu-genprokurora-otnositelno-korotko]. Дата доступа: 08.06.2016.

Телеканал «Дождь». Самое краткое содержание расследования Навального о семье генпрокурора. [https://tvrain.ru/articles/chaika-399312/]. Дата доступа: 09.06.2016.

Фонд Борьбы с Коррупцией. «Чайка». [https://www.youtube.com/watch?v=eXYQbgvzxdM]. Дата доступа: 25.31.2015.

Bastian M. et al. Gephi: an open source software for exploring and manipulating networks, *ICWSM*, 2009, 8: 361–362.

Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation. *The Journal of machine Learning research*, 2003, 3: 993–1022.

Boyd D., Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 2012, 15(5): 662–679.

Csardi G., Nepusz T. The igraph software package for complex network research. *InterJournal, Complex Systems 1695*, 2006. [http://igraph.sf.net]. Last accessed: 24.07.2016.

Bouman E. Youtube Comment Downloader. [https://github.com/egbertbouman/youtube-comment-downloader]. Last accessed: 28.31.2015.

Freeman L.C. A set of measures of centrality based on betweenness. *Sociometry*, 1977, 40(1): 35–41.

Freeman L.C. Centrality in social networks conceptual clarification. *Social networks*, 1978, 1(3): 215–239.

Hornik K., Grün B. topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 2011, 40(13): 1–30.

Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages, in: *Analysis of Images, Social Networks and Texts*. Springer International Publishing, 2015: 320–332.

Kotsiantis S.B., Kanellopoulos D., Pintelas P.E. Data preprocessing for supervised leaning. *International Journal of Computer Science*, 2006, 1(2): 111–117.

Light R. From Words to Networks and Back: Digital Text, Computational Social Science, and the Case of Presidential Inaugural Addresses. *Social Currents*, 2014: 111–129.

Lin D. Dependency-based evaluation of MINIPAR, in: *Treebanks*. Springer Netherlands, 2003: 317-329.

Loper E., Bird S. NLTK: The natural language toolkit, Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics. *Association for Computational Linguistics*, 2002, 1: 63–70.

Manning C.D., Schütze H. *Foundations of statistical natural language processing*. Cambridge: MIT Press, 1999.

Nerghes A. et al. The shifting discourse of the European Central Bank: Exploring structural space in semantic networks, in: *Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on IEEE*, 2014: 447–455.

Nivre J. et al. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 2007, 13(2): 95–135.

Singhal A. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull*, 2001, 24(4): 35-43.

YouTube Data API. [https://developers.google.com/youtube/v3/docs/comments/list#try-it]. Last accessed: 25.31.2015.